Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Youcheng Zhang, M.Sc.
Born in Guangdong, China
Oral-examination: 10 Sept. 2024

Factor-based approaches for molecular signature discovery in psychotic disorders

Referees: Prof. Dr. Carl Herrmann

PD Dr. Maïwen Caudron-Herger

Abstract

Schizophrenia is a complex mental disorder that has been extensively studied from various perspectives, emphasizing the complexity of factors influencing psychotic disorders. The unraveling of molecular genetics unshadowed a new epoch in the understanding of schizophrenia, redefining it as not only a neuropsychiatric abnormality but an complex disorder with a significant genetic underpinning. In addition, the nature of schizophrenia comorbidity covered a wide range of different conditions, of which the complexity involved emergent syndromes such as type 2 diabetes, cardiovascular disease, and bipolar disorders. In order to better understand the comorbidity of illness, it is important to explore various factors, including genetic variants, molecular profiles, and systematic biological networks.

The rapid development of high-throughput screening techniques has given rise to large-scale biomedical data, which offers the potential to characterize complex biological structures. One of the main challenges is to discover the molecular pattern from high-dimensional data. Secondly, the complexity of biomedical data introduced was compounded by effects of interest as well as different confounders, coming from the heterogeneity of the data itself, which required us to disentangle factors from various sources from the complex variance. And thirdly, multiple layers across various modalities to be analyzed simultaneously became feasible, resulting in the increased application of multi-modal integrative learning that utilizes multidimensional information simultaneously to improve understanding of biological systems.

To address this, factor-based techniques as one of the fundamental approaches in machine learning and data analysis, offer a chance to handle complex datasets to unveil hidden biological data structures. The landscape of factor-based techniques has evolved significantly for diverse datasets and applications, while the versatility of factor-based methods extends far beyond their original usage to current molecular data analysis. Therefore, in the thesis, we would further expand the landscape of the factor-based model and describe the development of our three different factor-based approaches applied in schizophrenia and psychotic disorder research. In Chapter 2, the development of a hierarchical factorization computation framework for signature-based disease comorbidity modeling for schizophrenia and type 2

diabetes was described. In Chapter 3, the deployment of a federated factorization approach was demonstrated, which projected the standard integrative factorization to a federated version based on the DataShield platform, thereby facilitating cross-modal and cross-institutional learning. In Chapter 4, a novel application based on the extended biological-informed interpretable factorized variational autoencoder was introduced for predicting genotype-to-function association in schizophrenia individuals, enabling the separation of the variance of diagnosis from the complex effects by controlling confounding factors as well as the prediction of biological processes with genetic variants. In the end, the prospects and outlook will be discussed in Chapter 5 regarding the further extension of the factor-based techniques and the research gap in schizophrenia and psychotic disorders.

Zusammenfassung

Schizophrenie ist eine komplexe psychische Störung, die aus verschiedenen Blickwinkeln eingehend untersucht wurde, wobei die Komplexität der Faktoren, die psychotische Störungen beeinflussen, im Vordergrund stand. Die Entschlüsselung der Molekulargenetik leitete eine neue Epoche im Verständnis der Schizophrenie ein und definierte sie nicht nur als eine neuropsychiatrische Anomalie, sondern als eine komplexe Störung mit einer bedeutenden genetischen Grundlage. Darüber hinaus umfasste die Art der Komorbidität der Schizophrenie ein breites Spektrum verschiedener Erkrankungen, zu deren Komplexität auch neu auftretende Syndrome wie Typ-2-Diabetes, Herz-Kreislauf-Erkrankungen und bipolare Störungen gehören. Um die Komorbidität von Krankheiten besser zu verstehen, ist es wichtig, verschiedene Faktoren zu untersuchen, darunter genetische Varianten, molekulare Profile und systematische biologische Netzwerke.

Die rasche Entwicklung von Hochdurchsatz-Screening-Techniken hat zu einer großen Menge biomedizinischer Daten geführt, die das Potenzial zur Charakterisierung komplexer biologischer Strukturen bieten. Eine der größten Herausforderungen besteht darin, die molekularen Muster aus den hochdimensionalen Daten zu entdecken. Zweitens wurde die Komplexität der biomedizinischen Daten durch interessierende Effekte und verschiedene Störfaktoren, die sich aus der Heterogenität der Daten selbst ergeben, verstärkt, so dass wir Faktoren aus verschiedenen Quellen aus der komplexen Varianz herauslösen mussten. Und drittens wurden mehrere gleichzeitig zu analysierende Ebenen über verschiedene Modalitäten hinweg möglich, was zu einer verstärkten Anwendung des multimodalen integrativen Lernens führte, das multidimensionale Informationen gleichzeitig nutzt, um das Verständnis biologischer Systeme zu verbessern.

In diesem Zusammenhang bieten faktorbasierte Verfahren als einer der grundlegenden Ansätze des maschinellen Lernens und der Datenanalyse die Möglichkeit, komplexe Datensätze zu verarbeiten, um verborgene biologische Datenstrukturen zu enthüllen. Die Landschaft der faktorbasierten Techniken hat sich für verschiedene Datensätze und Anwendungen erheblich weiterentwickelt, wobei die Vielseitigkeit der faktorbasierten Methoden weit über ihre ursprüngliche Verwendung für die aktuelle molekulare Datenanalyse

hinausgeht. Daher werden wir in dieser Arbeit die Landschaft der faktorbasierten Modelle weiter ausbauen und die Entwicklung unserer drei verschiedenen faktorbasierten Ansätze beschreiben, die in der Forschung zu Schizophrenie und psychotischen Störungen eingesetzt werden. In Kapitel 2 wurde die Entwicklung eines hierarchischen Faktorisierungsberechnungsrahmens für die signaturbasierte Krankheits-Komorbiditätsmodellierung für Schizophrenie und Typ-2-Diabetes beschrieben. In Kapitel 3 wurde der Einsatz eines föderierten Faktorisierungsansatzes demonstriert, der die standardmäßige integrative Faktorisierung auf eine föderierte Version auf Basis der DataShield-Plattform projiziert und dadurch modal- und institutionsübergreifendes Lernen erleichtert. In Kapitel 4 wurde eine neuartige Anwendung auf der Basis des erweiterten biologisch-informierten interpretierbaren faktorisierten Variations-Auto-Coders zur Vorhersage von Genotyp-Funktions-Assoziationen bei Schizophrenie-Personen vorgestellt, die die Trennung der Varianz der Diagnose von den komplexen Effekten durch die Kontrolle von Störfaktoren sowie die Vorhersage von biologischen Prozessen mit genetischen Varianten ermöglicht. Abschließend werden in Kapitel 5 die Aussichten und der Ausblick im Hinblick auf den weiteren Ausbau der faktorbasierten Techniken und die Forschungslücke bei Schizophrenie und psychotischen Störungen diskutiert.

Contents

List of Abbreviations

AE	Autoencoder
BPD	Bipolar disorders
CVD	Cardiovascular diseases
FA	Factor analysis
FL	Federated learning
fRMA	Frozen robust multiarray analysis
GO	Gene ontology
GSEA	Gene set enrichment analysis
GWAS	Genome-wide association study
HGNC	HUGO Gene Nomenclature Committee
ICA	Independent component analysis
iNMF	Integrative non-negative matrix factorization
jNMF	Joint non-negative matrix factorization
LD	Linkage disequilibrium
MDD	Major depressive disorders
MF	Matrix factorization
NB	Negative binomial
NMF	Non-negative matrix factorization
OR	Odd ratio
PaWAS	Pathway-wide association study
PCA	Principal component analysis
PGC	Psychiatric Genomics Consortium

PSY	Psychotic disorders
PWAS	Proteome-wide association study
SCZ	Schizophrenia
SNPs	Single nucleotide polymorphisms
SVD	Singular value decomposition
T2D	Type 2 diabetes
TWAS	Transcriptome-wide association study
VAE	Variational autoencoder

Chapter 1

Introduction

1.1 General concepts in psychiatry

1.1.1 Neurobiology of the nervous system

Psychiatry is a medical specialty field that focuses on researching, understanding, diagnosing, and treating diseases of the brain and dysfunctional mental conditions related to cognition, behavior, and emotion [1]. Biological psychiatry is one of the subspecialties of psychiatry with aims to explore the biological function of the nervous system while the study subject covers the biological basis of psychotics at different levels, including molecular, genetics, epigenetic, anatomy, and physiology in the fundamental research as well as diagnosis, intervention, drugs, and adverse effects from the clinical aspect (**Figure 1.1**) [2]. To actually understand how the underlying biological basis works, one important component to be specifically investigated that is closely involved in this topic is the nervous system. The human nervous system is composed of sub-system including the central nervous system and the peripheral nervous system, where the brain and spinal cord form the central nervous system, while the peripheral nervous system encompasses the rest of the nervous system [3]. Within these nervous systems, neurons and synapses are the two essential basic units of our brain [3]. Neuron or nerve cell is a type of electrically responsive cell to enable the electric signal (action potentials) run across the nervous system, while synapse is one of the most basic structures that allow the neuron to pass an electrical or chemical signal to another neuron (or to the effector cell). This cooperation between neurons and synapses constitutes the operation and activity of the nervous system, such as passing information and further processing information in our brain. There are two major types of synapses: electrical and chemical, of which the majority of synapses are chemical [3]. Chemical synapses utilize neurotransmitters, which are chemical substances released at the end of a neuron when the neuron receives an electrical impulse, to transfer signal and information across cell

membranes. Various types of neurotransmitters exist in the nervous system, specifically, several of them associated with psychosis that we are going to further investigate in the following Chapters of this thesis, including glutamate, Gamma-aminobutyric acid (GABA), dopamine, serotonin [4]. For instance, glutamate is an excitatory neurotransmitter that exerts its effects by binding the receptors and activating cascading effects involving multiple downstream biological functions such as regulating mitochondrial activities, changing levels of nitric oxide (NO) concentration, balancing oxygen reactive species (ROS), mediating transcription factors (TF) and further influencing specific gene expression. It has been observed that the changed levels of glutamate are associated with psychosis symptoms including schizophrenia which is also the disease subject to be discussed later [5]. In contrast with glutamate, gamma-aminobutyric acid (GABA) is the inhibitory neurotransmitter that reduces the excitability in the nervous system activities including stress, fear, clam and other feelings by controlling the chemical substances (mainly chloride) flow across the cell membrane [6]. Besides the central nervous system, GABA also exists and plays different roles such as mediating insulin signalling in insulin-producing beta cells (β-cells) of pancreas and regulating inflammatory responses in immune cells of multiple tissues [7,8]. Another important substance is dopamine that can have either excitatory or inhibitory in diverse brain functions by binding to different receptors (e.g. D1 to D5 dopamine receptors), and it activates downstream effects on learning, emotion, reward functions [9]. Dopamine also responds actively in other systems such as the immune, pancreas and kidneys [10–12]. It is also associated with several mental health conditions and has been frequently studied as the target of psychotropic medications in psychiatry [13]. Furthermore, serotonin plays multiple roles in multiple neuropsychological processes such as cognition, mood regulation [14]. It also participates in a wide range of activities and mechanisms including cardiovascular function, vascular system development, etc [15,16]. Numerous evidences believe that the nervous system has dysfunctions of these neurotransmitters, thereby resulting in changes in pathophysiology of psychosis regarding brain functions and further behavior, emotion, memory, cognition and so on [17]. However, the underlying facts of this abnormality have not been fully understood so far, the factors that contribute to these fluctuations could be complex and involve multiple underlying mechanisms.

1.1.2 Molecular genetics, genomics and transcriptomics

Molecular genetics is a merging field of biology that applies approaches from molecular biology and biotechnology to investigate the structure and function of genes, including aspects of genomic heritability, gene expression, gene regulation, and the molecular mechanism related to the biological processes [18]. Typically, molecular genetics aims to study the difference of gene patterns and activities associated with different phenotypes (e.g. patients with specific disease and healthy individuals, etc.), and these gene patterns can be in various forms such as mutations in a gene, or expression of a gene. Influenced by the Central Dogma from DNA, transcribed RNA, to translated proteins, molecular genetics gave rise to the specific fields including genomics, transcriptomics and proteomics, which aim to address the problems in genome, transcriptome and proteome, respectively (**Figure 1.1**) [19]. In this thesis, genomics and transcriptomics will be our main application fields, and be demonstrated in different following Chapters.

Genomics is an integrated field to provide comprehensive insight on genomes, with the purpose to characterize and quantify all the genes in an organism collectively [20]. One of the major applications is to implement large-scale comparison of the genome in populations and to investigate the relationship between the genetic variation and phenotypes (or traits), namely population genomics. Genetic variation can be expressed at different levels, of which single-nucleotide polymorphism (SNPs) is frequently used to describe the changes of a single nucleotide (a single DNA building block) at a location in the genome [21,22]. For instance, a SNP may substitute the nucleotide adenine (A) with the guanine (G) in a specific position of the genome. In general, the percentage of SNPs in a genome is 0.1%, and SNPs can be properly identified in at least 1% of the population [22]. Previous studies have observed that SNPs can be regarded as the biomarkers due to their association with traits such as disease condition and the relatively large effects on the diseases in particular psychosis. And for the SNPs that are located inside a gene region, they were found to have a more direct influence on the abnormal conditions by affecting the gene functions. Therefore, identifying SNPs that are associated with certain traits help predict the molecular functions, mechanisms, and biological processes. To approach this goal, it is common to perform Genome-wide association studies (GWAS) to determine the association and relationship between a genetic variant and a disease [23]. For example, GWAS is often conducted on two groups of individuals, including the case group (individuals with a specific disease) and control group (individuals who are healthy

control) to examine whether the frequency of the nucleotide in a certain location are significantly altered, and to quantify the effect size (how strong the association is), odds ratio is calculated as the ratio of the number of case individuals with a specific nucleotide and the number of control individuals with a specific nucleotide, divided by the ratio of the number of case without this nucleotide and the number of control without this nucleotide [24]. With the continuous development of GWAS methods, post-GWAS analysis (pGWAS) was implemented to move one step forward from studying the association of variants with phenotype to extensive genome exploration, such as 1) fine-mapping identification of lead SNPs with most significant functional influences on the phenotype; 2) calculating the polygenic risk score (PRS) aggregating all the SNPs in an individual using effect size from summary statistics as weights to summarize the heritable risk of a specific phenotype into a score; and 3) predicting gene expression or biological function based on the genetic variants and their corresponding information in summary statistics to enhance the interpretability as well as their annotation [25].

Transcriptomics, as another important approach, are applied to understand molecular genetics from the aspects of RNA transcripts, so-called transcriptomes, in either an individual or a single cell [26,27]. Gene expression is measured and quantified to evaluate the extent of the gene function as an approximation. Currently, there are two main categories of transcriptomics techniques to study the activity of transcriptome including microarrays and RNA-sequencing [28,29]. Both procedures include RNA isolation, extraction, enrichment and quantification [30]. Microarrays techniques utilize the chips that are composed of oligonucleotide probes with a pre-designed DNA sequence in the spot for matching the RNA of interest. To measure the level of gene expression, fluorescence intensity on every single spot of the chip will be recorded. The raw intensity information will be further preprocessed, normalized, summarized via data analysis tools, and finally transformed to a gene expression matrix where samples and features on rows and columns, whereas values correspond to the gene expression level. RNA sequencing (RNA-Seq) is another technique developed on the basis of next-generation sequencing technology. Instead of using the probes, RNA-Seq samples the transcriptome with short length fragments (reads) and aligns the reads to reconstruct the transcripts expression status. The reconstructed reads that have been mapped back to the specific location of the genome or gene region, will then be quantified. It allows the large-scale sequencing and data processing, and rapidly became more preferable in the recent years compared to the microarray due to the increased throughput, accuracy and

application such as sequencing on single cell level. Under the context of psychiatry, transcriptomics provides an in-depth insight to understand the gene expression variation in patients compared to the healthy individuals. It also contributes to molecular discovery and mechanism study that can explain the pathology and development of psychosis [26,31]. Overall, it serves as a complement to genomics and has been implemented to gain further understanding of the variations of transcriptome in the field of psychiatry.

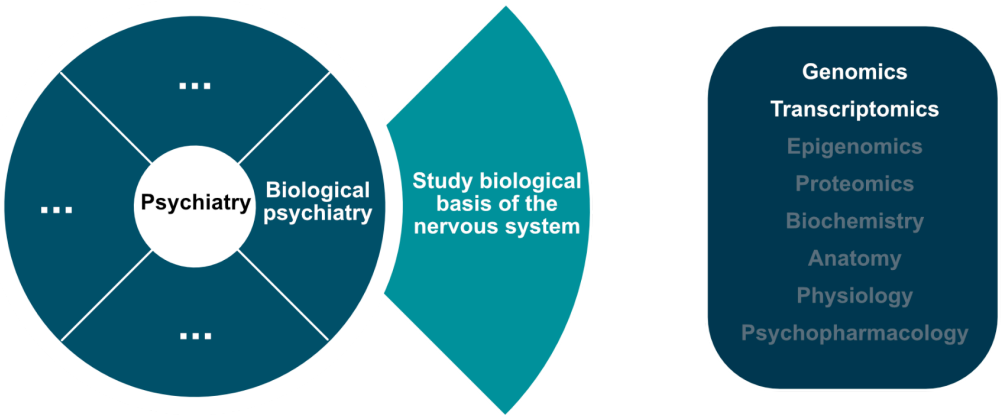


Figure 1.1 Illustration of the study fields and methods in psychiatry

1.2 Schizophrenia and psychotic disorders

1.2.1 The neurobiology of schizophrenia

Schizophrenia is a complex mental disorder that has been extensively studied from various perspectives [32–34]. Schizophrenia and related psychotic disorders encompass a diverse range of psychiatric conditions characterized by abnormalities in thinking and perception, including delusions, hallucinations, disorganized speech, and cognitive impairments [35–38]. These symptoms exist on a continuum, with each individual presenting a unique constellation of manifestations that define their place within the clinical spectrum [39]. The prevalence of schizophrenia is approximately in the range of 0.3% to 0.7% during the course of life,

impacting millions of individuals worldwide, signifying a substantial public health concern [40,41] . In terms of quality of life, these disorders often have a profound impact, impairing personal, social, and occupational functioning. The burden extends beyond the individual to family and community networks, where the consequences of disrupted relationships and reduced productivity are deeply felt [42–44]. The pervasive nature of psychotic disorders results not only in an ongoing endeavor for mental healthcare to provide treatment and support but also in extensive research efforts to understand the long-term effects on patients' lives.

Tracing back the history of psychosis studies, the cognitive neuropsychology of schizophrenia was explored to understand the cognitive deficits associated with the disorder [45]. Research has been conducted to search for the causes of schizophrenia, highlighting the importance of understanding the underlying factors contributing to the development of the disorder [45,46]. Studies have also investigated the distribution of body mass index (BMI) across schizophrenia patients to evaluate the weight changes induced by antipsychotic treatment [47]. Late-onset schizophrenia and very-late-onset schizophrenia-like psychosis have been subjects of interest, with efforts to clarify their positions in relation to schizophrenia [48]. The reformulation of the diagnosis of schizophrenia was proposed to improve the understanding of its causes and develop more accurate diagnostic criteria [49]. Anatomical asymmetries and language lateralization in schizophrenia have been analyzed through meta-analysis, revealing differences in brain structures in individuals with the disorder [50]. Various models, including psychological, social, and biological approaches, have been proposed to understand schizophrenia comprehensively [51]. Anatomical differences between first-episode and chronic schizophrenia have also been examined through meta-analysis, shedding light on the structural brain changes associated with the progression of the disorder [52].

In recent years, a systematic review has been performed on the overlap between psychotic symptoms, substance use, and adversity history, highlighting the complexity of factors influencing psychotic disorders [53]. The composition of mental traits was explored in major psychiatric disorders, including specific configurations for anxiety disorders, depression, and psychotic disorders [54]. In terms of the specific pathophysiology, the receptor affinity of lumateperone tosylate was profiled, highlighting the serotonergic, glutamatergic, and dopaminergic pathways, as well as showing efficacy in potentially treating various symptoms of schizophrenia in the study [55]. The antioxidant status was studied in patients with

paranoid schizophrenia, shedding light on the potential role of oxidative stress in the disorder [56]. A systematic review and meta-analysis were conducted on serum interleukin-6 levels in schizophrenia, suggesting a potential role for inflammatory markers in the pathophysiology of the disorder [57]. The shared mechanisms of neurodevelopmental and stress-related psychiatric disorders provided insights into transdiagnostic approaches [58]. Furthermore, a withdrawal of psychotic psychopathology for shared psychotic disorder was observed in a case study, emphasizing the importance of specific pharmacotherapy [59]. Another finding of altered dopaminergic state in the striatum consolidated the hyper-dopaminergic hypothesis in schizophrenia pathophysiology [60].

Overall, these studies shed light on the complex nature of schizophrenia and psychotic disorders, highlighting the importance of early diagnosis, appropriate treatment, and understanding the underlying biological mechanisms contributing to these conditions.

1.2.2 Molecular genetics of schizophrenia

The studies from the aspect of molecular genetics unshadowed a new epoch in the understanding of schizophrenia, redefining it as not only a neuropsychiatric abnormality but a complex disorder with a significant genetic underpinning.

The very early understanding of the genetic component of schizophrenia has been focused on in twin studies [61]. Substantial insights were provided by identifying susceptibility genes, along with variations in genomic regions known as copy number variants (CNVs), associated with schizophrenia, thereby suggesting an inherent predisposition within the genomic architecture of affected individuals [62,63]. Various alterations in the nervous system was found to contribute to the pathophysiology of the disorder, including dopaminergic, glutamatergic, and serotonergic systems — in the emergence of schizophrenic phenotypes, implicating multiple molecular targets for therapeutic intervention [64–66]. Later, a comprehensive overview of the key molecular and cellular mechanisms implicated in the pathogenesis of schizophrenia was provided [67]. Possible biomarkers as the factors that contributed to psychosis were discussed in several studies [68,69]. Novel mechanisms, such as hypoactivity of the glutamatergic system via NMDA receptors and abnormalities in GABAergic interneurons of neural circuits, were discovered that had an impact on cognitive

deficits and negative symptoms [70,71]. Disruptions in myelination and white matter integrity, as evidenced by alterations in oligodendrocyte-related genes, further implicate the importance of connectivity in the full manifestation of the disorder [72].

Genomic profiling was applied to reveal the risk variants associated with genes that had different functional roles in schizophrenia patients due to the fact that genetic architecture of schizophrenia were found to have high heritability in the population with relevant disease traits [73]. The common variants were identified associated with specific genes by GWAS. For example, the Schizophrenia Exome Sequencing Meta-analysis (SCHEMA) Consortium discovered a set of genes such as glutamatergic related genes *GRIN2A* and *GRIA3*, guanosine genes *DNM3*, guanylate associated genes *MAGI2*, etc. that contributed to synaptic organization and function directly or indirectly [74–76]. The variants of *GRM3* associated with mGlu3 metabotropic glutamate receptors in schizophrenia were identified, and impaired cognitive function was also observed [77]. In particular, risk variants related to genes enriched in the synaptic structure and function were highlighted. For instance, synapse formation and glycosylation of synaptic proteins were influenced by the regulation of copy number variation of *NRXN1*, *C4* and *SLC39A8 [78]*. Other functional analyses of schizophrenia demonstrated the roles of the risk variants in multiple ways, including *CACNA1G* and *HCN4* in ion channels and neuronal excitability [79]; *SP4* transcription factor; *STAG1* from the *SCC3* family, *KDM6B* demethylase in chromatin and transcription regulation [75]; as well as *AKAP11* in signaling scaffold and kinase regulation[80]. Processes related to synaptic structure were associated with molecular variations in the neuregulins *NRG1* and *DISC1 [81]*. To provide a comprehensive view of the population rather than specific genetic variants and genes, Legge et al. (2021) calculated the polygenic risk scores to explain approximately 7.7% of the variance in schizophrenia in a case and control study [82].

Other omics-specific studies such as transcriptomics and proteomics underpinned the expression pattern and signatures in schizophrenia. Transcriptomic profiles sequenced from cerebral organoids from schizophrenia patients were determined, with the aim of identifying aberrant cellular pathways in schizophrenia through gene expression studies [83]. Neuroinflammation pathways were identified by utilizing transcriptomics data to identify potential pharmacological targets for schizophrenia [84]. Similarly, RNA-sequencing was performed to characterize the transcriptomic profiles of lymphocytes in individuals with schizophrenia, suggesting the role of immune dysfunction in the incidence of the syndrome

while emphasizing the understanding of the transcriptomic landscape of different cell types and brain regions in schizophrenia [85]. Additionally, cell-type and region-of-interest-specific transcriptomic analysis was conducted on the human habenula in schizophrenia, shedding light on the cell-type-specific transcriptomic alterations in this brain region. These studies highlighted the importance of transcriptomics in unraveling the molecular underpinnings of schizophrenia, as well as how integrating transcriptomics with other omics approaches and investigating specific brain regions and cell types are crucial for advancing our understanding of schizophrenia at the molecular level [86]. Transcriptomics analysis for non-coding RNA was also done in several studies. The association of long non-coding RNAs (LncRNA) with regulatory molecular factors in the brain and their potential role in schizophrenia were further discussed [87]. MiRNA differences related to treatment-resistant schizophrenia were measured and used to identify a specific miRNA signature involved in the severe condition of schizophrenia [88]. Moreover, multiple studies on biomarker discovery at the transcriptomic and further proteomic levels were conducted. Potential biomarkers were identified for schizophrenia diagnosis based on gene expression changes and immune cell infiltration [87,89]. Proteomic analysis of the cerebellum in chronic schizophrenia subjects was implemented, identifying altered protein networks and highlighting the importance of understanding molecular changes in different brain regions [90]. Proteomic sequencing as well as metabolic profiling were performed in chronic schizophrenia patients undergoing a physical activity program, revealing variations in proteins linked to metabolites and further specific biological processes [90,91]. Multi-omic measurement identified subtypes of schizophrenia associated with dysregulated immune function, providing insights into potential individualized therapeutics [92].

Translational studies such as pharmacogenetic research further advanced the therapeutic investigation of schizophrenia and other psychotic disorders. Several studies have investigated the impact of polymorphisms on treatment outcomes in patients with schizophrenia, and comprehensive reviews were conducted on clozapine and other antipsychotics in schizophrenia treatment and highlighted potential pharmacogenetic biomarkers that may influence treatment response and side effects [93,94] . The associated role of genetic polymorphisms in *ABCB1*, *CYP1A2*, and *UGT1A4* in autonomic nervous system dysfunction was explored in schizophrenia patients treated with olanzapine [95]. The association between *DRD2* and *DAT1* gene variants and symptomatic remission in male schizophrenic patients receiving olanzapine monotherapy was investigated [95,96].

Furthermore, the pharmacogenetics of quetiapine, another antipsychotic used in schizophrenia treatment, and identified genetic markers such as *CYP1A2* and *DRD3* that may influence its efficacy and safety were tested [97]. The relationship between the *NRG1* mutation and cognition and the drug response curve in schizophrenia patients prescribed risperidone was examined [98]. Additionally, Elsheikh et al. (2022) provided an overview of the current state of pharmacogenetics in antipsychotic treatment for schizophrenia, emphasizing the potential of personalized medicine approaches in treatment selection [99].

Other studies involving gene-environment interactions explored the possible factors that complicated the molecular landscape of schizophrenia, where environmental factors may influence gene expression or trigger the manifestation of genetic risks. For instance, several studies have explored how gene-environment interactions react in schizophrenia conditions. One study analyzed genetic influence and associated them with environmental exposures in schizophrenic individuals, confirming the additive gene-environment interaction [100]. It was further observed that the polygenic risk was increased with specific phenotypes, such as endophenotypic expression, compared to the healthy siblings as negative controls, providing replicated evidence for the gene-environment effect [101]. Moreover, specific genes, such as *FKBP5* from the immunophilin family, were shown to be responsive to gene-environment interaction in a meta-analysis study [102]. Another systematic review of schizophrenia assessed the gene-environment correlation between the polygenic risk score and childhood adversity [103]. These studies collectively confirmed and illustrated the occurrence of the gene-environment effect in the etiology of schizophrenia.

The landscape of schizophrenia and psychotic disorders is consistently expanding with the advancements in different research techniques, however, accompanied by new challenges. Current challenges in treating these disorders stem from the heterogeneous nature of their clinical presentations and the variable responses patients have to existing pharmacotherapies. A review of etio-pathogenetics, which is a type of studies to investigate the cause of an abnormal condition, diagnostics, and treatment was discussed to explicate the potential for genotyping and biomarker analysis to identify specific subtypes of psychotic disorders, which could lead to more precise and effective interventions [104]. While such approaches are promising, there remain significant obstacles to their widespread implementation, including the need for robust, high-fidelity diagnostic tools and the integration of complex data sets into clinically actionable protocols. Therefore, personalized approaches hold the potential to not

only improve outcomes but also reduce the treatment-related burdens by targeting the unique pathophysiological processes underlying each individual's disorder. The embracement of individualized treatment strategies promises a new era in the management of psychotic disorders, with the aspiration that better outcomes and improved quality of life for patients are on the horizon.

1.2.3 Comorbidities of schizophrenia

The intersection of somatic comorbidities and their side effects associated with the occurrence and treatment of psychotic disorders added another layer of complexity to patient management and outcomes. Previous studies concurred, noting the bilateral relationship where psychotic disorders might be accompanied with somatic conditions, such as other psychosis, type 2 diabetes (T2D), cardiovascular diseases (CVD), and metabolic related syndrome [105–108]. The frequency and impact of comorbid somatic illnesses in individuals with psychotic disorders were later systematically evaluated, suggesting that the presence of physical health conditions can modify the course and prognosis of psychiatric illnesses [108]. The challenges posed by comorbid conditions were multifaceted. For example, diagnosis and treatment can be biased by the occurrence of comorbidity and the improper medication could further lead to extra aggravation of the symptoms. Consequently, knowing the underlying mechanism of comorbidity is vital to developing integrated treatment approaches, selecting antipsychotic medications with consideration of their somatic risk profiles, and improving the long-term prognosis and quality of life (QoL) for patients with coexisting symptoms.

To tackle the challenge above, one primary solution is to understand the molecular genetic basis of the comorbidity. To examine the molecular basis between schizophrenia and various cancer phenotypes (e.g. breast cancer, prostate cancer, etc.), signatures to segregate the comorbidity were identified using omics data [109,110]. In addition, the intersection of obsessive-compulsive disorder and schizophrenia has also been examined, with the emergence of the concepts of schizo-obsessiveness, obsessive-compulsive disorder, and schizophrenia spectrum disorders as dual diagnoses that present unique challenges in clinical practice [111,112]. Furthermore, genetically unrelated comorbidities of schizophrenia that may be modifiable were also observed, potentially improving the understanding of specific patients comorbid with complex disorders [113].

As shown above, the nature of schizophrenia comorbidity covers a wide range of different conditions. However, the complexity of comorbidity involved not only substance use disorders, cancers, etc. but also other emergent syndromes such as metabolic, vascular, and neuron-related conditions. To better address the fact of the comorbid illness, it is important to explore various factors, including genetic variants, molecular profiles, and systematic biological networks. Therefore, in our study, we would focus on the comorbidity with type 2 diabetes (T2D), cardiovascular disease (CVD) and bipolar disorders.

Type 2 diabetes was observed to be prevalent in patients with schizophrenia, particularly those with first-episode psychosis, as well as in patients who have been treated with antipsychotics [114]. The importance of central nervous system insulin was discussed in the pathogenesis of both schizophrenia and diabetes, highlighting common molecular dimensions in glucose metabolism, cognitive functioning, inflammation, and food preferences [115]. Ca2+/cAMP signaling were also found as the link during the occurrence of comorbidity between schizophrenia and diabetes [116]. Common patho-genetic processes involving inflammatory response processes and membrane transport functions were identified in peripheral blood mononuclear cells (PBMC) samples between schizophrenia and type 2 diabetes through systems biology analysis, and transcription factors such as *STAT1*, *RELA* and *ERG* were observed in regulation of the 28 comorbid genes [117]. Furthermore, a bidirectional two-sample mendelian randomization analysis was conducted to examine the insulin resistance in both diabetes and schizophrenia, suggesting that the inflammation-involved mechanism plays a role in establishing the association between cardiometabolic traits and schizophrenia conditions [118].

Cardiovascular diseases (CVD) were found to be comorbid with psychosis in recent studies that focused on exploring the shared molecular components between the two syndromes [119,120]. Overlaps were identified by characterizing 41 risk genes associated with neuronal morphology and three genes related to the neurotransmitter release pathway [121]. Coronary heart disease's (CHD) comorbidity with schizophrenia was explored suggesting changes in platelets and endothelial cells and further induction of hypertension as well as insulin resistance symptoms, ultimately leading to atherosclerosis as risk factors for the co-occurred situation [122]. Furthermore, common variants were discovered across psychiatric traits and cardiometabolic traits that involved factors such as glycemic control and adipokines [123]. Other findings highlighted and validated the existing related processes of neurodevelopment

and the immune system that contributed to the comorbidity between schizophrenia and cardio-metabolic conditions [124].

Another disease, bipolar disorder, which co-occurred with schizophrenia and other mental disorders, has been studied and well documented. For example, a high genetic correlation within different types of psychotics in the heritability study was detected with quantitative genetics approach [125]. Amygdala, a key brain region involved in emotional processing, were identified to show common but also diverse functional connectivity patterns of both schizophrenia and bipolar disorder [126]. Genomic studies further explored the genetic underpinnings of schizophrenia and bipolar disorder, which involved cognitive impairment [73]. One of the largest and most representative consortiums, CommonMind, provided valuable transcriptomic and epigenomic data from postmortem brains, revealing insights into the functional genomics of these disorders [73,127]. Additionally, white matter microstructural alterations were observed as a common pattern in patients with schizophrenia and bipolar disorder, particularly in the limbic system, distinguishing them from major depressive disorder [128]. Processes involved in sleep and circadian rhythm disturbances were also investigated in remitted schizophrenia and bipolar disorder, highlighting the commonality of these disturbances in both conditions [129]. Finally, genetic predictors of expression that embed the cis-effect of *DDHD2* and *XPNPEP3* in the human fetal brain were identified as associated with risk for neuropsychiatric disorders, including schizophrenia and bipolar disorder, providing insights into the molecular mechanisms underlying these conditions [130].

Overall, the literature on schizophrenia and other comorbidities highlights the complex interplay between genetic and neurobiological factors in the development and manifestation of these debilitating mental health conditions. Further research is in demand to better gain knowledge of the underlying mechanisms and to improve treatment outcomes for individuals affected by comorbid conditions.

1.3 Computational challenges in biomedical data

1.3.1 High dimensionality of data

The rapid development of high-throughput sequencing techniques in the past 20 years allowed for the generation of biomedical data in large-scale format with numerous features as well as a large pool of samples, which offered the possibility of *in silico* experimentation, which is poised to characterize complex biological processes. One of the objectives and applications that have benefited from high-throughput screening as well as large-scale data is molecular pattern discovery. Analysis of molecular expression patterns provides in-depth insights into the molecular data structure and condenses the original data into biological patterns that are difficult to understand at the original data structure. Usually, these biological patterns condensed and characterized from the complex data are regarded as signatures or factors. The concept of a biological signature has been explored in various medical contexts. For instance, it was earlier discussed in the context of immunosuppressive treatment study, within which several immune signatures were identified for autoinflammatory syndromes [131]. A biological signature of prenatal maternal stress was identified from a panel of circulating markers, linking psychological stress to low-grade inflammation and its impact on fetal development [132]. Severity-specific biological signatures associated with immune signaling networks and system mobilization were aggregated, utilizing an integrated computational approach to differentiate contagious disease severity with proteomic data at the single-cell level [133]. Another biological signature was identified for the inhibition of outer membrane lipoprotein biogenesis [134]. Winter et al. (2021) discussed the variable biological signature of refractory cytopenia of childhood in a retrospective study, showcasing the diverse nature of biological signatures in different medical conditions [135]. These studies collectively emphasized the significance of biological signatures in understanding disease mechanisms and treatment responses, while illustrating the rapid evolution in this field.

However, one major challenge in the signature discovery process is how to handle the high-dimensionality of the data due to several challenges. For instance, high-dimensional spaces might not be sufficient to break down completely when the dimensionality becomes very high using some of the traditional methods like principal component analysis (PCA) . In the realm of machine learning, signal processing or other relevant areas, clustering high dimensional complex datasets remains a challenging problem due to the curse of

dimensionality [136,137]. Previous evidence has discussed the relationship between bias, variance, and loss within the complex data structure, also highlighting the impact of high-dimensional data on classification tasks [138]. Therefore, the task for properly handling data with high-dimensionality becomes increasingly challenging. Secondly, the computational resources and time required for model learning have also increased as a result of the elevated dimensions. Different algorithms were compared, highlighting the issue of computational complexity [139]. Furthermore, the visualization of high-dimensional data makes it difficult to perform exploratory data analysis.

To summarize, it is a significant challenge to determine the dimensions of molecular data, especially in fields such as biomedical, genomic, and machine learning. And the high dimensionality of data requires innovative approaches to mitigate its effects and improve computational efficiency.

1.3.2 Covariates-confounding of data

Advances in high-throughput technologies, at the same time, also introduced additional complexity to the dataset. One major type of data complexity introduced was confounded by the heterogeneity of high-throughput technologies or platforms, whereas another type of complexity is usually associated with cohorts themselves such as different confounding effects. Various models and methods have been developed to enhance the analysis of high-dimensional biological data, such as molecular signature discovery mentioned in the previous section, with the aim of addressing the dimensionality of the data, allowing for the identification of the specific biological patterns in the lower dimensional structure within the data. However, the compressed data structure sometimes might not be able to exactly prevent the bias induced by unwanted variations inside the datasets, as previously mentioned as confounding effects. Confounding factors can have a significant impact on specific disease subjects such as generating biased results when analyzing the data. For example, when investigating the association between specific genetic variants or expression of a gene with certain phenotypes (e.g. diagnosis, drug response) in psychiatric disorders, confounding effects brought by factors such as gender, lifestyle, exercise, and smoking can cover the actual underlying relationship [140]. Similarly in other disciplines, Morrison et al. (1992) discussed in a more biological context how variations in trans-acting factor genes, such as the vitamin D receptor gene, confounded physiological studies, affecting the regulation of target genes

[141]. McGreevy et al. (2005) focused on the anthropometric differences of IGF-I and IGFBP-3 in plasma, both of which were insulin-like growth factors, emphasizing the need to control for confounding factors like age and height when analyzing racial differences in IGF levels [142]. Katz (2011) discussed the importance of including potential confounding variables in multivariable models to ensure accurate analysis of risk factors and outcomes [143]. Furthermore, Azevedo et al. (2016) explored how somatic symptoms of depression can confound the association between depression and medical prognosis, highlighting the need to control for these confounding factors in research studies [144]. Plenty of the previous studies suggested that confounding factors can significantly impact the results and interpretations of research studies in various fields. It emphasized the necessity to carefully identify, control for, and disentangle these confounds to ensure the validity and reliability of their findings.

To address the problem, attempts were made in previous studies, of which factor analysis (FA) approached the goal and were applied to distinguish and separate underlying factors that explain the relationships among observed variables [145,146]. Factor analysis is one of commonly used statistical techniques employed in biomedical research that lay in its ability to unearth latent variables, or factors, that orchestrate the co-variation observed in a multitude of measured variables. These factors, in other words, hidden effects, represent unobservable effects that influence the shared variance among the measured ones. Within the intricate tapestry of biomedicine, factorized models offer the chance to navigate complex datasets, streamline data structures, and gain insights into the underlying biological processes that dictate the observed phenomena [147]. One prominent application of factor analysis in biomedicine is in the field of omics data, involving diverse data modalities such as genomics, transcriptomics, proteomics, and metabolomics [148]. Factor modelling and factor analysis not only serve as a data dimensionality reduction champion, but also represent an effective approach to identify groups of factors that represent the most significant sources of variation within the data. The class of tools allowed researchers to discover key biological pathways and molecular signatures associated with specific diseases or treatment responses.

Nevertheless, challenges still exist, such as the fact that some of the factor models have not sufficiently addressed the limitation of interpretability. The methods were expected to disentangle underlying factors that specifically explain every source of variation in the dataset, such as diagnosis, gender variables, ethnicity, etc. The identified factors could be directly or indirectly associated with the given phenotypes or clinical index. However, the

factors are post-hoc linked to interpretable representation after the disentanglement, such as by performing enrichment on the factors to understand the biological meanings. So far, most of the previously developed models are facing problems like not only a lack of interpretability but also the additional burden of complicated external analysis to understand their factors [149,150]. Overall, factor analysis is a valuable tool in various research fields for identifying underlying factors that explain the relationships among observed variables. Studies have utilized factor analysis to uncover patterns related to genetics, activity outcomes, biological function and other issues. By advancing the factorized analytical model, it would be possible for researchers to interpret complex data and derive the actual real biological insights.

1.3.3 Integrative learning of cross modality data

With the rapid expansion of different sequencing assays applied, multiple layers across various modalities to be analyzed simultaneously became feasible [151]. The increased complexity is resulting in the increased application of multi-modal integrative learning that utilizes multi-dimensional information simultaneously to improve the understanding of biological systems [152–154]. Specifically, the integrative model was applied in different contexts, such as the integration across multimodality or the integration of multi-cohorts.

Multi-omic integration is a complex process that involves combining data from various 'omic views, such as genomics, transcriptomics, proteomics, and metabolomics, allowing for a comprehensive understanding of complex biological systems. A pilot integrative analysis was performed on colon and gut microbiota molecular expression data in inflammatory bowel disease (IBD) and revealed the gene regulatory networks involved in the regulation of bile acid [155]. In a comprehensive view of systems biology, Lancaster et al. (2020) developed a customized workflow for the integration of multi-omics analysis [156]. Gao et al. (2021) developed a novel algorithm for single-cell omic data integration in an iterative manner, demonstrating the effectiveness of the integrative approach in constructing omics atlas, whereas Chetrit et al. (2022) introduced spatial protein and transcriptome sequencing (SPOTS) methods, which were also factor-based techniques for the integration across spatial data and proteomics, offering a more comprehensive insight on biological activities with the support of location information [157,158]. Further combination of integrative analysis with advanced machine learning algorithms showed success in various studies. For instance, Yang et al. (2022) utilized similarity network fusion (SNF) to integrate multiple high-dimensional

multi-omic data modalities for detecting molecular subtypes of aging [159]. The multi-omic integrative analysis (MOTA) method was developed to leverage multiple molecular pieces of information to identify potential disease biomarkers utilizing the inter-omic connectivity network [160]. MORONET, a novel computational method was introduced for integrating multiple types of omics data using graph convolutional networks for biomedical classification [161]. Furthermore, efforts to develop algorithms were made to compute consensus clusters with multi-omics data for lung adenocarcinoma, demonstrating the utility of integrative analysis in precision medicine [162]. Lastly, MOSEGCN, a hybrid method, was implemented for integrating multi-omics data inspired by multi-head self-attention mechanisms from transformer architecture and deep graph convolutional networks (CNN) to improve the accuracy of disease classification [162,163].

For integrative learning across multiple cohorts, relevant studies were already implemented in fields, including genomics, clinical trials, and disease studies. Data integration approaches for harmonizing phenotypes from multiple GWAS cohorts were developed [164]. Clinical trial heterogeneity from multiple cohorts was discussed in the context of systemic lupus erythematosus patients [165]. In the context of disease studies, Titmuss et al. (2021) explored the use of a multivariate model combining genomics and transcriptomics from a tumor-agnostic cohort to discover various molecular markers for immune checkpoint inhibitors, providing evidence of the potential benefits of integrating data across multiple cohorts [166]. A cross-sectional study of young and older individuals was conducted to illustrate age-related changes in the functional architecture of the cerebral cortex [167]. Similar cross-cohort profiling studies were performed to discover various molecular features [168]. Furthermore, Murphy et al. (2022) highlighted the significance of data integration in assessing facility-level antiretroviral treatment patient status, emphasizing the potential for data integration to improve accuracy in system-level tracking and clinic switching [169]. Meanwhile, Aung et al. (2022) integrated data from two country-level longitudinal studies to investigate abnormal cognitive aging in people with HIV, highlighting the benefits of data integration in increasing sample size and study power [170].

However, studies also discussed the current situation in this field, such as facing difficulties with cross-sectional analyses under different disease scenarios, limited patient-generated data, computation platform development, specifically-designed algorithms for integrative studies. It urges the need for continuous development of new methods to accomplish different tasks for

different biomedical contexts. For instance, it was emphasized in the previous study that harmonizing data across institutions was essential for transforming data into a common representation to allow the data to be learnt in an integrative way [171]. For different target disease subject, different cross-cohort prognostic prediction models were designed and developed for multiple conditions such as osteoarthritis risk, neurodegeneration disease, choline degradation, coronary artery disease, etc, indicating the importance of integrating predictive models into the specific scenario [172–175]. Besides, Etingen et al. (2020) highlighted the need for institutional infrastructure to enhance data integration efforts [176]. In addition, patients' privacy issues are also aware of, in particular for the physically distributed clinical data that can not easily be transferred or even accessed by other institutions [177]. Overall, accessing and utilizing clinical index and genetic data from warehouses is becoming increasingly relevant for molecular studies, with a focus on data processing, algorithm design, and cross-institutional data integration [178].

To summarize, integrative learning across multi-modality and multi-cohorts was increasingly applied across other biological domains and the most advanced model implementation also enabled multi‑modal analyses at diverse molecular levels [179–181]. One of the primary challenges in multi-omic integration is the heterogeneity of the data, which poses a significant obstacle to integrating and analyzing complex datasets, hindering the comprehensive understanding of biological systems. On the other hand, the large-scale integration of data with high-throughput omics data was also limited by the platform utilized, which was not suitable for cross-sectional studies. To address these challenges, novel integration strategies and algorithms for multi-omics data were explored for proper integration to maximize the benefits of adding more omics data. Meanwhile, newly developed cross-institutional integrative analysis platforms are to be constructed. In conclusion, while multi-omics technology holds great promise for advancing our understanding of complex biological systems, current challenges in data integration and processing still need to be addressed to fully realize its potential in integrative molecular studies.

1.4 Factor-based solution for biomedical data analysis

1.4.1 Definition of factor-based technique

Factor-based techniques emerged as one of the fundamental tools in data analysis, machine learning, and signal processing, offering a sophisticated lens to analyze complex datasets to unveil hidden patterns in the biomedical context. Different factor-based methods shared common or very similar characteristics but also had their own specific characteristics regarding their fundamental hypothesis, implementations, functions, applications, etc.

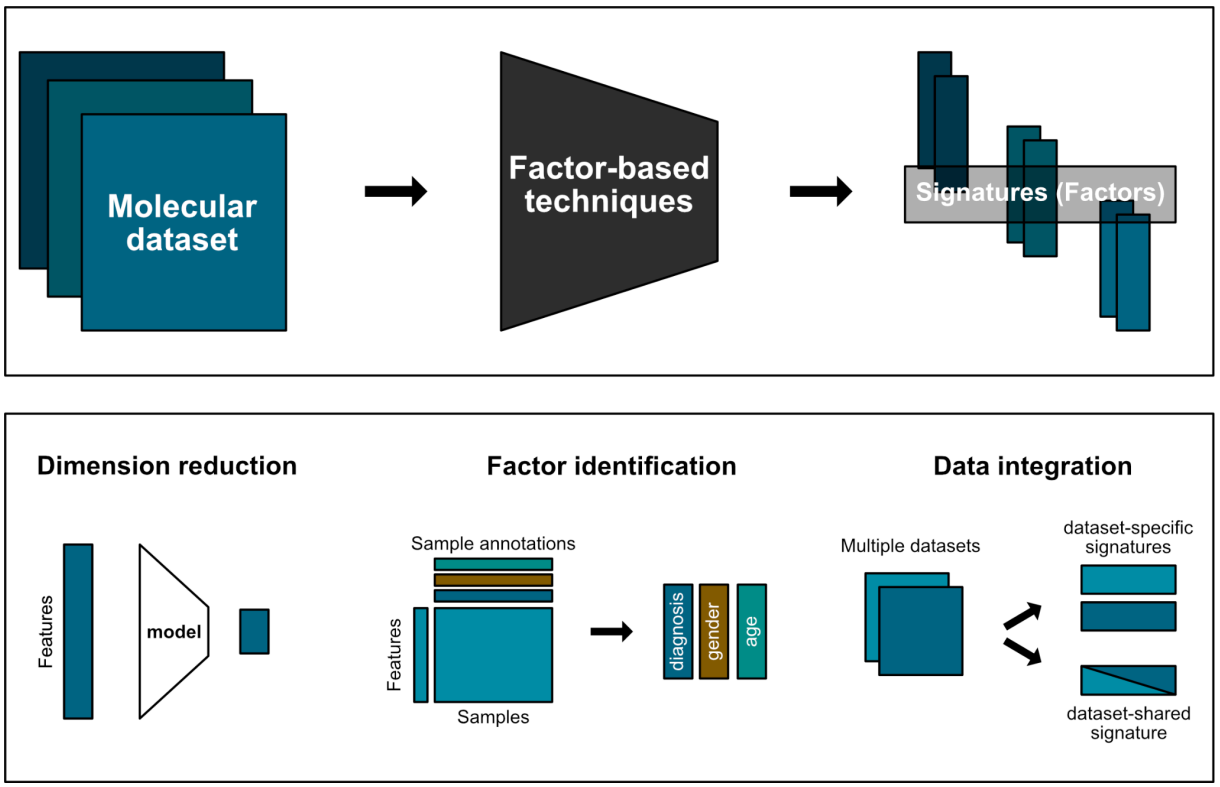


Figure 1.2 The schematic diagram of factor-based techniques and the applications

In a general view, factor-based techniques all involve factor extraction, estimation, and analysis that convert the complex data structure to a simpler or so-called hidden data structure (**Figure 1.2**). This transformed data structure is usually represented by several factors. In other words, using these factors, we can approximate the original data. Most standard factor-based approaches implement similar statistical procedures, including initialization of the model and dimensional reduction via decomposition of the target matrix into reduced features (factors) and their corresponding compressed values. However, differences exist in

several aspects including purposes, assumptions, objectives, ways of explaining variation, and methods to estimate the factor weights. For instance, one of the representative mathematical factor-based frameworks is matrix factorization, where a matrix is decomposed into a product of two or more matrices, typically facilitating the distillation of complex data into more interpretable forms. Within this type of method, a factor or component will be derived from new meta-features that can be approximated as independent against other factors. In practice, due to the complexity of input data, some of the derived factors might be correlated. Therefore, the proportion of variance a specific factor captured is usually used to evaluate the variation obtained in a factor. Different from matrix factorization, another similar approach is factor analysis in statistics, which seeks to uncover underlying variables that explain observed correlations by disentangling different variances of interest. Factor analysis assumed that the factors actually exist in the given input data and attempted to understand how these causal factors captured much of the information in a set of features in the dataset. In addition, factor-based analysis could also be applied across multiple cohorts in which integrative factor-based methods focused on data integration, standing for multi-modality or integrative factorization methods. Unlike most factor-based methods mentioned above that are linear models, aiming to capture linear relationships in the data, non-linear factor-based approaches were developed, of which deep learning architecture, autoencoder (AE) and variational autoencoder (VAE) are the two classic examples [182–184]. A standard autoencoder first passes the raw data through the encoder, which is composed of linear (e.g., linear regression) and nonlinear activation (e.g., sigmoid, tangent function) models, to the low-dimensional representation in the latent space (bottleneck of the model). This latent representation (or so-called latent embedding) can be seen as a factor that largely explains the variations in the original data. The model subsequently passed this representation through the decoder (which can be seen as the reverse process of the former-mentioned encoder) to reconstruct the output. Finally, the optimization of the model is to minimize the difference between the raw input and the reconstructed output. Similarly, a variational autoencoder, which is just a special case of an autoencoder, uses the probabilistic form to represent the data, learns the distribution in the latent space, and samples from this distribution to form the latent representation before passing it on to the decoder. From the perspective of the concepts, autoencoders and variational autoencoders are actually special types of factor models which add nonlinearity to the autoencoder, which holds the same model objectives as matrix factorization and factor analysis. Furthermore, the advanced factor-based architecture that combines deep learning

and conventional linear models is able to capture both linear and non-linear patterns within the same model simultaneously.

Here, I elucidated the landscape of factor-based techniques and summarized the current representative factor-based techniques based on their implementation and application on the diagram in **Figure 1.3**. The key features and connections among the different methods were illustrated with descriptions and arrows. Moreover, the general and primary application in actual data analysis of these approaches was indicated, where methods with similar implementation and practice tended to cluster together and were marked with the same color. We also highlighted the specific methods that have been applied to biomedical analysis, such as genomics, transcriptomics, or molecular studies, with the annotation of representative articles in the figure. In the thesis, I mainly focused on the specific types of factor-based models, including matrix factorization (**Chapter 2**), integrative matrix factorization (**Chapter 3**) and advanced VAE-based architecture (**Chapter 4**) and the technical details as well as mathematics would be demonstrated in each project chapters respectively.

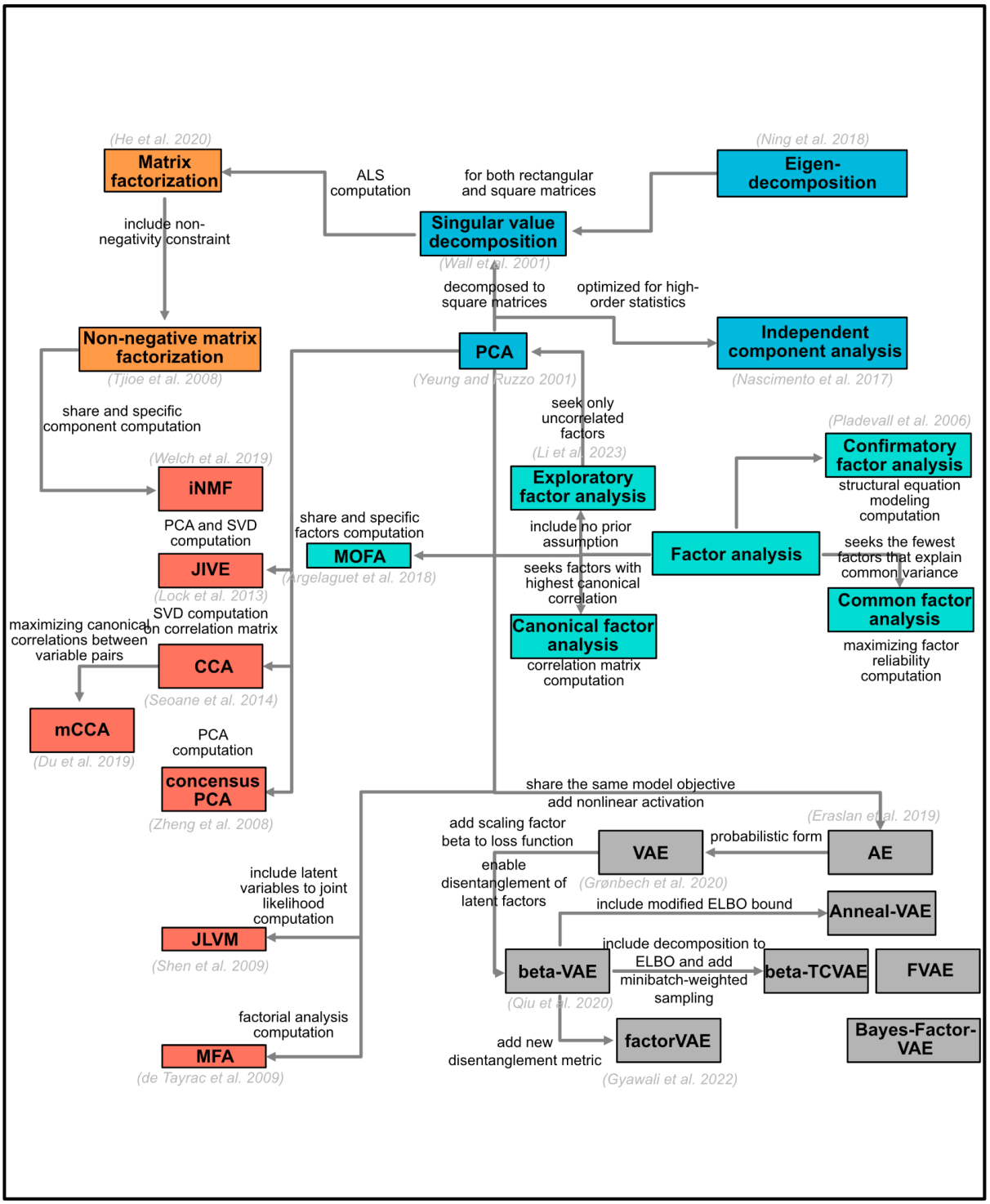

(He et al. 2020)
Matrix factorization
ALS computation
for both rectangular and square matrices
(Ning et al. 2018)
Eigen-decomposition
include non-negativity constraint
Singular value decomposition
(Wall et al. 2001)
decomposed to square matrices
optimized for high-order statistics
Non-negative matrix factorization
(Tjioe et al. 2008)
PCA
(Yeung and Ruzzo 2001)
Independent component analysis
(Nascimento et al. 2017)
share and specific component computation
seek only uncorrelated factors
(Li et al. 2023)
(Pladevall et al. 2006)
Confirmatory factor analysis
structural equation modeling computation
(Welch et al. 2019)
iNMF
Exploratory factor analysis
share and specific factors computation
PCA and SVD computation
include no prior assumption
Factor analysis
seeks the fewest factors that explain common variance
JIVE
(Lock et al. 2013)
MOFA
(Argelaguet et al. 2018)
seeks factors with highest canonical correlation
Common factor analysis
SVD computation on correlation matrix
maximizing canonical correlations between variable pairs
Canonical factor analysis
correlation matrix computation
maximizing factor reliability computation
CCA
(Seoane et al. 2014)
mCCA
(Du et al. 2019)
PCA computation
concensus PCA
(Zheng et al. 2008)
share the same model objective
add nonlinear activation
(Eraslan et al. 2019)
add scaling factor beta to loss function
VAE
probabilistic form
AE
enable disentanglement of latent factors
(Grønbech et al. 2020)
include latent variables to joint likelihood computation
include modified ELBO bound
Anneal-VAE
JLVM
(Shen et al. 2009)
include decomposition to ELBO and add minibatch-weighted sampling
beta-VAE
beta-TCVAE
FVAE
(Qiu et al. 2020)
factorial analysis computation
add new disentanglement metric
factorVAE
Bayes-Factor-VAE
MFA
(de Tayrac et al. 2009)
(Gyawali et al. 2022)

Figure 1.3 The map diagram of factor-based techniques. The color represented the different clusters based on similar method usage, including decomposition, matrix factorization, factor analysis, integrative analysis. The arrow with the description referred to the relationship between two connected methods. The citing annotation of the method name card showed in italics the example application article with that specific method . The detailed article was also provided and listed. Eigendecomposition [185], Singular value decomposition [186], Non-negative matrix factorization [187], Principal component analysis [188], Independent component analysis [189], Exploratory factor analysis [190], Confirmatory factor analysis [191], Factor regression model [192], Canonical Correlation Analysis (CCA) [193], Multi-Omics Factor Analysis (MOFA) [194], Joint and Individual Variation Explained (JIVE) [195], Joint Latent Variable Model (JLVM) [196], Multi-Set Canonical Correlation Analysis (mCCA) [197], Multiple Factor Analysis (MFA) [198], Consensus Principal Component Analysis (Consensus PCA) [199], Integrative Non-negative Matrix Factorization (iNMF) [200], Autoencoder (AE) [201], Variational autoencoder (VAE) [202], beta-VAE [203], factorVAE [204]

1.4.2 Landscape of factor-based techniques

The landscape of factor-based techniques has evolved significantly, witnessing a range of sophisticated methodologies tailored for diverse datasets and applications. The versatility of factorization extends far beyond the original usage of signal processing to current molecular data analysis. Here, four different types of factor-based methods will be introduced.

The first type of factor-based technique is matrix factorization. Matrix factorization could be categorized into two classes, including factorization based on solving linear systems and factorization based on eigenvectors and eigenvalues. Within the category of factorization based on solving linear systems, representative methods involve solving linear equation systems efficiently such as LU decomposition and QR decomposition [205][206]. The second class involved eigenvector and eigenvalue computations which will be the main focus of our studies. Eigendecomposition decomposes a square matrix into a set of eigenvalues and eigenvectors matrix [207]. This decomposition is fundamental in various applications such as principal component analysis. Singular value decomposition (SVD) factored the data matrix

into two different unitary matrices, and a diagonal matrix containing the singular values, which were used in data compression, noise reduction, and solving linear inverse problems [208,209]. Commonly-used matrix factorization methods include advanced forms based on the previously described methods such as singular value decomposition (SVD) and its numerous variants, each offering unique advantages in data deconstruction and interpretation. Principal component analysis (PCA) was one of them, utilizing the concept of eigen-based decomposition to factorize data matrices into orthogonal factors as principal components that explained the maximum variance in the data [210]. The standard steps to perform PCA included standardization of the data matrix, generation of the correlation matrix, decomposition of the correlation matrix into eigenvectors and their respective eigenvalues, and ranking of the vectors in descending order based on their eigenvalues. PCA is a special type of singular value decomposition that identifies the components that embed the observed features, which could explain all the variance in the original data matrix. Compared to PCA, independent component analysis is optimized for extracting independent components in high-order space and it isolates the mixed signal based on the separated sources. However, principal component analysis and singular value decomposition based methods allowed for negative values in the factors, which might not be biologically interpretable in some contexts. Therefore, among these, non-negative matrix factorization (NMF) gained particular attention for its ability to to distill and interpret large-scale genetic information by factorizing non-negative input into two matrices including sample-rank and rank-feature information, which holds promise for better understanding the complex interplay of biological systems and diseases under the constraint of non-negativity, such as identifying gene expression patterns, aggregating genetic signatures, or elucidating the regulatory networks [211]. Variant forms of non-negative matrix factorization were also developed for different purposes. For example, approximate non-negative matrix factorization is a variation of NMF that allows for slight deviations from strict non-negativity to improve convergence and computational efficiency by relaxing the non-negativity constraints during intermediate steps of the algorithm that can speed up the factorization process and is useful in large-scale applications where exact non-negativity is less critical than overall performance and speed [212].

The second type of factor-based technique is factor analysis [213]. Different from matrix factorization, factor analysis (FA) was developed to understand the cause, or, in other words, to capture the factors embedding information from a set of variables in the original data. It reduced data dimensions by identifying factors that were informative enough to represent the

variables with high correlation. The factors themselves are not directly measurable, and they do not explain all the variance in the data. And the loading of the factors is the magnitude of the correlation between the factors and the covariates, which means that covariates with a higher loading value are more likely to be closer to the specific factor. Factor analysis is also known as common factor analysis (CFA), with two variant forms, including exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) [214]. Exploratory factor analysis can be regarded as an advanced version of principal component analysis, by identifying the factors that explained the correlations among observed variables without any predefined regularization or already-informed structure on the outcome, whereas confirmatory factor analysis (CFA) is a more sophisticated and hypothesis-driven approach compared to EFA. It is used to test the data fitness of a hypothesized measurement model, which is typically based on previous theoretical knowledge. In general, confirmatory factor analysis is required to specify the number of factors and the pattern of loadings in advance, and then use the data to test the fit of this model. This method involves the evaluation of goodness-of-fit indices to confirm or reject the hypothesized structure. Multi-Omics Factor Analysis (MOFA) employed a probabilistic model to integrate multiple omics datasets by inferring latent factors that explain shared and dataset-specific variations [194].

Another type of factor-based technique is integrative factorization methods, developed for data integration and cross-cohort analysis. Canonical decomposition approaches, such as canonical correlation analysis (CCA) and canonical variate analysis (CVA), were used to identify relationships between sets of variables in biomedical datasets that embedded associations between molecular features and clinical outcomes by inferring information from cross-covariance matrices and finding linear relationships between two sets of variables [215–217]. Joint and Individual Variation Explained (JIVE) decomposed multi-view data into shared and individual components using a combination of PCA and singular value decomposition, enabling the separation of global and dataset-specific signals [195]. The Joint Latent Variable Model (JLVM) incorporated latent variables to model relationships between multiple datasets, capturing both common and dataset-specific factors through a joint likelihood framework [218]. Multi-Set Canonical Correlation Analysis (mCCA) identified canonical variates by maximizing correlation across datasets and solving for canonical correlations between paired sets of variables [219]. Multiple Factor Analysis (MFA) provides a holistic representation of multiple datasets by analyzing common and specific patterns across them through factor scores and factor loadings [220]. Consensus Principal Component

Analysis (Consensus PCA) utilized PCA results from diverse datasets to identify shared principal components, aiming to find robust and reproducible features [221,222]. Integrative non-negative matrix factorization (iNMF) uncovers shared and unique patterns across datasets by further dividing the original decomposed matrices in NMF to cohort-specific and cohort-shared matrices, where the cohort-specific matrices identify factors (signatures) that are responsive to each input cohort data while the cohort-shared matrix captures the homogeneous factors across all the cohort [223]. Other multi-view integration models that recognized common factors from multiple sources simultaneously were implemented such as iCluster, moCluster and ImWalkMF [195,224–226]. At last, Chauvel et. al. (2018) presented a benchmark and overview of several representative multi-view clustering approaches, including various matrix factorization approaches [227].

Further development on factor-based tools and applications lies in various and different directions, in particular towards advanced factor-based architecture and deep-learning based factor techniques [228,229]. Autoencoder (AE) and variational autoencoder (VAE) are the typical factor-based models with the ability to discover nonlinear patterns in the data. AE and VAE compressed the raw input to the latent embedding, which represented the factors with low dimensions that uncover the original high-dimensional data matrix through the encoder and reverse the data compression process to reconstruct the output with the decoder. The encoder and decoder are composed of a linear model and nonlinear activation functions, which enable the AE and VAE to learn nonlinear patterns of complex data. Advanced architecture allowed additional applications of the models, such as discovering both linear and nonlinear patterns simultaneously, by transforming the original factor models into some complicated form depending on the context. However, the latent embedding of both AE and VAE cannot distinguish the variations of the variables of interest from other covariates. Therefore, one classic development to address this limitation was the optimized autoencoder architecture beta-VAE, which attempted to separate the latent representation from multiple independent factors of variation without providing prior knowledge to the model [230]. The optimized model Anneal-VAE based on beta-VAE to enable robust training and relieve the trade-off between disentanglement performance and reconstruction accuracy was made [231]. A further improvement was made in beta-TCVAE (Total Correlation Variational Autoencoder) to remove the hyperparameter selection procedure during training [232]. With a similar idea, a refined model was developed called factorVAE, which combined a linear factor-based model with the standard variational autoencoder to disentangle different independent sources of

variation by delineating the representation distribution in the latent space [233]. Further development integrates other model structure (e.g. Bayesian autoencoder, dictionary learning model, deep neural network), such as Bayes-Factor-VAE, JNNSE, Factorized variational autoencoders (FVAE) and deep neural network factor model (DNN-FM) [234–237].

Overall, the survey of factor-based techniques illuminated a trajectory from theories to practical applications in a wide range of fields. Throughout the years, a constant development can be found in factorization methods, evolving from basic matrix decomposition approaches to sophisticated algorithms for complex data. However, growing needs for continuous development of the factor-based method exist to handle tasks under different biomedical contexts. For example, integrative analysis based on factor-based tools for cross-cohort and cross-disease analysis such as comorbidity modelling is required. Another direction is to enhance interpretability and predictive power, and the development of more diverse algorithms that handle various biological data more effectively. There is also a growing interest in extending these methods to encompass security and privacy considerations crucial for clinical data handling across research institutions.

1.4.3 General application of factor-based methods

Within the biomedical field, factor-based approaches have been broadly developed and applied in numerous previous studies for biomarker discovery, stratification of disease phenotypes, and functional analysis, via multiple ways and implementation such as dimension reduction, signature identification and data integration. Devarajan's study outlined the distinctive capability of factorization methods for capturing the inherent part-based representation of complex biological data, offering a valuable alternative to classical methods like principal component analysis that do not impose nonnegativity constraints [238]. Following the relevant foundational work, the factor-based model was progressively adopted across various omics fields, from genomics to proteomics.

In the context of genome-wide association studies (GWAS) datasets, researchers developed techniques to identify interacting single nucleotide polymorphisms (SNPs) that affect phenotype status. The imputation of missing data was completed by an effective factorization method [239]. Case studies on understanding the relationship between risk variants and disease diagnosis phenotypes via matrix factorization were performed [240]. Tissue-specific

regulatory networks were informed by a newly developed factorization approach, namely sn-spMF [241]. Network type data for the identification of disease-associated loci for genotype was used in cNMTF [242]. Variant effect predictions were achieved and computed as variant prediction scores via sNMTF [243]. An exploratory factor analysis was performed to identify both shared and specific genetic factors across depression and anxiety [190]. Similarly, confirmatory factor analysis was applied to discover specific genetic variant factors underlying metabolic syndromes [191]. These methods and applications not only identify genetic associations with greater precision but also infer cellular and molecular processes that drive phenotypic variation.

In the fields of transcriptomics, epigenomics and proteomics, Baldrian and López-Mondéjar demonstrated the efficacy of these techniques in dissecting the microbial activities and their regulatory patterns discoveries from transcriptomic datasets [244]. Meta-analytical based factorization methods were implemented such as jNMFMA [245]. Co-regulating transcriptomic patterns were explored via a new boolean matrix factorization tool [246]. Spatial transcriptomics were accessed in several studies for understanding expression signatures combined with morphology information [247,248]. SVDMAN was implemented utilizing singular value decomposition on microarray data for global gene expression analysis [186]. Cellrank was developed based on real schur decomposition for predicting the cell fate mapping at the single-cell level [249]. Tensor factorization extended matrix factorization to higher-order biomedical data arrays, such as multi-dimensional gene expression datasets [250,251]. The further in-depth analysis of epigenetic modifications across the genetic material of a cell, benefited greatly from the application of matrix factorization methods, which included analysis on DNA methylation, histone modification, and gene silencing. Previous work demonstrated that factorization methods can effectively decompose epigenomic data into signatures representing distinct epigenetic patterns or states [252]. Factorization methods in epigenomics also facilitated the stratification of cancer subtypes based on epigenetic profiles and the identification of potential therapeutic targets [253]. Deconvolution on epigenetic data in breast cancer patients was conducted with a novel tensor factorization method HOCMO [254]. The field of proteomics was largely enriched by the application of factor-based techniques. Protein interaction and protein complexes were predicted by applying collective matrix factorization and binary matrix factorization, respectively [255,256]. Moreover, to accommodate the sparsity problem, sparse matrix factorization (SMF) enforces sparsity in the resulting factors and was successfully applied for

protein sequence motifs identification to improve grouping performance [257]. Overall, this methodology enables the comprehensive analysis of protein expression patterns, post-translational modifications and protein-protein interactions, subsequently enhancing the opportunity of identifying cellular mechanisms and potential biomarkers and therapeutic targets, thereby streamlining the pathway from experimental data towards clinical applications.

Despite these advancements in single modality, there is still an ongoing demand for coming up with more factor-based algorithms, such as algorithms for multi-modality data integration, which can manage the ever-increasing complexity and scale of data. Factor-based approaches used for the integration of multifaceted data scenarios involve processes dealing with heterogeneous data. One class of data integration techniques existed that addressed the complexity brought by multiple varied sources by establishing connections based on empirical information. Some methods sought to handle the heterogeneity by mapping and projecting different data types to the common space, such as molecular expression and copy number variation data (CNV) to pathways [258]. Others applied the linear model to fit and regress out the difference across modalities by imposing weights or performing feature engineering [259–261]. In contrast to supervised approaches, data integration methods could also be applied without prior knowledge and directly accomplish the disentanglement of separating the signal of interest that is homogeneous across cohorts as well as heterogeneous effects that are varied across data sources. Many existing integration techniques adopt a similar idea by extending the Dirichlet process mixture modeling and principal component analysis [262,263]. Canonical correlation analysis (CCA) was also frequently applied to integrative analysis across genetic and other modalities [193]. Joint and Individual Variation Explained (JIVE) was examined on The Cancer Genome Atlas (TCGA) with diverse cancer genetic datasets simultaneously [195]. In addition, joint latent variable model (JLVM) was implemented on breast and lung tumors for subtype identification [196]. Moreover, the well-established method LIGER was extended from integrative non-negative matrix factorization (iNMF) to disentangle dataset-shared and dataset-specific features of brain cell identity [200].

1.4.4 Factor-based approaches in psychotic disorders

Factor-based tools have been increasingly used in the study of various psychotic disorders. In the context of schizophrenia, matrix factorization was applied to reveal resting-state cortical dynamics in first episode psychosis, shedding light on the pathophysiology of functional connectivity [264]. Factorization-based correlation learning, specifically matrix decomposition and canonical correlation analysis, has been applied to multi-modal MRI data to explore the dysconnectivity of large-scale neurocognitive networks in psychiatric disorders, including schizophrenia [265]. The relationships between brain morphology and schizophrenia on the aspect of genetics were also examined using modular decomposition on gene similarity matrices, suggesting the abnormality of neurodevelopmental in schizophrenia [266]. Furthermore, the molecular signature of extracellular matrix pathology was investigated in schizophrenia patients with focal cerebral ischemia, revealing dysfunctions and negative regulation of the extracellular matrix in the perineuronal net structure [267]. To disentangle the specific variance of interest from the overall complex variance, factor analysis was conducted to reveal brain morphometric changes at a network level in individuals with schizophrenia [268]. Kim et al. (2019) conducted a factor analysis on individuals with recently gained schizophrenia and those with high risk of psychosis, revealing multiple significant factors that explain variances from social-cognitive bias, reflective self, neurocognition, and pre-reflective self factors, respectively [269]. Fountoulakis et al. (2019) utilized exploratory factor analysis to establish a 5-factor solution for staging patients with schizophrenia based on the PANSS [270]. Sowunmi et al. (2019) employed principal component factor analysis to identify a four-factor solution for the medication adherence rating scale (MARS) among Nigerian patients [271]. Furthermore, exploratory factor analysis was conducted to identify positive and negative syndrome scale (PANSS) factor structure from a collection of multi-ethnic schizophrenia individuals [272].

Overall, factor-based analytic approaches have proven to be effective tools in the study of psychotic disorders, offering insights into neural dynamics, structural brain networks, dysconnectivity, genetic relationships, and comorbidities associated with conditions such as schizophrenia and bipolar disorder. These methods hold promise for furthering our understanding of the underlying mechanisms of psychiatric disorders and may pave the way for more targeted and effective interventions in the future.

1.5 Overview of the chapters

In the thesis, I have focused on the application of factor-based models summarized in Chapter 1.4 and describe the development of our three different factor-based methods applied in schizophrenia and psychotic disorder studies, which aimed to refine the molecular and genetic profiling of schizophrenia mentioned in **Chapter 1.2** while focusing on the different aspects of the dilemma of biomedical data analysis discussed in **Chapter 1.3**. In **Chapter 2**, the development of a factorization computation framework for signature-based disease comorbidity modeling will be described. A hierarchical factorization method that is able to alleviate the bias induced by the selection of hyperparameters is described. An application for investigating molecular factors that contributed to schizophrenia-type 2 diabetes comorbidity was performed, and a comprehensive comorbidity map linking the relevant mechanisms was constructed. In **Chapter 3**, the deployment of a distributed factorization approach is presented, which projects the standard integrative factorization to a federated version based on the DataShield platform, providing possibilities for both cross-modality and cross-institutional learning. Performance evaluation was made and simulated as well as real data case studies were performed. In **Chapter 4**, an extensive application based on factorized interpretable VAE hybrid architecture is presented for predicting genotype-to-function association in schizophrenia individuals. Biological-informed semi-supervised model structure enabled the separation of variance of diagnosis from the complex effects by controlling confounding factors, as well as the association of genetic variants with biological processes. Besides, the model illustrated the feasibility of the novel application on genomic data. In the end, the prospects and outlook will be discussed in **Chapter 5**, for the further extension of the factor-based techniques in this field.

Chapter 2

Project 1: Comorbidity modeling with hierarchical multi-rank matrix factorization

2.1 Introduction

Disease comorbidity refers to the co-occurrence of multiple conditions of an individual, wherein the presence of one condition may influence the development, progression, or prognosis of another. Mental illnesses, specifically schizophrenia and bipolar disorder, gain increasing attention due to the multi-faceted clinical manifestations. This is induced by not only their often chronic and recurrent nature but also an elevated risk for somatic comorbidities such as type 2 diabetes and other metabolic syndromes [273,274]. This special indication is of high importance in clinical practice and biomedical research as it can significantly impact patient management strategies, treatment outcomes, and healthcare resource occupation. For instance, psychiatric patients with two or more somatic comorbidities are significantly associated with the poorer treatment outcome, while the number of comorbidities are also associated with increase in hospitalizations [275]. Worse quality of life as well as increase risk of social-demographic behaviors such as suicide, remission, psychotic relapse were received in patients with comorbidities [276]. Meanwhile, bilateral relationships between psychotic diseases and somatic illnesses exist that both conditions might influence each other [273]. This complexity eventually led to the increasing risk of somatic disease with twice the incidence rate for individuals with severe psychotic disorders, compared to those without [273].

The overlap of molecular genetics across psychotic disorders and somatic comorbidities has been provided as evidence in many previous studies from the bottom to the top covering genotypes, transcriptomics, epi-genetics and proteomics. Several genotypic studies have illuminated a convergence of molecular profiles underlying somatic comorbidities and psychotic disorders [277]. It is observed that molecular linkages between and across schizophrenia, immune-dysregulated diseases, type 2 diabetes and cardiovascular disorders

have been identified. Specifically, risk variants associated with genes in immune systems were found in schizophrenia patients, strengthening the shared genetic evidence across psychotic disorders and immune-dysfunctional conditions. On the aspects of transcriptomic level, molecular markers as well as biological processes that were commonly dysregulated across psychotic disorders and their comorbidities in transcriptomic studies and integrated analysis respectively [278,279]. Epigenetic regulation uncovered genetic overlap and causal relationships were inferred, providing insights of the genetic variants colocalization on both schizophrenia and type 2 diabetes [280]. Common molecular pathway alterations were found in multiple functions including synaptic regulation, vesicle transport, mitochondrial and ribosomal proteins in schizophrenia and bipolar disorder patients as well as animal models under deep proteomics aspects [281]. Therefore, it is necessary to understand comorbidities, including shared risk factors, pathophysiological mechanisms, and treatment interactions, as well as to elucidate the relationships between coexisting conditions to advance preventive, diagnostic, and therapeutic approaches aimed at mitigating the burden of multimorbidity and optimizing patient care across diverse clinical contexts.

To investigate the underlying molecular genetics across psychotic disorders and somatic comorbidities, studies have been conducted using a wide range of techniques. One example was to conduct GWAS analysis to find the overlap modules. For instance, large-scale GWAS of schizophrenia and type 2 diabetes traits were applied to distinguish the common variants association [277]. Differential expressed genes (DEGs) that expressed in both conditions and the corresponding enriched processes were identified to describe the potential comorbid mechanism [9,10]. Analyzing from the significant biological processes, susceptibility genes were curated to understand the comorbid effect [282]. A Mendelian randomization analysis was implemented to investigate the exposure and outcome across two diseases and further inferred the causality using an inverse-variance weighted meta-analysis [283]. Hybrid model combining single-cell, bulk data with molecular quantitative trait loci summary statistics (molQTL) combining different levels, including eQTL, pQTL, caQTL, mQTL and sQTL was conducted on adult and fetal brain from schizophrenia and type 2 diabetes individuals as well as on other tissues such as islets, liver and adipose tissue from the patients [280]. Findings via genome-wide multi-trait colocalization suggested a possible connection between schizophrenia and metabolic symptoms, where linkage disequilibrium score regression (LDSC), genetic covariance analyzer (GNOVA), and heritability estimation (ρ-HESS) were applied in the study [284]. Other methods, such as polygenic risk score calculation, were

utilized to aggregate the SNPs into genes to better interpret the shared mechanisms and pathways [124]. Common molecular markers were discovered across cardiovascular disease and schizophrenia via genetic-pleiotropy-informed approaches to investigate the causal effect of risk factors in comorbidities on target conditions [285]. Risk evaluation was performed by computing the gene correlations with bi-directional Mendelian randomization (MR) model to understand the potential risk factors [274]. The deep proteomics technique was used to recognize shared molecular pathways in schizophrenia and bipolar disorder patients and animal models [281]. In addition, latent factors such as molecular patterns and signatures were identified to explore the molecular profiles in a lower dimensional view with independent component analysis (ICA) [286].

Nevertheless, few abovementioned modeling focused on the mechanism from aspects of molecular signatures. Therefore, in this Chapter, we attempted to approach this gap through systematic comorbidity modelling in schizophrenia and psychotic disorders on the signature levels. I developed an improved integrative unsupervised machine learning framework with multi-rank non-negative matrix factorization (mrNMF) applied on large-scale transcriptomic datasets. Unlike classical NMF approaches which delineate and aggregate the input transcriptomic datasets into a specific and pre-defined number of signatures, the proposed method alleviates the limitations without choosing the optimal number of signatures when initializing the model. It takes multiple ranks into consideration in the beginning. Hence, molecular signatures are defined at varying levels of granularity, thereby offering insights into potential shared comorbid mechanisms covering different hierarchical levels, from the top and wide signatures associated to broad molecular differences with more general biological concepts down to specific and detailed signatures representing less variance. To this end, a collection of 27 case-control microarray cohorts that covered diverse tissue sources was retrieved, including three main categories of major conditions: psychotic disorders (PSY), type 2 diabetes (T2D) and cardiovascular diseases (CVD). Then reciprocal best hit score matrices were computed to integratively connect signatures across different cohorts with specific conditions, and eventually a comorbid disease signature graph was built. Downstream interpretation studies and gene biomarker analysis on the level of comorbid signature pairs identified key genes and biological processes contributing to the comorbid effect across schizophrenia and co-occurred syndromes. Furthermore, genetic variants were retrieved from online resources to compare and validate the findings. In the end, a cross-mechanism map for linking different shared processes based on an external knowledge graph database was

established. It associates different key mechanisms along with molecular factors and provides additional insight for the comprehensive overview of the comorbid mechanism.

2.2 Method

2.2.1 Data and materials

27 transcriptomic cohorts (n_sample=1163, n_case=633, n_control=530) with case and control were collected from Gene Expression Omnibus (GEO). Conditions of the collected cohorts were categorized into three classes (in upper case letters), psychotic disorders (PSY), cardiovascular diseases (CVD), and type 2 diabetes (T2D). Cohorts from each classes were further annotated with their conditions (in lowercase letters) in details, including schizophrenia (scz), bipolar disorders (bd), and major depression (mdd) in PSY class; acute myocardial infarction (ami), coronary artery disease (cad), peripheral arterial disease (pad), acute coronary syndrome (acs), hypertension (hyp), and cardiomyopathy (cdm) in CVD class; and type 2 diabetes (t2d) in T2D class. The tissue source of each cohort was annotated as well, including blood, brain, liver, islet, etc. Other covariates including gender, age, BMI, etc, were also prepared for controlling confounding effects in the further pre-processing. All cohorts were built on the same platform GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array to control the biases induced by different platform batches. All cohort information above has been summarized in **Table 2.1**.

Table 2.1 Descriptive summary of cohorts used in the study

id	cohort	number of sample	number of control	number of case	condition	cohort	tissue
1	GSE21138	59	29	30	SCZ	PSY	Brain
2	GSE17612	49	22	27	SCZ	PSY	Brain
3	GSE27383	51	29	22	SCZ	PSY	Blood
4	GSE53987_2	70	18	52	BD	PSY	Brain
5	GSE53987_3	69	19	50	MDD	PSY	Brain

6	GSE74358	28	14	14	BD	PSY	Brain
7	GSE46449	88	39	49	BD	PSY	Blood
8	GSE73129	115	60	55	SCZ	PSY	OE
9	GSE7036	6	3	3	BD	PSY	Brain
10	GSE21935	42	19	23	SCZ	PSY	Brain
11	GSE76894	97	78	19	T2D	T2D	Islet
12	GSE76895	57	27	30	T2D	T2D	Islet
13	GSE23343	17	7	10	T2D	T2D	Liver
14	GSE15932	16	8	8	T2D	T2D	Blood
15	GSE71416	20	6	14	T2D	T2D	Adipose
16	GSE38396	8	4	4	T2D	T2D	Skin
17	GSE24752	6	3	3	HYP	CVD	Blood
18	GSE19339	8	4	4	ACS	CVD	Vessels
19	GSE48060	52	21	31	AMI	CVD	Blood
20	GSE66360	99	50	49	AMI	CVD	Vessels
21	GSE13985	10	5	5	ATH	CVD	Blood
22	GSE97320	6	3	3	AMI	CVD	Blood
23	GSE71226	6	3	3	CAD	CVD	Blood
24	GSE19303	48	8	40	CDM	CVD	Heart
25	GSE53987_1	66	18	48	SCZ	PSY	Brain
26	GSE161355	33	15	18	T2D	T2D	Brain
27	GSE27034	37	18	19	PAD	CVD	Blood

In detail, we listed the specific number and proportion of the gender, tissue and age by each condition in **Table 2.2**.

Table 2.2 Summary of demographic and clinical information

label	levels	bipolar disorder	control	cvd	major depressive disorder	schizophrenia	t2d
gender	female	23 (22.8)	136 (34.6)	13 (27.1)	22 (44.0)	63 (30.9)	33 (38.8)
	male	78 (77.2)	257 (65.4)	35 (72.9)	28 (56.0)	141 (69.1)	52 (61.2)
tissue	associative striatum	17 (16.8)	18 (5.4)		16 (32.0)	18 (8.8)	
	ba46	17 (16.8)	29 (8.7)		17 (34.0)	45 (22.0)	
	hippocampus	18 (17.8)	18 (5.4)		17 (34.0)	15 (7.3)	
	leukocytes from whole blood	49 (48.5)	39 (11.7)				
	ba10		22 (6.6)			27 (13.2)	
	ba22		19 (5.7)			23 (11.2)	
	blood		30 (9.0)	40 (58.8)			
	human temporal cortex		15 (4.5)				18 (64.3)
	liver		7 (2.1)				10 (35.7)
	lymphoblast from blood		41 (12.3)			36 (17.6)	
	oe		19 (5.7)			19 (9.3)	
	pbmc		47 (14.2)	19 (27.9)		22 (10.7)	
	peripheral blood		9 (2.7)	9 (13.2)			
	pre-frontal cortex		19 (5.7)				
age	Mean (SD)	43.4 (12.2)	50.9 (19.1)	49.2 (10.2)	45.6 (10.3)	50.7 (19.4)	68.3 (10.9)

2.2.2 Data preprocessing

The standard microarray preprocessing pipeline includes .CEL raw data retrieval, frozen robust multiarray analysis (fRMA) [287,288], followed by robust multichip average background correction, quantile normalization and robust weighted average summarization. Samples with missing values in expression matrix and metatable were removed. The mapping of probes to HGNC gene symbols with biomaRt [289] package was performed and rowmean normalization by aligning the mean value of each gene to 1 for each expression matrix was conducted in order to make signatures from different cohorts comparable.

2.2.3 Multi-rank non-negative matrix factorization (mrNMF)

Multi-rank non-negative matrix factorization (mrNMF) framework consists of four main steps. First, non-negative matrix factorization (NMF) was conducted on the gene expression matrix where genes are rows and samples are columns of every single cohorts, resulting in two decomposed matrices, W matrix (rows are genes while columns are ranks or signatures) and H matrix (columns are samples while rows are ranks or signatures). Here, the rank is defined as the "signature", where the signature is a weighted genes vector, or so-called meta-gene, whereas weights are the exposure values, referring to the positive contributions of each of the genes to this specific signature. The number of signatures is a hyperparameter that originally required users to manually select in the standard NMF algorithm, however, in mrNMF, an ensemble strategy was applied by directly using a continuous series of rank numbers from 2 to 20 during decomposition procedure iteratively. The decomposed matrices, including W matrices and H matrices were then concatenated into one large W matrix (Wc) and one large H matrix (Hc) along their signature dimension respectively. The loss function for every iteration of training with one specific rank number is shown below.

$$min_{WH}||X - W_{(k)}H_{(k)}||_F^2$$

$$s.t. W \geq 0, H \geq 0, k \geq 2$$

Second, a linear model was performed to identify significantly diagnosis-associated signatures (P-value < 0.05) while removing those that are insignificant. In the linear model, values of each signature in H matrices are used as independent variables x, whereas

diagnosis/group variables are dependent variables y and gender, age, and other available variables are selected as covariates. The corresponding insignificant signatures in Wc and Hc matrix were removed by this step.

Third, a similarity scoring matrix is computed to identify signature pairs with high correlation. Here, the reciprocal best hit metric (RBH) is applied. I defined a RBH signature pair as a pair of signatures, each from a different cohort, matched with each other with the highest score compared to the value of all the other paired signatures [290,291]. In each signature, only a few genes with high exposure would contribute to the signature, whereas most other genes are more likely to represent the random noise (**Figure 2.1**). In general, non-negative matrix factorization tends to emphasize the top exposed features (or genes) while putting less weights on the irrelevant features, which can be regarded as a denoising strategy. I computed the Jaccard similarity coefficients using the top N (N=1000) genes with top exposure values (**Figure 2.1**). The best match of signature pairs from a pair of different cohorts would be evaluated by the Jaccard similarity coefficients.

$$J(sig_i, sig_j) = \frac{|sig_i \cap sig_j|}{|sig_i \cup sig_j|}$$

With this scoring criterion, the output differences of using different values of N were also evaluated for control (**Appendix Supplementary Figure 2.1**).

Figure 2.1 Illustration of the matrix factorization, decomposed components and signatures, and Jaccard similarity

Fourth, permutation testing was performed to filter the signature pairs that occurred more frequently than the null condition, which is under randomized control. In this step, we

randomly shuffled the exposure values or so-called weights in the W matrices acquired during the multi-rank decomposition, and performed the following computing step such as reciprocal best hit matrices computing and graph construction. This permutation process was run 20 repetitions iteratively, and the permutation test was performed and only edges with a significance level of $P \leq 0.05$ were retained. Eventually, the final comorbid signature graph would be generated.

2.2.4 Biological inference on comorbid signature pairs

Gene set enrichment analysis was performed on each signature pair (edges) in the signature graph with the shared gene list from the signature pairs identified from the previous step as analysis input, using the ClusterProfiler package [292]. Gene Ontology resource (GO) from MsigDB was selected as gene set [293,294]. Significant GO terms were extracted with FDR adjusted P-value < 0.05.

2.2.5 Gene exposure analysis on comorbid signature pairs

Genes contributing to the enrichment of the four specific processes (*acute inflammation*, *angiogenesis*, *oxidative stress,* and *GABAergic synapse*) were extracted. For each of the enriched genes from each signature pair, two exposure values were acquired, including the value from the schizophrenia signature and the other from the comorbidity disease signature. Genes were then ranked based on the top 30 exposure values based on the geometric mean of the exposure values from the two individual signatures from the signature pair. Subsequently, these rankings were shown in scatter plots, with exposure values for schizophrenia and the comorbid disease represented on the x and y axis, respectively. Additionally, the gene types retrieved from the previous analysis were annotated on the plot as *comorbid, shared, schizophrenia specific,* and *somatic specific*.

2.2.6 Validation of the comorbid genes with the GWAS risk variants

Focusing on acute inflammation, angiogenesis, oxidative processes, and the GABAergic system, the genes found enrichment in these biological processes were further mapped and validated by comparing with genome-wide association study (GWAS) risk variants linked to

schizophrenia and type 2 diabetes traits. A total of 4988 associations from 142 studies with schizophrenia trait and 5263 associations from 218 studies with type 2 diabetes were retrieved from the GWAS catalog database, each containing corresponding mapped gene information. The common genetic variants shared across schizophrenia and type 2 diabetes in GWAS analyses were considered as comorbid GWAS variant genes. Subsequently, comorbid genes identified within the scz-t2d comorbid signature pairs were then categorized into three classes including *gwas_t2d*, *gwas_scz*, and *gwas_shared*.

2.2.7 Building cross-disease comorbidity connectivity knowledge graph

A literature-curated knowledge database was applied extensively to establish connectivity map from schizophrenia to type 2 diabetes, thereby validating the specific processes identified in the preceding analyses within the study, including acute inflammation, angiogenesis, oxidative processes, and the GABAergic system. Similar to the previous step, genes on the top 30 enriched gene list (see **Method 2.2.5**) associated with these processes were extracted. Using these biological processes and the corresponding genes as query input, a literature-curated knowledge graph database was used [295]. By selecting the hyperparameter set including maximal connectivity depth = 3 which referred to connected node degree, the initial large-scale connectivity knowledge graph was established. In order to better visualize the specific processes and genes, the shortest paths between schizophrenia node and type 2 diabetes node were computed by starting with the biological process nodes and performing the random walk algorithm. In the end, the schizophrenia - type 2 diabetes specific disease comorbidity connectivity graph was built.

2.3 Results

2.3.1 Generation of disease signature graph via factorization framework

We developed a signature level comorbidity modeling framework to understand the shared biological factors across multiple conditions via an improved integrative unsupervised learning technique based on multi-rank non-negative matrix factorization (mrNMF) algorithm. The novel approach is optimal, compared to the original standard NMF, mainly

due to the improvement on the algorithm that a series of factorization on the same data matrix is performed, with hyperparameters rank k from kmin to kmax. Generally, a higher rank k tends to identify signatures covering boarder biological meaning whereas a lower rank identifies signatures with more specific biological modules. This is similar to the hierarchical structure of the gene ontology tree, where on the top the GO terms represent more general processes while on the bottom represent more detailed processes (**Figure 2.2**). Thus, mrNMF can be considered an ensemble methodology that covers biological signatures from different granularities.

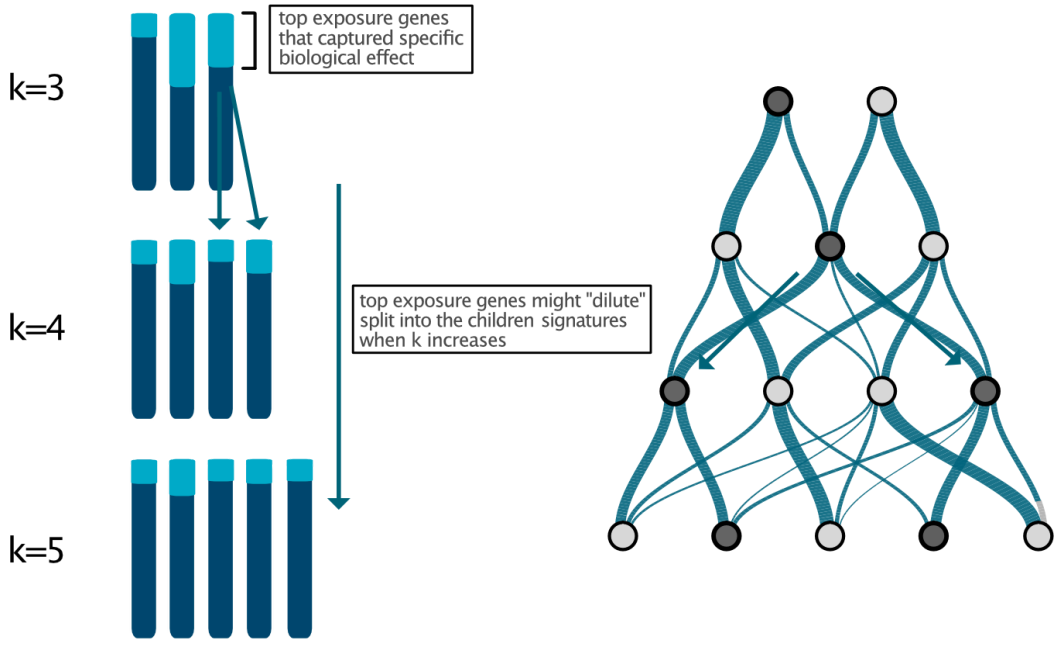


Figure 2.2 Illustration of the multi-rank factorization to identify factors from hierarchical levels

This study involved a comprehensive collection of 27 transcriptomic cohorts with 1163 individuals, within which 633 patients and 530 healthy individuals were included. Multiple conditions are included such as various psychotic disorders (e.g. schizophrenia, etc.), type 2 diabetes and cardiovascular disease (e.g. coronary artery disease, etc.) (**Table 2.1 and Table 2.2**). In **Figure 2.3**, the full procedure of mrNMF computational framework is illustrated (see 2.2 Methods). To start, each preprocessed gene expression matrix was factorized into rank-sample matrix *H* where the values in each row of the H were used to identify significant diagnosis-associated signatures (See **Method 2.2.3**). Next, a similarity matrix with reciprocal best-hit scoring criterion was calculated to extract signature pairs that capture common

biological factors, and permutation was conducted to generate the final signature graph. Then, comorbidity analysis was implemented, with our hypothesis that common molecular factors associated to shared biological processes can be found in both diseases. This would be revealed when two signatures from datasets of different diseases are found to be similar using a scoring metric. To compute the similarity of two signatures, shared gene lists were identified in the top 1000 genes of each signature. We verified that the selection of the number of top genes does not bias further analysis in this study (**Appendix Supplementary Figure 2.1**). Furthermore, enrichment analysis was performed to understand the actual processes that contributed to the comorbid effect. In the end, external sources such as GWAS risk variants and curated literature knowledge were adopted to confirm our findings of comorbidity across schizophrenia and its comorbidities.

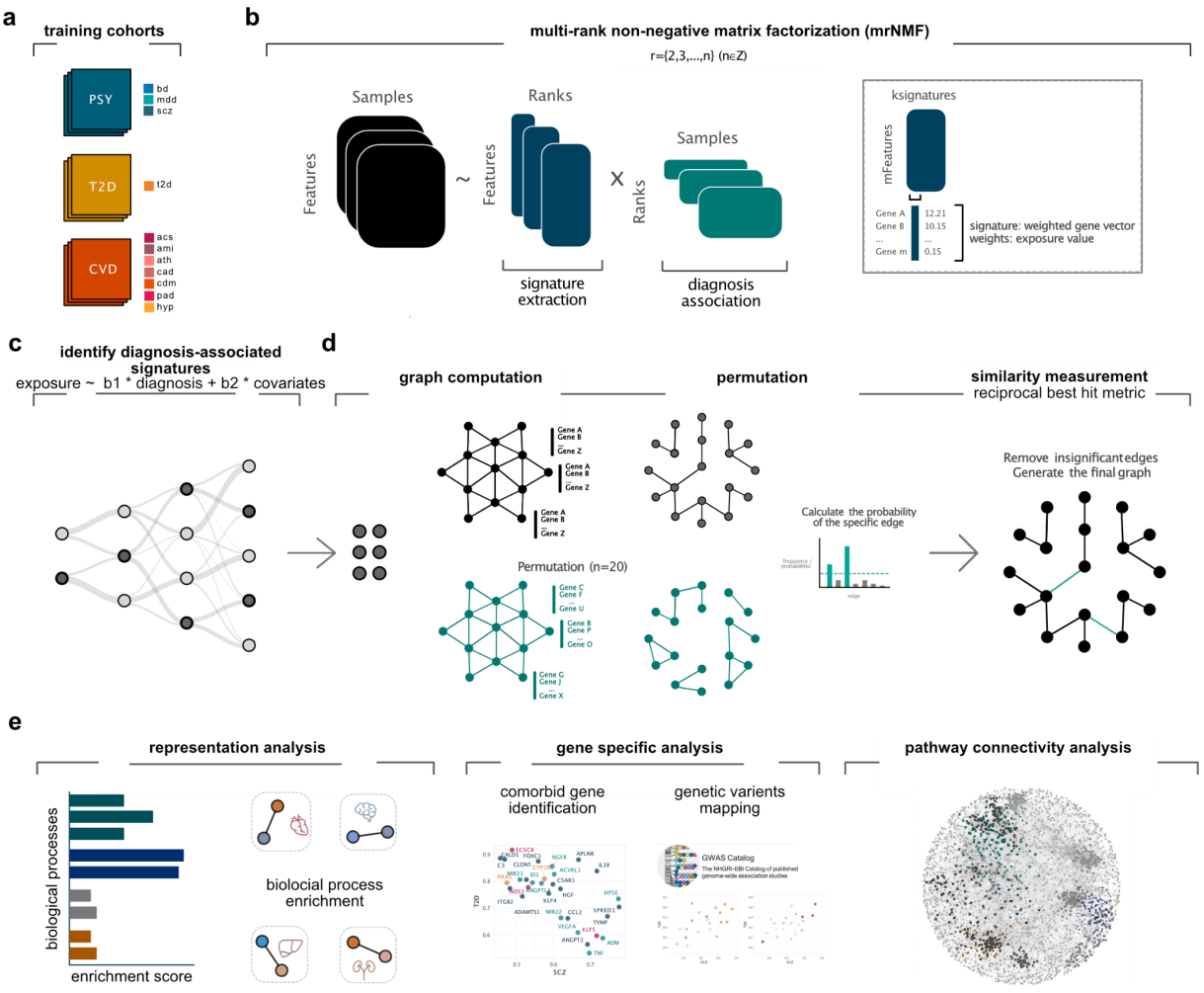


Figure 2.3 Workflow of the signature-level comorbidity modelling framework based on multi-rank non-negative matrix factorization (mrNMF). **a** Retrieval of transcriptomic dataset covering various diseases and tissue sources, specific conditions including scz (schizophrenia), bd (bipolar disorders), mdd (major depressive disorders), t2d (type 2 diabetes), acs (acute coronary syndrome), ami (acute myocardial infarction), ath (Atherosclerosis), cad (coronary artery disease), cdm (dilated cardiomyopathy), hyp (hypertension), pad (peripheral artery disease). **b** Algorithm of multi-rank non-negative matrix factorization. **c** Filtering of diagnosis-associated signatures with linear model. **d** Graph computation based on reciprocal best hit scoring matrices. **e** Comorbidity modelling including over representation analysis, gene analysis and cross-disease connectivity knowledge graph building on signature pairs level.

2.3.2 Comorbid signature pairs in the signature graph

305 signature pairs, represented as edges and 419 signatures, represented as nodes were generated and composed the signature graph (**Figure 2.4**). The signatures represent multiple disease classes (including PSY, T2D, and CVD), tissues (e.g. brain, blood, etc.), and encompass 58 types of edges and signature pairs (e.g., scz-scz, scz-t2d, bd-scz, etc.).

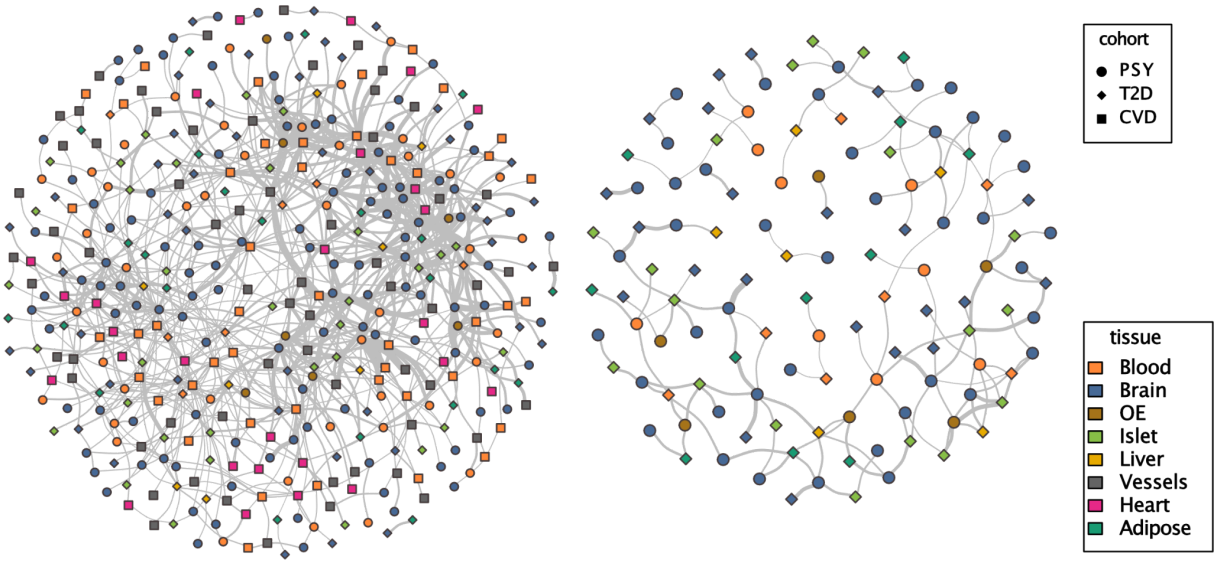


Figure 2.4 Computed signature graph. **right** Signature graph (full) within which each node referred to an individual signature and each edge indicated a signature pair with reciprocal best hit (RBH) matrices matched. The shape of nodes represented the disease class, whereas the color of nodes showed the tissue source of a signature extracted from. **left** Signature subgraph (schizophrenia - type 2 diabetes).

The descriptive summary was shown in **Figure 2.5**, where we categorized all signature pairs into four groups: single (involving signatures of identical disease, e.g., scz-scz), intra-class comorbid (involving signatures from the same disease class but from different specific conditions, e.g., scz-bd), inter-class comorbid (involving signatures from different classes but with one of the signature representing psychotic disorders, e.g., scz-t2d), and somatic (involving signatures from non-psychotic diseases but different classes). Notably, among the inter-class comorbid category, those linking psychotic disorders and somatic diseases, particularly schizophrenia - type 2 diabetes (scz-t2d) pairs, were most prevalent (**Figure 2.5**). Therefore, a scz-t2d subgraph containing only schizophrenia and type 2 diabetes signatures was specially isolated (**Figure 2.4**).

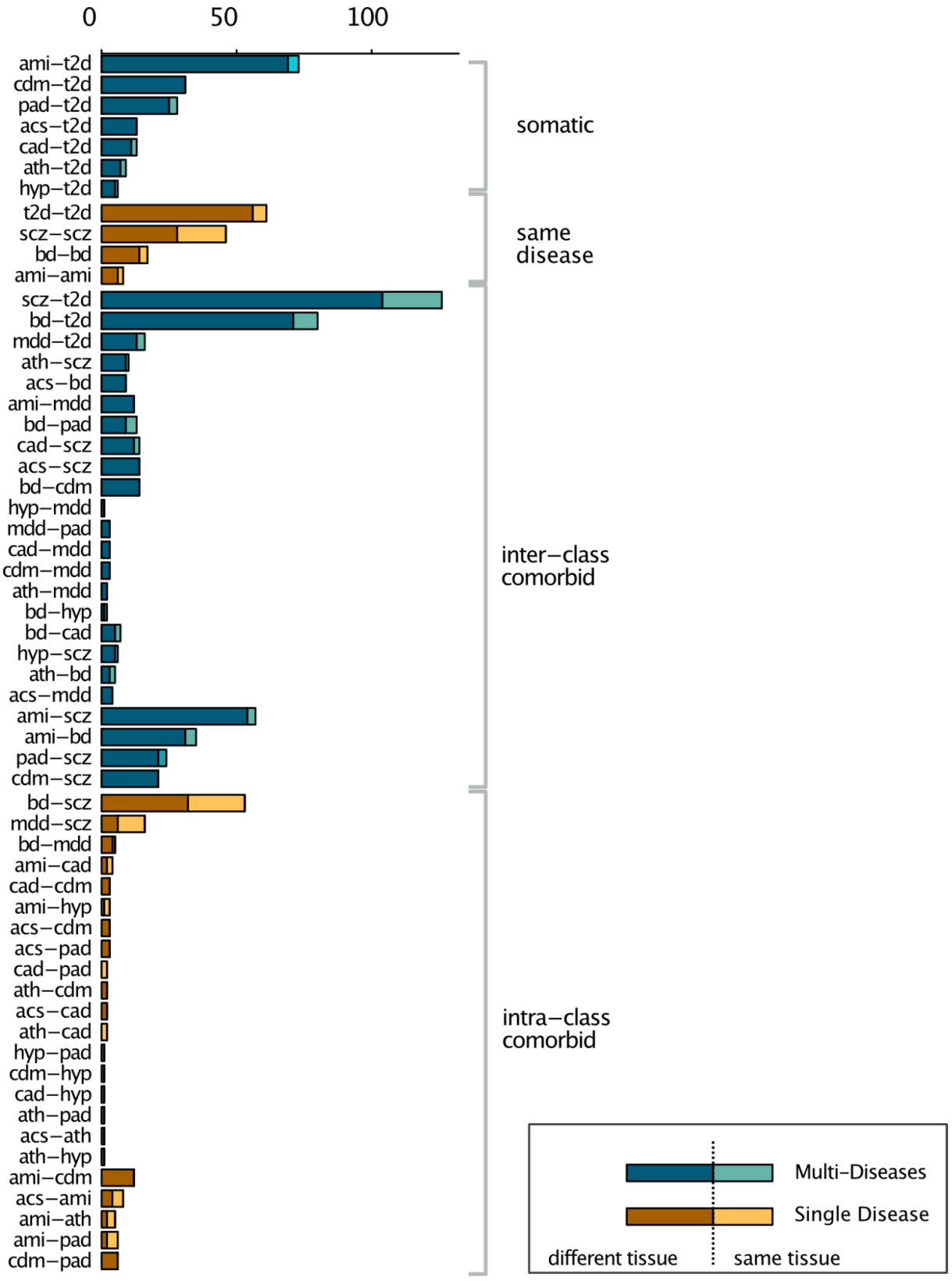

0
50
100
ami−t2d
cdm−t2d
pad−t2d
acs−t2d
cad−t2d
ath−t2d
hyp−t2d
t2d−t2d
scz−scz
bd−bd
ami−ami
scz−t2d
bd−t2d
mdd−t2d
ath−scz
acs−bd
ami−mdd
bd−pad
cad−scz
acs−scz
bd−cdm
hyp−mdd
mdd−pad
cad−mdd
cdm−mdd
ath−mdd
bd−hyp
bd−cad
hyp−scz
ath−bd
acs−mdd
ami−scz
ami−bd
pad−scz
cdm−scz
bd−scz
mdd−scz
bd−mdd
ami−cad
cad−cdm
ami−hyp
acs−cdm
acs−pad
cad−pad
ath−cdm
acs−cad
ath−cad
hyp−pad
cdm−hyp
cad−hyp
ath−pad
acs−ath
ath−hyp
ami−cdm
acs−ami
ami−ath
ami−pad
cdm−pad
somatic
same disease
inter−class comorbid
intra−class comorbid
Multi−Diseases
Single Disease
different tissue
same tissue

Figure 2.5 Statistics of the computed signature graph on signature pair level (edges). X-axis represented the signature pair categories (e.g. scz-t2d, scz-bd, etc.) whereas the y-axis referred to the count of the signature pairs in the corresponding specific conditions. The single disease pair category contained signature pairs within which the individual signature came from the same disease cohort such as either PSY, CVD or T2D.The comorbid disease pair category is composed of signature pairs from different cohorts such as PSY-CVD and PSY-T2D. The somatic disease category refers to signature pairs that represent CVD and T2D.

Nevertheless, cross-tissue cohorts were utilized in our study which raised our concern of a possible tissue confounding effect, considering that signatures from the same tissues tend to be associated. However, we observed in our case that the mrNMF framework connected signatures from different diseases and tissues combinations (**Figure 2.4**). To assess the tissue effect, we investigated the count of cross-tissue signature pairs. Predominantly, signature pairs linking signatures from different diseases at the same different tissues were observed, suggesting that confounding effect brought by tissue source was not the primary factor that drove the clustering of the signatures (**Figure 2.4 and Figure 2.5**). Additionally, we tracked the source of individual signatures in the full graph identified from different ranks in the factorization model in **Figure 2.6**. Moreover, within the category of "multiple disease multiple tissue", the top connections with most counts involved signature pairs from the tissue source combination of brain-blood, brain-islets, and brain-vessels. Specifically, for schizophrenia - type 2 diabetes (scz-t2d) signature pairs, the top 3 major sources were brain-islets, brain-adipose, and brain-liver respectively, indicating no substantial bias leading to the linkage of signatures from the same tissue (**Appendix Supplementary Figure 2.2**). In summary, the generated signature graph revealed cross-disease and cross-tissues signature pairs, with the majority of computed signature pairs identified from schizophrenia and type 2 diabetes cohorts.

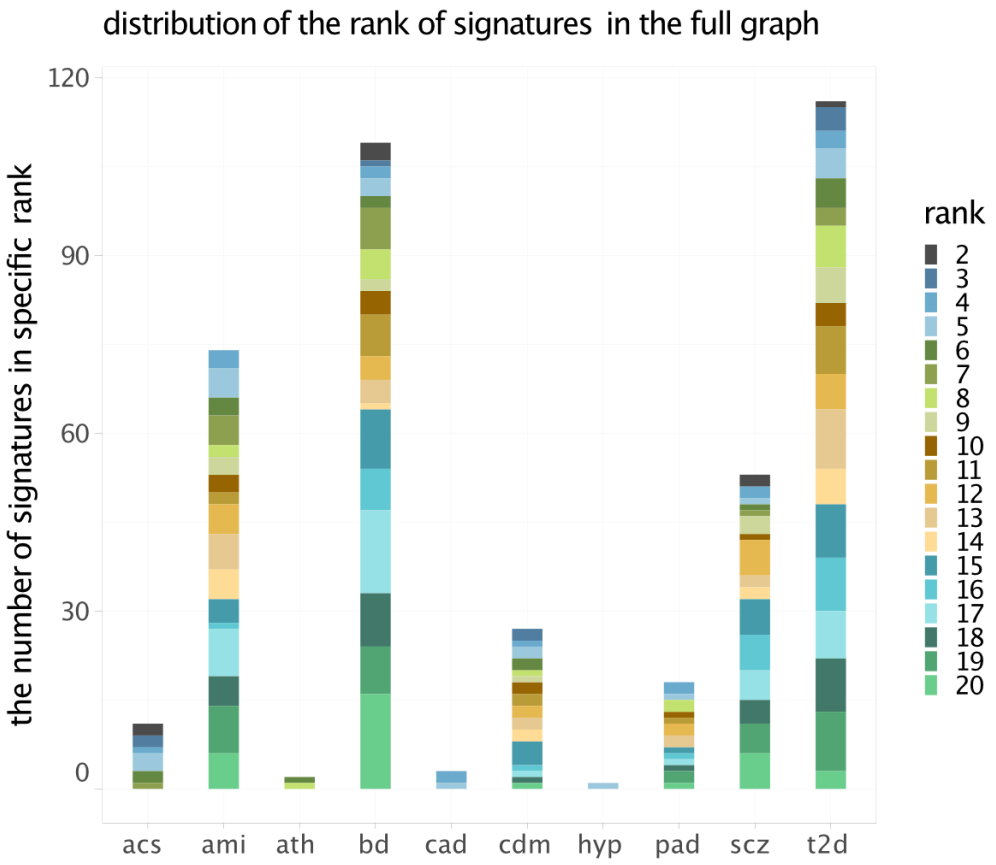


Figure 2.6 Tracking the source of signatures decomposed with different rank levels during the factorization procedure.

2.3.3 Investigating gene composition variety in comorbid signature pairs

To better understand the difference across various signature pairs, linking schizophrenia signatures to those signatures from somatic comorbidity diseases (referred to as "comorbid pairs"), a detailed comparison was made between the shared genes associated with the comorbid signature pairs (e.g. scz-t2d, scz-ami, etc.) and the genes identified in schizophrenia pairs (e.g. scz-scz) and somatic pairs (e.g. t2d-t2d, ami-ami, etc.), respectively (**Figure 2.7**). The rationale was to verify if these comorbid pairs capture genes that are not found in the single disease signature pairs. The gene sets were categorized as follows: genes specific to schizophrenia i.e. found solely in the schizophrenia signature pair (e.g., scz-scz), those specific to somatic disorders (e.g., t2d-t2d or ami-ami), genes shared between schizophrenia and somatic diseases (found in both scz-scz and t2d-t2d or scz-scz and ami-ami), and genes exclusive to the comorbid signature pairs, not present in any one of the diseases involved in

the signature pair (**Figure 2.7**). Focusing on comorbid signature pairs of schizophrenia and type 2 diabetes, as well as schizophrenia and acute myocardial infarction, 39% in scz-t2d and 48% in scz-ami the comorbid genes were not observed in the single-disease categories, implying that underlying comorbid mechanism might be due to specific molecular processes. These genes had not been identified in either schizophrenia or type 2 diabetes and acute myocardial infarction signature pairs (scz-scz or t2d-t2d and scz-scz or ami-ami), suggesting their novelty as additional comorbid gene markers (**Figure. 2.7**).

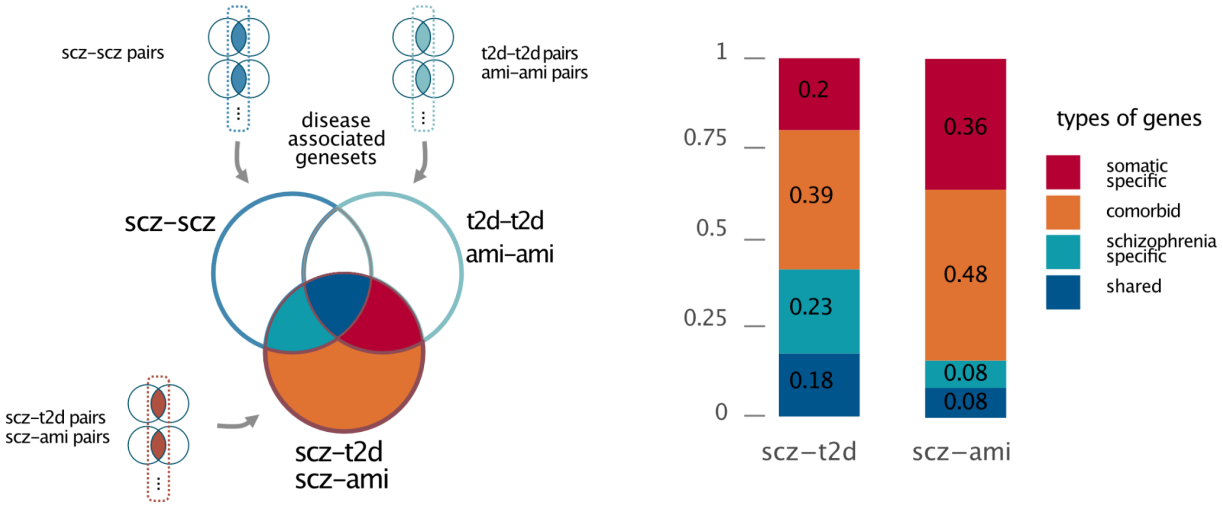


Figure 2.7 Diverse variety of comorbid signature pairs and the corresponding gene sets. **left** Illustrated scheme of categorizing the type of genes identified from signature pairs. **right** Types of genes in comorbid signature pairs including genes annotated as "schizophrenia specific" that were found in the schizophrenia signature pairs (e.g. scz-scz signature pairs), genes as "somatic specific" in scz-t2d signature pairs that were also found in the pure somatic comorbidity signature pairs (e.g. t2d-t2d signature pairs), genes as "shared" that were observed in both the scz-scz and t2d-t2d/ami-ami signature pair as well and the distinct genes as "comorbid" that were not identified in any previous mentioned type of signature pairs.

2.3.4 Elucidating the central roles of inflammatory response in scz-t2d comorbid signature pairs

To investigate the comorbid mechanism further, a major focus was put on the schizophrenia and type 2 diabetes signature pairs, given our previous finding that these two conditions

exhibited the most connections in the signature graph (**Figure 2.3**). To probe the underlying mechanisms associated with this comorbidity, functional enrichment analysis was performed on the scz-t2d signature pairs with the overlapping gene lists as model input (see **Method 2.2.4**). In total, 590 different GO terms were enriched and highlighted, and it was also observed that many terms were identified to show frequent enrichment across multiple signature pairs, indicating the robustness of the results. I performed clustering on these terms for the following analyses. Among all the clusters, I prioritized several of the largest clusters, covering a broad range of immune systems involved in the production, activation, migration, and regulation of cells contributing to the inflammatory response, such as cytokines and leukocytes (**Figure 2.8**).

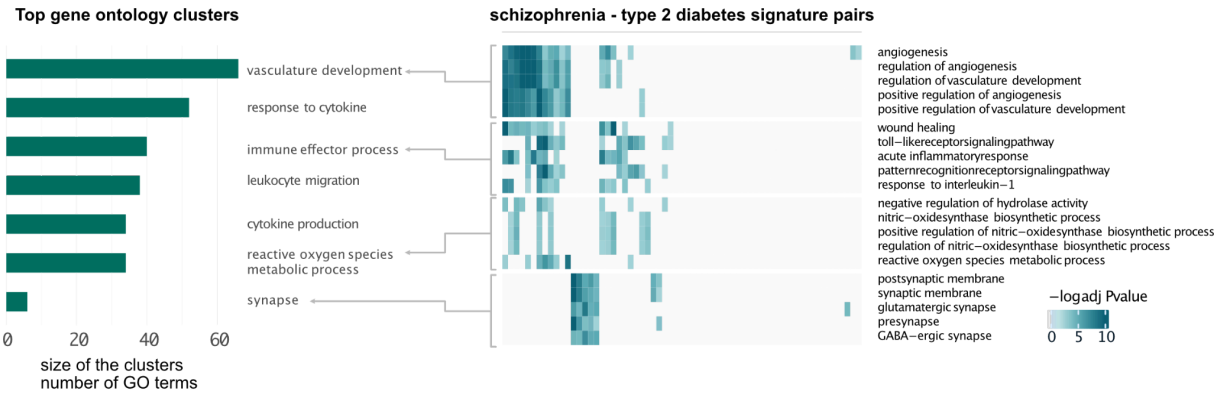


Figure 2.8 Biological interpretation of signature pairs with the corresponding embed gene set. **left** Largest clusters with the most enriched gene ontology term members observed in schizophrenia - type 2 diabetes signature pairs. **right** Specific and representative processes from the three largest clusters on the left subfigure vasculature development, immune effector process and reactive oxygen species and one minor cluster synapse, that showed most frequent enrichment based on the number of scz-t2d signature pairs. Each column along the x-axis represented one scz-t2d signature pair whereas the y-axis indicated the detailed GO terms in the corresponding clusters linked with the arrow.

This observation might suggest the pivotal role played by immunological processes and inflammatory responses, contributing to the comorbidity between schizophrenia and type 2 diabetes. This finding aligns with many previous studies elucidating potential molecular mechanisms that explain schizophrenia comorbidity with diabetes [118,284], thereby confirming the validity of our approach. Meanwhile, some clusters that were less expected

were found in the study, such as vasculature development. Notably, angiogenesis was identified frequently in 19 scz-t2d signature pairs' enrichment. Previous research indicated that dysregulation of the microvascular environment induced angiogenic molecules as well as neuro-inflammatory response in schizophrenia patients [296,297]. Similarly, abnormalities of vascular implicated in the symptoms of type 2 diabetes, including diabetic nerve damage, entail alterations in the microvascular environment [298,299]. Furthermore, biological processes related to reactive oxygen species metabolism and oxidase activities were identified that are interdependent with inflammatory processes. The altered function of the oxidative process was implicated in the dysregulation of lipid transportation, exacerbating diabetes [297,300]. In addition, other processes of interest were revealed in smaller clusters. Dopamine systems were identified in 8 scz-t2d pairs that involved 4 different schizophrenia and 3 diabetes datasets (**Figure 2.12**). Another cluster associated with synapse processes, including the GABAergic system function, exhibited enrichment in five signature pairs. These findings also documented that the GABAergic system might have interactions with pro-inflammatory molecular processes that dysregulated immune signaling [301].

To sum up, our analysis confirmed the already-known comorbid processes identified in previous studies that contributed to the comorbidity mechanism, such as inflammatory processes, while also unveiling additional less-seen mechanisms such as angiogenesis, reactive oxygen species metabolism, and GABAergic systems.

2.3.5 Elucidating and quantify the comorbid genes across schizophrenia and type 2 diabetes

To investigate as well as to quantify the specific genes associated with the previous findings, the exposure values of the genes contributing to the individual signatures were extracted from the comorbid signature pairs (see **Method 2.2.5**). Here, we focused on the former identified processes that were also frequently enriched in the largest clusters, namely angiogenesis, acute inflammatory response, and reactive oxygen species, as well as the GABAergic synapse, which were specifically investigated. After retrieving all the scz-t2d comorbid signature pairs that showed enrichment in these terms, the top gene markers were ranked based on their highest exposure (**Figure 2.9** and **Method 2.2.5**). Within the signature pairs

enriched for angiogenesis, we noted high exposure genes presented in schizophrenia and type 2 diabetes signatures, such as *APLNR*, *NGFR*, *IL18*, *SPRED1*, *TYMP* and *HPSE*, among which *APLNR*, *IL18*, *NGFR*, and *SPRED1*, have been previously implicated in relevant processes involved in energy metabolism, peripheral neuropathy, cell plasticity as well as inflammatory response in both schizophrenia and type 2 diabetes patients [302–306]. For example, *APLNR*, part of the apelin APJ system, was associated with neuropathy activities in chronic schizophrenia patients but at the same time served as a key biomarker for type 2 diabetes therapy due to its involvement in regulation of adipokines in energy metabolism, oxygen metabolism and resistance to insulin [306,307]. *TYMP*, known to promote angiogenesis and stimulate endothelial cell growth, has been found to be related to dysfunction in endothelial systems and metabolic symptoms in type 2 diabetes studies[308], but it has not specifically been described in the schizophrenia context from previous findings. Instead, mitochondrial neuro-gastrointestinal like symptoms (MNGI), associated with *TYMP* variants, were found in schizophrenia [309,310]. Furthermore, a GABAergic system related gene, *GABRA5*, identified in scz-t2d comorbid genes, showed changes in the regulation of neurotransmitters in psychosis and in the regulation of the endocrine system in diabetics, respectively, impacting insulin secretion, episodic memory, and cognition [311]. Neuronal cell-adhesion molecules such as *CDH10* and the cadherin family were identified, which were already shown in diabetes mellitus and neuropsychiatry such as autism spectrum disorders but were not specifically illustrated in schizophrenia [312,313].

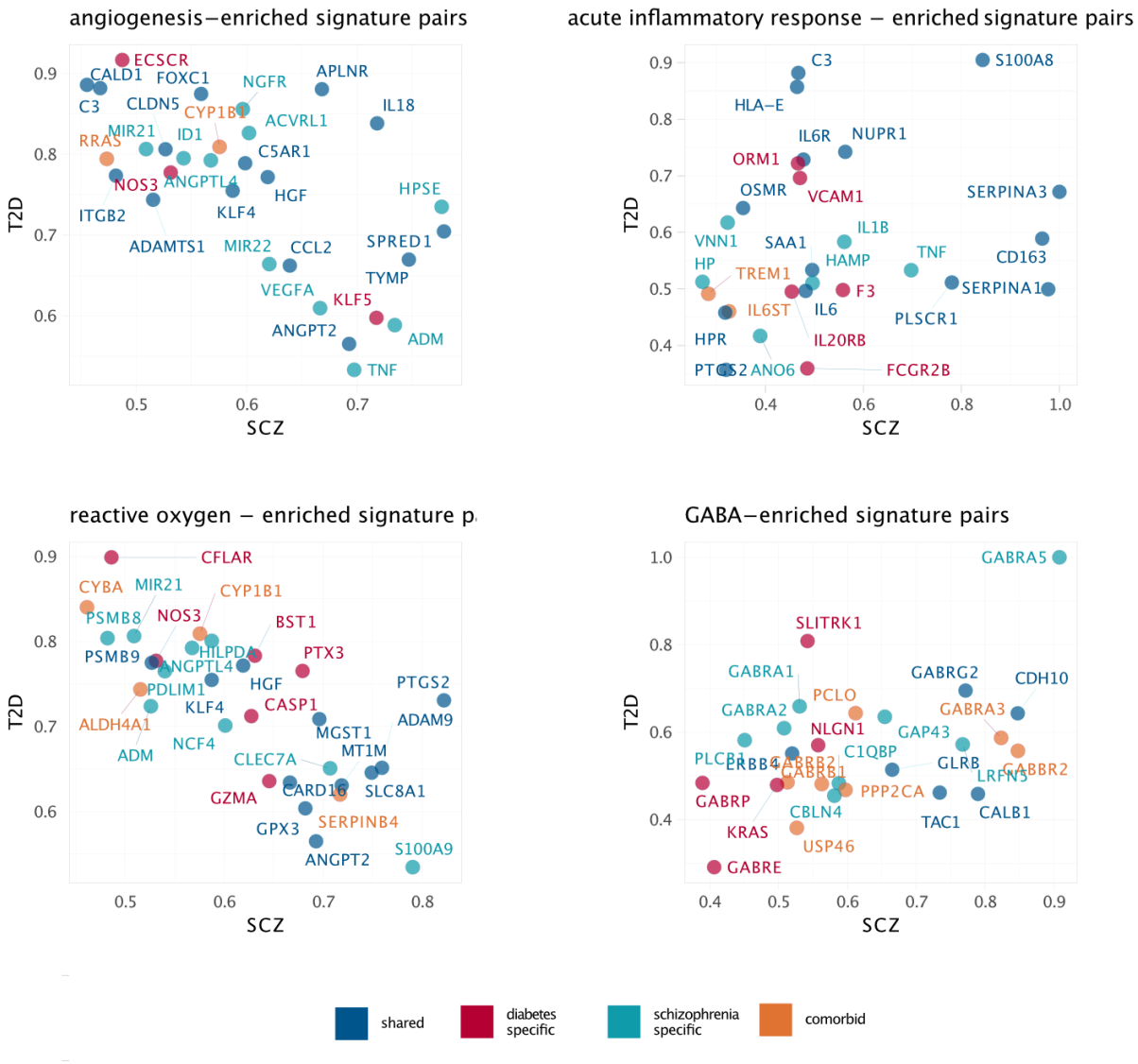


Figure 2.9 Exposed gene analysis of the signature pairs. The x-axis referred to specific gene exposure in scz signature from the scz-t2d signature pair while y-axis showed the specific gene exposure in t2d signature from the scz-t2d signature pair. The type of genes were annotated with different colors which adopt the same criterion as in **Figure 2.7**. The genes with the highest geometric mean exposure values associated with the enriched pathways were included in the figure.

The identified biomarker genes were further compared with the risk variants associated with schizophrenia and diabetes traits summarized from GWAS (**Figure 2.10**). 33 common genes were recognized, including *VEGFA*, implicated in angiogenesis and also one of the risk variants and key biomarkers specifically responded to type 2 diabetes, and *PSMB9*, the member of immunoproteasome and oxidative processes. Meanwhile, *ERBB4*, implicated in

schizophrenia and crucial in the development of cortical inhibitory GABAergic circuits, was also identified.

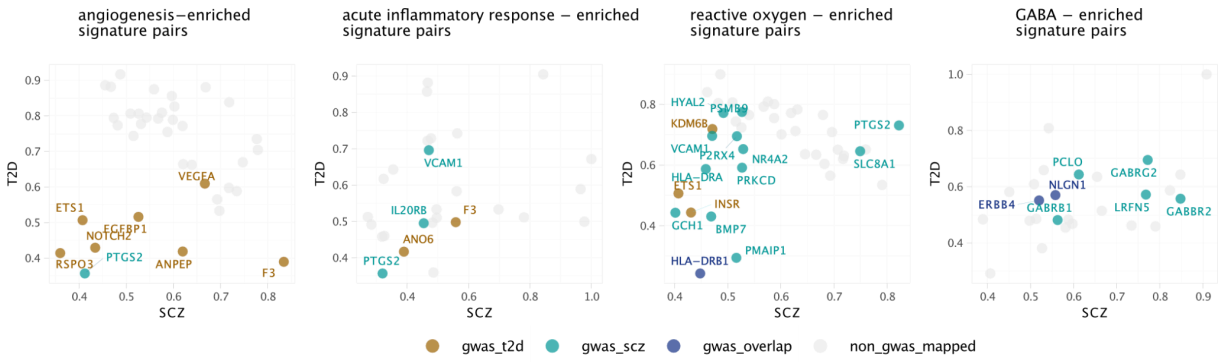


Figure 2.10 GWAS variants mapping between identified genes with SNPs associated with schizophrenia and type 2 diabetes traits from GWAS catalog database

In summary, the analysis illustrated the comorbid signature pairs linking schizophrenia and type 2 diabetes, including acute inflammatory response, angiogenesis, reactive oxygen species metabolism, and the GABAergic system, as well as revealed the potential representative gene markers contributing to these processes. It is frequently observed that many of these processes are, to some extent, associated with the immune system and inflammatory response, suggesting they could be a central mechanism in schizophrenia and type 2 diabetes comorbidity. These mechanisms and processes have been individually investigated in the context of either schizophrenia or type 2 diabetes solely. My analysis consolidated the hypothesis that the shared molecular mechanisms play a central role in the increased susceptibility to the incidence of type 2 diabetes comorbidity in schizophrenia.

2.3.6 Establishing the schizophrenia - type 2 diabetes comorbidity mechanism connectivity graph

To connect the putative mechanism in the schizophrenia - type 2 diabetes comorbidity, connectivity graph was constructed utilizing a curative knowledge database developed by our collaborator Vinay S. Bharadhwaj (see **Method 2.2.6**). We linked the previous identified pivotal processes including inflammatory response, angiogenesis, reactive oxygen species, and GABAergic system, to illustrate the specific connection and paths linking schizophrenia and type 2 diabetes (**Figure 2.11**).

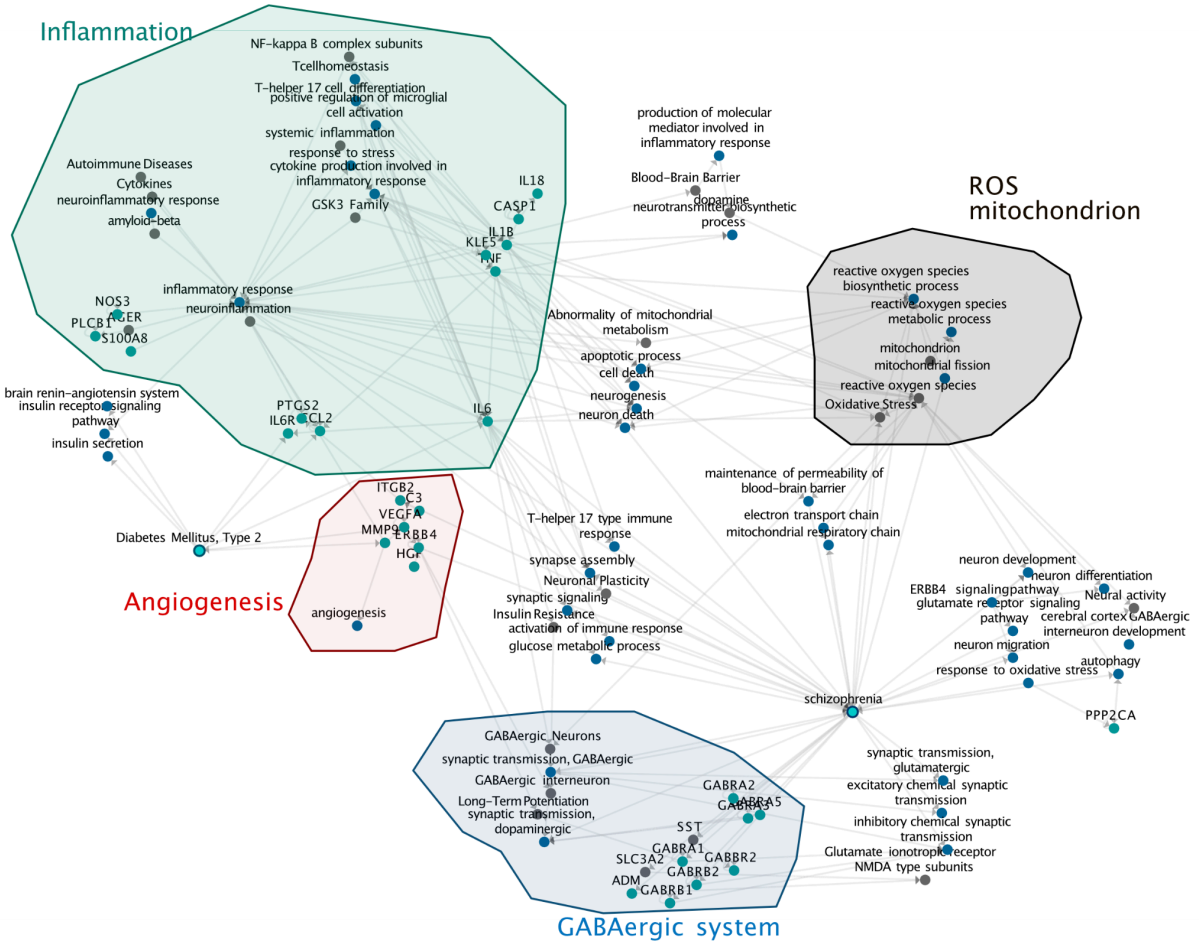


Figure 2.11 The comorbidity connectivity graph linking schizophrenia and type 2 diabetes based on representative biological processes in the previous analysis including angiogenesis, inflammatory response, reactive oxygen species and GABA-ergic system, as query on a literature curative knowledge database. Connections were established with hyperparameter setting (maximal depth = 5) of the input processes. Nodes in the connectivity graph corresponded to either HGNC genes, biological processes, other ontology terms or diseases. Edges indicated the existence of interaction between the linked nodes.

In this map, inflammation, along with related terms such as neuroinflammation and autoimmune, emerges as a central interconnected hub node. It is shown that the path to type 2 diabetes passed via the insulin receptor signaling pathway and multiple nodes, through *IL6*, to schizophrenia. Notably, the direct linkage between schizophrenia and type 2 diabetes can be formed via *IL6* and its receptor, IL6R, suggesting its important roles. Alternatively, schizophrenia and diabetes are also connected through multiple other nodes involved in

processes including GABAergic neurons, oxidative stress, and neuronal plasticity. Consistent results of genes and enriched pathways can also be found in the graph, such as *TNF* and *IL6*, associated with the inflammatory response (**Figure 2.9**).

The connectivity of reactive oxygen species involved mitochondrial respiratory chains, electron transport chains, and autophagy, which were highly related to energy metabolism. Neuron death, cell death, and regulation of the blood-brain barrier could also be identified in the route to schizophrenia. Angiogenesis was linked to genes such as *MMP9* and *VEGFA*, and further passed through the dopaminergic system, long-term potentialization and other synapse terms to the schizophrenia destination. Interestingly, GABAergic systems were shown to have frequent connections with other processes that are less likely to be associated with neuro-related disorders, such as insulin resistance, which suggested their extensive role in the etiology of type 2 diabetes [311].

Expanding upon the comorbidity connectivity graph, additional mechanisms that might play an intermediate, but important, role were recognized. For instance, the development of the nervous system involved regulation of neuronal plasticity and neuronal signal transduction, and blood-brain barrier systems that included the brain renin-angiotensin module as well as the insulin system with insulin receptor signaling regulation were observed. Combining these findings strengthened our understanding of a complex mechanism with the co-occurrence of these processes and connecting not only oxidative response and cell recovery, suggesting damage and repair mechanisms as another aspect of evidence that could support the significant roles of inflammatory processes in schizophrenia and type 2 diabetes comorbidity.

2.4 Discussion

To summarize, a new factorization approach for comorbidity modeling at the signature level with multi-rank non-negative matrix factorization (mrNMF) was presented. The novel framework alleviated the limitations of standard NMF by introducing ensemble-like algorithms to identify biological signatures at different levels of rank decomposition. I used a reciprocal best hit approach to identify signature-level-based comorbidity patterns in a large collection of cross-condition transcriptomic cohorts. A substantial number of signature pairs

came from pairs of different diseases, emphasizing the overlapped molecular dimension between different diseases. The comorbidity analysis, specializing in schizophrenia and type 2 diabetes, illustrated the special roles of the inflammatory response in the comorbid incidence. Moreover, it elucidated the close interconnection among angiogenesis, reactive oxygen species metabolism, and the GABAergic systems with other relevant processes.

One limitation of the study is the inability to distinguish the causative effect within a signature pair. Hence, the computation cannot evaluate the risk in a quantitative way or the proportion of contribution to the comorbidity from each disease in the signature pair. However, the GWAS mapping analysis in the study confirmed several gene biomarkers as shared risk variants in type 2 diabetes and schizophrenia. For instance, *NLGN1* plays an interesting role in the extracellular matrix (ECM) of angiogenesis processes in type 2 diabetes. Moreover, *NLGN2*, as well as the variants of *NLGN1*, were demonstrated as cell-cell adhesion factors in the post-synaptic membrane in schizophrenia patients. Additionally, this variant was further associated with the formation of GABAergic synapses, which inspired our findings on the GABA-ergic systems. Moveover, as an already well-known control of our comorbidity modelling, dopaminergic system was indicated in our analysis, covering basically all the relevant important processes such as synthesis, secretion, transport, and metabolism (**Figure 2.12**).

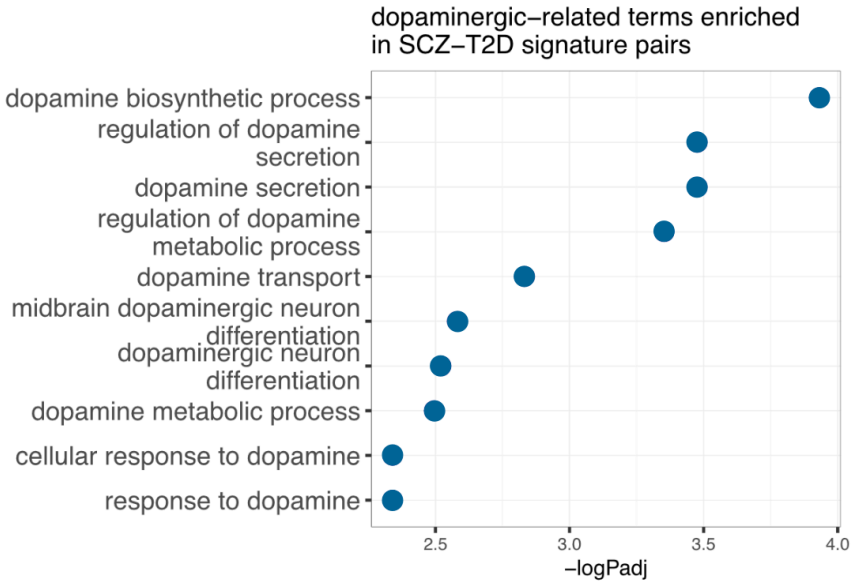


Figure 2.12 Dopaminergic related processes identified in schizophrenia - type 2 diabetes comorbid signature pairs

Interestingly, the serotonergic system was not found in scz-t2d comorbidity modelling, though it was earlier mentioned in one of the hypotheses of schizophrenia occurrence. However, the further investigation suggested that this could be because inconsistent results were observed regarding the role of serotonin from different studies in schizophrenia. In contrast, we observed that serotonin and major depressive disorder tended to be more related to each other compared to schizophrenia, and it was found that transport, secretion, and signaling of serotonin can be identified in the comorbidity modelling between major depressive disorder and cardiovascular diseases (**Figure 2.13**).

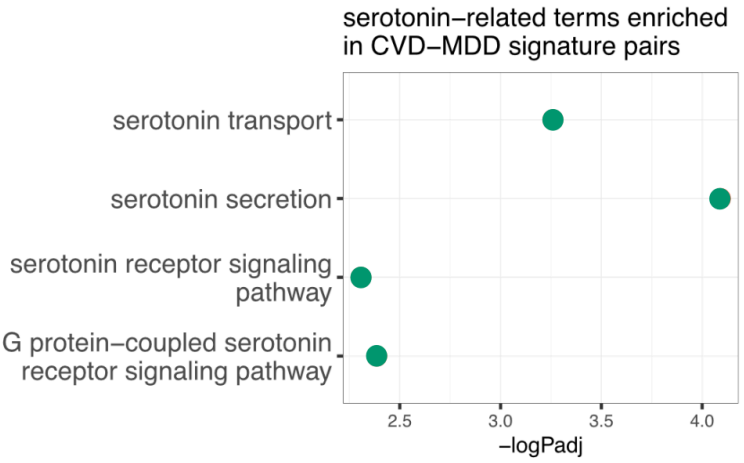


Figure 2.13 Serotonergic related processes identified in cardiovascular disease - major depressive disorder somatic signature pairs.

The clinical implication of our framework might help in several aspects. For instance, the identified molecular factors in comorbidity modeling could be further translated to clinical markers for specifically distinguishing schizophrenia patients with and without diabetes or other metabolic syndrome. In addition, molecular factors associated with side effects of the antipsychotic medications, could be identified and their potentiality on contributing to next-generation antipsychotic drug discovery that were able to alleviate the comorbid symptoms.

Overall, the hierarchical factorization computational framework on the signature level provides an alternative for cross-disease comorbidity modeling. Using the ensemble strategy with multiple rank implementation simultaneously, it shows its capability of delineating transcriptomic cohorts and identifying molecular signatures across different granularity, illustrating potential contributing molecular mechanisms that underlie disease and their comorbidity syndromes.

Chapter 3

Project 2: Cross-cohort distributed data integration with federated factorization learning

3.1 Introduction

Innovative applications of artificial intelligence led to substantial progress in genomics, pathology, clinical diagnosis and many other fields; however, the current data-driven models generally accompanied the need for sufficiently large training input data to achieve sufficient performance accuracy to further translate to practical clinical use. For example, training disease predictive models required a large database as well as complete training data covering biological layers involving anatomies, pathologies, and various data types [314–316]. Currently, the constant evolution of high-throughput technologies and the storage capacity and processing power have allowed data science in industrial engineering to become more realistic. Hence, the demand for cross-institutional collaboration in artificial intelligence is increasing to support the explosive development of data-driven machine learning [317]. Nevertheless, two major challenges in the field of machine learning development still raised awareness. Firstly, data governance and the privatization of data, driven by legal concerns, stood out as critical aspects [317]. Health data, especially clinical data, was challenging to

obtain because of its high sensitivity and was therefore tightly regulated. Despite the potential effectiveness of data anonymization in circumventing these constraints, the removal of clinically sensitive data, such as patient ID entities like their names and births, could sometimes also be de-censored. Notably, computed tomography (CT) or magnetic resonance imaging (MRI) and other imaging modalities could yield a large amount of information to reconstruct a patient's facial features, which also challenged the work of preserving the privacy of the participants. Moreover, the reluctance towards data sharing in healthcare could also be attributed to commercial reasons due to the investments of time, resources, and financial capital required for the collection, curation, and maintenance of high-quality datasets. Secondly, the issue of lack of advanced co-analytic techniques, platforms, and standards presented another obstacle to the advancement of this field. Though access to more extensive training data could substantially enhance training performance, in reality, problems such as data processing, efficient learning across distributed data, specific algorithms to handle cross-institutional cohorts, etc.

One potential solution to address the above-mentioned challenges simultaneously is the federated learning (FL) technique. Federated learning represents a specific class of machine learning methods developed to address the challenge of data fragmentation while preserving data privacy. It involved multiple client servers and users, such as research institutions or local machines, collaborating with specific central servers in decentralized machine learning setups. This concept was first introduced in 2016, initially applied to enable collaborative learning from multiple devices [318]. With further extending to applications, federated learning, to be applicable to any edge device, held promise for revolutionizing critical domains such as healthcare. A notable example was the collaborative development of a distributed model for COVID-19 disease diagnosis with scan data by researchers and medical practitioners worldwide [319]. With further extension of federated learning infrastructure, fusion of bioinformatics and medical informatics tools with a federated learning framework also allows collaborative machine learning in many fields, such as electronic health records, molecular identification, drug discovery, disease diagnosis, and GWAS [320–324]. Another study investigated the feasibility of the federated learning approach for molecular genetic studies, such as performing differential gene expression (DEG) analysis across multiple data sources. Additional practice also highlighted its capability of handling class imbalance data issues with a specific-designed swarm-learning (SL) model, which incorporated blockchain

techniques in distributed ML scenarios and was then applied to large-scale COVID-19 data analyses while ensuring data security [325].

Though the current federated learning approach has been versatile regarding their methodology and application, the general procedure for a typical federated learning model involves several main steps [326]. First, each server initializes a standardized global model for subsequent local training. Second, the global model parameters were broadcasted to the client servers or machines. The global model undergoes refinement through multiple local updates using server-specific data, with each device executing updates independently and then uploading encrypted gradient information to the cloud. Third, local model training where clients will receive and keep the model's summary statistics learned from the training data. The data server aggregates the averaged updates from the local models and disseminates them back to the iterative updated global model. Lastly, model aggregation that client server will send back the model parameters to the server and updated parameters are aggregated and form the new global model. The above steps are iterated and repeated for the given times or until the stopping criterion is met. The advent of federated learning technology holds promise for reconciling the tension between data privacy and data sharing across distributed devices. Given that data remains localized and is not exposed to a central server, federated learning is particularly suitable for applications involving privacy-sensitive data, such as in healthcare or on mobile devices where legal constraints prohibit data aggregation.

Despite numerous applications of federated machine learning developed in several critical scenarios across a range of scenarios, it is still leaving massive gaps to be filled. For example, a common assumption of most machine learning approaches, federated or otherwise, is that all observed data points (e.g., individuals affected by illnesses) represent the same underlying distribution and population, also known as the IID assumption. However, in the actual biomedical data analysis scenario, this assumption rarely holds true since biological and technical factors often induce modality-specific confounding effects that are challenging to capture using federated machine learning alone. Some machine learning algorithms, such as integrative non-negative matrix factorization (iNMF), offer a solution to this issue by simultaneously learning patterns associated with outcomes (e.g., diagnosis) across datasets, considering both cohort-specific and shared effects. MTL holds promise for various applications, including comorbidity modeling, and has already demonstrated success in disease progression analysis.

In this Chapter, I will describe my contribution to the development of distributed integrative non-negative matrix factorization (dsMTL_iNMF) based on the federated learning strategy. The goal is to identify the signal that is shared between cohorts and the one that is cohort-specific. The algorithm is integrated into a statistical software R package called dsMTL (Federated Multi-Task Learning for DataSHIELD) co-developed by Dr. Han Cao (**Figure 3.1**), for integrative learning. As the platform that dsMTL was built upon, DataSHIELD is a software infrastructure supporting the federated analysis of distributed data that holds the need of privacy preservation and limited assessment under restriction of data barriers from different research institutions throughout analysis. dsMTL package has been deployed and divided into two repositories, including the server-side package: https://github.com/transbioZI/dsMTLBase and the client-side package: https://github.com/transbioZI/dsMTLClient. First, for the design and implementation of the algorithm, the theoretical fundamentals of the approaches, the computational workflow as well as the deployment of the infrastructure were derived and established in order to fulfill the prerequisites of the model practice. Next, dsMTL_iNMF from the R-based dsMTL package was developed based on the DataSHIELD platform, which is a distributed learning solution for biomedical research to handle sensitive data. Subsequently, the feasibility and efficiency of the model algorithms, including dsMTL_iNMF in the packages, were examined in the scenario of cross-cohort comorbidity modelling. Specifically, simulated data and real-world data analysis were performed in different trial settings on the aspects of data heterogeneity, disease variety and runtime. In addition, validation of the cohort-specific and cohort-shared biological signatures was explored across the schizophrenia and bipolar disorder cohorts.

3.2 Methods

3.2.1 Datashield infrastructure

DataSHIELD is a platform designed to facilitate federated data analysis while at the same time preserving individual-level data privacy and preventing further leakage. It comprises two primary modules: the R programming environment and the OPAL data warehouse deployed at each institution. The R environment was applied as the basic computational language for the

training and analysis to initiate across the client server and central data server. Therefore, DataSHIELD, the well-designed infrastructure and platform that was specifically developed for R computation, facilitated the full procedure of the model implementation. With the approach to bring the analysis to the data rather than the other way around, the operating mechanism of the DataSHIELD involves three general steps. First, analysis requests are submitted by the clients from client machines, or so-called analytical machines, to data servers where the harmonized data is stored. In the second step, the analysis results, usually the non-disclosive summary statistics of the data, is returned in parallel to the client servers. The platform excels in importing and exporting large datasets, accommodating diverse data types, and efficiently handling extensive datasets, such as those encountered in genome-wide association studies that required tens of gigabytes. To support various data types effectively, DataSHIELD incorporated various extensive-developed algorithm resources in its community, which also facilitated the seamless communication of extensive data with the specific-designed compressed forms.

In terms of security, to sufficiently control the risk of data disclosure, DataSHIELD considers two different aspects, including the optimization of the software architecture itself and their computational methods for statistics. The architecture of DataSHIELD provides several non-disclosure mechanisms to improve the system's security, such as the firewall setting behind it, fixed IP addresses for a specific set of clients, restrictions on communication, SSL protocol protection of the network, and an internal R environment with limited command and disclosure control on the server to restrict privacy information. For administrators, permissions can be set to control access to sensitive data, but together with a limited degree of internal function availability on the server, such as user registration for data usage, In some cases, users could be permitted to obtain the summary results but not the data due to the control of usable functions. With these settings, DataSHIELD can be customized to meet different requirements for better data management. In the aspect of statistics, DataSHIELD allowed the summary statistics to be shared based on its safety assumptions, similar to the context that GWAS summary data are generally accessible while raw genotypes on an individual level are not. This assumption aligned with common practices in the biomedical field.

3.2.2 dsMTL package

Common mathematical form shared across the methods in the dsMTL package,

$$L(\theta) + \lambda S(\theta) + C\aleph(\theta)$$

where the first term $L(\theta)$ is the cost function (loss function) for the determinant of the model training and testing solutions and the inside argument θ represents the model parameters. The other two terms refer to the restriction implementation of θ, which could be further customized by the users depending on the needs, such as adding prior knowledge and information to constrain the parameter sets. For example, here, $S(\theta)$ is smoothing regularization to stabilize the training while $\aleph(\theta)$ is for sparsity constraint. And λ and C are the hyperparameters to tune for the magnitude of the regularization effect.

Built on the above consensus fundamental mathematical form, four federated machine learning algorithms were developed and optimized which includes three supervised learning and one unsupervised learning algorithm. The algorithms were designed and improved from former well-known non-federated machine learning approaches such as Lasso, NMF, etc. Here we specifically introduced the factorization based federated learning technique **dsMTL_iNMF** which was originally integrative non-negative matrix factorization (iNMF) applied to integratively factorize multi-view data into various latent factors, in other word, signatures , including cohort-specific and cohort-shared signatures. The integrative non-negative matrix factorization was used for data integration, clustering and biomarker discovery for multiple cohorts simultaneously. However, as other integrative algorithms that required data pooling in the same environment, integrative non-negative matrix factorization was not capable of handling physically distributed data. Therefore, a federated optimized version of the algorithm is on the demand. In the next section, the mathematical basis, the algorithm practice and novel training mechanism would be introduced.

3.2.3 Federated integrative factorization

Integrative non-negative matrix factorization (iNMF) is an unsupervised machine learning algorithm for discovering the hidden structure in heterogeneous as well as homogenous in high-dimensional biological data. We developed dsMTL_iNMF algorithm by adapting the federated analysis with DataShield infrastructure on the iNMF with the aim of collaborative

learning while preserving the data privacy. Currently this distributed algorithm has been integrated in a R package namely dsMTL. In **Figure 3.1,** the architecture of dsMTL_iNMF was shown, with the example of three data matrices stored in three warehouse servers being decomposed and factorized simultaneously. During the implementation, the shared decomposed matrix H (rows are features while columns are ranks), cohort-specific decomposed matrices V (rows are features while columns are ranks), and W (columns are samples while rows are ranks) would be generated. The loss function was shown below.

$$\sum_{k=1}^{t} ||X_k - (H + V_k)W_k||_F^2 + \lambda \sum_{k=1}^{t} ||V_k W_k||_F^2 + \lambda_s \sum_{k=1}^{t} |W_k|_1$$

The linear addition of H and V enabled the model to delineate cohort-specific and cohort shared after iterative training and optimization, in which H identified the cohort shared homogenous factors across cohorts whereas V distinguished the cohort-specific heterogeneous factors solely responsive to each cohort respectively. The regularization $\sum_{k=1}^{t} ||V_k W_k||_F^2$ allowed the magnitude of heterogeneous factors, on the other hand, also the degree of homogeneous factors to be identified in the model. Additionally, $\sum_{k=1}^{t} |W_k|_1$ was applied to tune the coefficients of sparsity in the model.

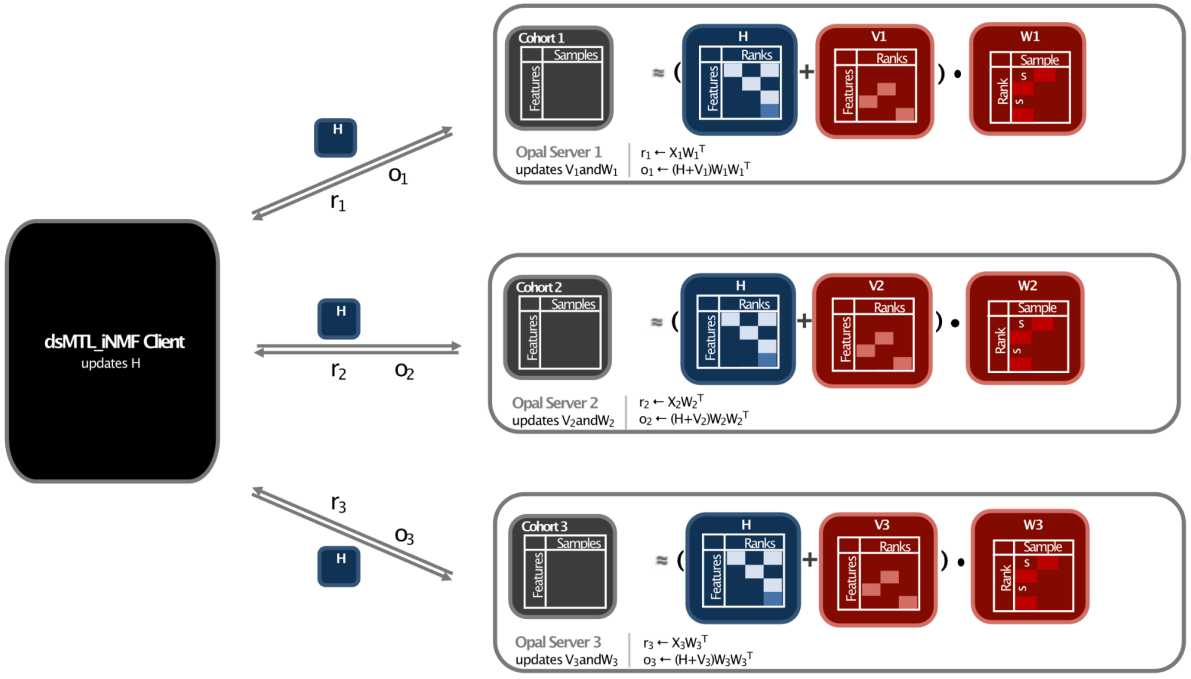


Figure 3.1 The communication across different physically distributed servers (client-data and client-client) of dsMTL_iNMF algorithm built on integrative learning architecture

In the context of distributed learning, the cohort-specific variables W_k and V_k were updated on server k. The updating equation of W_k and V_k in dsMTL_iNMF is the same as the multiplicative update functions in the original non-distributed version of integrative NMF, shown in the below formulas [223]. The homogeneous matrix H is updated on the client after receiving summary statistics from all servers, where these statistics were non-disclosure and calculated on the server. Followed by the aggregation of the intermediate matrices, the client updated H and continued the training on a new iteration. To preserve data privacy from leakage during the training, only the homogeneous matrix H is returned. Below we summarized and demonstrated the algorithm solver for dsMTL_iNMF and the update mechanism.

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

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

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

As the formulas demonstrated, the parameters and matrices were updated iteratively based on the above operations. In the federated integrative factorization approach, the cohort-specific heterogeneous matrices W_k and V_k were updated on data server k respectively while the cohort-shared homogeneous matrix H was optimized on the client-side after retrieving summary statistics from all data servers. Based on the updating function for matrices above, we showed the final updating formulas for H with server information informed below, which is consistent with the illustration of communications across clients and data servers in **Figure 3.1**.

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

Therefore, the solver of the dsMTL_iNMF algorithm was illustrated by each step in Table 3.1 as follows.

Table 3.1 Schematic illustration of dsMTL_iNMF solver

Solver of dsMTL_iNMF algorithm
input: λ>0, λ_s>0, maxIter>0, H, W_1…,W_t, V…,V_t
output: H
for i=1 to maxIter **do**
update H on client send H to all servers update W_1…,W_t on data server 1,…,k update V_1…,V_t on data server 1,…,k send summary statistics H back client server
If termination rule met,

return

end for

During the model optimization of dsMTL_iNMF, a specific termination rule from the ButchR package was applied to evaluate the convergent state of the algorithm, as shown in **Figure 3.2** [327]. The new method assigned each sample to a clustering membership by $G = \arg arg\,|H_{i,j}|$ after every iteration of training. The optimization of the model training would stop when clustering membership assignments for samples remained constant for a given number of additional iterations. In the default setting of our algorithm, the number of iterations for the model to be seen converged is 10. At this point, the assignment of clustering memberships were regarded as stable. The underlying assumption is that as an integrative unsupervised learning method, integrative factorization aimed to achieve a state that the clusters succeeded in capturing sufficient small variance within the same cluster while maximizing the variance across different clusters.

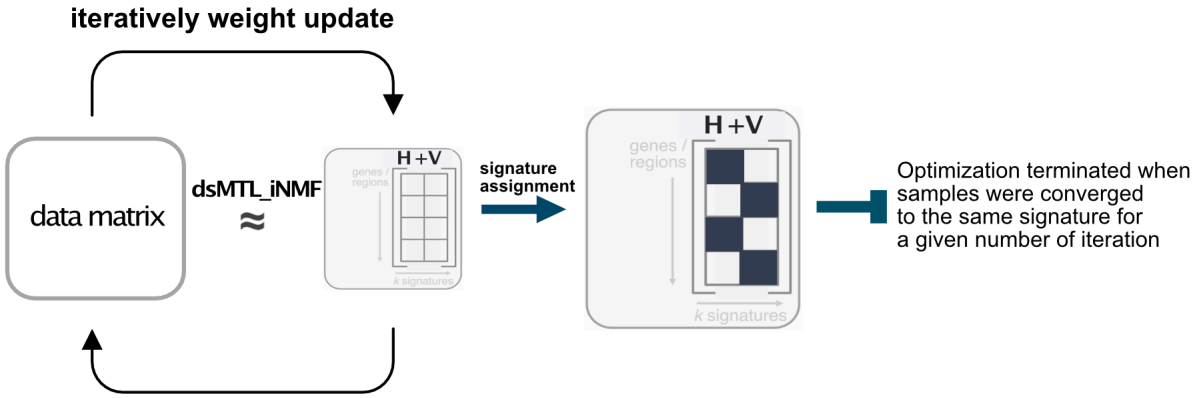


Figure 3.2 The optimization and termination rules based on stable cluster membership assignment.

Overall, the full training process of dsMTL_iNMF is provided below in **Table 3.2**. Distributed solver algorithms were illustrated for within every single epoch or iteration while the distributed model training demonstrated the model training process.

Table 3.2 Full training procedure of dsMTL_iNMF algorithm

Full training procedure of dsMTL_iNMF

input: λ>0, λ_s>0, maxIter>0, nInitialization, rank, {X_1, …X_k, …X_t }

output: H_1, H_2, …

for i= 1 to nInitialization **do**

initialize H_i~U_(n×rank) (0,1)

for each k **do**

initialize V_k~U_(n×rank) (0,1)
initialize W_k~U_(rank×p_k) (0,1)

H_i = **dsMTL_iNMF Solver** (λ=λ, λ_s=λ_s, H=H_i, {W_1, …,W_t }, {V_1, …,V_t })

end for

3.2.4 Data key mechanism

The DataSHIELD infrastructure provides the general security mechanisms for the framework and machine learning application. For our model dsMTL_iNMF, additionally, we implemented and examined an extensive algorithm developed by the collaborators to potential machine-learning-specific privacy leakage issues. For instance, membership inference attempts to select or de-select specific individuals in the training dataset using the model whereas inverse attacks were able to retrieve the meta information on individual level from the models [328,329]. Since dsMTL_iNMF might return multiple matrices in the intermediate phase of the model training, which could expose the model to these attacks. Therefore, dsMTL_iNMF only allows the homogenous matrix H to be retrieved in the intermediate step, whereas the heterogeneous matrices $V$and W stay in the remote data server, to avoid the exposure of the full model. Another purpose is that, among all the decomposed matrices, only one intermediate parameter matrix would be returned. The proportion of accessible information would decrease dramatically with the increased number of cohorts, e.g. 1 out of 5 in two servers setting, and 1 out of 9 in four servers setting. The incompleteness of the model

made the inverse inference of the input data difficult, so that the cost of inverse inference and other attacks increased in the end.

In addition to the previously mentioned fact that existed internally in the model algorithm, a new data key mechanism was developed as an additional barrier for the situation that specific authorized clients were allowed to actually access the data in the server (**Figure 3.3**). There are several reasons to add this mechanism given that DataSHIELD has already provided a relatively safe environment for distributed learning. For example, the new data key mechanism extended the utilization of original DataSHIELD that self-defined functions are not able to obtain identity information from the clients. Secondly, it added another gate to improve IP security when specifying the client identity in DataSHIELD and ensured the safety when special client requests were made such as accessing all results during biomedical analysis. This mechanism as shown in **Figure 3.3** allows the authorized clients to safely receive the factorized matrices identified by the federated integrative factorization from the server upon their request. For the implementation in practice, two separated keys are generated by the administrator, of which one, so-called local key, is kept in the key database whereas the other one, namely the remote key, will be authorized to the specific clients in the list. Clients who own the remote key are regarded as the data provider that 'owns' the data in the data server, so that they can access the originally limited data generated from federate learning methods.

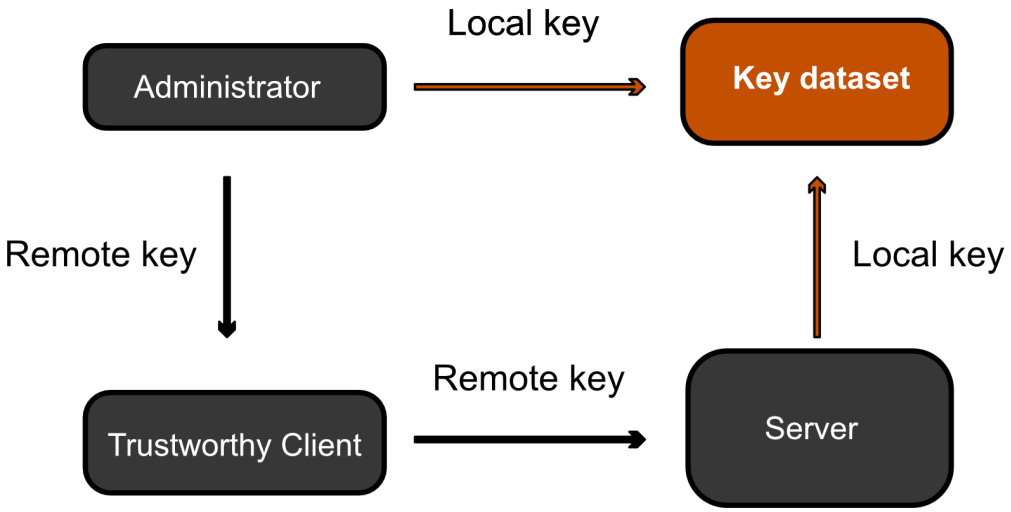


Figure 3.3 The data key mechanism for providing additional access to the specific clients

3.2.5 Generation of simulated RNA-seq count data

The structure of general RNA-seq count data followed a Negative Binomial distribution (NB distribution) pattern [330]. To simplify in our following study, simulation was performed under the two servers settings. Various degrees of data heterogeneity were assigned to the test simulated data (20%, 40%, 60% and 80%). To specify, the magnitude of data heterogeneity represented the proportion of diagnosis-associated genes that were shared across different cohorts. The data simulation is composed of the following procedures with the configuration and parameters in **Table 3.3**.

Table 3.3 Configurations of data simulation for each cohort including proportion of sample size, genes, and the sampling distribution parameters

		cohort 1	**cohort 2**
samples	diagnosis-case	0.5	0.5
	diagnosis-control	0.5	0.5
genes	background	0.5, $NB_{fb\ x\ n1}(r=2, p=0.3)$	0.5, $NB_{fb\ x\ n2}(r=2, p=0.3)$
	homogeneous	P_{fc}, $NB_{fc\ x\ nd1}(r=2, p=0.002)$	P_{fc}, $NB_{fc\ x\ nd2}(r=2, p=0.002)$
	heterogeneous	$1\text{-}P_{fc}$, $NB_{fs\ x\ nd1}(r=2, p=0.002)$	$1\text{-}P_{fc}$, $NB_{fs\ x\ nd2}(r=2, p=0.002)$

In the first step, two simulated data matrices (each server stored one single cohort) would be initialized, and background data was sampled from the negative binomial distribution with r = 2, p = 0.3. In each simulated data matrix, rows are the gene dimension whereas columns are the different sample size. Then, a specific proportion of genes and samples was selected as diagnosis-related genes and diagnosis-case samples. Here, the percentage of diagnosis-associated genes and diagnosis-case samples were set to 0.5 for all cohorts. In the next step, within the diagnosis-associated genes, cohort-specific heterogeneous genes and cohort-shared homogeneous genes were set based on the configuration provided in **Table 3.3**

where the proportion of homogeneous genes represented the different degree of data heterogeneity. For the genes marked as either homogeneous and heterogeneous, the gene effects were sampled from the negative binomial distribution with r = 2, p = 0.0.002, and further added to the background data. To note, the heterogeneous genes in different cohorts would not be overlapped, while the homogeneous genes were the same. Overall, a simplified illustration was provided as below in **Figure 3.4**.

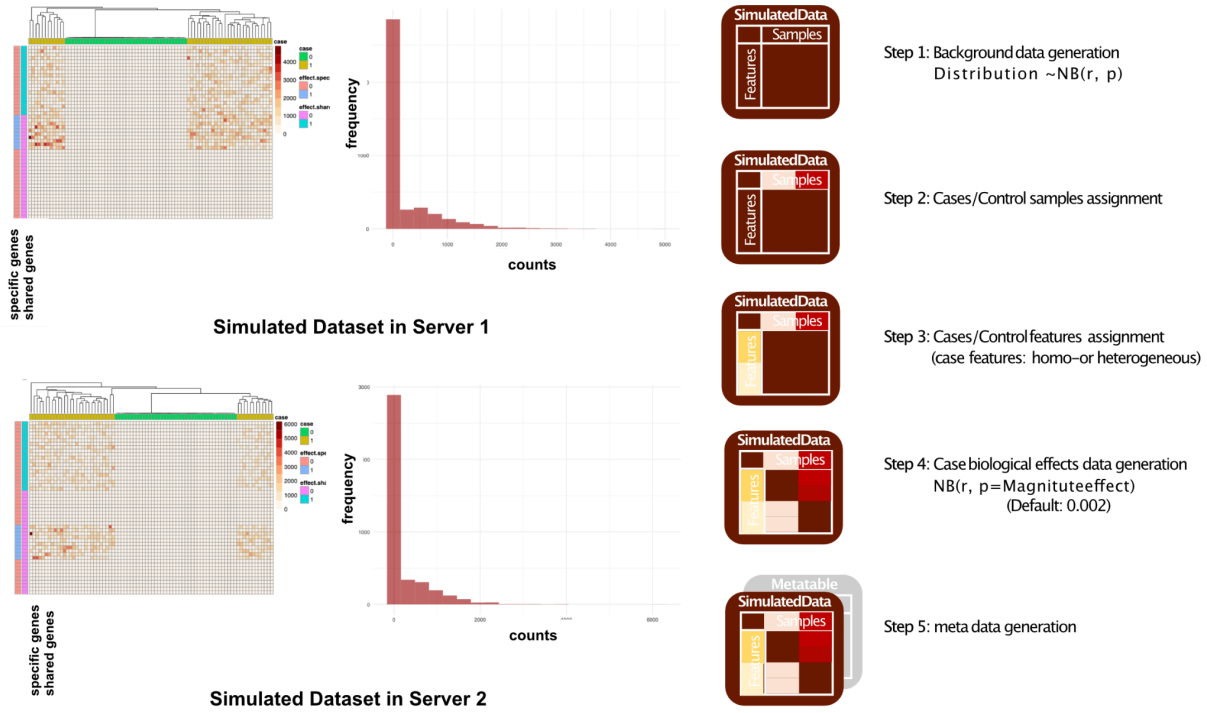


Figure 3.4 Scheme of simulated RNAseq count data generation

3.2.6 Simulated data analysis

To evaluate the feasibility and utility of dsMTL_iNMF, cross-cohort model implementation was performed by sampling simulated data. To better testify the performance of unsupervised algorithm dsMTL_iNMF on recovering the cohort-specific and cohort-shared factors embed in the simulated data matrices, assumption was made that diagnosis-associated factors was the addition of the cohort-shared factor and two cohort-specific factors in the following analysis.

Following the data generation steps described in **Method 3.2.5**, two simulated cohorts with RNAseq count like data were sampled. In the data, diagnosis-unrelated genes and

diagnosis-related genes were marked, within which cohort-specific and cohort-shared factors, or so-called signatures were specifically generated, together with annotation of each gene in the meta table. The comparison was made between dsMTL_iNMF with the standard local NMF and the ensemble of local NMF, using the identification accuracy of the diagnosis-related genes, including cohort-specific genes and cohort-shared genes as criterion under various settings shown in **Table 3.4.**

In detail, the cohorts were initialized with 200 samples and 1000 genes, 0.5 proportion of the the genes were chosen as diagnosis-unrelated genes while the other 0.5 proportion were selected diagnosis-related genes. Depending on the proportion of homogeneous effect in **Table 3.4**, the number of cohort-shared and cohort-specific genes and signatures varied, in this test, from 0.2 to 0.8 were examined.

Table 3.4 Configurations of simulated data structure in the evaluation

index	proportion of homogeneous genes/signatures	number of samples	number of genes	number of diagnosis-related genes/signatures	proportion of patients
1	20%	200	1000	500	50%
2	40%				
3	60%				
4	80%				

As demonstrated in the **Method 3.2.3**, three factorized matrices would be computed in the result including cohort-shared gene signature H, cohort-specific gene signature V and rank-sample matrices W. Utilizing the rank-sample matrices W, we evaluated the association of each gene signature in H as well as V with the diagnosis effect and outputs the weights for each of the gene signatures. By conducting weighted sum on the genes across all the signatures in the factorized matrices, the predicted gene values would be obtained which represented the gene effect to the signature. Then the predicted genes in the signature were categorized by being binarized compared to the mean value. The gene values were converted to 1 if exposure values of that gene are larger than the mean, or 0 if their values are smaller.

The responsive genes within the signature would be highlighted by this conversion. To measure the gene identification, we computed the accuracy between these predicted genes with the ground truth, which was annotated in the mata table when simulating the data matrices.

With the above metric, assessment was performed on the factorized matrices from dsMTL_iNMF, including dsMTL_iNMF-H, dsMTL_iNMF-V1 and dsMTL_iNMF-V2, which referred to the cohort-shared gene signature, cohort-1-specific gene signature and cohort-2-specific gene signature respectively. Similar operation was performed on results obtained by applying the standard local non-negative matrix factorization (NMF), which included three matrices as well, local-NMF1, local-NMF2 and NMF-bagging. Different from the output from dsMTL_iNMF that the homogeneous matrix was predicted directly, NMF-bagging identified the cohort-shared gene signature by aggregating cohort-specific gene signature and then binarizing and converting to the cohort-shared genes.

3.2.7 RNA-seq data for dsMTL_iNMF

To verify the actual performance of dsMTL_iNMF on the real-world biomedical data, two RNA sequencing cohorts were used, including one schizophrenia (GSE164376) and one bipolar disorders (GSE134497) dataset, retrieved from the GEO database for the analysis [331,332,333]. Feasibility as well as the computational time of the model were evaluated using two data servers, and the personal laptop was used as the local client server. This analysis aimed at identifying the cohort-shared and cohort-specific gene signatures across the two cohorts. As shown in **Table 3.5**, the two data servers were set up in Mannheim and Heidelberg respectively, with all the necessary DataSHIELD implementation as well as algorithms deployed. We applied dsMTL_iNMF to train the model based on the two datasets and record the time-consumption.

Table 3.5 The configurations of the servers used for real data analysis in details.

Server number		**Server 1**	**Server 2**
Location		Heidelberg, Germany	Mannheim, Germany
Hardware	**CPU**	Intel Xeon 2.4 GHz	I7-4790 3.6GHz

Memory (RAM)	4G	4G
GEO ID	GSE164376	GSE134497
Size of subjects	17	16
Size of genes	15215	15215

3.2.8 Biological interpretation with enrichment analysis

To further explore the identified gene signatures, the pathway enrichment analysis was applied to interpret both the shared and the disease-specific genes. From the DisGeNET database, the disease-relevant genes were retrieved for schizophrenia and bipolar disorders and then overlapped with our selected gene signatures [334]. The resulting genes were used as gene enrichment analysis with R package clusterProfiler [335].

3.3 Results

3.3.1 Federated learning infrastructure

Research in the field of biomedicine increasingly depended on large-scale data analysis or the integrative analysis across multiple large-scale cohorts simultaneously. However, the access to the individual level data that is sensitive, might raise severe legal issues due to the diversity of the increasing data amount which also involved names, birth, facial id and other private data. Therefore, techniques that allowed the researcher to query on as well as utilize the data in a restricted mode were constantly being developed. One example is DataSHIELD, which offered a novel solution to facilitate the availability of data training by sending the analysis from the client to the data, rather than sending data to the analysis. We established a new environment that can optimally fuse large-scale data that is physically distributed while ensuring the data security during cross institutional collaboration. We applied a federated learning architecture (**Figure 3.5**), together with distributed machine learning algorithms that can be applied across multiple institutions. In this study, I developed dsMTL_iNMF combining a federated machine learning infrastructure with integrative non-negative matrix factorization (iNMF), which enables us to perform cross-cohorts and multimodalities

factorization to further identify shared and specific signatures. dsMTL_iNMF was part of the algorithms in dsMTL R-based packages. The federated algorithm based on the architecture not only fulfilled the requirement of data privacy preserving, but also outperformed its job at finding homogeneous factors across heterogeneous data. The implementation followed the idea that instead of sending the data to the analysis, the analysis would be sent to the data server from the clients. Summary statistics is returned for downstream analysis while preserving safety of data from being accessed. Specifically, for an integrative learning algorithm to discover heterogeneous and homogeneous factor across multiple cohorts dsMTL_iNMF, weights matrices are initialized respectively in the remote data server, while during training and updates, only the intermediate shared matrix is transferred to learn the common factor in multi-cohorts. The novel data key mechanism adds another barrier to restrict action related model training. Moreover, to increase the computation speed, the number of digits of parameters was limited to relieve the data transfer burden as well as additional termination rules were applied. By implementing, the training procedure can be relaxed while the regularization depth can be minimized. After training, the three output matrices are learned and generated, including two cohort-specific matrices and one cohort-shared matrix. These matrices discovered the heterogeneous and homogeneous factor embedding from the input dataset. Additionally, matrices are stored in the data server without permission of retrieval by clients. However, further analysis requests could be sent, such as enrichment analysis, of which the results and outcomes would be returned instead. Overall, dsMTL_iNMF and dsMTL facilitate research in various aspects. The federate implementation enables the integrative analysis of sensitive data. It shows promising usage for research groups and consortiums to easily handle the data without processing additional steps involving privacy, security, regulations, and administration. Furthermore, it was also useful in the situation that cohorts, especially the collection of multiple large-scale multi-omics data, were over-size and time-consuming to transfer or distribute to other collaborators and institutions.

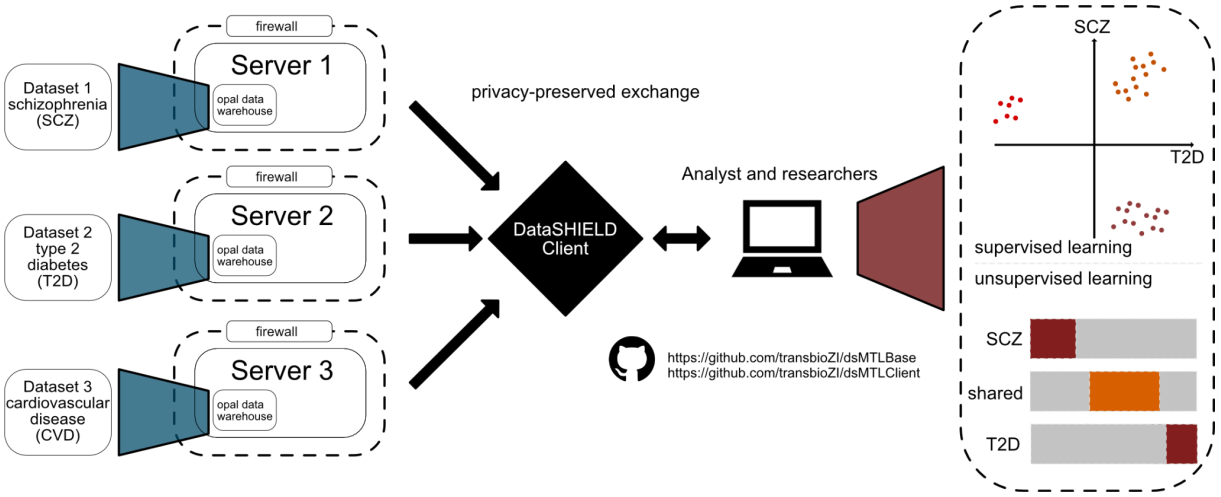


Figure 3.5 The overview of the federated learning package for distributed data, dsMTL, and the illustration of the internal unsupervised algorithm dsMTL_iNMF.

3.3.2 Simulated data analysis and performance evaluation

To evaluate the dsMTL_iNMF performance on the standardized RNA-seq data without bias, we introduced a data simulation strategy that generated an RNA-seq dataset sampled from the negative binomial distribution that followed the data pattern of the RNA-seq count. Our data simulation protocol contains four steps. First, the background data for each subject i and each gene j was sampled from the negative binomial distribution, since the RNAseq count data was observed to follow this pattern. Second, 50% of subjects were randomly selected and annotated as patients for perturbation. Similarly, 50% of genes were randomly marked as outcome-unrelated genes, while the rest were recognized as heterogeneous and homogeneous genes (the proportion varied in the test), where homogeneous genes represent shared signatures across cohorts, whereas heterogeneous genes correspond to cohort-specific signatures. In the next step, the effect of heterogeneous and homogeneous signatures were modulated by adding the signature effect as sampled from $NB(r = 2, p = 0.002)$ to the background data of specific groups. In the end, meta table was generated by summarizing the annotations of genes and individuals above.

With the above data simulation strategy, a two-server scenario was simulated. For every server data, 200 subjects and 1000 features were generated; 0.5 proportion of features were diagnosis-unrelated, while the other 0.5 were diagnosis-related signatures. The homogeneous

gene signatures were commonly chosen across data on different servers, and the heterogeneous signatures were not overlapped. 0.5 proportion of the samples were identified as diagnosis control samples, while the other half were diagnosis cases. Under four different settings regarding the proportion of homogeneous signatures, from 20% to 80%, evaluation was performed. Details are shown in **Table 3.1**.

In this case study, the dsMTL_iNMF and local standard NMF, namely ensembled NMF were compared regarding the accuracy of hidden structure identification on heterogeneous RNA-seq data. To assess the performance of the signature stratification of homogeneous and heterogeneous signatures in parallel, we applied the following comparison approach. The implementation of dsMTL_iNMF predicted three factorized matrices: the homogeneous gene signature matrix H, heterogeneous gene signature matrices V and rank-sample matrix W, where H and V represented the stratified profile of homogeneous and heterogeneous signatures. To identify the diagnosis-associated signatures, the association was computed using rank-sample exposure matrix W and the diagnosis group in the meta table and was maximized, leading to identification of the diagnosis-associated gene signatures. The gene exposure linked by this hidden factor was identified as the final profile of heterogeneous and homogeneous signatures. To quantify the set of signatures, the profile was binarized following the conversion described in **Method 3.2.6**. Using this metric, factorized outputs from dsMTL_iNMF, including dsMTL_iNMF-H, dsMTL_iNMF-V1 and dsMTL_iNMF-V2, referring to the homogeneous, cohort 1 specific and cohort 2 specific gene signatures. A similar strategy was applied for the ensembled NMF, where NMF-bagging, local-NMF1 and local-NMF2, representing the shared and specific gene signature respectively (see **Method 3.2.6**).

I implemented performance evaluation on the gene signature accuracy of dsMTL_iNMF and local standard NMF methods for disentangling cohort-shared and cohort-specific signatures from multiple cohorts, with different magnitudes of data heterogeneity, as demonstrated in **Figure 3.6**. The results illustrated the performance in separating homogeneous signatures from heterogeneous signatures in cohorts using the federated integrative factorization approach dsMTL_iNMF. For both homogeneous and heterogeneous signatures, dsMTL_iNMF received a better accuracy compared to the local NMF models regarding the gene signature accuracy, showing its capability of recovering the correct types of cohort-specific and cohort-shared genes. Especially, along the increment of the

heterogeneity's severity, the accuracy obtained by applying dsMTL_iNMF stayed relatively stable, illustrating the robustness of dsMTL_iNMF against the data heterogeneity. For the ensemble of local NMF, along the increment of the heterogeneity's severity, the accuracy of homogenous identification was continuously decreased until ~50% (20% homogenous signatures), while the accuracy of heterogeneous identification was continuously increased to 75% (20% homogenous signatures). This suggests that ensembled NMF was only able to jointly learn the feature pattern directly in a limited manner whereas dsMTL_iNMF can learn the specific as well as shared feature pattern directly from the heterogeneous matrices (V) and homogeneous matrix (H) in a collaborative as well as federated way across cohorts in multiple servers.

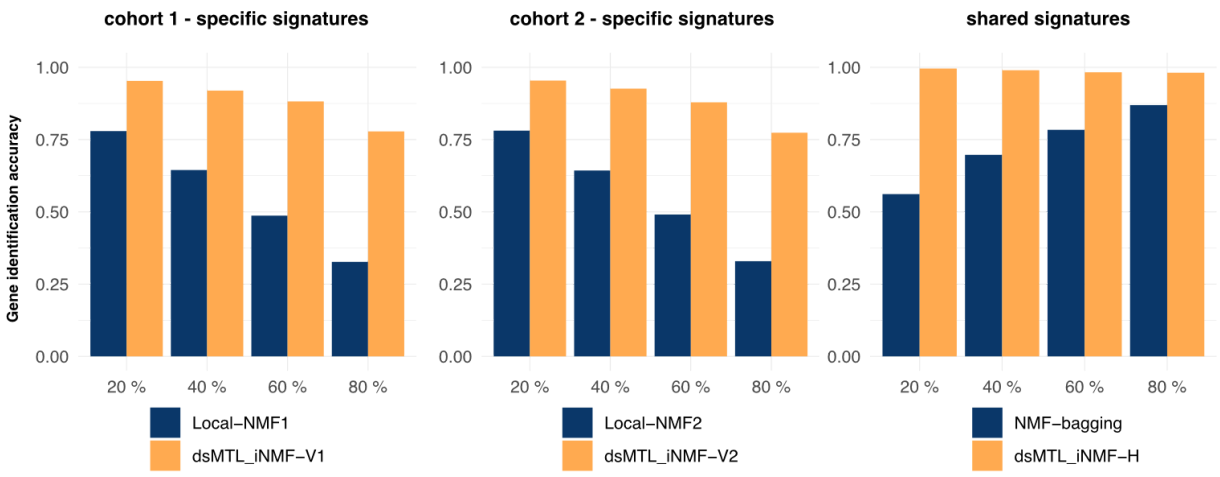


Figure 3.6 Evaluation of model performance with gene identification accuracy on predicting the cohort-shared and cohort-specific gene signatures. **left** The predicted gene accuracy of cohort 1 specific signatures. **middle** The predicted gene accuracy of cohort 2 specific signatures. **right** The predicted gene accuracy of shared signatures. Local-NMF1, Local-NMF2 and NMF-bagging showed the cohort-specific predicted gene signatures and cohort-shared gene signatures, whereas dsMTL_iNMF-V1, dsMTL_iNMF-V2 and dsMTL_iNMF-H represented predicted respective results by applying dsMTL_iNMF algorithm. The x-axis indicated the proportion of shared genes in the simulated data from 20% to 80%, showing the gene heterogeneity. The model parameter rank = 4 was used in the above evaluation.

Meanwhile, the performance of dsMTL_iNMF under different hyperparameter settings was also examined with rank k=5 and k=6. Similar results were obtained as **Figure 3.7** illustrated below.

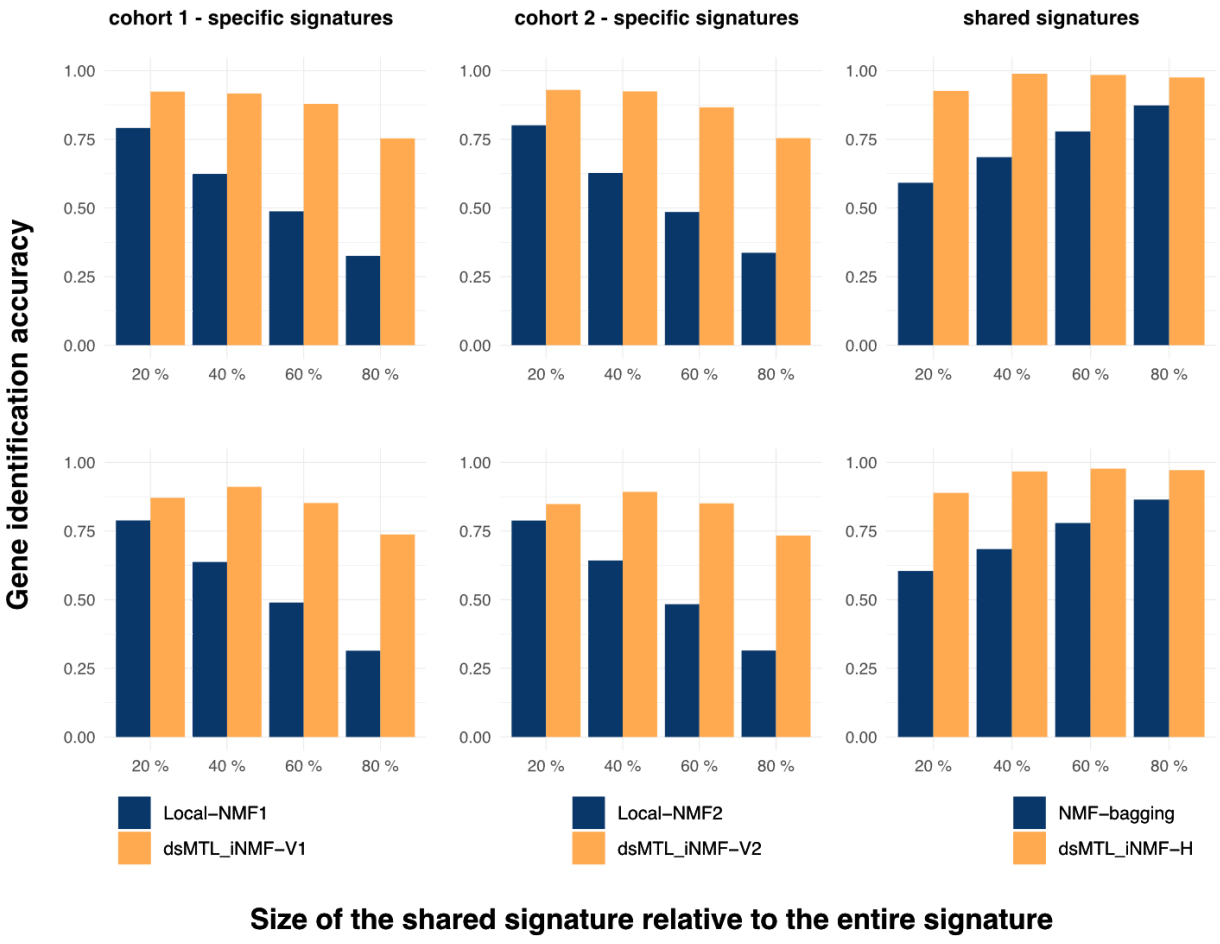


Figure 3.7 Evaluation of model performance with gene identification accuracy on predicting the shared and specific signatures using simulated data under different hyperparameter settings. **top** The training model parameter rank = 5 was used. **bottom** The training model parameter rank = 6 was applied.

3.3.3 Real-world data analysis and signature discovery

For real-world case study, I implement an evaluation of the computational speed as well as further biological enrichment analysis. Cross cohort integrative analysis for identifying heterogeneous and homogeneous gene signatures was conducted on schizophrenia and bipolar disorder. The computational efficiency test indicated that the dsMTL_iNMF ran in 34.9 minutes with 1,003 network accesses in total to infer the factorized matrices, with 7 min for each round of initialization with the hyperparameter rank number k=4 applied. As for the biological interpretation of the identified gene signatures, I performed an over-representation analysis. Shared gene signatures with 473 comorbid genes across schizophrenia and bipolar disorders were identified, while schizophrenia-specific signatures with 37 genes and bipolar disorder-specific signatures with 152 genes were extracted, respectively. In the result of enrichment, the shared gene signatures were significantly linked to pathways relevant across psychiatric disorders, such as transmission across chemical synapses and neuronal systems (**Figure 3.8 and Supplementary Table 3.1**). Previous evidence showed consistency with the neurobiological mechanisms found in this study, involving synaptic neurochemical systems, not only in schizophrenia and bipolar disorders, but in various major psychotic disorders such as depression [336–339].

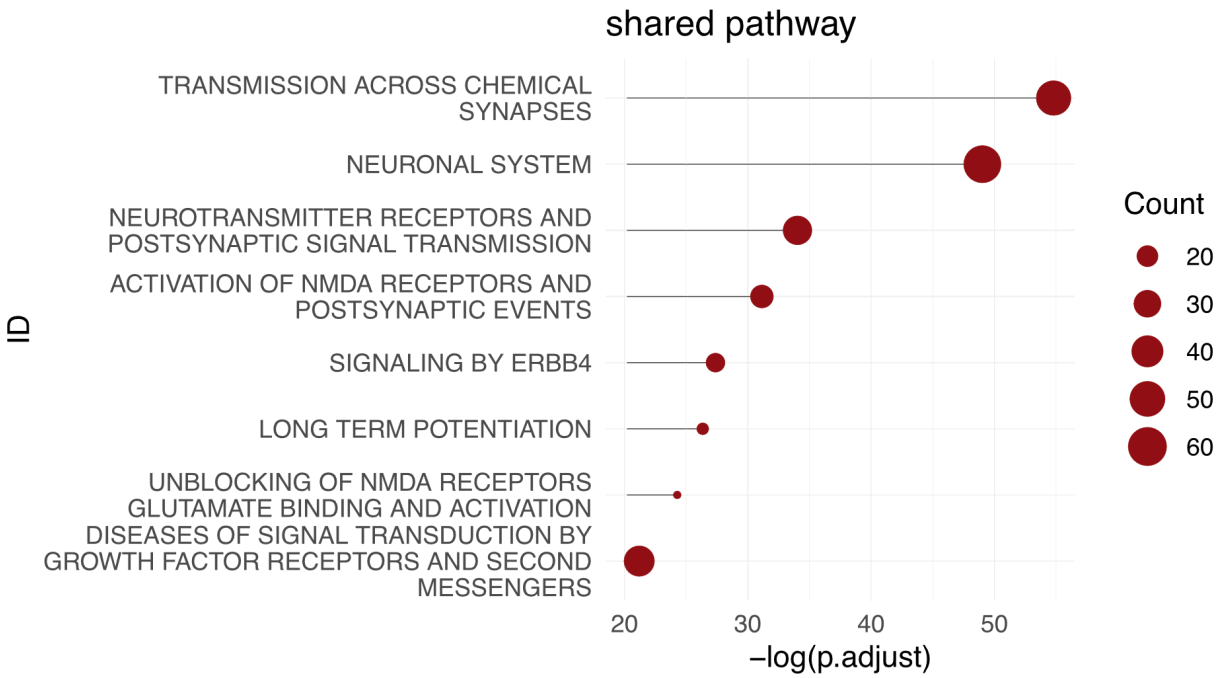


Figure 3.8 shared pathways across schizophrenia and bipolar disorder

For the schizophrenia-specific signature, nonsense-mediated decay (nmd) factor, and others. Nonsense-mediated decay factors were shown to be associated with various neuro-development disorders by influencing regulation of neural development, the differentiation of neural stem cells, cognition processes and synaptic plasticity (**Figure 3.9 and Supplementary Table 3.2**). The dysfunction of nonsense-mediated decay impaired differentiation decisions that led to nerve-related disorders. Previous studies observed that the impairment of NMD showed correlation with schizophrenia, but also amyotrophic lateral sclerosis, intellectual disorder (ID) and autism spectrum disorder (ASD). To our best knowledge, the nonsense-mediated RNA decay was hardly identified in the bipolar disorder research, compared to schizophrenia [340]. We also found that EIF2AK4 senses amino acid deficiency specifically responded to schizophrenia, though some of the mendelian randomization analysis also suggested that EIF2AK4 could be a druggable target to regulate stress conditions and alter the cellular homeostasis [341,342].

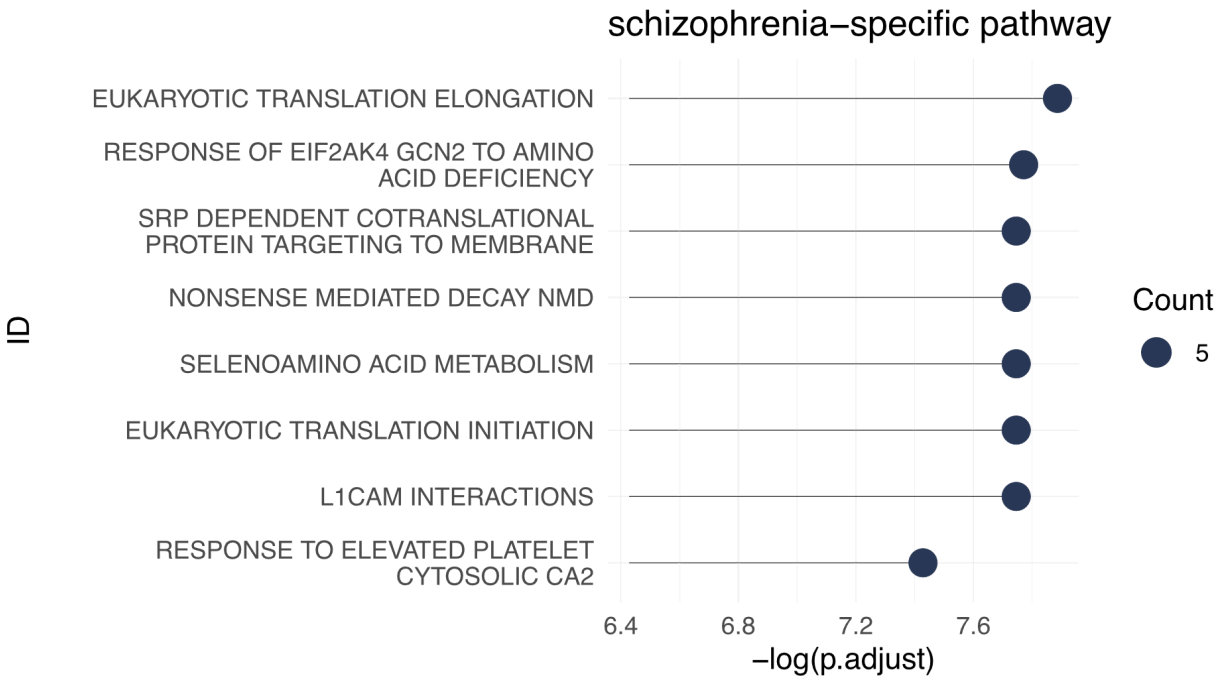


Figure 3.9 Specific pathways in schizophrenia against bipolar disorder

Additionally, the pathways linked to bipolar-specific signature included ion homeostasis and chaperon related processes, which have previously been associated with the illness (**Figure 3.10 and Supplementary Table 3.3**). Specifically, it was found that over three-fourth of the genes associated with bipolar disorders are related to ion dysregulation [343,344], however way less evidence could be found regarding the role in schizophrenia. Clinical trial evidence suggested that chaperones proteins from the endoplasmic reticulum stress response system in bipolar I patients were treated as putative targets as mood stabilizers [345].

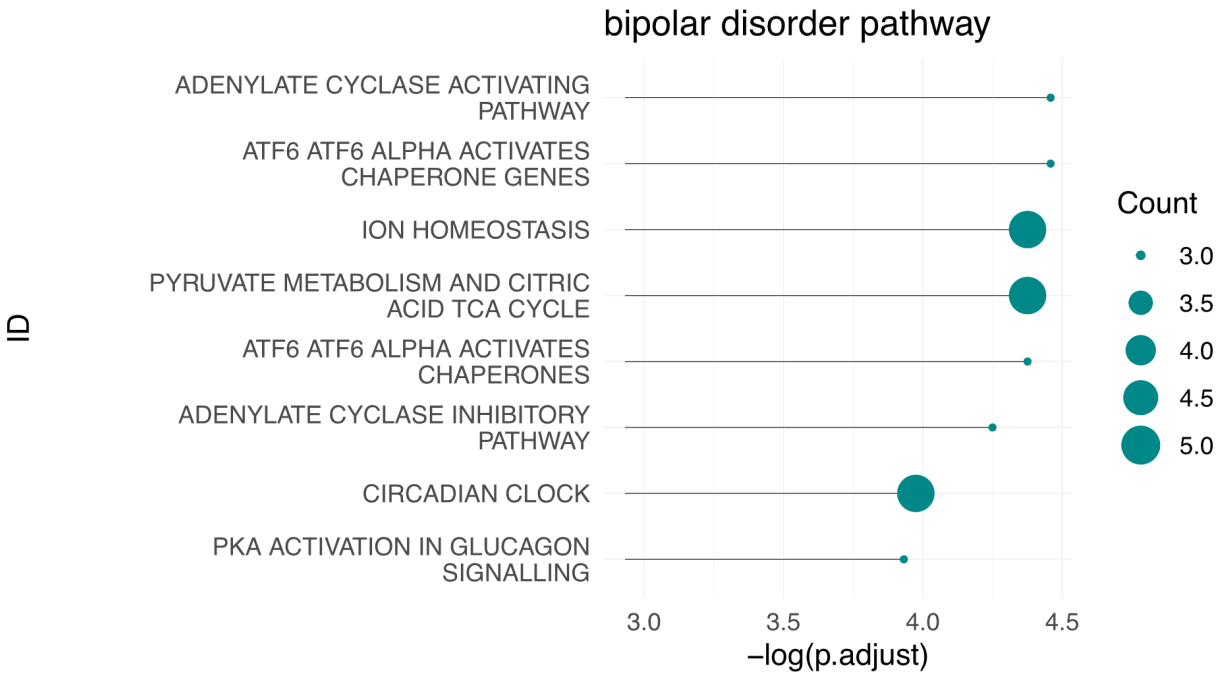


Figure 3.10 Specific pathways in bipolar disorder against schizophrenia

3.4 Discussion

In this chapter, an extended version of factorization-based method dsMTL_iNMF was introduced, that has also been integrated into a security-preserving, federated learning package in R, dsMTL, with DataSHIELD as an underlying framework for distributed data analysis. dsMTL facilitates research inquiries that pose challenges to conventional machine learning methodologies, particularly in identifying heterogeneous patterns and signatures across multiple cohorts. The specific federated designed framework further allowed the approaches in the packages to handle tasks across cohorts that were physically distributed in different locations. The deployment of a privacy protective infrastructure for the extensive and adaptive application of dsMTL_iNMF is crucial for its widespread adoption on a large scale. Through the utilization of such a distributed server configuration, our study exemplified the application and efficacy of dsMTL_iNMF in discerning biomarker and signatures under both simulation and real-data context. In scenarios where the biomarker patterns under investigation vary while sharing an overlapping set of features, conventionally employed machine learning algorithms may lack relevance. Notably, one application of dsMTL_iNMF

lies in comorbidity modeling, where the identified signatures delineate the shared biological underpinnings among multiple clinically comorbid conditions, complementing the other previously released approaches based on univariate statistics. Additionally, our examination revealed the actual communication of dsMTL_iNMF across multiple servers which were located in Heidelberg and Mannheim separately. In our analysis, it was observed that an increasing subject population correlated with a reduction in communication costs for federated MTL methodologies. This phenomenon may be attributed to the enhanced efficiency in identifying shared effects among tasks within a larger sample. Furthermore, our empirical analysis illustrated the ability of dsMTL_iNMF to capture biological meaningful gene signatures with cohort-specific and cohort-shared disentanglement. It was demonstrated that dsMTL_iNMF held promise in disentangling shared effects from cohort-specific effect, thereby suggesting its potential utility in comorbidity analyses and other applications, such as exploring biological patterns across diverse single-cell measurements. In terms of the language, R was used as the primary environment for dsMTL_iNMF and other approaches in the dsMTL package to ensure compatibility with DataSHIELD platform, but also for future extension with other practices in the DataSHIELD community. This concept allows flexible implementation for clients as well as future developments. Given the vast array of published packages available on the comprehensive R archive network, our development strategy favors a broad adoption of dsMTL.

Being a distributed system, dsMTL_iNMF faces several challenges. These limitations also commonly exist in the field of federated learning. In the first place, the overheads during the data transfer and data communication are the bottlenecks in most of the federated learning methods. Previous work tried to tackle this limitation by either developing novel algorithms involving data compression or by optimizing the data transmission between client and remote server [346,347]. A further issue is the heterogeneity of the systems and data stored in the server that might affect the performance of the federated learning models, where solutions such as FedAvg-based methods might solve these limitations [348,349]. Moreover, the attacks on central server or local client server or even by any user inside the system could still occur even with the well-establishment of the infrastructure. In this study, the new server key mechanism was introduced to prevent the situation to a large extent. However, awareness and attention should still be made in the further development. For instance, one solution is to complicate the privacy protection strategy in servers by providing different safety protocols on different servers [350]. On the other hand, defense on data poisoning attacks should be

prepared, where this type of attack imposes the contaminated data to the data server or other client servers [351].

Future developments of the federated learning approach address different aspects. For efficient computation, the implementation of asynchronous gradient descent could provide a faster convergence to the model approximation. Alternative computing language and machine learning framework could be considered such as python based tensorflow and pytorch that largely extend the potential application of DataSHIELD system, as well as bottom level C++ based framework which substantially increase the computation efficiency. Regarding the adaption of the federate learning to other algorithms, recent interest guided the researcher towards the field of one-shot federated learning, a method where the final model can be tuned and optimized in only one single round of data transfer and communication which largely released the potentiality of federated learning application due to the previous mentioned overheads limitation [352]. A distilled one-shot distributed learning approach was proposed to address the communication overheads by aggregating the gradients to bulk [353]. Factorization approaches inspired by this concept of one-shot learning were developed, such as FedMF with stand matrix factorization and Fedsplit with joint NMF on federated recommendation systems and [354,355]. Incentive mechanisms are also crucial for encouraging device participation in federated learning. While current federated learning approaches operate under the assumption that devices will cooperate without considering rewards, practical scenarios necessitate economic compensation for participation. Furthermore, asynchronous federated learning with training and data transfer running asynchronously is essential to accommodate systems and magnitudes of heterogeneous data, including already published works such as VAFL and SAFA [356]. Unlike FedAvg, which operates synchronously, asynchronous averaging techniques can adapt more client servers and perform parameter updates with time gaps. Asynchronous factorization models such as ADMF (asynchronous distributed matrix factorization) was also developed for collaborative filtering [357]. To address the constraints on widespread adoption due to the asynchronous arrival of parameters from client servers, blockchain-based frameworks for federated learning have been proposed by various works, including BAFL, PPFchainm, HBFL and block-chain based matrix factorization BMF [358–361]. These frameworks leveraged the decentralized nature of blockchain networks, enabling devices to collaboratively learn without the need for a central aggregator.

To conclude, the dsMTL_iNMF and dsMTL federated learning package was developed to provide a versatile and extensible framework for privacy-preserving and distributed analysis across multiple data institutions that were physically located remotely. dsMTL_iNMF benefited from the standard integrative factorization, with the ability to distinguish cohort-specific and cohort-shared effects across multiple cohorts. The case study proved its practical application on comorbidity modeling and clinical outcome prediction from different sources, as well as the future potential application on large-scale biomedical collaborations.

Chapter 4

Project 3: Genotype-to-function association prediction based on extended application with interpretable factorized autoencoder architecture

4.1 Introduction

Genotype data embed the comprehensive genetic information characterizing an individual's genome, detailing the allelic composition at various loci regions across their DNA. It serves as a comprehensive record of the genetic variations present within an organism's genome, encompassing a diverse array of alterations ranging from single nucleotide polymorphisms (SNPs) to larger structural variants [21,362]. These variations, scattered across the genome, contribute to the genetic architecture that brought heritable influence as well as phenotypic patterns of human populations, including patterns of genetic diversity, population structure, and evolutionary history [25,363,364].

Generally, genotype data was used to find associations between genetic variations and traits, thereby unraveling the complex genetic architecture governing biological diversity and disease predisposition [365,366]. Specifically, genome-wide association studies (GWAS) represent a powerful analytic tool for uncovering genomic associations by systematically comparing the genotype data of individuals with a particular phenotype (e.g., disease status, trait, etc.) to that of healthy controls [367]. By identifying genetic variants susceptibility, GWAS facilitates the downstream analysis such as the understanding of transcriptomics, proteomics or biological processes, the discovery of potential targets and biomarkers that can be further developed for disease diagnosis, prognosis, and treatment. Genome-wide association studies were broadly used in a wide range of trait specific studies, such as those complex disorders like type 2 diabetes, cardiovascular diseases, metabolic syndrome and psychotic disorders [82,368–370]. One commonly-used tool for GWAS is called PLINK [371]. To perform, GWAS generally involves three major steps including, quality control to filter out the SNPs and individuals, association test and post-GWAS analysis. Quality control cleaned the original dataset due to issues related to existence of missing values, linkage disequilibrium (LD) where redundant SNPs at the same chromosome with high correlation required to be pruned, etc. The subsequent association test performed linear regression to estimate association between SNPs and specific phenotype, followed by multiple testing correction and generated the summary statistics that record the results for each SNPs such as effect size that represents the magnitude of genotype-phenotype association, P value, etc. After that, the post era of genomics research, also known as post-GWAS analysis (pGWAS) was conducted advancing pace from discovering association variants with phenotype to additional objectives, including fine-mapping identification of lead SNPs that are with most significant functional influences on the phenotype, calculating the polygenic risk score (PRS) using effect size for summary statistics to quantify the heritable risk of a specific phenotype into a score. In addition, another important type of post-GWAS analysis to build further association to enhance the interpretability as well as annotation of the variants, such as inferring genotype-transcriptomics-phenotype association in transcriptome-wide association studies (TWAS), etc. (**Figure 4.1**).

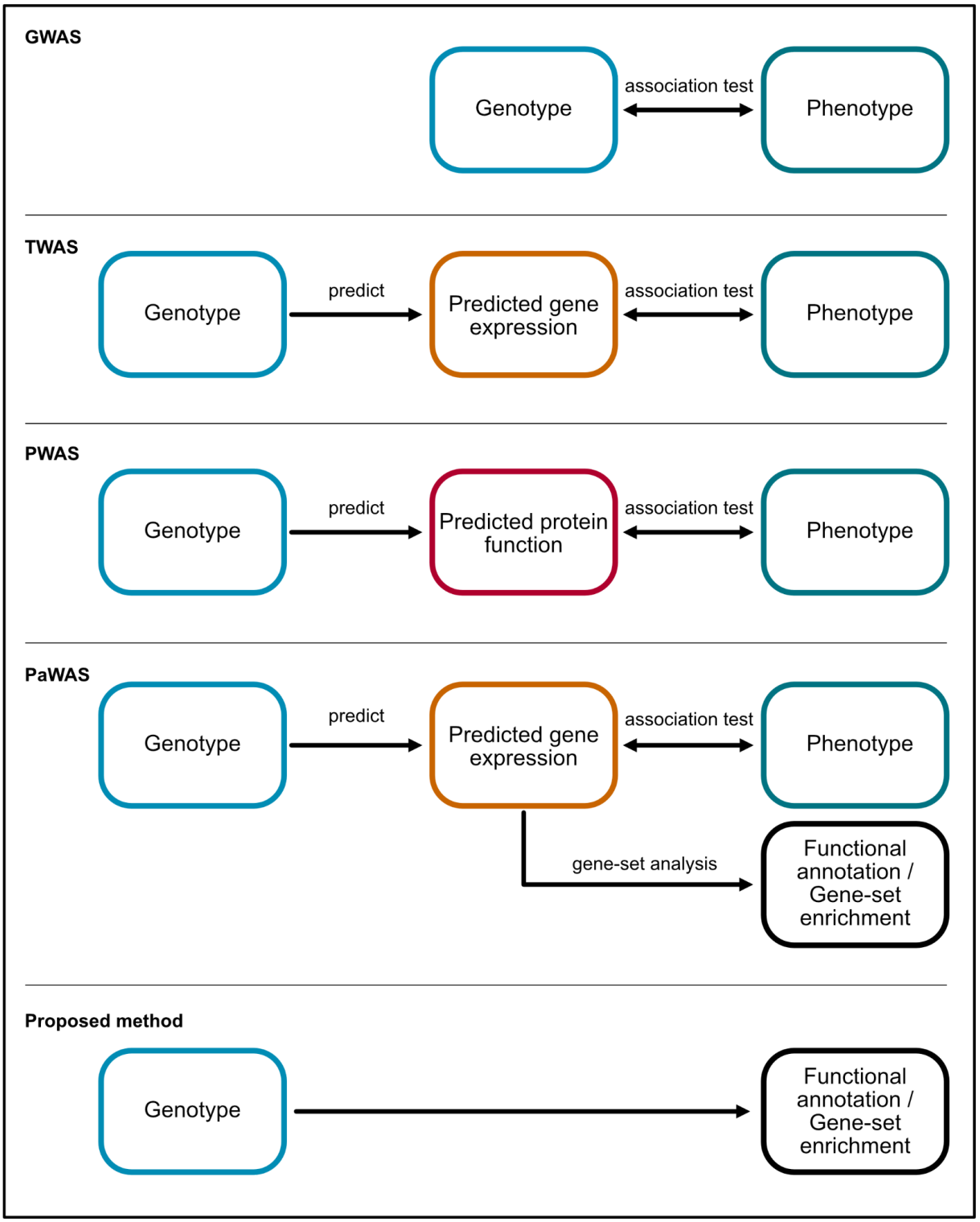


Figure 4.1 Examples of post-GWAS analysis to build further association to enhance the interpretability as well as annotation of the variants, including GWAS, TWAS, PWAS, PaWAS (gene-set analysis on GWAS) and applied method in this study

Association studies in post-GWAS analysis have been developed to combine the original genotype-to-phenotype prediction with other modalities such as transcriptomics and proteomics, in order to improve the annotation of the variants as well as to better understand the relationship between genotype and phenotype. For instance, transcriptome-wide association studies (TWAS) and proteomics-wide association studies (PWAS) represent two innovative approaches in the field of genomics and bioinformatics, aimed at elucidating the linkage between transcriptomic profiles and complex traits and phenotype [372,373]. TWAS leverages transcriptomic data to uncover associations between molecular expression patterns with phenotypic traits of interest. Unlike genome-wide association studies (GWAS) which focus on genetic variants, TWAS integrates information from gene expression data obtained from diverse tissues or cell types to identify genes whose expression levels are correlated with specific traits or diseases. By leveraging large-scale transcriptomic datasets, TWAS enables researchers to prioritize candidate genes and pathways underlying complex phenotypes, thereby providing valuable insights into the molecular mechanisms driving disease susceptibility or phenotype variability. Similar to transcriptome-wide association studies (TWAS), proteomics-wide association studies (PWAS) aim to elucidate the relationship between protein abundance or modifications and complex traits or diseases. PWAS harnesses high-throughput proteomic technologies to systematically interrogate the proteome across diverse biological samples, such as tissues, cells, or biofluids, in association with phenotypic traits or clinical outcomes. In addition, another novel concept of pathway-wide association studies PaWAS (or so-called gene-set analysis of GWAS) extends the concept of TWAS and PWAS by considering the collective behavior of genes organized into biological pathways or functional modules [374]. Gene set enrichment analysis (GSEA) and over-representation analysis were employed in PaWAS to identify pathways that are significantly associated with the trait of interest and controlling for multiple testing [375,376]. PaWAS extended the GWAS and TWAS to a more complex and advanced genotype-function-phenotype association study, which has emerged as a valuable tool in identifying the activity or dysregulation of biological pathways and phenotypic traits, offering a holistic view of the molecular underpinnings of complex traits or diseases. By aggregating information from multiple genes within a pathway, PaWAS enhances statistical power and enables the identification of biological processes that are perturbed in disease states. In practice, PaWAS have shown promise in elucidating shared genetic pathways and identifying potential therapeutic targets for a range of complex traits and diseases [374,377]. PaWAS have been extended to pathway-based analysis to identify potential therapeutic targets for schizophrenia with

symptom persistence after treatment with antipsychotics [378,379]. Additionally, PaWAS have been applied to plasma proteome analyses in individuals, revealing insights into target genes, variants, tissues, and transcriptional pathways [380]. Association in the complement systems related to coronary artery disease was also explored via PaWAS [381]. Furthermore, by integrating TWAS, PWAS and PaWAS, researchers have identified novel drug targets for epilepsy and depression, shedding light on the shared pathways between these conditions [382].

However, most of the previously mentioned methods had their own limitations, for instance, the analysis of association studies were complicated by the mixture effect including both linear and nonlinear, which were embedded in the genomic data. The method for association studies were designed for detecting simple but direct signals from the genetic architecture, which were useful but not sufficiently sensitive for identifying the genetic components in the complex syndromes. On the other hand, GWAS was developed for computing the association between genotype to phenotype. Though it was further extended to other modalities such as transcriptomics, proteomics, etc, the fundamental processing were similar, generally by first mapping other modalities to genes and then back to genotypes. This step-by-step mapping procedure might be less straightforward and involve many manual operations, compared to the alternative strategy to directly build the association between the genotype to the modality of interest (**Figure 4.1**). Recently, deep learning based models incorporating various architecture were developed to study the molecular to function association. One of the solutions was to introduce additional interpretability into the model due to the fact that the standard deep learning model provided limited insight for associating the learnt feature to other more easily understandable modalities. Efforts were then made to improve the interpretability of the model. For instance, linear modification on the original VAE structure allowed the researcher to better understand the feature association to pre-defined knowledge such as VEGA, DCell, and gene ontology autoencoder (GOAE) [383–385]. A recent more advanced model, namely OntoVAE, was developed to integrate hierarchical biological ontologies as prior information with variational autoencoder, to guide the model to establish direct associations between the molecular expression and biological processes as well as to enable the interpretation ability that was originally not available in the standard autoencoder structure [386]. As for the other limitation, the unsupervised characteristics hindered the path of standard VAE to factorize and disentangle the specific effect embedded together to complex signals respectively. One successful attempt in a different area of drug screening,

compositional perturbation autoencoder (CPA) was developed to predict cell level responses to cellular perturbations, such as genetic, environmental or drug perturbations, which also an representative implementation to add the interpretability to the restricted unsupervised learning approaches [387].

In this chapter, the main goal is to extend the novel application on formerly-developed supervised machine learning architecture - factorized interpretable variational autoencoder, namely COBRA, to enable direct prediction of the genotype-to-function association. To approach this goal, my four contributions were made, including 1) development of the pipeline to preprocess the genotype data to model-adaptive format; 2) integration of summary statistics information into the model by applying the weighted loss mechanism; 3) genotype specific domain adaptation by training and fine-tuning the original model architecture; 4) analysis and evaluation on the genotype-to-function prediction results. This study was performed from the aspect of genotype and biological function, to improve our understanding of schizophrenia which is a complex disorder complicated by many potential effects besides diagnosis, such as lifestyle, gender, etc. Therefore, the model architecture implemented here inherited ideas from ontoVAE which is a variational autoencoder with biological functional interpretability and CPA which is able to isolate the contributions of covariates and remove the unwanted effect. The hybrid model adopted linear structure in the latent space and applied a flexible deep learning structure to learn factorized embedding representing different sources of variations. Since the raw genotype data only contained information such as what allele existed in the location rather than the type of the allele such as whether risk allele or not as well as the effect size or other ratio to quantify, additional reference summary statistics would be needed, same as the GWAS analysis. Therefore, in our study, by introducing a novel weight loss function, the significance of every genetic variant was weighted based on their effect size from the reference GWAS summary statistics. With genotype data instead of gene expression as model input, along with variables of interest as well as covariates, the model decomposed the data into a collection of factorized embeddings associated with the diagnosis effect, gender effect, residual effect or other possible covariates depending on the cohort. By means of the discriminator classifier, the approach was able to ensure these embeddings were independent of each other, and further recombined these disentangled effects to the total effect. We demonstrated the usefulness and extended feasibility of this factorized model on direct genotype-to-function prediction with two reference GWAS summary statistics, PGC

and FinnGen, as well as on identifying different sources of variation such as diagnosis groups and covariates during model training and testing.

4.2 Methods

4.2.1 Material and datasets

The model received two genotype inputs including target (query) dataset that contained the raw genotyping allele data and reference dataset (GWAS summary statistics) with effect size for each SNPs provided. The target (query) dataset Genome-Wide Association Study of Schizophrenia (dbGaP Study Accession: phs000021:phg000013, phs000021:phg000014) was selected to find susceptibility genes for schizophrenia, retrieved from The Genetic Association Information Network (GAIN) [388], which originally composed of more than 90 studies covering multiple diseases from oncology, neurology to complex disorders with measurement of 500,000 single nucleotide polymorphisms (SNPs) accounting for four-fifth of the risk variants in human genome. The target dataset was composed of European-American (EA) ancestry with General research use (GRU) cohort with 2659 individuals including 1217 EA cases and 1442 EA healthy controls. For the reference dataset, two summary statistics specifically on schizophrenia traits were used, GWAS summary statistics from the Psychiatric Genomics Consortium (PGC) and from FinnGen [389,390].

4.2.2 Genotype data preprocessing

The preprocessing of genotype data in target dataset was performed with the specialized tool PLINK 1.9, a package for genome association linkage analysis, which involved several stages [371,391]. In the first step, the variant IDs were converted to rs_id (src.sh with PLINK) due to the inconsistency of annotation across different cohorts, platforms and methods. Linkage disequilibrium corresponded to the commonly-seen condition that linkage of different genetic variants at different locations of the genome existed, bringing the confounding effect to the whole genomic analysis as well as difficulty of distinguishing the real causal variants responsive to the phenotype [392]. To control the collinear effect, clumping was performed to filter SNPs with the linkage disequilibrium (LD) information ($r2<0.1$) and kept the remaining

independent variants in the second step. Next, we removed duplicated and undefined SNPs based on PLINKs and converted the odd ratio (OR) to log format. To be noted, in the PGC reference dataset, odd ratio was used originally to measure the effect size whereas in the FinnGen reference dataset, beta values were applied. Thus to stay consistent, we performed additional transformation on the beta values in FinnGen as beta is equivalent to log-transformation of odd ratio [393,394]. In the next step, we mapped the SNPs to the gene symbol using PLINK's internal function. In the end, we processed the .tped files to align each participant with the allele status and calculated the score for each SNPs. Here, we mapped the heterozygous, homozygous with no major alleles and homozygous with two major alleles to 0.5, 1.0 and 0.0 respectively.

4.2.3 Variational autoencoder (VAE)

The variational autoencoder is the advanced development of autoencoder in probabilistic form, which composes of an encoder to compress the high-dimensional input data to latent embedding with a lower dimension and a decoder as the 'reverse' implementation of the encoder to reconstruct the latent embedding to the output with the same dimension as original input data matrix (**Figure 4.2**).

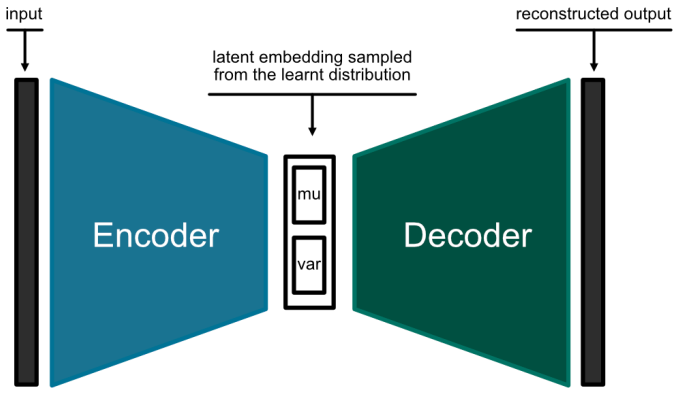


Figure 4.2 The model architecture of a standard VAE

Different from the autoencoder that transforms input to the latent embedding in a numeric form, variational autoencoder learns the distribution parameters (e.g. mean and variance in Gaussian distribution) in the latent space and samples from this trained distribution to latent embedding for the decoding process. To compress the data, encoder and decoder generally

contain linear models for dimension reduction, followed by activation functions (e.g. rectified linear unit ReLU, hyperbolic tangent Tanh, sigmoid, etc.) for learning nonlinear patterns of the data. The objective of variational autoencoder is to maximize evidence of lower bound (ELBO) as the loss function shown below, which contains two parts including reconstruction loss and the Kullback-Leibler (KL) Divergence [395].

$$L(X) = L_{MSE} + D_{KL}$$

To be specific, the above loss function can be transformed into evidence of lower bound form as the below formula where the first term is used to optimize the likelihood of the input data given the model parameters whereas the second term represents the distance between the true posterior distribution (Gaussian in this model) and the predicted posterior distribution. In order to recast the backpropagation properly, given the existence of sampling steps in the training process, reparameterization trick is used [182].

$$L(X) = E_{q(Z|X,\theta)}[\, logp(X|Z,\theta)\,] - KL(\, q(Z|X,\phi) \,||\, p(Z|\theta)\,)$$

4.2.4 Model architecture of interpretable factorized VAE

The COBRA model combined two main model structures, including a ontoVAE interpretable VAE using prior ontologies knowledge with an encoder and a factorized additive latent space to decouple covariate-specific effects from the biological effect of interest (such as disease effect). In **Figure 4.3**, we illustrated the basic structure of OntoVAE on the left, which is a VAE with prior biological knowledge using a fully connected encoder and hierarchical ontology-like structure decoder and the CPA model on the right that has the capability to isolate variation of different sources in the latent space. In OntoVAE, the decoder was built as a directed acyclic graph representing the Gene ontology structure or Reactome, where the root terms (nodes closed to the latent space side) represented broad biological concepts and child terms (nodes closed to the reconstructed output side) represented more specific and detailed terms [396–404]. In CPA, the latent space is represented as a linear additive embedding which corresponds to the residuals, the variable of interest and the other covariates provided. And in the interpretable factorized VAE used in this study (**Figure 4.4**), following the standard encoder where genotype data input passed through the encoding layers, the latent space was

represented as a series of linear additive embeddings, which was achieved by introducing a discriminator classifier to disentangle against the overall embedding encoded by the model. After that, we summarized by adding all factorized embeddings into one overall embedding that represented the total effect and passed the overall embedding to the decoder of which each layer was designed to represent one depth level of the GO ontology. During the process of decoding, embedding would walk through each ontology term and the corresponding weights would be learnt, from the processes with board concepts (e.g. system development) to the terms with increasingly specific and detailed biological meanings (e.g. peripheral nervous system neuron differentiation). Full details of the architecture (e.g. layer structure, the number of nodes, etc.) can be found in **Appendix Supplementary Table 4.1**. The objective function of the interpretable factorized VAE has two components including the VAE loss L_{vae} and adversarial loss L_{adv} as the below formula showed, where L_{adv} is the cross entropy loss to measure the prediction (or classification) on diagnosis labels d_i or covariates labels c_i (e.g. gender, etc.) with residual embedding $z_i^{residual}$ in the adversarial way.

$$L(X) = L_{vae} + \lambda L_{adv}$$

$$where \quad L_{adv} = CrossEntropy(f_d^{adv}(z_i^{residual}),\ d_i) + CrossEntropy(f_{c_i}^{adv}(z_i^{residual}),\ c_i)$$

The optimization of the model contains two parts, involving VAE optimizer and adversarial optimizer. Two step optimization was implemented at every iteration, first the optimization of VAE updated the weights on encoder, decoder and the covariates embedding, and then the adversarial optimizer updated the weights in the discriminator classifiers.

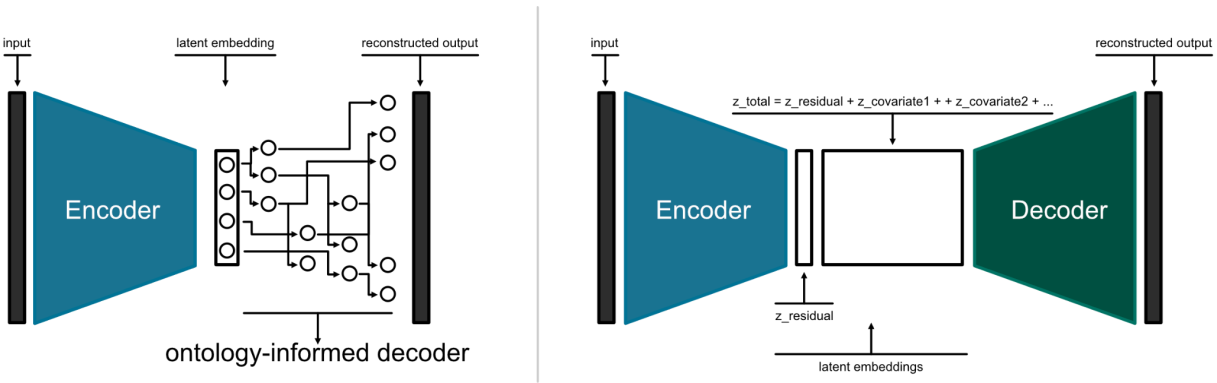


Figure 4.3 Model architecture of ontoVAE (**left**) and CPA (**right**)

4.2.5 Weighted loss objective function

In general, weights could be incorporated to sample level in the model training to assign the importance of different training entities. However, contribution on feature levels was easily neglected in previous studies. The objective was to allow the inclusion of input node weightings for the reconstruction loss of training based on the hypothesis that the meaningfulness of the latent values generated by the model could be improved by weighting the contribution of each input node in the original loss function. The weighted loss was implemented in a previous validated study on four subsets of the full GTEx RNA-seq expression dataset by our collaborator David Hirst and further explored the extensive application on genomic dataset in this study [405]. To implement, a curated weight file that included each feature name (SNPs) and the corresponding weight values were generated. The log OR indicates the importance of a SNP in defining the phenotype of interest. Hence, our idea was to use this measure of importance to weight the contribution of each input feature in the loss function. Here, weights for each SNPs were retrieved from reference GWAS summary statistics, where log-transformed odd ratio (effect size of each genetic variant) were applied. The log-transformed odd ratio indicates the importance of a SNP in defining the phenotype of interest. Hence, the idea was to use this measure of importance to weight the contribution of each input feature in the loss function. In practice, during the training, the weighted reconstruction loss was computed at every iteration by conducting a matrix product on the original reconstruction loss matrix with the weights vector before summing up the loss as the formula shown below. The first formula is the standard loss function of a variational autoencoder with reconstruction loss in the first part and Kullback-Leibler (KL) Divergence in the lateral, in which *D* refers to the dimension of the features. In the weighted loss function indicated in the second formula, weights w_d was incorporated into the reconstruction loss calculation and assigned the importance to each feature.

$$L(x^{(i)}) = \sum_{d=1}^{D} (x_d^{(i)} - x'^{(i)}_d)^2 + D_{KL}$$

$$WL(x^{(i)}) = \sum_{d=1}^{D} w_d (x_d^{(i)} - x'^{(i)}_d)^2 + D_{KL}$$

To evaluate the difference between the standard loss and weighted loss, we compared the interpretation results and model consistency with and without the weighted loss implementation in further analysis.

4.2.6 Mapping between genotype and biological processes

The preparation of ontoVAE required us to provide a mapping file between the input modality and the modality that is used for interpretation. In the original ontoVAE, the input modality and the interpretable modality were gene and GO term biological processes, while in our study, genotype and GO term biological processes are needed instead. Therefore, the genotype-ontology mapping was modified based on the ontologies file from the ontoVAE which is originally a gene-to-ontology mapping. Ontology file (go-basic.obo) was retrieved from geneontology.org together with the GO annotation file (goa-human.gaf) [406]. Terms under the 'biological process' category were kept. A threshold between 30 on the bottom and 1000 on the top was applied to filter terms with too many or too few connections, followed by mapping and annotating the Ensembl IDs with HGNC symbol, leading to the generation of the gene - gene ontology mapping. To further align the genotype with ontologies, we utilized the SNPs - gene mapping created from the previous step by using PLINK internal function. After matching the SNPs and genes, the genotype - gene ontology alignment was generated.

4.2.7 Model implementation, training and testing

The whole model implementation involved two parts, the model preparation and model training. For model preparation, genotype - ontology mapped object was first created as .ontobj file format following the previous section **Method 4.2.4**, genotype data with sample-wise clinical information was packed as .h5ad format, and reference dataset that contained feature-wise SNP effect size was prepared.

For model training, given a transformed genotype data matrix, variable of interest (disease diagnosis, either schizophrenia or healthy), the covariate information (gender), and the effect size for each SNP from the reference dataset as inputs, model training and testing were implemented. Considering a genotype dataset with SNPs as features and individuals as

samples, training and testing samples were divided. For the training phase, the data matrix was passed through the encoder to the latent space. Within the latent space, different embedding vectors were learnt respectively. The diagnosis group embeddings and covariate embeddings were projected from the one-hot encoding transformation, meanwhile the residual embedding (also known as basal state embedding) were learnt through a discriminator distinguishing the other signals from the diagnosis and covariate effects. Here, the residual embedding represents the embedding that has captured variations that contribute to neither diagnosis effect nor all the covariates effect. Therefore, in order to obtain this residual embedding, a discriminator or so-called classifier was trained in an adversarial manner that held the objective to not be able to classify the diagnosis labels and covariate labels correctly with the residual embedding as input. Thereby, the learnt residual embedding would be independent of all the other provided variable embedding vectors. In this phase, the process can be regarded as a decomposition or factorization to linearly separate the variance of interest as well as to denoise in a supervised way. Finally, the simple linear addition of all the factorized embeddings was computed and passed to the decoder layers, which was an ontology-like structure illustrated in the ontoVAE method. For the test phase, the weights of the architecture would be fixed while test data was forwarded from the input encoder to the reconstructed output in the decoder. The collections of latent embedding represented different variables, including diagnosis, gender, residuals, etc.. The exploration of both the latent space and the node output in the ontological decoders could be further made in the downstream interpretation stage.

4.2.8 Retrieval of the responsive biological processes activity from the decoder

To quantitatively interpret the functional association learnt from the genotypic input, the pathway activation values were computed following the algorithms provided in ontoVAE.

The pathway activation was used to represent the strength or extent of the response of biological process terms; in other words, to evaluate what and how the processes were activated in the group of individuals in a genomic cohort. Given the ontology-like structure of the decoder, we were able to associate each node of the decoder with a specific GO term. Therefore, the pathway activity scores were calculated from the activated output of the neurons retrieved by the pytorch attach hooks function when the model ran through samples

from a train or test set. To specify, here each factorized latent embedding would correspond to one pathway activity score matrix where rows are samples and columns are GO terms. In this study, four pathway activity score matrices were generated that represented the pathway activation induced by different variations, including diagnosis, gender, residual, and total effect. Therefore, in order to find out which biological process contributes most to explaining the variation of interest, the pathway activity score matrix that corresponded to the diagnosis effect was extracted. In the next step, a random forest classification was performed on this pathway activity matrix using sklearn for each term separately given the test set and computed the impurity-based feature importance, also known as Gini importance. In the end, all the terms were ranked based on their importance values, where the terms at the top represented the most significant and contributing processes that could explain the input genotypic cohort.

4.3 Results

4.3.1 Novel extended application to genotype-to-function prediction based on former-introduced interpretable factorized model

In this study, an advanced deep learning model was used, combining factorized latent space and the variational autoencoder, with the novelty of combining ideas from factorized autoencoder such as CPA, factorVAE and beta-VAE (introduced in **Chapter 1.4** and **Chapter 4.2.4**) to identify disentangled representations on complex data with multiple sources of variation and the interpretable autoencoder such as ontoVAE and GOAE (**Chapter 4.1** and **Chapter 4.2.4**). Here, the application of the method was further extended to build up the genotype-to-function linkage and to directly predict the association between genetic variants and biological processes. Given a genotypic dataset with covariates, the factorized interpretable autoencoder first passes the genotype data through the encoder into the compressed-dimensional state. Then, the lower-dimension data is linearly factorized into a series of embeddings simultaneously, corresponding to its diagnosis-associated, covariate-associated, and residual-associated effect.

Specifically, the residual embedding was learnt via a discriminator with diagnosis- and covariates information provided, resulting in a latent representation in which the influence of

all variables and covariates has been removed. By factorizing the total effect, variable-specific effects could be extracted for further investigation. After the summation of the latent vectors, a total embedding is built and passed to the decoder, which integrates the ontology structure in which every node represents a specific biological process term. The weights of each node in the interpretable decoder correspond to the strength or magnitude of that process. Using the prior knowledge, the transformation from genetic variants to functions that were more easily understandable was achieved directly during the training. In addition, the constraint on residual embedding also allowed the improvement of the prediction performance for complex signals compared to the previous method that used a classifier for competition to disentangle the variance of interest [407].

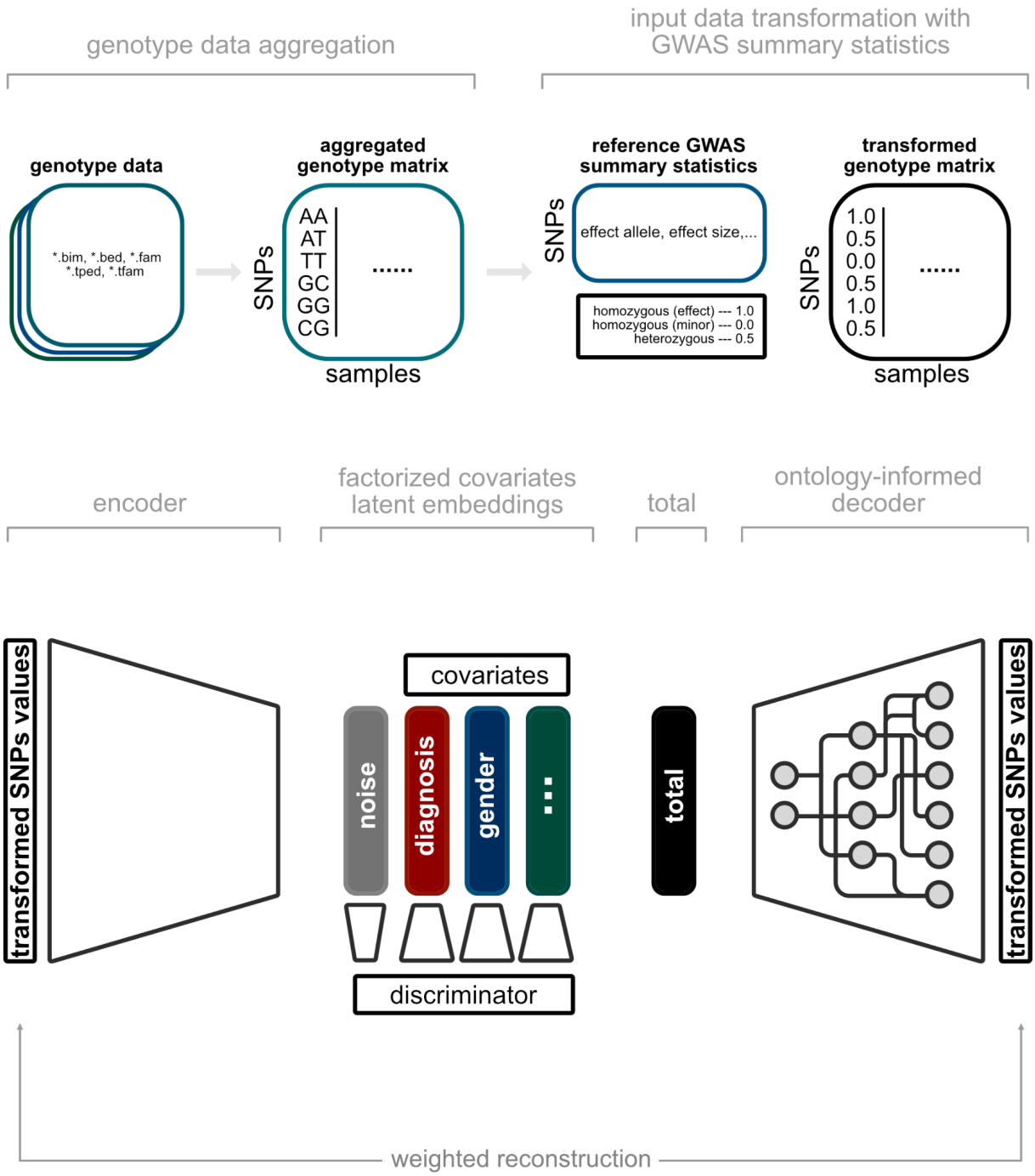


Figure 4.4 The preprocessing and application of genotype data on the model architecture of interpretable factorized hybrid variational autoencoder

4.3.2 Elucidation of factorized latent embeddings learnt by the model

To evaluate the performance of the disentanglement, we explored the outcome of decoupling the diagnosis and the covariates (age, gender,...) from the total effect given the schizophrenia

genotype dataset. We trained our model and passed the test set through the architecture with the learnt weights fixed. Then, to assess the actual performance of disentanglement of variance from different sources including diagnosis, gender, and residuals, we performed a UMAP projection to reduce the dimension for visualization. In schizophrenia, gender effects were frequently observed, and numerous studies have demonstrated the difference in symptoms, behavioral patterns, as well as molecular factors, introducing complexity in diagnosis [408–411]. As shown in **Figure 4.5** using the trained model with weights from PGC as an example, a set of learnt embeddings including four latent representations of different factors, including z_residual, z_diagnosis, z_gender, and z_total, were observed. Specifically, each dot represents one individual in the cohort and is labeled with variables in each subplot. It is clear that each latent representation succeeds in distinguishing the corresponding variance from the complex effects underlying the genotype dataset. For instance, the z_diagnosis embedding and z_gender disentangle the variable-specific effect and separate it from other unrelated variances. The additive effect could be observed in the z_total embedding, in which each cluster contains one combination of the input variables. This shows the denoising effect in this model architecture, in which the z_residual captures the residual variance specifically. Overall, the factorized model identified the compositional effect from sources of interest and, at the same time, was capable of extrapolating and removing unexpected residuals from the data.

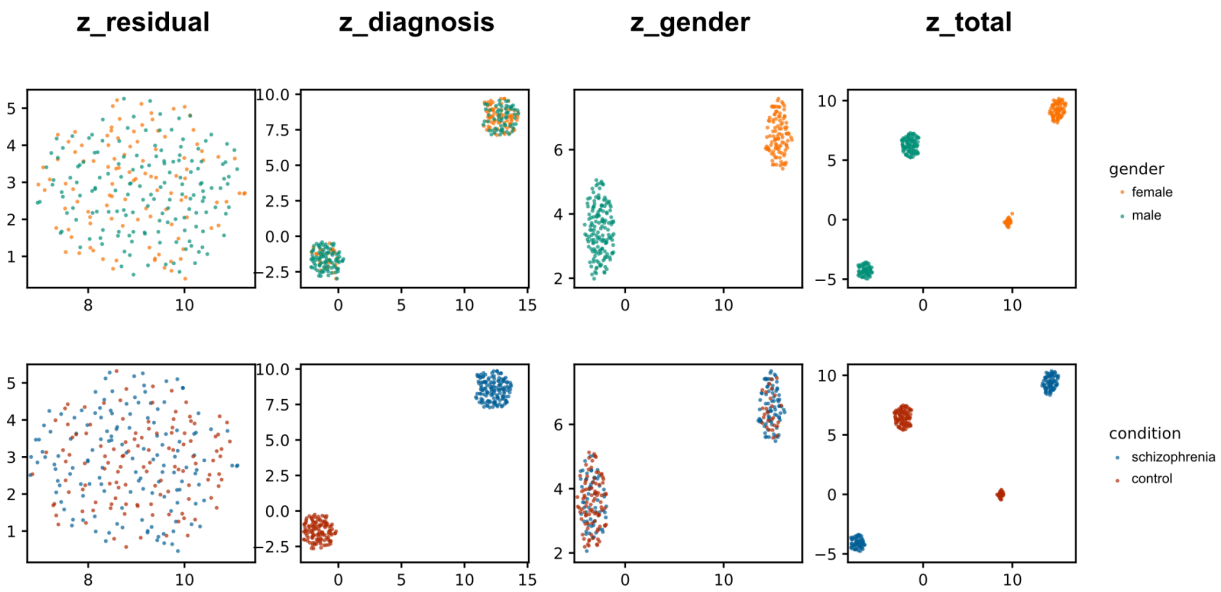


Figure 4.5 UMAP illustration of the factorized embeddings including residual embedding (z_residual), diagnosis embedding (z_diagnosis), covariate embedding(z_gender) and total embedding (z_total)

4.3.2 Interpretation and quantification of the biological process activities from trained decoder

To further evaluate the interpretability of the hybrid model, we investigated the genotype-to-function predicted outcome by computing the pathway activity scores (see **Method 4.2.6**). Random forest classifiers were further trained and importance scores of each term were calculated to rank the biological processes with high contribution to explain the diagnosis-associated variance (**Figure 4.6**). From the prediction results with the model trained with PGC weights, general processes related to the nervous system process were identified, including regulation of postsynaptic membrane potential, regulation of calcium ion transmembrane transport and behavior [338,412,413]. Other contributing processes such as regulation of interleukin 2 production were identified, which was found to be significantly relevant to cognitive deficit and other negative symptoms in schizophrenia [414]. Interestingly, it was also observed that immune response processes were recognized which were found in our previous comorbidity modeling study in Chapter 2. For the model trained with FinnGen weights, nervous system related processes were consistently observed. For instance, as one of the representative mechanisms in the pathology of schizophrenia, synaptic processes were highlighted as well (**Figure 4.6**). In addition, neurogenesis, hippocampus development and system process regulation were found on the top of the list, among which neurogenesis, in particularly, the neurogenesis in hippocampus area was found as one of the predominant role of learning and memory impairment in schizophrenia with numerous evidence of support [415–417]. Other factors such as levels of cytokine and mitotic cell cycle activity in cell dynamics were also observed altered in several studies [418–420]. To compare the results obtained from model with different summary statistics as weights (PGC and FinnGen), consistent but also distinguishable activation of specific biological processes were identified, which is to some extent similar with the conventional genome-wide summary statistics study (GWAS) analysis where the query data and reference data were required to predict the association between genotype and phenotype. Depending on the summary

statistics incorporated, the factorized interpretable VAE optimized for genotype-to-function prediction, regularized and constrained the training towards the specific optimal states based on the given prior weights information. Altogether, our analysis demonstrated the preliminary results of the factorized interpretable model by leveraging genotypic data to predict the functional outcome. It provided the evidence of feasibility of using factorized interpretable model to generate genotype-to-function association straightforwardly rather than the traditional gene-set analysis of GWAS with multiple steps (**Figure 4.1**), by utilizing the variational autoencoder structure as the backbone, along with the linear additive factorized latent space for isolation of variations of interest as well as ontology-informed decoder to enhance model interpretability.

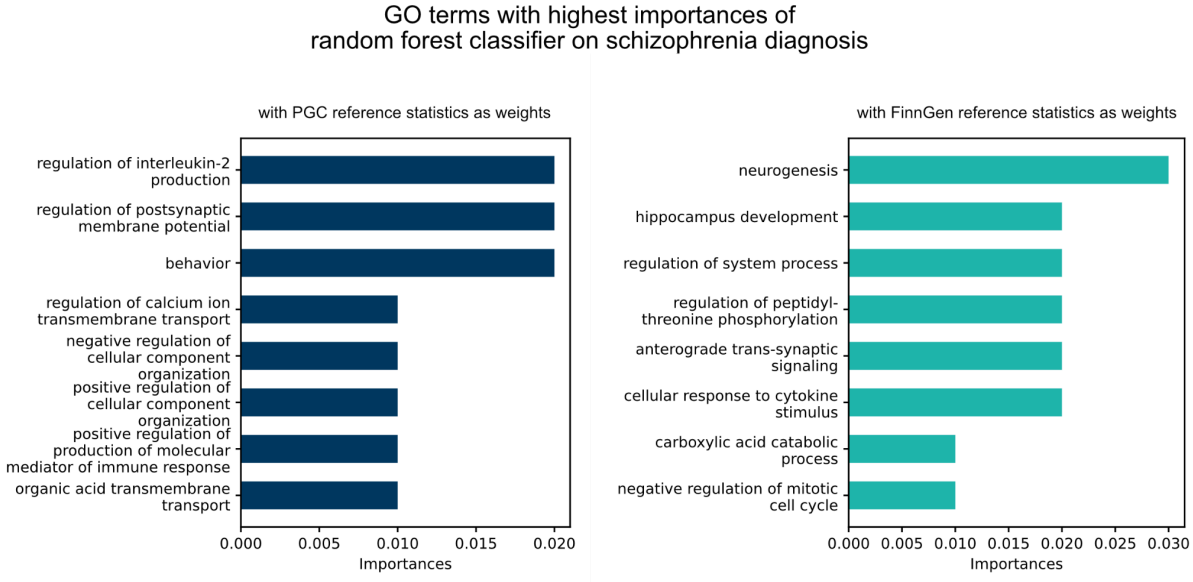


Figure 4.6 Top biological processes with highest importances contributing to random forest classification on schizophrenia disease diagnosis. **left** Prediction from the trained model with PGC summary statistics as weights. **right** Prediction from the trained model with FinnGen summary statistics as weights

4.4 Discussion

In this Chapter, a interpretable factorized semi-supervised approach was extended and applied to model genotype to function association. The model architecture was capable of directly learning the functional annotation from the genomic variants. At the same time, the factorized

latent space in the architecture allowed us to specifically isolate each source of variance as well as further combine them again to represent the total effect, which made the application entirely flexible and versatile for downstream analysis. The prototype of the model used in this study was originally developed by Daria Docenvic and was implemented as a interpretable autoencoder trained with AdamW optimizer, scaling up to complex datasets with large numbers of features in single-cell transcriptomic cohorts [421]. The continuous modification and application was made to the field of post-GWAS analysis. Weighted features learning in the novel loss function enabled the integration of summary statistics information from the reference dataset, which held the similar idea as the standard protocol when performing a conventional genome-wide association study to associate variants to the particular trait. Furthermore, comparison between models applied with and without weighted loss implementation was made. First, with the weight loss implemented, models were trained with the same hyperparameters settings (**Appendix Supplementary Table 4.2**), and used the same protocol to extract the pathway activity score matrix (**Method 4.2.8**). The training was repeated three times and thereby three different trained models as well as the corresponding pathway activity matrices would be obtained. Then, the Pearson correlation for each GO term across different repetitions of the model was calculated (so-called GO term wise Pearson correlation in **Figure 4.7**). And next the mean of Pearson correlation was computed, which referred to the model consistency. We implemented the same procedure for the model without weight loss function. Therefore in **Figure 4.7**, the model consistency of the model with weighted loss was observed much higher compared to the model without loss weighting, suggesting the weighting in the loss function based on the SNPs effect size from GWAS summary statistics improved the reproducibility of the analysis whereas models without weights where less reproducible and more noisy. Application on a variety of datasets and tasks were implemented in this study, including verification of the ability of capturing distinct variations from complex signals, exploring the learnt latent embeddings, validation of the functional association and examination with different reference summary statistical cohorts. In the study, we applied the target genomic dataset with schizophrenia and healthy individuals from the GAIN dataset, together with two reference GWAS summary statistics datasets used as prior weighing for loss calculation. We applied the model to extract diagnosis-associated effects in the diagnosis embedding while separating covariate-associated and residual-associated signals in other embeddings respectively and validating in the UMAP plot. Furthermore, we illustrated the actual interpretation with the learnt biological process patterns

from the genotype input and ranked the terms with the highest contribution to characterize the cohort.

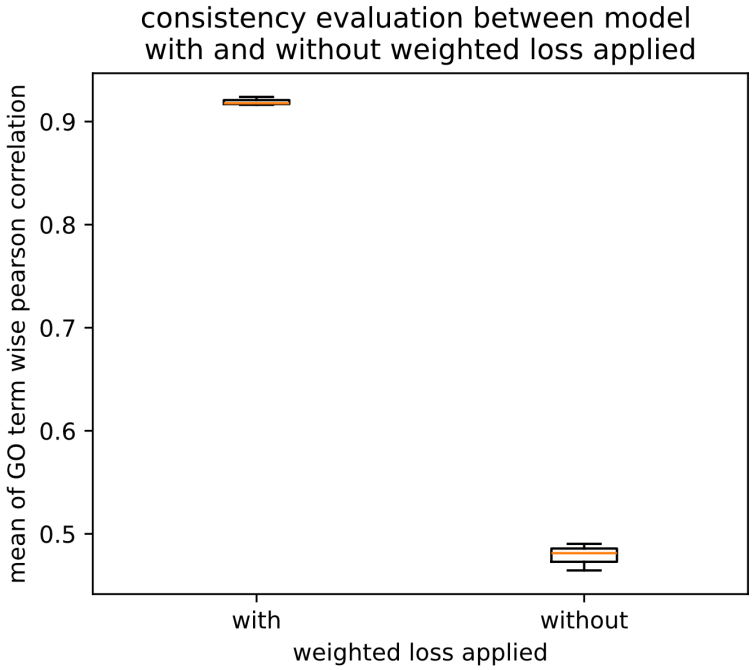


Figure 4.7 Consistency evaluation between model with weighted reconstruction loss and without weighted reconstruction loss applied

The training of the factorized latent embeddings integrated into the nonlinear training structure could be found in several previous studies. Previous models also utilizing the discriminator classifier to enhance the capability of distinguishing variance of interest, such as factorVAE which was proposed by Kim et al. (2018) to learn disentangled representations by encouraging the latent representation to separate independent embedding responsive to different variations [233]. Ansari et al. (2019) introduced an approach that outperforms beta-VAE and FactorVAE in terms of model factorizing of latent representations learned via deep generative models [422]. Braithwaite et al. (2020) proposed the variance-constrained autoencoder (VCAE) as a more principled approach to learning disentangled representations, showing equivalent performance to FactorVAE on 3D-Shapes [423]. Shao et al. (2020) presented DynamicVAE, which decouples disentanglement and reconstruction accuracy by maintaining a different β at different training stages [424]. Jiang et al. (2021) introduced Inference-InfoGAN, a GAN-based disentanglement framework that achieves higher disentanglement scores compared to FactorVAE in an unsupervised manner [425]. Different

from most of the GANs-based models described above, a cross-contrastive learning structure, namely Harmony, was also developed to disentangles semantic content from multiple parameterized transformations. Despite these successful attempts, disentangling by factorizing faces several challenges. For instance, former research pointed out the limitations and discussed the trade-off between factorizing disentanglement and reconstruction performance of variational autoencoder [426]. They emphasize the need for a new metric that can accurately evaluate the disentanglement process without the limitations of existing metrics.

Limitations could also be observed in our study. Due to the fact that the model structure was partly inherited from other architecture, the issues originally in the source model structure such as CPA and ontoVAE could also be observed. For example, the ontology merely represents one type of relationship across the terms compared to knowledge graphs with complex relationships. Secondly, the prior information might influence the prediction from varied aspects. The selection and usage of prior information and the modification of the decoder connections between the nodes could also have an impact on the prediction outcome, given a situation that the specific terms relevant to a condition might not exist in the compiled ontology file. Another crucial issue that should be kept in mind was the model uncertainty. After our hypertunning, we controlled the correlation of the pathway activation values per term between every independent round of training to a sufficiently high value with 0.91, shown in **Figure 4.8**.

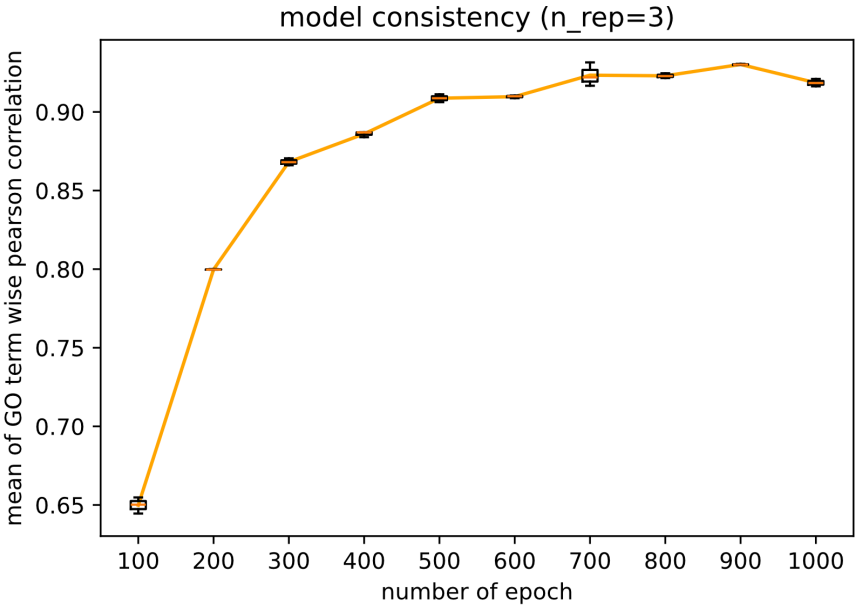


Figure 4.8 GO term wise correlation across different repetition of training given different epoch

However, compared to the ontoVAE with the 0.99 correlation, it showed a slight decrease possibly due to the introduction of the discriminator classifier. Additionally, one frequently discussed point in the VAE with the ability of factorizing latent embeddings, such as beta-VAE, factorVAE and CPA is the trade-off balance between the disentanglement performance and reconstruction quality. This suggests that incorporating the additional disentanglement capability into the model might also bring the uncertainty to the model. Given that part of the COBRA was inspired from CPA, we would assume that the same concerns should be considered in our case. Currently, the model is not able to capture higher order effects, such as combinatorial effects across different SNPs. Furthermore, the weighting metrics for the loss function should be carefully selected. In our application, the SNP effect size from the reference summary statistics helps the model to learn meaningful information. However, for the modality other than genomics, such as transcriptomics, epigenomics and proteomics, further examination would be necessary to transfer and extend the weighting model. Optimal hyperparameter sets might differ for different types of data, such as gene expression and protein expression. However, in order to achieve the best performance, systematic hyperparameter testing was required if it is available. We observed that different parameters were used in genotype and single-cell transcriptomic data in order to achieve the best factorzing performance, where in some of the cases the separation visualized in the UMAP did not receive satisfying results.

Further extension of the factorized interpretable model could also easily be developed in the future. The hybrid architecture could also be used to associate the genotype with transcriptomics, gene regulatory networks, or other molecular modality, especially due to the fact that ontoVAE, which is the precedent version of COBRA, was designed for bulk data, showing outstanding performance in the relevant task. The key feature of this interpretable factorized model was its capability of separating different source variances, benefiting from the factorized latent space while directly associating the genomics to the function. Therefore, the decoder layers actually embed the linkage between biological processes, which can be used for mechanistic modelling. Another possibility lies in the potential prediction of synergistic regulation or other interactions across features. The optimization could also be

done to the decoder part, for instance, by replacing the ontology-like structure with a graph network to infer a more complex relationship across terms or pathway structure. Hypergraph graph structure on the decoder might also enable the model to learn the attribute in a multivariable way rather than using a single node value for each term.

In conclusion, the interpretable factorized model facilitated the prediction of functional associations with genetic variants from the very preliminary results shown in the previous sections. So far, the study is still ongoing, and there are further analyses to be done. However, the elucidation of genotype-to-function associations holds profound implications for clinical practice and personalized medicine. By identifying genetic variants associated with functional annotations, further pharmacogenomic studies, for example, aim to optimize the functional outcome from population genomics, thereby improving patient outcomes and reducing healthcare costs.

Chapter 5

Prospect

5.1 Potential development of factor-based techniques

Factor-based technique is a fundamental method in statistical data analysis that has been applied for machine learning, dimension reduction, data compression, feature extraction, and sample stratification when processing high dimensional complex molecular data. The novel factorization algorithms would be extensively developed and applied in various fields. These novel algorithms might focus on various aspects of improvement, including by adding different constraints, hybriding with other algorithms, tailored to specific problems to solve respective tasks, working as the auxiliary module to improve the efficiency of other frameworks, and so on. In my thesis, three different factorization methods for signature

discovery were introduced 1) a hierarchical factorization computation framework for signature-based comorbidity modelling for schizophrenia and type 2 diabetes; 2) a distributed factorization approach which implementing the standard integrative matrix factorization to a federated version based on the DataShield platform for privacy-preserving and cross-institutional collaboration; and 3) extended application based on a biological-informed interpretable semi-supervised autoencoder to predict genotype-to-function association in schizophrenia that enabled the separation of variance of diagnosis from the confounding effects, as well as the linkage between genetic variants with biological processes. For the three different works demonstrated, further extensions and applications can still be made. For instance, multi-rank matrix factorization could be further applied on single-cell data denoising to provide insights on heterogeneous signatures across cell types illustrated in a hierarchical level. The interpretable factorized model on predicting genotype-to-function association could be modified to build the linkage of genotype with other modality such as signatures, drug targets and drug response patterns. Meanwhile, both of the above factor based algorithms could be further integrated into the dsMTL federated learning framework, enabling large-scale cross-institutional collaboration.

For future development, one of the potential fields is combining factor-based models with transformers and other large-scale natural language models [427]. Language models have been gradually optimized and applied to the molecular field due to its capability of providing a more accurate representation and prediction in a range of tasks including not just the original natural language processing but also transcriptomics profiling. However, the drawbacks of transformer-based language models are also obvious, such as resource-intensive, over-number of parameters and features included in the model learning, and lack of interpretability [428,429]. To address the issue, one possible strategy is implementing feature extraction and compression to optimize the original expensive model, enabling efficient learning. Matrix factorization model structures such as NMF have been integrated for model compression to improve effectiveness. It is noted that structure pruning of large models and factorized transformers for language modeling were implemented in multiple domains, which implicates the potentiality of further application on the biomedical field [430,431]. Another attempt that can be made is to further develop model-driven learning architecture such as the physical-informed and prior-knowledge embed model, of which the interpretable factorized model in **Chapter 4** that establishes genotype-to-function prediction falls into this category. Thus combination inspired the possibility of combining supervised

learning involving label classification with unsupervised algorithms such as factorization and factor analysis, to finally delineate the complex data into a reduced but more informative structure. These examples showed the flexibility of the factor based approaches as a versatile structure to be integrated in the architecture to improve the learning capability of the model.

5.2 Future application of factor-based approach in psychotic disorders

With increasing new evidence focusing on various aspects of schizophrenia and other psychotic disorders, it is necessary to revisit the discovered facts and plan what gap we should fill in the next step. A comprehensive review given by Tandon et al. (2024) summarized what we knew so far about the nature of schizophrenia [432]. As a genetically relevant syndrome with multiple dimensional symptoms regarding cognitive, mood, and motor manifestations, schizophrenia exhibits complicated conditions that frequently induce significant impairment of system regulation in individuals. Molecular factors are involved, such as common genetic variants with a small impact on the incidence risk at the individual level and rare gene variants with a relatively larger risk at the population level. Functional impacts were primarily brain-targeted; however, the same variants could also influence the risk of other conditions, so-called disease comorbidity, among which psychiatric disorders such as bipolar disorder and other metabolic-related syndromes like type 2 diabetes and cardiovascular disease. It is observed that no single and straightforward pathological factor could comprehensively explain this complex syndrome. Meanwhile, the boundaries of schizophrenia with other schizophrenia-like conditions such as catatonia are still to be demonstrated, suggesting the urgent need for continuous method development of distinguishing significant factors to separate the similar symptoms as well as of aggregating factors to summarize the spectrum of the syndrome [433]. Specifically, several facts about schizophrenia remained unclear and yet to be fully explored. For instance, there is inconsistency between different genders, though some studies suggest earlier onset ages and higher mortality in the male population [434]. To address this, the concept of using integrative factorization to find the cohort-specific and cohort-shared effects could be borrowed and applied to gender-specific datasets to find the heterogeneous factors. Understanding gene-gene as well as gene-environment interactions across various levels associated with disease risk is unclear [435]. One possible solution that

is promising is to implement hierarchical factorization to disentangle effects on different levels. Besides, pharmacological perturbation on cholinergic and serotonergic systems that leads to changes of molecular profile was to be further investigated due to inconsistent findings, while other key questions related to different antipsychotic treatment were still to be answered, such as biological processes that treatment modulates, the exact therapeutic effect directly relevant to the treatment and whether the treatment explicitly tackled specific disease pathology [436–439]. To address, interpretable factorized models associated with different modalities that were originally designed for gene-to-function perturbation could be further extended and applied for molecular-to-drug as well as pharmacogenomics-to-function prediction.

To conclude all the chapters, factor-based techniques have shown great potential in various fields in the biomedical domains particularly in schizophrenia and other psychotic disorders. With the increasing availability of molecular data, new factor-based models could be developed, optimized and utilized to address more complex tasks in molecular signature discovery. Continued exploration and development in this area will lead to more efficient algorithms and novel insights for applications, ultimately driving progress in the field of factor-based machine learning and molecular genetics.

Bibliography

[1] Borges NJ, Backes KA, Binder B, Roman B. First-year medical student objective structured clinical exam performance and specialty choice. J Int Assoc Med Sci Educ 2013;4:38–40.
[2] Vetulani J. [Biological basis of psychiatry]. Psychiatr Pol 2001;35:911–9.
[3] Kandel ER, Schwartz JH, Jessell T. Principles of Neural Science, Fourth Edition. McGraw-Hill Companies,Incorporated; 2000.
[4] Smelser NJ, Baltes PB. International Encyclopedia of the Social & Behavioral Sciences. Elsevier; 2001.
[5] Lisman JE, Coyle JT, Green RW, Javitt DC, Benes FM, Heckers S, et al. Circuit-based framework for understanding neurotransmitter and risk gene interactions in schizophrenia. Trends Neurosci 2008;31:234–42.
[6] Li K, Xu E. The role and the mechanism of gamma-aminobutyric acid during central nervous system development. Neurosci Bull 2008;24:195–200.
[7] Soltani N, Qiu H, Aleksic M, Glinka Y, Zhao F, Liu R, et al. GABA exerts protective and regenerative effects on islet beta cells and reverses diabetes. Proc Natl Acad Sci U S A 2011;108:11692–7.
[8] Tian J, Chau C, Hales TG, Kaufman DL. GABA(A) receptors mediate inhibition of T cell responses. J Neuroimmunol 1999;96:21–8.
[9] Neve K. The Dopamine Receptors. Springer Science & Business Media; 2009.
[10] Buttarelli FR, Fanciulli A, Pellicano C, Pontieri FE. The dopaminergic system in peripheral blood lymphocytes: from physiology to pharmacology and potential applications to neuropsychiatric disorders. Curr Neuropharmacol 2011;9:278–88.
[11] Rubí B, Maechler P. Minireview: new roles for peripheral dopamine on metabolic control and tumor growth: let's seek the balance. Endocrinology 2010;151:5570–81.
[12] Hussain T, Lokhandwala MF. Renal dopamine receptors and hypertension. Exp Biol Med 2003;228:134–42.
[13] Moncrieff J. The Myth of the Chemical Cure: A Critique of Psychiatric Drug Treatment. Palgrave Macmillan UK; 2007.
[14] Young SN. How to increase serotonin in the human brain without drugs. J Psychiatry Neurosci 2007;32:394–9.
[15] McDuffie JE, Motley ED, Limbird LE, Maleque MA. 5-hydroxytryptamine stimulates phosphorylation of p44/p42 mitogen-activated protein kinase activation in bovine aortic endothelial cell cultures. J Cardiovasc Pharmacol 2000;35:398–402.
[16] Noguchi M, Furukawa KT, Morimoto M. Pulmonary neuroendocrine cells: physiology, tissue homeostasis and disease. Dis Model Mech 2020;13. https://doi.org/10.1242/dmm.046920.
[17] Luvsannyam E, Jain MS, Pormento MKL, Siddiqui H, Balagtas ARA, Emuze BO, et al. Neurobiology of Schizophrenia: A Comprehensive Review. Cureus 2022;14:e23959.
[18] Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, et al. Molecular Biology of the Cell. Garland Science; 2014.
[19] Crick FH. On protein synthesis. Symp Soc Exp Biol 1958;12:138–63.
[20] Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. Nature 1953;171:740–1.
[21] Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms

and other classes of minor genetic variation. Genome Res 1999;9:677–9.
[22] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature 2015;526:68–74.
[23] Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med 2010;363:166–76.
[24] Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. Nat Protoc 2011;6:121–33.
[25] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 2017;101:5–22.
[26] Fundamentals of Advanced Omics Technologies: From Genes to Metabolites. Elsevier Science; 2014.
[27] Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 1991;252:1651–6.
[28] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57–63.
[29] Taub FE, DeLeo JM, Thompson EB. Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs. DNA 1983;2:309–27.
[30] Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, et al. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. Med Sci Monit Basic Res 2014;20:138–42.
[31] Cellerino A, Sanguanini M. Transcriptome analysis: Introduction and examples from the neurosciences. 1st ed. Pisa, Italy: Scuola Normale Superiore; 2019.
[32] Wong AHC, Van Tol HHM. Schizophrenia: from phenomenology to neurobiology. Neurosci Biobehav Rev 2003;27:269–306.
[33] Koenen KC, Rudenstine S, Susser E, Galea S. A Life Course Approach to Mental Disorders. Oxford University Press; 2013.
[34] Freudenreich O. Psychotic Disorders: A Practical Guide. Lippincott Williams & Wilkins; 2007.
[35] Irtelli F, Vincenti E. Psychosis: Biopsychosocial and Relational Perspectives. BoD – Books on Demand; 2018.
[36] Marcsisin MJ, Rosenstock JB, Gannon JM. Schizophrenia and Related Disorders. Oxford University Press; 2016.
[37] Rössler W, Salize HJ, van Os J, Riecher-Rössler A. Size of burden of schizophrenia and psychotic disorders. Eur Neuropsychopharmacol 2005;15:399–409.
[38] Ellis MB, Walters AS. Schizophrenic Spectrum and Other Psychotic Disorders. Mason Crest Publishers; 2022.
[39] Stentebjerg-Olesen M, Pagsberg AK, Fink-Jensen A, Correll CU, Jeppesen P. Clinical Characteristics and Predictors of Outcome of Schizophrenia-Spectrum Psychosis in Children and Adolescents: A Systematic Review. J Child Adolesc Psychopharmacol 2016;26:410–27.
[40] Liu J. Balancing therapeutic safety and efficacy to improve clinical and economic outcomes in schizophrenia: a managed care perspective. Am J Manag Care 2014;20:S174–83.
[41] Schizophrenia fact sheet. Nasnewsletter 2002;17:22.
[42] Thornicroft G, Tansella M. Mental Health Outcome Measures. RCPsych Publications; 2010.
[43] Pinikahana J, Happell B, Hope J, Keks NA. Quality of life in schizophrenia: a review of the literature from 1995 to 2000. Int J Ment Health Nurs 2002;11:103–11.

[44] Katschnig H, Freeman H, Sartorius N. Quality of Life in Mental Disorders. John Wiley & Sons; 2006.
[45] Frith C. Cognitive Neuropsychology of Schizophrenia. 1992.
[46] Häfner H, Nowotny B, Löffler W, an der Heiden W, Maurer K. When and how does schizophrenia produce social deficits? Eur Arch Psychiatry Clin Neurosci 1995;246:17–28.
[47] Allison DB, Mentore JL, Heo M, Chandler LP, Cappelleri JC, Infante MC, et al. Antipsychotic-induced weight gain: a comprehensive research synthesis. Am J Psychiatry 1999;156:1686–96.
[48] Howard MA, Cowell PE, Boucher J, Broks P, Mayes A, Farrant A, et al. Convergent neuroanatomical and behavioural evidence of an amygdala hypothesis of autism. Neuroreport 2000;11:2931–5.
[49] Tsuang M. Schizophrenia: genes and environment. Biol Psychiatry 2000;47:210–20.
[50] Sommer KL, Williams KD, Ciarocco NJ, Baumeister RF. When silence speaks louder than words: Explorations into the intrapsychic and interpersonal consequences of social ostracism. Basic Appl Soc Psych 2001;23:225–43.
[51] Read J, van Os J, Morrison AP, Ross CA. Childhood trauma, psychosis and schizophrenia: a literature review with theoretical and clinical implications. Acta Psychiatr Scand 2005;112:330–50.
[52] Ellison-Wright I, Glahn DC, Laird AR, Thelen SM, Bullmore E. The anatomy of first-episode and chronic schizophrenia: an anatomical likelihood estimation meta-analysis. Am J Psychiatry 2008;165:1015–23.
[53] Patterson V, Pencer A, Tibbo P. M92. The overlap between psychotic symptoms, substance use, and adversity history: A systematic review. Schizophr Bull 2020;46:S169–70.
[54] Bhargav H, Eiman N, Jasti N, More P, Kumar V, Holla B, et al. Composition of yoga-philosophy based mental traits () in major psychiatric disorders: A trans-diagnostic approach. Front Psychol 2023;14:1075060.
[55] Vyas P, Hwang BJ, Brašić JR. An evaluation of lumateperone tosylate for the treatment of schizophrenia. Expert Opin Pharmacother 2020;21:139–45.
[56] Sokolova SV, Sozarukova MM, Khannanova AN, Grishina NK, Portnova GV, Proskurnina EV. [Antioxidant status in patients with paranoid schizophrenia and Alzheimer disease]. Zh Nevrol Psikhiatr Im S S Korsakova 2020;120:82–7.
[57] Zhou X, Tian B, Han H-B. Serum interleukin-6 in schizophrenia: A system review and meta-analysis. Cytokine 2021;141:155441.
[58] van Eijndhoven P, Collard R, Vrijsen J, Geurts DEM, Vasquez AA, Schellekens A, et al. Measuring Integrated Novel Dimensions in Neurodevelopmental and Stress-Related Mental Disorders (MIND-SET): Protocol for a Cross-sectional Comorbidity Study From a Research Domain Criteria Perspective. JMIRx Med 2022;3:e31269.
[59] Ivanovic Kovacevic S, Sobot V, Vejnovic AM, Knezevic V. Shared psychotic disorder - a case study of folie à famille. Eur Rev Med Pharmacol Sci 2022;26:5362–6.
[60] Nakata Y, Kanahara N, Iyo M. Dopamine supersensitivity psychosis in schizophrenia: Concepts and implications in clinical practice. J Psychopharmacol 2017;31:1511–8.
[61] Kendler KS. Overview: a current perspective on twin studies of schizophrenia. Am J Psychiatry 1983;140:1413–25.
[62] Harrison PJ, Owen MJ. Genes for schizophrenia? Recent findings and their pathophysiological implications. Lancet 2003;361:417–9.
[63] van den Bree MBM, Owen MJ. The future of psychiatric genetics. Ann Med 2003;35:122–34.
[64] Niznik HB, Van Tol HH. Dopamine receptor genes: new tools for molecular psychiatry. J

Psychiatry Neurosci 1992;17:158–80.
[65] Schmidt W, Reith MEA. Dopamine and Glutamate in Psychiatric Disorders. Springer Science & Business Media; 2010.
[66] Kroeze WK, Roth BL. The molecular biology of serotonin receptors: therapeutic implications for the interface of mood and psychosis. Biol Psychiatry 1998;44:1128–42.
[67] Lang UE, Puls I, Muller DJ, Strutz-Seebohm N, Gallinat J. Molecular mechanisms of schizophrenia. Cell Physiol Biochem 2007;20:687–702.
[68] Pratt J, Hall J. Biomarkers in Psychiatry. Springer; 2019.
[69] Yadav M, Kumar N, Kumar A, Jindal DK, Dahiya M. Possible biomarkers and contributing factors of psychosis: A review. Curr Pharmacol Rep 2021;7:123–34.
[70] Neill JC, Barnes S, Cook S, Grayson B, Idris NF, McLean SL, et al. Animal models of cognitive dysfunction and negative symptoms of schizophrenia: focus on NMDA receptor antagonism. Pharmacol Ther 2010;128:419–32.
[71] Le Magueresse C, Monyer H. GABAergic interneurons shape the functional maturation of the cortex. Neuron 2013;77:388–405.
[72] Davis KL, Stewart DG, Friedman JI, Buchsbaum M, Harvey PD, Hof PR, et al. White matter changes in schizophrenia: evidence for myelin-related dysfunction. Arch Gen Psychiatry 2003;60:443–56.
[73] Owen MJ, Legge SE, Rees E, Walters JTR, O'Donovan MC. Genomic findings in schizophrenia and their implications. Mol Psychiatry 2023;28:3638–47.
[74] Singh T, Poterba T, Curtis D, Akil H, Al Eissa M, Barchas JD, et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature 2022;604:509–16.
[75] Molecular mechanisms of schizophrenia: Insights from human genetics. Curr Opin Neurobiol 2023;81:102731.
[76] Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 2022;604:502–8.
[77] Dogra S, Stansley BJ, Xiang Z, Qian W, Gogliotti RG, Nicoletti F, et al. Activating mGlu3 Metabotropic Glutamate Receptors Rescues Schizophrenia-like Cognitive Deficits Through Metaplastic Adaptations Within the Hippocampus. Biol Psychiatry 2021;90:385–98.
[78] Hall J, Bray NJ. Schizophrenia genomics: Convergence on synaptic development, adult synaptic plasticity, or both? Biol Psychiatry 2022;91:709–17.
[79] Becker F, Reid CA, Hallmann K, Tae H-S, Phillips AM, Teodorescu G, et al. Functional variants in and may contribute to genetic generalized epilepsy. Epilepsia Open 2017;2:334–42.
[80] Palmer DS, Howrigan DP, Chapman SB, Adolfsson R, Bass N, Blackwood D, et al. Exome sequencing in bipolar disorder identifies AKAP11 as a risk gene shared with schizophrenia. Nat Genet 2022;54:541–7.
[81] Mei L, Xiong W-C. Neuregulin 1 in neural development, synaptic plasticity and schizophrenia. Nat Rev Neurosci 2008;9:437–52.
[82] Legge SE, Santoro ML, Periyasamy S, Okewole A, Arsalan A, Kowalec K. Genetic architecture of schizophrenia: a review of major advancements. Psychol Med 2021;51:2168–77.
[83] Kathuria A, Lopez-Lengowski K, Jagtap SS, McPhie D, Perlis RH, Cohen BM, et al. Transcriptomic Landscape and Functional Characterization of Induced Pluripotent Stem Cell-Derived Cerebral Organoids in Schizophrenia. JAMA Psychiatry 2020;77:745–54.
[84] Koole L, Martinez-Martinez P, van Amelsvoort T, Evelo CT, Ehrhart F. Interactive neuroinflammation pathways and transcriptomics-based identification of drugs and chemical compounds for schizophrenia. World J Biol Psychiatry 2024;25:116–29.

[85] Gatta E, Saudagar V, Drnevich J, Forrest MP, Auta J, Clark LV, et al. Concordance of Immune-Related Markers in Lymphocytes and Prefrontal Cortex in Schizophrenia. Schizophr Bull Open 2021;2:sgab002.
[86] Yalcinbas EA, Ajanaku B, Nelson ED, Garcia-Flores R, Montgomery KD, Stolz JM, et al. Transcriptomic analysis of the human habenula in schizophrenia. bioRxiv 2024. https://doi.org/10.1101/2024.02.26.582081.
[87] Mishra P, Kumar S. Association of lncRNA with regulatory molecular factors in brain and their role in the pathophysiology of schizophrenia. Metab Brain Dis 2021;36:849–58.
[88] Pérez-Rodríguez D, Penedo MA, Rivera-Baltanás T, Peña-Centeno T, Burkhardt S, Fischer A, et al. MiRNA Differences Related to Treatment-Resistant Schizophrenia. Int J Mol Sci 2023;24. https://doi.org/10.3390/ijms24031891.
[89] Identification of potential biomarkers and their correlation with immune infiltration cells in schizophrenia using combinative bioinformatics strategy. Psychiatry Res 2022;314:114658.
[90] Vera-Montecinos A, Rodríguez-Mias R, MacDowell KS, García-Bueno B, Bris ÁG, Caso JR, et al. Analysis of Molecular Networks in the Cerebellum in Chronic Schizophrenia: Modulation by Early Postnatal Life Stressors in Murine Models. Int J Mol Sci 2021;22. https://doi.org/10.3390/ijms221810076.
[91] Vallejo-Curto MDC, Rodrigues-Amorim D, Jardón-Golmar L, Blanco-Formoso M, Rivera-Baltanás T, Rodriguez-Jamardo C, et al. Proteomic and metabolic profiling of chronic patients with schizophrenia induced by a physical activity program: Pilot study. Rev Psiquiatr Salud Ment 2021;14:125–38.
[92] Luo C, Pi X, Hu N, Wang X, Xiao Y, Li S, et al. Subtypes of schizophrenia identified by multi-omic measures associated with dysregulated immune function. Mol Psychiatry 2021;26:6926–36.
[93] Lerer B. Pharmacogenetics of Psychotropic Drugs. Cambridge University Press; 2011.
[94] Gorwood P, Hamon MD. Psychopharmacogenetics. Springer Science & Business Media; 2006.
[95] Hattori S, Suda A, Miyauchi M, Shiraishi Y, Saeki T, Fukushima T, et al. The association of genetic polymorphisms in CYP1A2, UGT1A4, and ABCB1 with autonomic nervous system dysfunction in schizophrenia patients treated with olanzapine. BMC Psychiatry 2020;20:72.
[96] Zivkovic M, Mihaljevic-Peles A, Muck-Seler D, Sagud M, Ganoci L, Vlatkovic S, et al. Remission Is not Associated with DRD2 rs1800497 and DAT1 rs28363170 Genetic Variants in Male Schizophrenic Patients after 6-months Monotherapy with Olanzapine. Psychiatr Danub 2020;32:84–91.
[97] Cabaleiro T, López-Rodríguez R, Román M, Ochoa D, Novalbos J, Borobia A, et al. Pharmacogenetics of quetiapine in healthy volunteers: association with pharmacokinetics, pharmacodynamics, and adverse effects. Int Clin Psychopharmacol 2015;30:82–8.
[98] Yang J, Kang C, Wu C, Lin Y, Zeng L, Yuan J, et al. Pharmacogenetic associations of NRG1 polymorphisms with neurocognitive performance and clinical symptom response to risperidone in the untreated schizophrenia. Schizophr Res 2021;231:67–9.
[99] Elsheikh SSM, Müller DJ, Pouget JG. Pharmacogenetics of Antipsychotic Treatment in Schizophrenia. Methods Mol Biol 2022;2547:389–425.
[100] Guloksuz S, Pries L-K, Delespaul P, Kenis G, Luykx JJ, Lin BD, et al. Examining the independent and joint effects of molecular genetic liability and environmental exposures in schizophrenia: results from the EUGEI study. World Psychiatry 2019;18:173–82.
[101] van Os J, Pries L-K, Delespaul P, Kenis G, Luykx JJ, Lin BD, et al. Replicated evidence that endophenotypic expression of schizophrenia polygenic risk is greater in healthy siblings of patients compared to controls, suggesting gene-environment

interaction. The EUGEI study. Psychol Med 2020;50:1884–97.
[102] Hertzberg L, Zohar AH, Yitzhaky A. Gene Expression Meta-Analysis of Cerebellum Samples Supports the FKBP5 Gene-Environment Interaction Model for Schizophrenia. Life 2021;11. https://doi.org/10.3390/life11030190.
[103] Woolway GE, Smart SE, Lynham AJ, Lloyd JL, Owen MJ, Jones IR, et al. Schizophrenia Polygenic Risk and Experiences of Childhood Adversity: A Systematic Review and Meta-analysis. Schizophr Bull 2022;48:967–80.
[104] Orsolini L, Pompili S, Volpe U. Schizophrenia: A Narrative Review of Etiopathogenetic, Diagnostic and Treatment Aspects. J Clin Med Res 2022;11. https://doi.org/10.3390/jcm11175040.
[105] Buckley PF, Miller BJ, Lehrer DS, Castle DJ. Psychiatric comorbidities and schizophrenia. Schizophr Bull 2009;35:383–402.
[106] Hennekens CH, Hennekens AR, Hollar D, Casey DE. Schizophrenia and increased risks of cardiovascular disease. Am Heart J 2005;150:1115–21.
[107] Leucht S, Burkard T, Henderson J, Maj M, Sartorius N. Physical illness and schizophrenia: a review of the literature. Acta Psychiatr Scand 2007;116:317–33.
[108] Newcomer JW. Medical risk in patients with bipolar disorder and schizophrenia. J Clin Psychiatry 2006;67 Suppl 9:25–30; discussion 36–42.
[109] Hernández-Huerta D, Morillo-González J. Dopamine D partial agonists in the treatment of psychosis and substance use disorder comorbidity: a pharmacological alternative to consider? CNS Spectr 2021;26:444–5.
[110] Gonda X, Tarazi FI. Dopamine D3 Receptors: From Bench to Bedside. Neuropsychopharmacol Hung 2021;23:272–80.
[111] Buesa-Lorenzo JB-L, Rojo-Bofill LMR-B, Plumed-Domingo JP-D, Rubio-Granero T, Rojo-Moreno L. Marchiafava-Bignami disease in a patient with schizophrenia and alcohol use disorder. Actas Esp Psiquiatr 2021;49:228–31.
[112] Cavaco TB, Ribeiro JS. Drawing the Line Between Obsessive-Compulsive Disorder and Schizophrenia. Cureus 2023;15:e36227.
[113] Vessels T, Strayer N, Lee H, Choi KW, Zhang S, Han L, et al. Integrating Electronic Health Records and Polygenic Risk to Identify Genetically Unrelated Comorbidities of Schizophrenia That May Be Modifiable. Biol Psychiatry Glob Open Sci 2024;4:100297.
[114] Nielsen J, Skadhede S, Correll CU. Antipsychotics Associated with the Development of Type 2 Diabetes in Antipsychotic-Naïve Schizophrenia Patients. Neuropsychopharmacology 2010;35:1997–2004.
[115] Agarwal SM, Caravaggio F, Costa-Dookhan KA, Castellani L, Kowalchuk C, Asgariroozbehani R, et al. Brain insulin action in schizophrenia: Something borrowed and something new. Neuropharmacology 2020;163:107633.
[116] Bergantin LB. A link among schizophrenia, diabetes, and asthma: Role of Ca2/cAMP signaling. Brain Circ 2020;6:145–51.
[117] Rahman MR, Islam T, Nicoletti F, Petralia MC, Ciurleo R, Fisicaro F, et al. Identification of Common Pathogenetic Processes between Schizophrenia and Diabetes Mellitus by Systems Biology Analysis. Genes 2021;12. https://doi.org/10.3390/genes12020237.
[118] Perry BI, Burgess S, Jones HJ, Zammit S, Upthegrove R, Mason AM, et al. The potential shared role of inflammation in insulin resistance and schizophrenia: A bidirectional two-sample mendelian randomization study. PLoS Med 2021;18:e1003455.
[119] Rødevand L, Rahman Z, Hindley GFL, Smeland OB, Frei O, Tekin TF, et al. Characterizing the Shared Genetic Underpinnings of Schizophrenia and Cardiovascular Disease Risk Factors. Am J Psychiatry 2023. https://doi.org/10.1176/appi.ajp.20220660.
[120] Chen C-J, Liao W-Y, Chattopadhyay A, Lu T-P. Exploring the genetic correlation of

cardiovascular diseases and mood disorders in the UK Biobank. Epidemiol Psychiatr Sci 2023;32:e31.
[121] Rosato M, Stringer S, Gebuis T, Paliukhovich I, Li KW, Posthuma D, et al. Combined cellomics and proteomics analysis reveals shared neuronal morphology and molecular pathway phenotypes for multiple schizophrenia risk genes. Mol Psychiatry 2021;26:784–99.
[122] De Hert M, Detraux J, Vancampfort D. The intriguing relationship between coronary heart disease and mental disorders. Dialogues Clin Neurosci 2018;20:31–40.
[123] Strawbridge RJ, Graham N. Dissecting the Genetic Relationship Between Schizophrenia and Cardiovascular Disease. Am J Psychiatry 2023;180:785–6.
[124] So H-C, Chau K-L, Ao F-K, Mo C-H, Sham P-C. Exploring shared genetic bases and causal relationships of schizophrenia and bipolar disorder with 28 cardiovascular and metabolic traits. Psychol Med 2019;49:1286–98.
[125] Cardno AG, Owen MJ. Genetic relationships between schizophrenia, bipolar disorder, and schizoaffective disorder. Schizophr Bull 2014;40:504–15.
[126] Liu H, Tang Y, Womer F, Fan G, Lu T, Driesen N, et al. Differentiating patterns of amygdala-frontal functional connectivity in schizophrenia and bipolar disorder. Schizophr Bull 2014;40:469–77.
[127] Hoffman GE, Bendl J, Voloudakis G, Montgomery KS, Sloofman L, Wang Y-C, et al. CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. Sci Data 2019;6:180.
[128] Koshiyama D, Fukunaga M, Okada N, Morita K, Nemoto K, Usui K, et al. White matter microstructural alterations across four major psychiatric disorders: mega-analysis study in 2937 individuals. Mol Psychiatry 2020;25:883–95.
[129] Meyer N, Faulkner SM, McCutcheon RA, Pillinger T, Dijk D-J, MacCabe JH. Sleep and Circadian Rhythm Disturbance in Remitted Schizophrenia and Bipolar Disorder: A Systematic Review and Meta-analysis. Schizophr Bull 2020;46:1126–43.
[130] Hall LS, Pain O, O'Brien HE, Anney R, Walters JTR, Owen MJ, et al. Cis-effects on gene expression in the human prenatal brain associated with genetic risk for neuropsychiatric disorders. Mol Psychiatry 2021;26:2082–8.
[131] Gottlieb J, Madrange M, Gardair C, Sbidian E, Frazier A, Wolkenstein P, et al. PAPASH, PsAPASH and PASS autoinflammatory syndromes: phenotypic heterogeneity, common biological signature and response to immunosuppressive regimens. Br J Dermatol 2019;181:866–9.
[132] Keane JM, Khashan AS, McCarthy FP, Kenny LC, Collins JM, O'Donovan S, et al. Identifying a biological signature of prenatal maternal stress. JCI Insight 2021;6. https://doi.org/10.1172/jci.insight.143007.
[133] Feyaerts D, Hédou J, Gillard J, Chen H, Tsai ES, Peterson LS, et al. Integrated plasma proteomic and single-cell immune signaling network signatures demarcate mild, moderate, and severe COVID-19. Cell Rep Med 2022;3:100680.
[134] Lehman KM, Smith HC, Grabowicz M. A Biological Signature for the Inhibition of Outer Membrane Lipoprotein Biogenesis. MBio 2022;13:e0075722.
[135] de Winter DTC, Langerak AW, Te Marvelde J, Dworzak MN, De Moerloose B, Starý J, et al. The variable biological signature of refractory cytopenia of childhood (RCC), a retrospective EWOG-MDS study. Leuk Res 2021;108:106652.
[136] Bellman R, Kalaba RE. Dynamic Programming and Modern Control Theory: By Richard Bellman and Robert Kalaba. 1965.
[137] Berisha V, Krantsevich C, Hahn PR, Hahn S, Dasarathy G, Turaga P, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med 2021;4:153.
[138] A high-bias, low-variance introduction to Machine Learning for physicists. Phys Rep

2019;810:1–124.
[139] Daum F. Nonlinear filters: beyond the Kalman filter n.d. https://ieeexplore.ieee.org/document/1499276 (accessed May 1, 2024).
[140] Barajas A, Ochoa S, Obiols JE, Lalucat-Jo L. Gender differences in individuals at high-risk of psychosis: a comprehensive literature review. ScientificWorldJournal 2015;2015:430735.
[141] Morrison NA, Yeoman R, Kelly PJ, Eisman JA. Contribution of trans-acting factor alleles to normal physiological variability: vitamin D receptor gene polymorphism and circulating osteocalcin. Proc Natl Acad Sci U S A 1992;89:6665–9.
[142] McGreevy K, Hoel B, Lipsitz S, Bissada N, Hoel D. Racial and anthropometric differences in plasma levels of insulin-like growth factor I and insulin-like growth factor binding protein-3. Urology 2005;66:587–92.
[143] Hajian Tilaki K. Methodological issues of confounding in analytical epidemiologic studies. Caspian J Intern Med 2012;3:488–95.
[144] Azevedo R, Silva-Cavalcante MD, Gualano B, Lima-Silva AE, Bertuzzi R. Effects of caffeine ingestion on endurance performance in mentally fatigued individuals. Eur J Appl Physiol 2016;116:2293–303.
[145] Kim J-O, Mueller CW. Factor Analysis: Statistical Methods and Practical Issues. Beverly Hills, Calif. : Sage Publications; 1978.
[146] Kim J-O, Mueller CW. Introduction to Factor Analysis: What it is and how to Do it. 1982.
[147] Yong AG, Pearce S. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. Tutor Quant Methods Psychol 2013;9:79–94.
[148] Yamada R, Okada D, Wang J, Basak T, Koyama S. Interpretation of omics data analyses. J Hum Genet 2021;66:93–102.
[149] Meng M, He J, Guan Y, Zhao H, Yi J, Yao S, et al. Factorial Invariance of the 10-Item Connor-Davidson Resilience Scale Across Gender Among Chinese Elders. Front Psychol 2019;10:1237.
[150] Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics 2019;35:3055–62.
[151] Dai X, Shen L. Advances and Trends in Omics Technology Development. Front Med 2022;9:911861.
[152] Urbanski AH, Araujo JD, Creighton R, Nakaya HI. Integrative Biology Approaches Applied to Human Diseases. In: Husi H, editor. Computational Biology, Brisbane (AU): Codon Publications; 2019.
[153] Tanay A, Steinfeld I, Kupiec M, Shamir R. Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. Mol Syst Biol 2005;1:2005.0002.
[154] Haider S, Pal R. Integrated analysis of transcriptomic and proteomic data. Curr Genomics 2013;14:91–110.
[155] Quraishi MN, Acharjee A, Beggs AD, Horniblow R, Tselepis C, Gkoutos G, et al. A Pilot Integrative Analysis of Colonic Gene Expression, Gut Microbiota, and Immune Infiltration in Primary Sclerosing Cholangitis-Inflammatory Bowel Disease: Association of Disease With Bile Acid Pathways. J Crohns Colitis 2020;14:935–47.
[156] Lancaster SM, Sanghi A, Wu S, Snyder MP. A Customizable Analysis Flow in Integrative Multi-Omics. Biomolecules 2020;10. https://doi.org/10.3390/biom10121606.
[157] Gao C, Liu J, Kriebel AR, Preissl S, Luo C, Castanon R, et al. Iterative single-cell multi-omic integration using online learning. Nat Biotechnol 2021;39:1000–7.
[158] Ben-Chetrit N, Niu X, Swett AD, Sotelo J, Jiao MS, Stewart CM, et al. Integration of

whole transcriptome spatial profiling with protein markers. Nat Biotechnol 2023;41:788–93.
[159] Yang M, Matan-Lithwick S, Wang Y, De Jager PL, Bennett DA, Felsky D. Multi-omic integration via similarity network fusion to detect molecular subtypes of ageing. Brain Commun 2023;5:fcad110.
[160] Fan Z, Zhou Y, Ressom HW. MOTA: Network-Based Multi-Omic Data Integration for Biomarker Discovery. Metabolites 2020;10. https://doi.org/10.3390/metabo10040144.
[161] Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun 2021;12:3445.
[162] Ruan X, Ye Y, Cheng W, Xu L, Huang M, Chen Y, et al. Multi-Omics Integrative Analysis of Lung Adenocarcinoma: An Profiling for Precise Medicine. Front Med 2022;9:894338.
[163] Wang J, Liao N, Du X, Chen Q, Wei B. A semi-supervised approach for the integration of multi-omics data based on transformer multi-head self-attention mechanism and graph convolutional networks. BMC Genomics 2024;25:86.
[164] Luningham JM, McArtor DB, Hendriks AM, van Beijsterveldt CEM, Lichtenstein P, Lundström S, et al. Data Integration Methods for Phenotype Harmonization in Multi-Cohort Genome-Wide Association Studies With Behavioral Outcomes. Front Genet 2019;10:1227.
[165] Le Sueur H, Bruce IN, Geifman N. The challenges in data integration – heterogeneity and complexity in clinical trials and patient registries of Systemic Lupus Erythematosus. BMC Med Res Methodol 2020;20:1–5.
[166] Pender A, Titmuss E, Pleasance ED, Fan KY, Pearson H, Brown SD, et al. Genome and Transcriptome Biomarkers of Response to Immune Checkpoint Inhibitors in Advanced Solid Tumors. Clin Cancer Res 2021;27:202–12.
[167] Chong JSX, Ng KK, Tandi J, Wang C, Poh J-H, Lo JC, et al. Longitudinal Changes in the Cerebral Cortex Functional Organization of Healthy Elderly. J Neurosci 2019;39:5534–50.
[168] Jiang M-Z, Aguet F, Ardlie K, Chen J, Cornell E, Cruz D, et al. Uncovering Cross-Cohort Molecular Features with Multi-Omics Integration Analysis. bioRxiv 2022:2022.11.10.515908. https://doi.org/10.1101/2022.11.10.515908.
[169] Murphy JP, Shumba K, Jamieson L, Nattey C, Pascoe S, Fox MP, et al. Assessment of facility-level antiretroviral treatment patient status utilizing a national-level laboratory cohort: Toward an understanding of system-level tracking and clinic switching in South Africa. Front Public Health 2022;10:959481.
[170] Aung HL, Gates TM, Mao L, Brew BJ, Rourke SB, Cysique LA. Abnormal cognitive aging in people with HIV: evidence from data integration between two countries' cohort studies. AIDS 2022;36:1171–9.
[171] Mate S, Kampf M, Rödle W, Kraus S, Proynova R, Silander K, et al. Pan-European Data Harmonization for Biobanks in ADOPT BBMRI-ERIC. Appl Clin Inform 2019;10:679–92.
[172] Bron EE, Klein S, Papma JM, Jiskoot LC, Venkatraghavan V, Linders J, et al. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. Neuroimage Clin 2021;31:102712.
[173] Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Author Correction: Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 2019;25:1948.
[174] Genders TSS, Steyerberg EW, Myriam Hunink MG, Nieman K, Galema TW, Mollet

NR, et al. Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. BMJ 2012;344. https://doi.org/10.1136/bmj.e3485.
[175] Prediction models for the risk of total knee replacement: development and validation using data from multicentre cohort studies. The Lancet Rheumatology 2022;4:e125–34.
[176] Etingen B, Amante DJ, Martinez RN, Smith BM, Shimada SL, Richardson L, et al. Supporting the Implementation of Connected Care Technologies in the Veterans Health Administration: Cross-Sectional Survey Findings from the Veterans Engagement with Technology Collaborative (VET-C) Cohort. J Particip Med 2020;12:e21214.
[177] Aouedi O, Sacco A, Piamrat K, Marchetto G. Handling Privacy-Sensitive Medical Data With Federated Learning: Challenges and Future Directions. IEEE J Biomed Health Inform 2023;27:790–803.
[178] Arnold CG, Sonn B, Meyers FJ, Vest A, Puls R, Zirkler E, et al. Accessing and utilizing clinical and genomic data from an electronic health record data warehouse. Transl Med Commun 2023;8. https://doi.org/10.1186/s41231-023-00140-0.
[179] Danese A, Richter ML, Chaichoompu K, Fischer DS, Theis FJ, Colomé-Tatché M. EpiScanpy: integrated single-cell epigenomic analysis. Nat Commun 2021;12:5228.
[180] Statistical single cell multi-omics integration. Current Opinion in Systems Biology 2018;7:54–9.
[181] Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu, Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell 2017;32:185–203.e13.
[182] Kingma DP, Welling M. Auto-Encoding Variational Bayes 2013.
[183] Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AIChE J 1991;37:233–43.
[184] Kramer MA. Autoassociative neural networks. Comput Chem Eng 1992;16:313–28.
[185] Ning C, Wang D, Zheng X, Zhang Q, Zhang S, Mrode R, et al. Eigen decomposition expedites longitudinal genome-wide association studies for milk production traits in Chinese Holstein. Genet Sel Evol 2018;50:1–10.
[186] Wall ME, Dyck PA, Brettin TS. SVDMAN—singular value decomposition analysis of microarray data. Bioinformatics 2001;17:566–8.
[187] Tjioe E, Berry M, Homayouni R. Using a literature-based NMF model for discovering gene functional relationships. 2008 IEEE International Conference on Bioinformatics and Biomeidcine Workshops, IEEE; 2008. https://doi.org/10.1109/bibmw.2008.4686234.
[188] Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics 2001;17:763–74.
[189] Nascimento M, Silva FFE, Sáfadi T, Nascimento ACC, Ferreira TEM, Barroso LMA, et al. Independent Component Analysis (ICA) based-clustering of temporal RNA-seq data. PLoS One 2017;12:e0181195.
[190] Exploratory factor analysis of shared and specific genetic associations in depression and anxiety. Prog Neuropsychopharmacol Biol Psychiatry 2023;126:110781.
[191] Pladevall M, Singal B, Williams LK, Brotons C, Guyer H, Sadurni J, et al. A Single Factor Underlies the Metabolic Syndrome: A confirmatory factor analysis. Diabetes Care 2006;29:113–22.
[192] Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. J Am Stat Assoc 2008;103:1438–56.
[193] Seoane JA, Campbell C, Day INM, Casas JP, Gaunt TR. Canonical Correlation Analysis for Gene-Based Pleiotropy Discovery. PLoS Comput Biol 2014;10:e1003876.
[194] Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi‐Omics

Factor Analysis—a framework for unsupervised integration of multi‐omics data sets. Mol Syst Biol 2018. https://doi.org/10.15252/msb.20178124.
[195] Lock EF, Hoadley KA, Marron JS, Nobel AB. JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. Ann Appl Stat 2013;7:523–42.
[196] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 2009;25:2906–12.
[197] Du L, Liu K, Yao X, Risacher SL, Han J, Saykin AJ, et al. Multi-Task Sparse Canonical Correlation Analysis with Application to Multi-Modal Brain Imaging Genetics n.d. https://ieeexplore.ieee.org/abstract/document/8869839 (accessed May 22, 2024).
[198] de Tayrac M, Lê S, Aubry M, Mosser J, Husson F. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. BMC Genomics 2009;10:1–17.
[199] Zheng C-H, Huang D-S, Kong X-Z, Zhao X-M. Gene Expression Data Classification Using Consensus Independent Component Analysis. Genomics Proteomics Bioinformatics 2008;6:74–82.
[200] Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell 2019;177:1873–87.e17.
[201] Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun 2019;10:1–14.
[202] Grønbech CH, Vording MF, Timshel PN, Sønderby CK, Pers TH, Winther O. scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics 2020;36:4415–22.
[203] Qiu YL, Zheng H, Gevaert O. Genomic data imputation with variational auto-encoders. Gigascience 2020;9. https://doi.org/10.1093/gigascience/giaa082.
[204] Gyawali PK, Murkute JV, Toloubidokhti M, Jiang X, Horacek BM, Sapp JL, et al. Learning to Disentangle Inter-Subject Anatomical Variations in Electrocardiographic Data. IEEE Trans Biomed Eng 2022;69:860–70.
[205] 1995A&AS..110..405S Page 405 n.d. https://adsabs.harvard.edu/full/1995A&AS..110..405S (accessed May 19, 2024).
[206] 13 Computation using the QR decomposition 1993;9:467–508.
[207] Salkind NJ. Encyclopedia of Measurement and Statistics. SAGE; 2007.
[208] Demmel J, Kahan W. Accurate singular values of bidiagonal matrices. SIAM J Sci Stat Comput 1990;11:873–912.
[209] Golub G, Kahan W. Calculating the singular values and pseudo-inverse of a matrix. J Soc Ind Appl Math Ser B Numer Anal 1965;2:205–24.
[210] Pearson K. LIII. On lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci 1901;2:559–72.
[211] Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet 2018;34:790–805.
[212] Algorithms and applications for approximate nonnegative matrix factorization. Comput Stat Data Anal 2007;52:155–73.
[213] Lord FM. Principals of Modern Psychological Measurement: A Festschrift for Frederic M. Lord. Routledge; 1983.
[214] Polit DF, Beck CT. Nursing Research: Generating and Assessing Evidence for Nursing Practice. Lippincott Williams & Wilkins; 2012.
[215] Thompson. Canonical Correlation Analysis. n.d.
[216] Thompson B. Canonical Correlation Analysis: Uses and Interpretation. SAGE; 1984.

[217] Weinberg SL. Large Variable Canonical Analysis: A Proposed Variant of Canonical Variate Analysis. 1971.
[218] Adel T, Ghahramani Z, Weller A. Discovering Interpretable Representations for Both Deep Generative and Discriminative Models 2018.
[219] Parra LC. Multi-set Canonical Correlation Analysis simply explained 2018.
[220] Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. Wiley Interdiscip Rev Comput Stat 2013;5:149–79.
[221] Smilde AK, Westerhuis JA, de Jong S. A framework for sequential multiblock component methods. J Chemom 2003;17:323–37.
[222] Xu Y, Goodacre R. Multiblock principal component analysis: an efficient tool for analyzing metabolomics data which contain two influential factors. Metabolomics 2011;8:37–51.
[223] Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics 2016;32:1–8.
[224] Meng C, Helm D, Frejno M, Kuster B. moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets 2015. https://doi.org/10.1021/acs.jproteome.5b00824.
[225] Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, et al. Integrative Subtype Discovery in Glioblastoma Using iCluster. PLoS One 2012;7:e35236.
[226] Zhang C, Yu L, Zhang X, Chawla N. ImWalkMF: Joint matrix factorization and implicit walk integrative learning for recommendation n.d. https://ieeexplore.ieee.org/abstract/document/8258001 (accessed May 2, 2024).
[227] Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. Brief Bioinform 2019;21:541–52.
[228] Rendle S, Krichene W, Zhang L, Anderson J. Neural collaborative filtering vs. Matrix factorization revisited. Fourteenth ACM Conference on Recommender Systems, New York, NY, USA: ACM; 2020. https://doi.org/10.1145/3383313.3412488.
[229] Ferrari Dacrema M, Boglio S, Cremonesi P, Jannach D. A troubling analysis of reproducibility and progress in recommender systems research. ACM Trans Inf Syst Secur 2021;39:1–49.
[230] Higgins I, Matthey L, Pal A, Burgess CP, Glorot X, Botvinick M, et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. International Conference on Learning Representations 2016.
[231] Burgess CP, Higgins I, Pal A, Matthey L, Watters N, Desjardins G, et al. Understanding disentangling in β-VAE 2018.
[232] Chen RTQ, Li X, Grosse R, Duvenaud D. Isolating Sources of Disentanglement in Variational Autoencoders 2018.
[233] Kim H, Mnih A. Disentangling by Factorising 2018.
[234] Kim M, Wang Y, Sahu P, Pavlovic V. Bayes-Factor-VAE: Hierarchical Bayesian Deep Auto-Encoder Models for Factor Disentanglement 2019.
[235] Fyshe A, Talukdar PP, Murphy B, Mitchell TM. Interpretable Semantic Vectors from a Joint Model of Brain- and Text-Based Meaning. Proceedings of the Conference Association for Computational Linguistics Meeting 2014;2014. https://doi.org/10.3115/v1/p14-1046.
[236] Lai J, Wang X, Xiang Q, Li R, Song Y. FVAE: a regularized variational autoencoder using the Fisher criterion. Applied Intelligence 2022;52:16869–85.
[237] Lee H, Ha ILDO, Lee Y. Deep Neural Networks for Semiparametric Frailty Models via H-likelihood 2023.
[238] Devarajan K. Nonnegative Matrix Factorization: An Analytical and Interpretive Tool

in Computational Biology. PLoS Comput Biol 2008;4:e1000029.
[239] Jiang B, Ma S, Causey J, Qiao L, Hardin MP, Bitts I, et al. SparRec: An effective matrix completion framework of missing data imputation for GWAS. Sci Rep 2016;6:35534.
[240] Zhao J, Feng Q, Wu P, Warner JL, Denny JC, Wei W-Q. Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein(a) (LPA). PLoS One 2019;14:e0212112.
[241] He Y, Chhetri SB, Arvanitis M, Srinivasan K, Aguet F, Ardlie KG, et al. sn-spMF: matrix factorization informs tissue-specific genetic regulation of gene expression. Genome Biol 2020;21:235.
[242] Leal LG, David A, Jarvelin M-R, Sebert S, Männikkö M, Karhunen V, et al. Identification of disease-associated loci using machine learning for genotype and network data integration. Bioinformatics 2019;35:5182–90.
[243] Arani AA, Sehhati M, Tabatabaiefar MA. Genetic variant effect prediction by supervised nonnegative matrix tri-factorization. Mol Omics 2021;17:740–51.
[244] Baldrian P, López-Mondéjar R. Microbial genomics, transcriptomics and proteomics: new discoveries in decomposition research using complementary methods. Appl Microbiol Biotechnol 2014;98:1531–7.
[245] Wang H-Q, Zheng C-H, Zhao X-M. jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. Bioinformatics 2015;31:572–80.
[246] Liang L, Zhu K, Lu S. BEM: Mining Coregulation Patterns in Transcriptomics via Boolean Matrix Factorization. Bioinformatics 2020;36:4030–7.
[247] Walter FC, Stegle O, Velten B. FISHFactor: a probabilistic factor model for spatial transcriptomics data with subcellular resolution. Bioinformatics 2023;39. https://doi.org/10.1093/bioinformatics/btad183.
[248] Shang L, Zhou X. Spatially aware dimension reduction for spatial transcriptomics. Nat Commun 2022;13:7203.
[249] Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, et al. CellRank for directed single-cell fate mapping. Nat Methods 2022;19:159–70.
[250] Kuleshov V, Chaganty A, Liang P. Tensor Factorization via Matrix Factorization. Artificial Intelligence and Statistics, PMLR; 2015, p. 507–16.
[251] Batmanghelich N, Dong A, Taskar B, Davatzikos C. Regularized tensor factorization for multi-modality medical image classification. Med Image Comput Comput Assist Interv 2011;14:17–24.
[252] Cieślik M, Bekiranov S. Combinatorial epigenetic patterns as quantitative predictors of chromatin biology. BMC Genomics 2014;15:76.
[253] Hamamoto R, Takasawa K, Machino H, Kobayashi K, Takahashi S, Bolatkan A, et al. Application of non-negative matrix factorization in oncology: one approach for establishing precision medicine. Brief Bioinform 2022;23. https://doi.org/10.1093/bib/bbac246.
[254] Lu C, Sherpa R, Klindziuk L, Kriel S, Mollah S. HOCMO: A Tensor-based Higher-Order Correlation Model to Deconvolute Epigenetic Microenvironment in Breast Cancer. bioRxiv 2024:2020.12.01.406249. https://doi.org/10.1101/2020.12.01.406249.
[255] Xu Q, Xiang EW, Yang Q. Protein-protein interaction prediction via Collective Matrix Factorization. 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE; 2010. https://doi.org/10.1109/bibm.2010.5706537.
[256] Tu S, Chen R, Xu L. A binary matrix factorization algorithm for protein complex prediction. Proteome Sci 2011;9:1–8.
[257] Sparse nonnegative matrix factorization for protein sequence motif discovery. Expert Syst Appl 2011;38:13198–207.

[258] Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, et al. Methods and approaches in the topology-based analysis of biological pathways. Front Physiol 2013;4:278.
[259] Li W, Zhang S, Liu C-C, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics 2012;28:2458–66.
[260] Srihari S, Ragan MA. Systematic tracking of dysregulated modules identifies novel genes in cancer. Bioinformatics 2013;29:1553–61.
[261] Zhang X, Liu C-T. Information-incorporated sparse convex clustering for disease subtyping. Bioinformatics 2023;39. https://doi.org/10.1093/bioinformatics/btad417.
[262] Coleman S, Kirk PDW, Wallace C. Consensus clustering for Bayesian mixture models. BMC Bioinformatics 2022;23:290.
[263] Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi‐Omics Factor Analysis—a framework for unsupervised integration of multi‐omics data sets. Mol Syst Biol 2018. https://doi.org/10.15252/msb.20178124.
[264] Phalen H, Coffman BA, Ghuman A, Sejdić E, Salisbury DF. Non-negative Matrix Factorization Reveals Resting-State Cortical Alpha Network Abnormalities in the First-Episode Schizophrenia Spectrum. Biol Psychiatry Cogn Neurosci Neuroimaging 2020;5:961–70.
[265] Shu H, Wang X, Zhu H. D-CCA: A Decomposition-based Canonical Correlation Analysis for High-Dimensional Datasets. J Am Stat Assoc 2020;115:292–306.
[266] Stauffer E-M, Bethlehem RAI, Dorfschmidt L, Won H, Warrier V, Bullmore ET. The genetic relationships between brain structure and schizophrenia. Nat Commun 2023;14:1–15.
[267] Pantazopoulos H, Katsel P, Haroutunian V, Chelini G, Klengel T, Berretta S. Molecular signature of extracellular matrix pathology in schizophrenia. Eur J Neurosci 2021;53:3960–87.
[268] Cai M, Ma J, Wang Z, Zhao Y, Zhang Y, Wang H, et al. Individual-level brain morphological similarity networks: Current methodologies and applications. CNS Neurosci Ther 2023;29:3713–24.
[269] Kim HK, Park HY, Seo E, Bang M, Song YY, Lee SY, et al. Factors Associated With Psychosocial Functioning and Outcome of Individuals With Recent-Onset Schizophrenia and at Ultra-High Risk for Psychosis. Front Psychiatry 2019;10:459.
[270] Fountoulakis KN, Dragioti E, Theofilidis AT, Wikilund T, Atmatzidis X, Nimatoudis I, et al. Staging of Schizophrenia With the Use of PANSS: An International Multi-Center Study. Int J Neuropsychopharmacol 2019;22:681–97.
[271] Sowunmi OA, Onifade PO. Psychometric evaluation of medication adherence rating scale (MARS) among Nigerian patients with schizophrenia. Niger J Clin Pract 2019;22:1281–5.
[272] Lim K, Peh O-H, Yang Z, Rekhi G, Rapisarda A, See Y-M, et al. Large-scale evaluation of the Positive and Negative Syndrome Scale (PANSS) symptom architecture in schizophrenia. Asian J Psychiatr 2021;62:102732.
[273] Dornquast C, Tomzik J, Reinhold T, Walle M, Mönter N, Berghöfer A. To what extent are psychiatrists aware of the comorbid somatic illnesses of their patients with serious mental illnesses? – a cross-sectional secondary data analysis. BMC Health Serv Res 2017;17. https://doi.org/10.1186/s12913-017-2106-6.
[274] Veeneman RR, Vermeulen JM, Abdellaoui A, Sanderson E, Wootton RE, Tadros R, et al. Exploring the Relationship Between Schizophrenia and Cardiovascular Disease: A Genetic Correlation and Multivariable Mendelian Randomization Study. Schizophr Bull 2022;48:463–73.
[275] Simunovic Filipcic I, Igor F, Marijana B, Matic K, Ena I, Antonija V, et al. Somatic

comorbidities are associated with poorer treatment outcome in schizophrenia spectrum disorders, independently of psychiatric comorbidities and other clinical factors. Eur Psychiatry 2017;41:S384–S384.
[276] Abdullah HM, Azeb Shahul H, Hwang MY, Ferrando S. Comorbidity in Schizophrenia: Conceptual Issues and Clinical Management. Focus 2020;18:386–90.
[277] Hackinger S, Prins B, Mamakou V, Zengini E, Marouli E, Brčić L, et al. Evidence for genetic contribution to the increased risk of type 2 diabetes in schizophrenia. Translational Psychiatry 2018;8. https://doi.org/10.1038/s41398-018-0304-6.
[278] Liu H, Sun Y, Zhang X, Li S, Hu D, Xiao L, et al. Integrated Analysis of Summary Statistics to Identify Pleiotropic Genes and Pathways for the Comorbidity of Schizophrenia and Cardiometabolic Disease. Front Psychiatry 2020;11:256.
[279] Mizuki Y, Sakamoto S, Okahisa Y, Yada Y, Hashimoto N, Takaki M, et al. Mechanisms Underlying the Comorbidity of Schizophrenia and Type 2 Diabetes Mellitus. Int J Neuropsychopharmacol 2021;24:367–82.
[280] Arruda AL, Khandaker GM, Morris AP, Smith GD, Huckins LM, Zeggini E. Genomic insights into the comorbidity between type 2 diabetes and schizophrenia. Schizophrenia 2024;10:1–12.
[281] Aryal S, Bonanno K, Song B, Mani DR, Keshishian H, Carr SA, et al. Deep proteomics identifies shared molecular pathway alterations in synapses of schizophrenia and bipolar disorder patients and mouse model. bioRxiv 2022:2022.09.21.508852. https://doi.org/10.1101/2022.09.21.508852.
[282] Liu Y, Li Z, Zhang M, Deng Y, Yi Z, Shi T. Exploring the pathogenetic association between schizophrenia and type 2 diabetes mellitus diseases based on pathway analysis. BMC Med Genomics 2013;6:1–14.
[283] Li Z, Chen P, Chen J, Xu Y, Wang Q, Li X, et al. Glucose and Insulin-Related Traits, Type 2 Diabetes and Risk of Schizophrenia: A Mendelian Randomization Study. EBioMedicine 2018;34:182–8.
[284] Perry BI, Bowker N, Burgess S, Wareham NJ, Upthegrove R, Jones PB, et al. Evidence for Shared Genetic Aetiology Between Schizophrenia, Cardiometabolic, and Inflammation-Related Traits: Genetic Correlation and Colocalization Analyses. Schizophr Bull Open 2022;3:sgac001.
[285] Andreassen OA, Djurovic S, Thompson WK, Schork AJ, Kendler KS, O'Donovan MC, et al. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. Am J Hum Genet 2013;92:197–209.
[286] Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing 1991;24:1–10.
[287] McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics 2010;11:242–53.
[288] McCall MN, Jaffee HA, Irizarry RA. fRMA ST: frozen robust multiarray analysis for Affymetrix Exon and Gene ST arrays. Bioinformatics 2012;28:3153–4.
[289] Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 2009;4:1184–91.
[290] Cantini L, Kairov U, de Reyniès A, Barillot E, Radvanyi F, Zinovyev A. Assessing reproducibility of matrix factorization methods in independent transcriptomes. Bioinformatics 2019;35:4307–13.
[291] Greco A, Sanchez Valle J, Pancaldi V, Baudot A, Barillot E, Caselle M, et al. Molecular Inverse Comorbidity between Alzheimer's Disease and Lung Cancer: New Insights from Matrix Factorization. Int J Mol Sci 2019;20.

https://doi.org/10.3390/ijms20133114.
[292] clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov J 2021;2:100141.
[293] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
[294] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015;1:417–25.
[295] Bharadhwaj VS, Mubeen S, Sargsyan A, Jose GM, Geissler S, Hofmann-Apitius M, et al. Integrative analysis to identify shared mechanisms between schizophrenia and bipolar disorder and their comorbidities. Prog Neuropsychopharmacol Biol Psychiatry 2023;122:110688.
[296] Casas BS, Vitória G, Prieto CP, Casas M, Chacón C, Uhrig M, et al. Schizophrenia-derived hiPSC brain microvascular endothelial-like cells show impairments in angiogenesis and blood–brain barrier function. Mol Psychiatry 2022;27:3708–18.
[297] Schmidt-Kastner R, van Os J, Esquivel G, Steinbusch HWM, Rutten BPF. An environmental analysis of genes associated with schizophrenia: hypoxia and vascular factors as interacting elements in the neurodevelopmental model. Molecular Psychiatry 2012;17:1194–205. https://doi.org/10.1038/mp.2011.183.
[298] Tahergorabi Z, Khazaei M. Imbalance of Angiogenesis in Diabetic Complications: The Mechanisms. Int J Prev Med 2012;3:827.
[299] Fadini GP, Albiero M, Bonora BM, Avogaro A. Angiogenic Abnormalities in Diabetes Mellitus: Mechanistic and Clinical Aspects. J Clin Endocrinol Metab 2019;104:5431–44.
[300] Bryll A, Skrzypek J, Krzyściak W, Szelągowska M, Śmierciak N, Kozicz T, et al. Oxidative-Antioxidant Imbalance and Impaired Glucose Metabolism in Schizophrenia. Biomolecules 2020;10. https://doi.org/10.3390/biom10030384.
[301] Shan Y, Zhao J, Zheng Y, Guo S, Schrodi SJ, He D. Understanding the function of the GABAergic system and its potential role in rheumatoid arthritis. Front Immunol 2023;14:1114350.
[302] Li R, Wang B, Wu C, Li D, Wu Y, Ye L, et al. Acidic fibroblast growth factor attenuates type 2 diabetes-induced demyelination via suppressing oxidative stress damage. Cell Death Dis 2021;12:1–17.
[303] Zaharieva E, Kamenov Z, Velikova T, Tsakova A, El-Darawish Y, Okamura H. Interleukin-18 serum level is elevated in type 2 diabetes and latent autoimmune diabetes. Endocrine Connections 2018;7:179.
[304] Syed AAS, He L, Shi Y, Mahmood S. Elevated levels of IL-18 associated with schizophrenia and first episode psychosis: A systematic review and meta-analysis. Early Interv Psychiatry 2021;15. https://doi.org/10.1111/eip.13031.
[305] Moreno M, Lanni A. Hormonal and Neuroendocrine Regulation of Energy Balance. Frontiers Media SA; 2016.
[306] Lv S-Y, Chen W-D, Wang Y-D. The Apelin/APJ System in Psychosis and Neuropathy. Front Pharmacol 2020;11. https://doi.org/10.3389/fphar.2020.00320.
[307] Sahpolat M, Ari M, Kokacya MH. Plasma Apelin, Visfatin and Resistin Levels in Patients with First Episode Psychosis and Chronic Schizophrenia. Clin Psychopharmacol Neurosci 2020;18:109.
[308] Marei I, Chidiac O, Thomas B, Pasquier J, Dargham S, Robay A, et al. Angiogenic content of microparticles in patients with diabetes and coronary artery disease predicts networks of endothelial dysfunction. Cardiovasc Diabetol 2022;21:17.

[309] Colijn MA. The Co-occurrence of Gastrointestinal Symptoms and Psychosis: Diagnostic Considerations. Prim Care Companion CNS Disord 2022;24. https://doi.org/10.4088/PCC.22nr03236.
[310] Pacitti D, Levene M, Garone C, Nirmalananthan N, Bax BE. Mitochondrial Neurogastrointestinal Encephalomyopathy: Into the Fourth Decade, What We Have Learned So Far. Front Genet 2018;9:669.
[311] Wan Y, Wang Q, Prud'homme GJ. GABAergic system in the endocrine pancreas: a new target for diabetes treatment. Diabetes Metab Syndr Obes 2015;8:79–87.
[312] Anney RJL. Common Genetic Variants in Autism Spectrum Disorders. The Neuroscience of Autism Spectrum Disorders 2013:155–67. https://doi.org/10.1016/b978-0-12-391924-3.00010-7.
[313] Tao Y, Wei X, Yue Y, Wang J, Li J, Shen L, et al. Extracellular vesicle-derived AEBP1 mRNA as a novel candidate biomarker for diabetic kidney disease. J Transl Med 2021;19. https://doi.org/10.1186/s12967-021-03000-3.
[314] Bhattacharyya S, Bhaumik H, Mukherjee A, De S. Machine Learning for Big Data Analysis. Walter de Gruyter GmbH & Co KG; 2018.
[315] Hassanien AE, Darwish A. Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges. Springer Nature; 2020.
[316] Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, Taha K. Efficient Machine Learning for Big Data: A Review 2015.
[317] A review of applications in federated learning. Comput Ind Eng 2020;149:106854.
[318] McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-Efficient Learning of Deep Networks from Decentralized Data 2016.
[319] Yu C-S, Chang S-S, Chang T-H, Wu JL, Lin Y-J, Chien H-F, et al. A COVID-19 Pandemic Artificial Intelligence-Based System With Deep Learning Forecasting and Automatic Statistical Data Acquisition: Development and Implementation Study. J Med Internet Res 2021;23:e27806.
[320] Pan W, Xu Z, Rajendran S, Wang F. An adaptive federated learning framework for clinical risk prediction with electronic health records from multiple hospitals. Patterns (N Y) 2024;5:100898.
[321] Bebortta S, Tripathy SS, Basheer S, Chowdhary CL. FedEHR: A Federated Learning Approach towards the Prediction of Heart Diseases in IoT-Based Electronic Health Records. Diagnostics (Basel) 2023;13. https://doi.org/10.3390/diagnostics13203166.
[322] Danek BP, Makarious MB, Dadu A, Vitale D, Lee PS, Singleton AB, et al. Federated learning for multi-omics: A performance evaluation in Parkinson's disease. Patterns (N Y) 2024;5:100945.
[323] Linardos A, Kushibar K, Walsh S, Gkontra P, Lekadir K. Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. Sci Rep 2022;12:3551.
[324] Mendelsohn S, Froelicher D, Loginov D, Bernick D, Berger B, Cho H. sfkit: a web-based toolkit for secure and federated genomic analysis. Nucleic Acids Res 2023;51:W535–41.
[325] Saldanha OL, Quirke P, West NP, James JA, Loughrey MB, Grabsch HI, et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. Nat Med 2022;28:1232–9.
[326] Mammen PM. Federated Learning: Opportunities and Challenges 2021.
[327] Quintero A, Hübschmann D, Kurzawa N, Steinhauser S, Rentzsch P, Krämer S, et al. ShinyButchR: Interactive NMF-based decomposition workflow of genome-scale datasets. Biol Methods Protoc 2020;5:bpaa022.
[328] Shokri R, Stronati M, Song C, Shmatikov V. Membership Inference Attacks against

Machine Learning Models 2016.
[329] Nguyen N-B, Chandrasegaran K, Abdollahzadeh M, Cheung N-M. Re-thinking Model Inversion Attacks Against Deep Neural Networks 2023.
[330] Ren X, Kuan P-F. Negative binomial additive model for RNA-Seq data analysis. BMC Bioinformatics 2020;21:1–15.
[331] GEO Accession viewer n.d. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164376 (accessed May 5, 2024).
[332] GEO Accession viewer n.d. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134497 (accessed May 5, 2024).
[333] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.
[334] Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 2017;45:D833–9.
[335] Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.
[336] Tomasetti C, Iasevoli F, Buonaguro EF, De Berardis D, Fornaro M, Fiengo ALC, et al. Treating the Synapse in Major Psychiatric Disorders: The Role of Postsynaptic Density Network in Dopamine-Glutamate Interplay and Psychopharmacologic Drugs Molecular Actions. Int J Mol Sci 2017;18:135.
[337] Region and diagnosis-specific changes in synaptic proteins in schizophrenia and bipolar I disorder. Psychiatry Res 2010;178:374–80.
[338] Obi-Nagata K, Temma Y, Hayashi-Takagi A. Synaptic functions and their disruption in schizophrenia: From clinical evidence to synaptic optogenetics in an animal model. Proc Jpn Acad Ser B Phys Biol Sci 2019;95:179–97.
[339] Torrey EF, Barci BM, Webster MJ, Bartko JJ, Meador-Woodruff JH, Knable MB. Neurochemical markers for schizophrenia, bipolar disorder, and major depression in postmortem brains. Biol Psychiatry 2005;57. https://doi.org/10.1016/j.biopsych.2004.10.019.
[340] Jaffrey SR, Wilkinson MF. Nonsense-mediated RNA decay in the brain: emerging modulator of neural development and disease. Nat Rev Neurosci 2018;19:715–28.
[341] Ting Z. Druggable causal genes of bipolar disorder identified through Mendelian Randomization analysis offer a route to intervention in integrated stress response. medRxiv 2023:2023.12.20.23300345. https://doi.org/10.1101/2023.12.20.23300345.
[342] Aryal S, Bonanno K, Song B, Mani DR, Keshishian H, Carr SA, et al. Deep proteomics identifies shared molecular pathway alterations in synapses of patients with schizophrenia and bipolar disorder and mouse model. Cell Rep 2023;42. https://doi.org/10.1016/j.celrep.2023.112497.
[343] El-Mallakh RS, Yff T, Gao Y. Ion Dysregulation in the Pathogenesis of Bipolar Illness n.d. https://austinpublishinggroup.com/depression-anxiety/fulltext/depression-v3-id1076.php (accessed May 7, 2024).
[344] Ion homeostasis and the mechanism of action of lithium. Clin Neurosci Res 2004;4:227–31.
[345] Bengesser SA, Fuchs R, Lackner N, Birner A, Reininghaus B, Meier-Allard N, et al. Endoplasmic Reticulum Stress and Bipolar Disorder - Almost Forgotten Therapeutic Drug Targets in the Unfolded Protein Response Pathway Revisited. CNS Neurol Disord Drug Targets 2016;15. https://doi.org/10.2174/1871527315666160321104613.

[346] Chen M, Shlezinger N, Poor HV, Eldar YC, Cui S. Communication-efficient federated learning. Proceedings of the National Academy of Sciences 2021;118:e2024789118.
[347] Luo B, Li X, Wang S, Huang J, Tassiulas L. Cost-Effective Federated Learning Design 2020.
[348] Sun T, Li D, Wang B. Decentralized Federated Averaging 2021.
[349] Xia Y, Yang D, Li W, Myronenko A, Xu D, Obinata H, et al. Auto-FedAvg: Learnable Federated Averaging for Multi-Institutional Medical Image Segmentation 2021.
[350] Wei K, Li J, Ding M, Ma C, Yang HH, Farokhi F, et al. Federated Learning With Differential Privacy: Algorithms and Performance Analysis n.d. https://ieeexplore.ieee.org/abstract/document/9069945 (accessed May 6, 2024).
[351] Tolpegin V, Truex S, Gursoy ME, Liu L. Data Poisoning Attacks Against Federated Learning Systems. Computer Security – ESORICS 2020 2020:480–501.
[352] Guha N, Talwalkar A, Smith V. One-Shot Federated Learning 2019.
[353] Zhou Y, Pu G, Ma X, Li X, Wu D. Distilled One-Shot Federated Learning 2020.
[354] Eren ME, Richards LE, Bhattarai M, Yus R, Nicholas C, Alexandrov BS. FedSPLIT: One-Shot Federated Recommendation System Based on Non-negative Joint Matrix Factorization and Knowledge Distillation 2022.
[355] Chai D, Wang L, Chen K, Yang Q. Secure Federated Matrix Factorization n.d. https://ieeexplore.ieee.org/abstract/document/9162459 (accessed May 6, 2024).
[356] Chen T, Jin X, Sun Y, Yin W. VAFL: a Method of Vertical Asynchronous Federated Learning 2020.
[357] Bikash Joshi University of Grenoble Alpes, Grenoble, France, Franck Iutzeler University of Grenoble Alpes, Grenoble, France, Massih-Reza Amini University of Grenoble Alpes, Grenoble, France. Asynchronous Distributed Matrix Factorization with Similar User and Item Based Regularization n.d. https://doi.org/10.1145/2959100.2959161.
[358] Feng L, Zhao Y, Guo S, Qiu X, Li W, Yu P. BAFL: A Blockchain-Based Asynchronous Federated Learning Framework n.d. https://ieeexplore.ieee.org/abstract/document/9399813 (accessed May 6, 2024).
[359] PPFchain: A novel framework privacy-preserving blockchain-based federated learning method for sensor networks. Internet of Things 2023;22:100781.
[360] HBFL: A hierarchical blockchain-based federated learning framework for collaborative IoT intrusion detection. Comput Electr Eng 2022;103:108379.
[361] Cai W, Du X, Xu J. A Personalized QoS Prediction Method for Web Services via Blockchain-Based Matrix Factorization. Sensors 2019;19:2749.
[362] Brookes AJ. The essence of SNPs. Gene 1999;234:177–86.
[363] Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet 2009;10:387–406.
[364] Bochner BR. New technologies to assess genotype-phenotype relationships. Nat Rev Genet 2003;4:309–14.
[365] Mickle JE, Cutting GR. Genotype-phenotype relationships in cystic fibrosis. Med Clin North Am 2000;84:597–607.
[366] Pérusse L, Bouchard C. Genotype-environment interaction in human obesity. Nutr Rev 1999;57:S31–7; discussion S37–8.
[367] Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers 2021;1:1–21.
[368] Chen Z, Schunkert H. Genetics of coronary artery disease in the post-GWAS era. J Intern Med 2021;290:980–92.
[369] Imamura M, Maeda S. Genetics of type 2 diabetes: the GWAS era and future perspectives [Review]. Endocr J 2011;58:723–39.

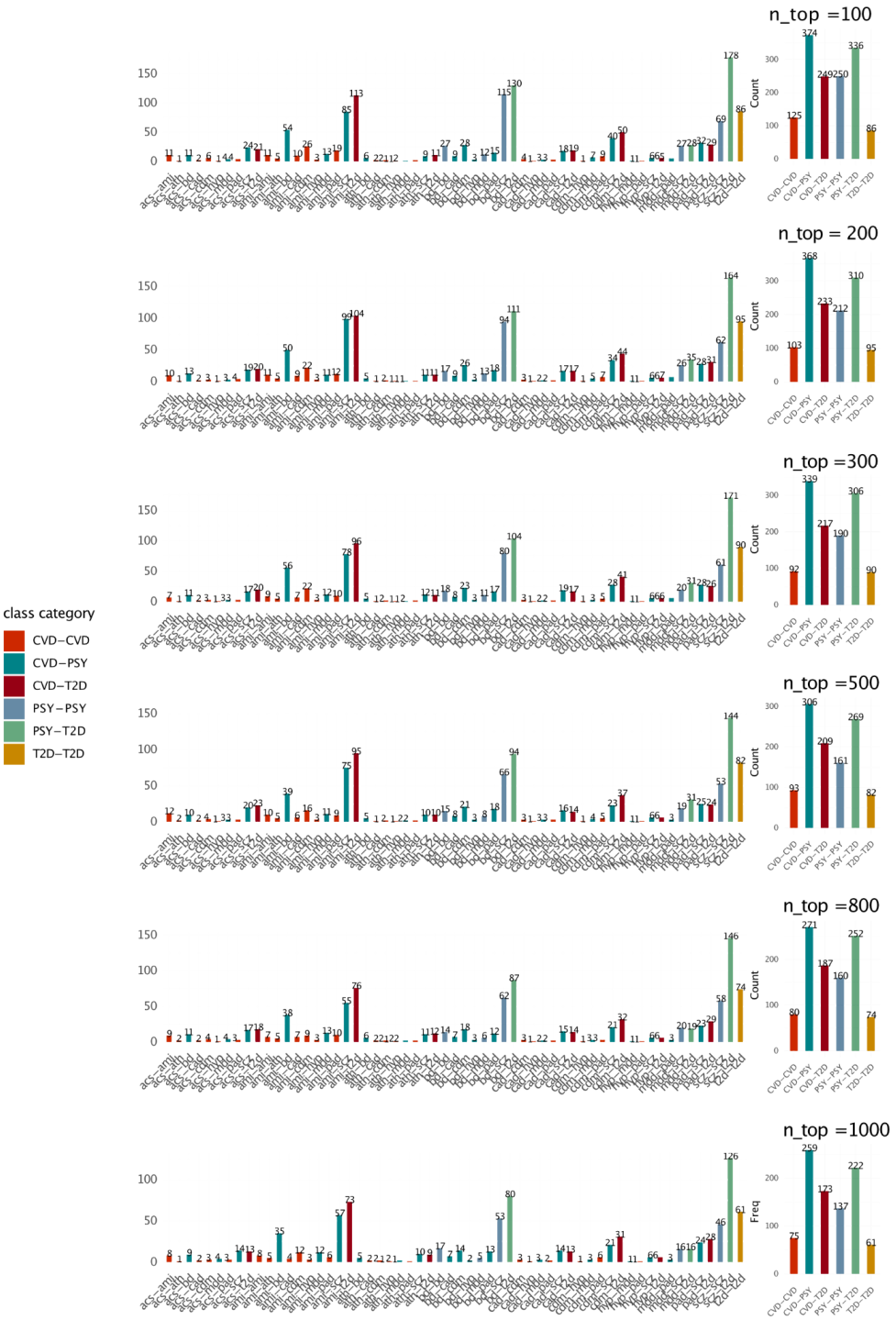
[370] Horwitz T, Lam K, Chen Y, Xia Y, Liu C. A decade in psychiatric GWAS research. Mol Psychiatry 2019;24:378–89.
[371] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:s13742–015 – 0047–8.
[372] Mai J, Lu M, Gao Q, Zeng J, Xiao J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. Commun Biol 2023;6:899.
[373] Brandes N, Linial N, Linial M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. Genome Biol 2020;21:1–22.
[374] Liu C, Bousman CA, Pantelis C, Skafidas E, Zhang D, Yue W, et al. Pathway-wide association study identifies five shared pathways associated with schizophrenia in three ancestral distinct populations. Transl Psychiatry 2017;7:e1037–e1037.
[375] Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc 2019;14:482–517.
[376] Paczkowska M, Barenboim J, Sintupisut N, Fox NS, Zhu H, Abd-Rabbo D, et al. Integrative pathway enrichment analysis of multivariate omics data. Nat Commun 2020;11:735.
[377] Wang K, Edmondson AC, Li M, Gao F, Qasim AN, Devaney JM, et al. Pathway-Wide Association Study Implicates Multiple Sterol Transport and Metabolism Genes in HDL Cholesterol Regulation. Front Genet 2011;2:41.
[378] Pistis G, Vázquez-Bourgon J, Fournier M, Jenni R, Cleusix M, Papiol S, et al. Gene set enrichment analysis of pathophysiological pathways highlights oxidative stress in psychosis. Mol Psychiatry 2022;27:5135–43.
[379] Sonnenschein SF, Grace A. Emerging therapeutic targets for schizophrenia: a framework for novel treatment strategies for psychosis. Expert Opin Ther Targets 2021;25:15–26.
[380] Wang Q, Shi Q, Wang Z, Lu J, Hou J. Integrating plasma proteomes with genome-wide association data for causal protein identification in multiple myeloma. BMC Med 2023;21:377.
[381] Xu C, Yang Q, Xiong H, Wang L, Cai J, Wang F, et al. Candidate Pathway-Based Genome-Wide Association Studies Identify Novel Associations of Genomic Variants in the Complement System Associated With Coronary Artery Disease. Circ Cardiovasc Genet 2014. https://doi.org/10.1161/CIRCGENETICS.114.000738.
[382] Lu M, Feng R, Zhang C, Xiao Y, Yin C. Identifying Novel Drug Targets for Epilepsy Through a Brain Transcriptome-Wide Association Study and Protein-Wide Association Study with Chemical-Gene-Interaction Analysis. Mol Neurobiol 2023;60:5055–66.
[383] Seninge L, Anastopoulos I, Ding H, Stuart J. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. Nat Commun 2021;12:5684.
[384] Using deep learning to model the hierarchical structure and function of a cell - Nature Methods. Nature n.d. https://www.nature.com/articles/nmeth.4627 (accessed May 3, 2024).
[385] Peng J, Wang X, Shang X. Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq data. BMC Bioinformatics 2019;20:284.
[386] Doncevic D, Herrmann C. Biologically informed variational autoencoders allow predictive modeling of genetic and drug-induced perturbations. Bioinformatics 2023;39:btad387.
[387] Lotfollahi M, Susmelj AK, De Donno C, Hetzel L, Ji Y, Ibarra IL, et al. Predicting cellular responses to complex perturbations in high‐throughput screens. Mol Syst Biol

2023. https://doi.org/10.15252/msb.202211517.
[388] GAIN Collaborative Research Group, Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Collaborative Association Study of Psoriasis, et al. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. Nat Genet 2007;39:1045–51.
[389] Sullivan PF, Agrawal A, Bulik CM, Andreassen OA, Børglum AD, Breen G, et al. Psychiatric Genomics: An Update and an Agenda. Am J Psychiatry 2017. https://doi.org/10.1176/appi.ajp.2017.17030283.
[390] Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature 2023;613:508–18.
[391] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
[392] Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet 2008;9:477–85.
[393] Lloyd-Jones LR, Robinson MR, Yang J, Visscher PM. Transformation of Summary Statistics from Linear Mixed Model Association on All-or-None Traits to Odds Ratio. Genetics 2018;208:1397–408.
[394] Ghosh A, Zou F, Wright FA. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. Am J Hum Genet 2008;82:1064–74.
[395] Kingma DP, Welling M. An Introduction to Variational Autoencoders. MAL 2019;12:307–92.
[396] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9.
[397] Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. Genetics 2023;224. https://doi.org/10.1093/genetics/iyad031.
[398] Thomas PD, Hill DP, Mi H, Osumi-Sutherland D, Van Auken K, Carbon S, et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat Genet 2019;51:1429–33.
[399] Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics 2009;25:288–9.
[400] Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, et al. The Reactome Pathway Knowledgebase 2024. Nucleic Acids Res 2024;52:D672–8.
[401] Griss J, Viteri G, Sidiropoulos K, Nguyen V, Fabregat A, Hermjakob H. ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. Mol Cell Proteomics 2020;19:2115–25.
[402] Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2020;48:D498–503.
[403] Fabregat A, Korninger F, Viteri G, Sidiropoulos K, Marin-Garcia P, Ping P, et al. Reactome graph database: Efficient access to complex pathway data. PLoS Comput Biol 2018;14:e1005968.
[404] Wu G, Haw R. Functional Interaction Network Construction and Analysis for Disease Discovery. Methods Mol Biol 2017;1558:235–53.
[405] onto-vae/weighted_loss_analysis/weighted_loss_analysis.md at main · david-hirst/onto-vae. GitHub n.d. https://github.com/david-hirst/onto-vae/blob/main/weighted_loss_analysis/weighted_loss_analysis.md (accessed May 29, 2024).

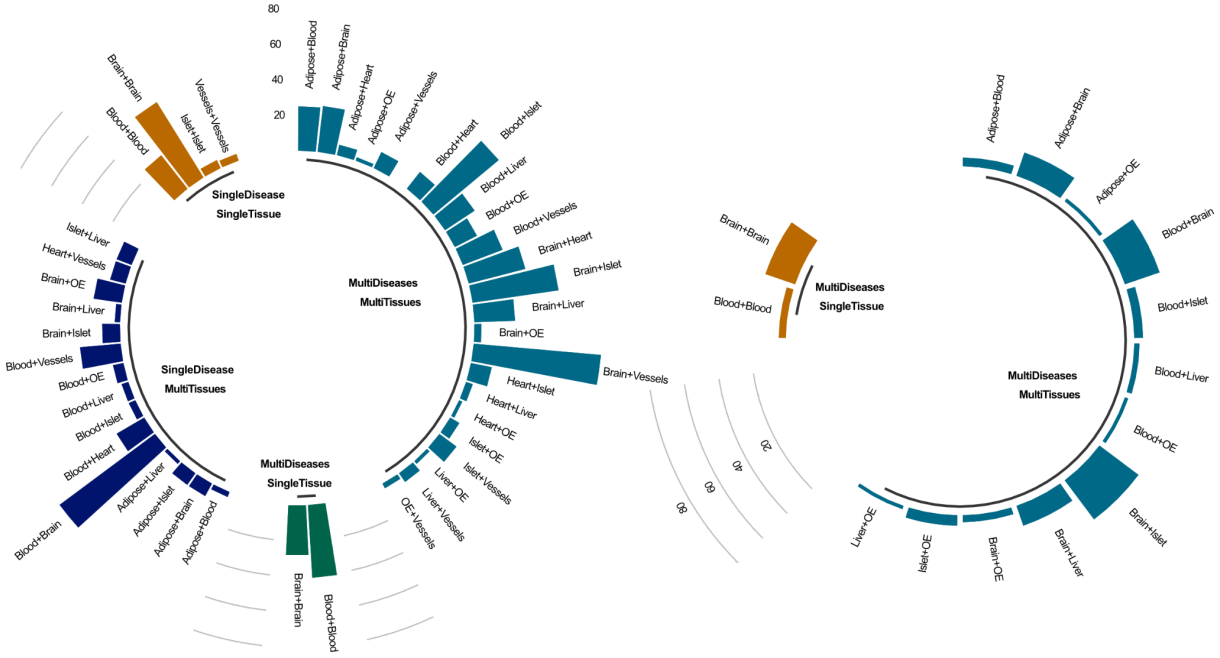
[406] The Gene Ontology Consortium. Gene Ontology Annotations and Resources. Nucleic Acids Res 2012;41:D530–5.
[407] Lample G, Conneau A, Denoyer L, Ranzato M 'aurelio. Unsupervised Machine Translation Using Monolingual Corpora Only 2017.
[408] Deng F, Ma J. Gender Differences in Prevalence and Associated Factors of Dyslipidemia in Initial-Treatment and Drug-Naïve Schizophrenia Patients. Neuropsychiatr Dis Treat 2024;20:957–66.
[409] Piccinelli M, Homen FG. Gender Differences in the Epidemiology of Affective Disorders and Schizophrenia. 1997.
[410] Gender differences in schizophrenia. Psychoneuroendocrinology 2003;28:17–54.
[411] Falkenburg J, Tracy DK. Sex and schizophrenia: a review of gender differences. Psychosis 2014. https://doi.org/10.1080/17522439.2012.733405.
[412] Calcium as a Trojan horse in mental diseases—The role of PMCA and PMCA-interacting proteins in bipolar disorder and schizophrenia. Neurosci Lett 2018;663:48–54.
[413] Maes M, Plaimas K, Suratanee A, Noto C, Kanchanatawan B. First Episode Psychosis and Schizophrenia Are Systemic Neuro-Immune Disorders Triggered by a Biotic Stimulus in Individuals with Reduced Immune Regulation and Neuroprotection. Cells 2021;10. https://doi.org/10.3390/cells10112929.
[414] Huang Y, Zhang X, Zhou N. The Interrelation between Interleukin-2 and Schizophrenia. Brain Sci 2022;12. https://doi.org/10.3390/brainsci12091154.
[415] Kang E, Wen Z, Song H, Christian KM, Ming G-L. Adult Neurogenesis and Psychiatric Disorders. Cold Spring Harb Perspect Biol 2016;8. https://doi.org/10.1101/cshperspect.a019026.
[416] Sheu J-R, Hsieh C-Y, Jayakumar T, Tseng M-F, Lee H-N, Huang S-W, et al. A Critical Period for the Development of Schizophrenia-Like Pathology by Aberrant Postnatal Neurogenesis. Front Neurosci 2019;13:466459.
[417] Weissleder C, North HF, Bitar M, Fullerton JM, Sager R, Barry G, et al. Reduced adult neurogenesis is associated with increased macrophages in the subependymal zone in schizophrenia. Mol Psychiatry 2021;26:6880–95.
[418] Momtazmanesh S, Zare-Shahabadi A, Rezaei N. Cytokine Alterations in Schizophrenia: An Updated Review. Front Psychiatry 2019;10:892.
[419] Kim H, Baek S-H, Kim J-W, Ryu S, Lee J-Y, Kim J-M, et al. Inflammatory markers of symptomatic remission at 6 months in patients with first-episode schizophrenia. Schizophrenia 2023;9:1–7.
[420] Dawidowski B, Górniak A, Podwalski P, Lebiecka Z, Misiak B, Samochowiec J. The Role of Cytokines in the Pathogenesis of Schizophrenia. J Clin Med Res 2021;10. https://doi.org/10.3390/jcm10173849.
[421] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization 2017.
[422] Ansari AF, Soh H. Hyperprior Induced Unsupervised Disentanglement of Latent Representations 2018.
[423] Braithwaite DT, O'Connor M, Kleijn WB. Variance Constrained Autoencoding 2020.
[424] Shao H, Lin H, Yang Q, Yao S, Zhao H, Abdelzaher T. DynamicVAE: Decoupling Reconstruction Error and Disentangled Representation Learning 2020.
[425] Jiang H, Yin J, Luo X, Wang F. Inference-InfoGAN: Inference Independence via Embedding Orthogonal Basis Expansion 2021.
[426] Szabó A, Hu Q, Portenier T, Zwicker M, Favaro P. Understanding Degeneracies and Ambiguities in Attribute Transfer. Computer Vision – ECCV 2018 2018:721–36.
[427] Bowman SR. Eight Things to Know about Large Language Models 2023.
[428] Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a

foundation model for single-cell multi-omics using generative AI. Nat Methods 2024:1–11.
[429] Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A Survey of Large Language Models 2023.
[430] Wang Z, Wohlwend J, Lei T. Structured Pruning of Large Language Models 2019. https://doi.org/10.18653/v1/2020.emnlp-main.496.
[431] Ma X, Zhang P, Zhang S, Duan N, Hou Y, Song D, et al. A Tensorized Transformer for Language Modeling 2019.
[432] The schizophrenia syndrome, circa 2024: What we know and how that informs its nature. Schizophr Res 2024;264:1–28.
[433] Fink M, Taylor MA. The catatonia syndrome: forgotten but not gone. Arch Gen Psychiatry 2009;66:1173–7.
[434] Aleman A, Kahn RS, Selten J-P. Sex differences in the risk of schizophrenia: evidence from meta-analysis. Arch Gen Psychiatry 2003;60:565–71.
[435] Colodro-Conde L, Couvy-Duchesne B, Whitfield JB, Streit F, Gordon S, Kemper KE, et al. Association Between Population Density and Genetic Risk for Schizophrenia. JAMA Psychiatry 2018;75:901–10.
[436] Paul SM, Yohn SE, Popiolek M, Miller AC, Felder CC. Muscarinic acetylcholine receptor agonists as novel treatments for schizophrenia. Am J Psychiatry 2022;179:611–27.
[437] Vaidya S, Guerin AA, Walker LC, Lawrence AJ. Clinical Effectiveness of Muscarinic Receptor-Targeted Interventions in Neuropsychiatric Disorders: A Systematic Review. CNS Drugs 2022;36:1171–206.
[438] Brannan SK, Sawchak S, Miller AC, Lieberman JA, Paul SM, Breier A. Muscarinic Cholinergic Receptor Agonist and Peripheral Antagonist for Schizophrenia. N Engl J Med 2021;384:717–26.
[439] Yohn SE, Weiden PJ, Felder CC, Stahl SM. Muscarinic acetylcholine receptors for psychotic disorders: bench-side to clinic. Trends Pharmacol Sci 2022;43:1098–112.

Appendix


n_top =100
n_top = 200
n_top =300
n_top =500
n_top =800
n_top =1000
class category
CVD−CVD
CVD−PSY
CVD−T2D
PSY−PSY
PSY−T2D
T2D−T2D
Count
Freq

Supplementary Figure 2.1 Comparison of the signature pairs types under different top N settings**. left** The number of condition class (e.g. scz-t2d, scz-bd) of the signature pairs ; **right** The number of disease class (e.g. PSY-T2D, PSY-CVD) of signature pairs



Supplementary Figure 2.2 **left** tissue sources of all the signature pairs**; right** tissue sources of only the schizophrenia - type 2 diabetes signature pairs

Supplementary Table 3.1 Top 20 shared pathways across schizophrenia and bipolar disorder

ID	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count
REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES	53/412	269/13198	8,3765E-28	1,5865E-24	8,6851E-25	53
REACTOME_NEURONAL_SYSTEM	61/412	410/13198	5,4163E-25	5,1293E-22	2,8079E-22	61
REACTOME_NEUROTRANSMITTER_RECEPTORS_AND_POSTSYNAPTIC_SIGNAL_TRANSMISSION	37/412	205/13198	2,641E-18	1,6673E-15	9,1276E-16	37

REACTOME_ACTIVATION_OF_NMDA_RECEPTORS_AND_POSTSYNAPTIC_EVENTS	25/412	94/13198	6,3071E-17	2,9864E-14	1,6349E-14	25
REACTOME_SIGNALING_BY_ERBB4	19/412	58/13198	5,079E-15	1,2882E-12	7,0519E-13	19
REACTOME_LONG_TERM_POTENTIATION	13/412	23/13198	1,911E-14	3,6195E-12	1,9815E-12	13
REACTOME_UNBLOCKING_OF_NMDA_RECEPTORS_GLUTAMATE_BINDING_AND_ACTIVATION	12/412	21/13198	1,6668E-13	2,87E-11	1,5711E-11	12
REACTOME_DISEASES_OF_SIGNAL_TRANSDUCTION_BY_GROWTH_FACTOR_RECEPTORS_AND_SECOND_MESSENGERS	41/412	386/13198	5,6106E-12	6,2509E-10	3,4219E-10	41
REACTOME_CREB1_PHOSPHORYLATION_THROUGH_NMDA_RECEPTOR_MEDIATED_ACTIVATION_OF_RAS_SIGNALING	12/412	28/13198	1,4145E-11	1,41E-09	7,719E-10	12
REACTOME_IONOTROPIC_ACTIVITY_OF_KAINATE_RECEPTORS	8/412	10/13198	3,594E-11	2,8363E-09	1,5527E-09	8
REACTOME_NCAM_SIGNALING_FOR_NEURITE_OUT_GROWTH	16/412	63/13198	5,8217E-11	3,6354E-09	1,9901E-09	16
REACTOME_SIGNALING_BY_ERBB2_IN_CANCER	11/412	26/13198	1,2178E-10	6,9896E-09	3,8264E-09	11
REACTOME_SIGNALING_BY_ERBB2	14/412	50/13198	2,2555E-10	1,1867E-08	6,4962E-09	14
REACTOME_PI3K_AKT_SIGNALING_IN_CANCER	19/412	102/13198	3,0347E-10	1,4738E-08	8,0681E-09	19
REACTOME_INTRACELLULAR_SIGNALING_BY_SECOND_MESSENGERS	33/412	304/13198	4,1506E-10	1,8282E-08	1,0008E-08	33

REACTOME_FLT3_SIGNALING	32/412	292/13198	6,069E-10	2,4989E-08	1,368E-08	32
REACTOME_NEGATIVE_REGULATION_OF_THE_PI3K_AKT_NETWORK	19/412	110/13198	1,1628E-09	4,4045E-08	2,4112E-08	19
REACTOME_MAPK_FAMILY_SIGNALING_CASCADES	33/412	327/13198	2,7161E-09	9,025E-08	4,9406E-08	33
REACTOME_ACTIVATION_OF_BAD_AND_TRANSLOCATION_TO_MITOCHONDRIA_	8/412	15/13198	4,4776E-09	1,3047E-07	7,1424E-08	8
REACTOME_L1CAM_INTERACTIONS	19/412	121/13198	6,0767E-09	1,6926E-07	9,2656E-08	19

Supplementary Table 3.2 Top 20 specific pathways in schizophrenia against bipolar disorder

ID	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count
REACTOME_EUKARYOTIC_TRANSLATION_ELONGATION	5/33	94/13198	3,34E-06	0,000376	0,000235	5
REACTOME_RESPONSE_OF_EIF2AK4_GCN2_TO_AMINO_ACID_DEFICIENCY	5/33	102/13198	4,99E-06	0,000422	0,000264	5
REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	5/33	113/13198	8,25E-06	0,000433	0,000271	5
REACTOME_NONSENSE_MEDIATED_DECAY_NMD_	5/33	116/13198	9,38E-06	0,000433	0,000271	5
REACTOME_SELENOAMINO_ACID_METABOLISM	5/33	118/13198	1,02E-05	0,000433	0,000271	5
REACTOME_EUKARYOTIC_TRANSLATION_INITIATION	5/33	120/13198	1,11E-05	0,000433	0,000271	5
REACTOME_L1CAM_INTERACTIONS	5/33	121/13198	1,15E-05	0,000433	0,000271	5

REACTOME_RESPONSE_TO_ELEVATED_PLATELET_CYTOSOLIC_CA2_	5/33	132/13198	1,76E-05	0,000594	0,000372	5
REACTOME_INFLUENZA_INFECTION	5/33	157/13198	4,05E-05	0,001245	0,00078	5
REACTOME_SELECTIVE_AUTOPHAGY	4/33	82/13198	4,94E-05	0,001391	0,000871	4
REACTOME_REGULATION_OF_EXPRESSION_OF_SLITS_AND_ROBOS	5/33	172/13198	6,26E-05	0,001623	0,001016	5
REACTOME_RHO_GTPASES_ACTIVATE_IQGAPS	3/33	32/13198	6,72E-05	0,001623	0,001016	3
REACTOME_RRNA_PROCESSING	5/33	205/13198	0,000143	0,003232	0,002023	5
REACTOME_AGGREPHAGY	3/33	44/13198	0,000176	0,003716	0,002326	3
REACTOME_SIGNALING_BY_ROBO_RECEPTORS	5/33	218/13198	0,000191	0,003801	0,002379	5
REACTOME_RECYCLING_PATHWAY_OF_L1	3/33	49/13198	0,000243	0,0041	0,002567	3
REACTOME_GAP_JUNCTION_TRAFFICKING_AND_REGULATION	3/33	51/13198	0,000273	0,004399	0,002754	3
REACTOME_METABOLISM_OF_AMINO_ACIDS_AND_DERIVATIVES	6/33	374/13198	0,000288	0,004426	0,00277	6
REACTOME_DERMATAN_SULFATE_BIOSYNTHESIS	2/33	11/13198	0,000329	0,004816	0,003015	2

Supplementary Table 3.3 Top 20 specific pathways in bipolar disorder against schizophrenia

ID	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count
REACTOME_ADENYLATE_CYCLASE_ACTIVATING_PATHWAY	3/116	10/13198	7,5901E-05	0,01158254	0,0103545 4	3

REACTOME_ATF6_ATF6_ALPHA_ACTIVATES_CHAPERONE_GENES	3/116	10/13198	7,5901E-05	0,01158254	0,01035454	3
REACTOME_ION_HOMEOSTASIS	5/116	54/13198	0,00010789	0,01258193	0,01124797	5
REACTOME_PYRUVATE_METABOLISM_AND_CITRIC_ACID_TCA_CYCLE	5/116	55/13198	0,00011785	0,01258193	0,01124797	5
REACTOME_ATF6_ATF6_ALPHA_ACTIVATES_CHAPERONES	3/116	12/13198	0,00013738	0,01258193	0,01124797	3
REACTOME_ADENYLATE_CYCLASE_INHIBITORY_PATHWAY	3/116	14/13198	0,00022439	0,01426772	0,01275503	3
REACTOME_CIRCADIAN_CLOCK	5/116	70/13198	0,00036927	0,01878333	0,01679189	5
REACTOME_PKA_ACTIVATION_IN_GLUCAGON_SIGNALLING	3/116	17/13198	0,00041121	0,01960948	0,01753044	3
REACTOME_DAG_AND_IP3_SIGNALING	4/116	41/13198	0,00044642	0,01968929	0,01760179	4
REACTOME_PKA_MEDIATED_PHOSPHORYLATION_OF_CREB	3/116	20/13198	0,00067625	0,02579907	0,02306381	3
REACTOME_CITRIC_ACID_CYCLE_TCA_CYCLE_	3/116	22/13198	0,00090192	0,03128005	0,02796368	3
REACTOME_G_PROTEIN_MEDIATED_EVENTS	4/116	54/13198	0,00127704	0,03694178	0,03302514	4
REACTOME_CARDIAC_CONDUCTION	6/116	137/13198	0,00130725	0,03694178	0,03302514	6
REACTOME_ION_TRANSPORT_BY_P_TYPE_ATPASES	4/116	55/13198	0,00136796	0,03727692	0,03332475	4
REACTOME_GABA_RECEPTOR_ACTIVATION	4/116	60/13198	0,00189118	0,04955024	0,04429683	4
REACTOME_NEUROTRANSMITTER_RECEPTORS_AND_POSTSYNAPTIC_SIGNAL_TRANSMISSION	7/116	205/13198	0,00220215	0,05287087	0,04726547	7
REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES	8/116	269/13198	0,00255928	0,05579233	0,04987712	8

REACTOME_GLUCAGON_SIGNALING_IN_METABOLIC_REGULATION	3/116	33/13198	0,00297864	0,06313056	0,05643734	3
REACTOME_CA_DEPENDENT_EVENTS	3/116	37/13198	0,00413529	0,08303221	0,07422898	3
REACTOME_REGULATION_OF_INSULIN_LIKE_GROWTH_FACTOR_IGF_TRANSPORT_AND_UPTAKE_BY_INSULIN_LIKE_GROWTH_FACTOR_BINDING_PROTEINS_IGFBPS_	5/116	124/13198	0,00472323	0,09240577	0,08260873	5

Supplementary Table 4.1 model architecture details (encoder, ontodecoder, covariate_embeddings)

encoder:

```
(encoder): Encoder(
        (encoder): ModuleList(
        (0): Sequential(
        (0): Linear(in_features=8600, out_features=128, bias=True)
        (1): BatchNorm1d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
        (2): None
        (3): ReLU()
        (4): Dropout(p=0.2, inplace=False)
        )
        (1): Sequential(
        (0): Linear(in_features=128, out_features=128, bias=True)
        (1): BatchNorm1d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
        (2): None
        (3): ReLU()
        (4): Dropout(p=0.2, inplace=False)
        )
        (2): Sequential(
        (0): Linear(in_features=128, out_features=128, bias=True)
        (1): BatchNorm1d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
        (2): None
        (3): ReLU()
        (4): Dropout(p=0.2, inplace=False)
        )
        )
        (mu): Sequential(
        (0): Linear(in_features=128, out_features=128, bias=True)
        (1): Dropout(p=0.5, inplace=False)
        )
        (logvar): Sequential(
        (0): Linear(in_features=128, out_features=128, bias=True)
        (1): Dropout(p=0.5, inplace=False)
        )
)
```

ontodecoder:

```
(decoder): OntoDecoder(
        (decoder): ModuleList(
        (0): Sequential(
        (0): Linear(in_features=128, out_features=343, bias=True)
        (1): BatchNorm1d(343, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
        (2): None
        (3): ReLU()
        (4): None
        )
        (1): Sequential(
        (0): Linear(in_features=343, out_features=19, bias=True)
        (1): BatchNorm1d(19, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
        (2): None
        (3): ReLU()
        (4): None
        )
        (2): Sequential(
        (0): Linear(in_features=362, out_features=111, bias=True)
        (1): BatchNorm1d(111, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
        (2): None
        (3): ReLU()
        (4): None
        )
        (3): Sequential(
        (0): Linear(in_features=473, out_features=210, bias=True)
        (1): BatchNorm1d(210, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
        (2): None
        (3): ReLU()
        (4): None
        )
        (4): Sequential(
        (0): Linear(in_features=683, out_features=320, bias=True)
        (1): BatchNorm1d(320, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
        (2): None
        (3): ReLU()
        (4): None
```

```
)
(5): Sequential(
(0): Linear(in_features=1003, out_features=362, bias=True)
(1): BatchNorm1d(362, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): None
)
(6): Sequential(
(0): Linear(in_features=1365, out_features=341, bias=True)
(1): BatchNorm1d(341, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): None
)
(7): Sequential(
(0): Linear(in_features=1706, out_features=240, bias=True)
(1): BatchNorm1d(240, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): None
)
(8): Sequential(
(0): Linear(in_features=1946, out_features=133, bias=True)
(1): BatchNorm1d(133, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): None
)
(9): Sequential(
(0): Linear(in_features=2079, out_features=75, bias=True)
(1): BatchNorm1d(75, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): None
)
```

(10): Sequential(
(0): Linear(in_features=2154, out_features=29, bias=True)
(1): BatchNorm1d(29, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): None
)
(11): Sequential(
(0): Linear(in_features=2183, out_features=5, bias=True)
(1): BatchNorm1d(5, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): None
)
(12): Sequential(
(0): Linear(in_features=2188, out_features=8600, bias=True)
(1): Sigmoid()
)
)
)

covars_embeddings:

(covars_embeddings): ModuleDict(
(condition): Embedding(3, 128)
(gender): Embedding(2, 128)
)
(covars_classifiers): ModuleDict(
(condition): Classifier(
(classifier): ModuleList(
(0): Sequential(
(0): Linear(in_features=128, out_features=64, bias=True)
(1): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): Dropout(p=0.2, inplace=False)
)

```
(1): Sequential(
(0): Linear(in_features=64, out_features=64, bias=True)
(1): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): Dropout(p=0.2, inplace=False)
)
(2): Sequential(
(0): Linear(in_features=64, out_features=3, bias=True)
(1): BatchNorm1d(3, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): Softmax(dim=1)
)
)
)
(gender): Classifier(
(classifier): ModuleList(
(0): Sequential(
(0): Linear(in_features=128, out_features=64, bias=True)
(1): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): Dropout(p=0.2, inplace=False)
)
(1): Sequential(
(0): Linear(in_features=64, out_features=64, bias=True)
(1): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): ReLU()
(4): Dropout(p=0.2, inplace=False)
)
(2): Sequential(
(0): Linear(in_features=64, out_features=2, bias=True)
(1): BatchNorm1d(2, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(2): None
(3): Sigmoid()
```

```
        )
        )
        )
  )
```

Supplementary Table 4.2 model hyperparameter settings

weighted loss	recact layer	actdec layer	adv coefficient	learning rate vae	learning rate adv
True	torch.nn.Sigmoid	torch.nn.ReLU	100.0	1e-3	1e-3

neuron number per node	batchnorm decoder	activation decoder	pos weights	number of epoch	batch size
1	True	True	True	700	128

Acknowledgement

First and foremost, I would like to express my gratitude to my supervisor, Prof. Dr. Carl Herrmann for his patience, invaluable advice, continuous support and constant encouragement during my doctoral study. I feel fortunate to be in the Biomedical Genomics Group and grateful to Carl for giving me this opportunity 4 years ago, to be able to join and learn in this field integrating biomedicine and machine learning. I would like to thank all the brothers, sisters and 亲人 as well as the interns to build and maintain the unique and greatest atmosphere in the group, Aaron, Albert, Ana, Andres, Ashwini, Carlos, Carolina, Daria, David, Eugenia, Han, Jean, Kersten, Lin, Nelly, Nils, Pablo, Qianwu, Robin, Saifullah, Sofia, Tiago, Wangjun, Wenjun, Yimin, Youcheng and their families. I would also like to thank my thesis committee members Prof. Dr. Julio Saez-rodriguez, Prof. Dr. Benedikt Brors and Dr. Jürgen Pahle for their suggestions during the TAC meeting, as well as Prof. Emanuel Schwarz and Prof. Maïwen Caudron-Herger for joining the doctoral examination committee. In addition, I would like to thank COMMITMENT for supporting my Ph.D with the doctoral fellowship, collaborators in the consortium and the staff of BioQUANT as well as IPMB for providing me with valuable advice during the study. In the end, I want to thank my parents and Xin for everything.