Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Medicine 2093

Biomarkers for depression: genetic, epigenetic, and expression evidence

ALEKSANDR V. SOKOLOV

ACTA UNIVERSITATIS
UPSALIENSIS
2024

ISSN 1651-6206
ISBN 978-91-513-2270-4
urn:nbn:se:uu:diva-540129

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in room A1:111a, Uppsala biomedicinska centrum (BMC), Husargatan 3, Uppsala, Wednesday, 4 December 2024 at 13:00 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Docent Daniel Lindqvist (Unit for Biological and Precision Psychiatry, Lund University, Sweden).

Abstract

Sokolov, A. V. 2024. Biomarkers for depression: genetic, epigenetic, and expression evidence. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 2093. 98 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2270-4.

Depression is a very prevalent disorder affecting between 2 to 21% of the world population. This thesis extends the knowledge on the biological aspects of depression, aiming to identify and validate markers of genetic, epigenetic, and gene expression origin.

In Study I, the main focus was depression-related gene MAD1L1 that was previously linked to depression by SNPs and frequently mentioned as a stress-related marker. We identified that depression-related SNPs in MAD1L1 affect DNA methylation levels at cg02825527, cg18302629, and cg19624444 that were associated with depressive phenotypes in independent cohorts.

In Study II, we investigated whether GWAS catalog depression SNPs located in Olink-detectable genes could be replicated in a UKBiobank cohort and whether these associations are supported by DNA methylation and transcriptome. We validated eight depression SNPs and found very weak evidence that TNXB may be related to depression.

Study III was based on comparison of different depression -OMIC layers, including genetics, DNA methylation, and transcriptome. We explored how the identified genes from different -OMICs overlap, are functionally related and if they could show patterns in drugs and clinical trials. Only three genes were supported by evidence at all three -OMIC levels and included: FOXP1, VPS41, and AKTIP. Different -OMIC levels showed involvement of multiple systems in depression.

In Study IV, we used the Neuro Exploratory panel (Olink) to identify depression proteomic changes in blood. We took antidepressant intake into the account and validated associations in the independent datasets. We identified several proteins that showed nominally different levels between depression risk groups in the adolescent cohort. Validation of identified markers yielded that only PPP3R1 was also differentially expressed in prefrontal cortex and whole blood in the independent open-access cohorts with matching association directions.

In Study V, we used the entire blood DNA methylation as a depression marker. We investigated stability of DNA-methylation in eight independent datasets with meta-analysis and compared common machine learning and deep learning strategies for the depression detection purposes. We found 1987 CpG sites related to depression in both mega- and meta-analysis at the nominal level. Random forest classifiers achieved the best performance in identifying depression based on DNA methylation data in blood (AUC 0.73 and 0.76) in CV and hold-out tests respectively on the batch-level processed data.

Overall, the thesis supports multiple depression genetic, epigenetic, and expression markers. However, identified individual and systemic depression changes show high variability, which is in agreement with previous studies and observations.

Keywords: Depression, Biomarkers, Epigenetics, DNA methylation, Proteomics, Transcriptomics, Genetics, Machine learning

Aleksandr V. Sokolov, Functional Pharmacology and Neuroscience, 593, Uppsala University, SE-75124 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-513-2270-4
URN urn:nbn:se:uu:diva-540129 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-540129)

Dedicated to my family and friends

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Sokolov, A. V.**, Manu, D.-M., Nordberg, D. O. T., Boström, A. D. E., Jokinen, J., & Schiöth, H. B. (2023) Methylation in MAD1L1 is associated with the severity of suicide attempt and phenotypes of depression. *Clinical Epigenetics*, 15(1), 1

II Lafta M. S, **Sokolov, A. V.**, Rukh G., & Schiöth, H. B. (2023) Identification and validation of depression-associated genetic variants in the UK Biobank cohort with transcriptome and DNA methylation analyses in independent cohorts (Manuscript)

III **Sokolov, A. V.**, Arloth J, & Schiöth, H. B. (2023) Cross-omics cross-cohort analysis indicates multiple biological systems implicated in depression (Manuscript)

IV **Sokolov, A. V.**, Lafta M.S., Nordberg D.O.T., Jonsson J., & Schiöth, H. B. (2024) Depression proteomic profiling in adolescents with transcriptome analyses in independent cohorts. *Frontiers in Psychiatry* (15:1372106)

V **Sokolov, A. V.** & Schiöth, H. B. (2024) Decoding depression: a comprehensive multi-cohort exploration of blood DNA methylation using machine learning and deep learning approaches. *Translational Psychiatry*. 14(1):287

Additional publications

This list includes extra publications and manuscripts written by the author during studies. These articles are not a part of main thesis work.

1 Attwood, M. M., Fabbro, D., **Sokolov, A. V.**, Knapp, S., & Schiöth, H. B. (2021). Trends in kinase drug discovery: Targets, indications and inhibitor design. Nature Reviews Drug Discovery, 20(11), 839–861

2 **Sokolov, A. V.**, Dostdar, S. A., Attwood, M. M., Krasilnikova, A. A., Ilina, A. A., Nabieva, A. Sh., Lisitsyna, A. A., Chubarev, V. N., Tarasov, V. V., & Schiöth, H. B. (2021). Brain Cancer Drug Discovery: Clinical Trials, Drug Classes, Targets, and Combinatorial Therapies. Pharmacological Reviews, 73(4), 1172–1203

3 Mälarstig, A., Grassmann, F., Dahl, L., Dimitriou, M., McLeod, D., Gabrielson, M., Smith-Byrne, K., Thomas, C. E., Huang, T.-H., Forsberg, S. K. G., Eriksson, P., Ulfstedt, M., Johansson, M., **Sokolov, A. V.**, Schiöth, H. B., Hall, P., Schwenk, J. M., Czene, K., & Hedman, Å. K. (2023). Evaluation of circulating plasma proteins in breast cancer using Mendelian randomisation. Nature Communications, 14(1), 7680

4 Smith-Byrne K., Hedman Å., Dimitriou M., Desai T., **Sokolov, A. V.**, Schioth H.B., Koprulu M., Pietzner M., Langenberg C., Atkins J., Penha R.C., McKay J., Brennan P., Zhou S., Richards B.J., Yarmolinsky J., Martin R.M., Borlido J., Mu X.J., Butterworth A., Shen X., Wilson J., Assimes T.L., Hung R.J., Amos C., Purdue M., Rothman N., Chanock S., Travis R.C., Johansson M., Mälarstig A. (2024) Identifying therapeutic targets for cancer among 2074 circulating proteins and risk of nine cancers. Nat Commun. 15(1):3621

5 Desai T. A., Hedman Å.K., Dimitriou M., Koprulu M., Figiel S., Yin W., Johansson M., Watts E.L., Atkins J.R., **Sokolov, A. V.**, Schiöth H.B., Gunter M.J., Tsilidis K.K., Martin R.M., Pietzner M., Langenberg C., Mills I.G., Lamb A.D., Mälarstig A., Key T.J.; The PRACTICAL Consortium; Travis R.C., Smith-Byrne K. (2024) Identifying proteomic risk factors for overall, aggressive, and early onset prostate cancer using Mendelian Randomisation and tumour spatial transcriptomics. EBioMedicine. 15(1):3621

6 Lafta, M. S., **Sokolov, A. V.**, Landtblom, A., Ericson, H., Schiöth, H. B., & Abu Hamdeh, S. (2023). Exploring biomarkers in trigeminal neuralgia patients operated with microvascular decompression: A comparison with multiple sclerosis patients and non-neurological controls. European Journal of Pain, ejp.2231

7 Lafta, M. S., Rukh G., Hamdeh S. A., Molero Y., **Sokolov, A. V.**, Rostami E., Schiöth H. B. Genomic Validation in the UK Biobank Cohort Suggests a Role of C8B and MFG-E8 in the Pathogenesis of Trigeminal Neuralgia. (2024) J Mol Neurosci. 74(4):91

8 Andreoli M.F., Kruger A.L., **Sokolov, A. V.**, Rukh G., De Francesco P.N., Perello M., & Schiöth, H. B. LEAP2 is associated with impulsivity and reward sensitivity depending on the nutritional status and decreases with protein intake in humans. (2024) Diabetes Obes Metab. Epub ahead of print

9 Namiot, E. D., Smirnovová, D., **Sokolov, A. V.**, Chubarev, V. N., Tarasov, V. V., & Schiöth, H. B. (2023). The international clinical trials registry platform (ICTRP): Data integrity and the trends in clinical trials, diseases, and drugs. Frontiers in Pharmacology, 14, 1228148

10 Namiot E.D., Smirnovová D., **Sokolov, A. V.**, Chubarev V.N., Tarasov V.V., & Schiöth H.B.. Depression clinical trials worldwide: a systematic analysis of the ICTRP and comparison with ClinicalTrials.gov. (2024) Transl Psychiatry. 14(1):315

11 Nazarova, V. A., **Sokolov, A. V.**, Chubarev, V. N., Tarasov, V. V., & Schiöth, H. B. (2022). Treatment of ADHD: Drugs, psychological therapies, devices, complementary and alternative methods as well as the trends in clinical trials. Frontiers in Pharmacology, 13, 1066988

12 Niemi, J. V. L., **Sokolov, A. V.**, & Schiöth, H. B. (2022). Neoantigen Vaccines; Clinical Trials, Classes, Indications, Adjuvants and Combinatorial Treatments. Cancers, 14(20), 5163

13 Namiot, E. D., **Sokolov, A. V.**, Chubarev, V. N., Tarasov, V. V., & Schiöth, H. B. (2023). Nanoparticles in Clinical Trials: Analysis of Clinical Trials, FDA Approvals and Use for COVID-19 Vaccines. International Journal of Molecular Sciences, 24(1), 787

Contents

1 Introduction 13
1.1 Mental health, depression, suicide 13
1.1.1 Depression and its characterization 13
1.1.2 Suicidal behavior and its characterization 14
1.2 Biological factors in depression and suicide 15
1.2.1 Overview of genetic variation 16
1.2.2 Genetic factors in depression 17
1.2.3 Genetic factors in suicide 18
1.2.4 Epigenetic factors 18
1.2.5 DNA methylation biology 19
1.2.6 DNA methylation in depression and suicide 19
1.2.7 Transcriptomic and proteomic factors 21
1.3 Advances in high-throughput screening studies 22
1.3.1 Array-based analysis of DNA methylation 22
1.3.2 High-throughput transcriptomics and proteomics 24
1.3.3 Limitations of high-throughput biomarker discovery 27
1.4 Data analysis, modeling, machine learning and deep learning in -OMICs 27
1.4.1 Association modeling in the -OMICs field 28
1.4.2 Machine learning in the depression -OMICs field 30
1.4.3 Brief intro to deep learning 31
1.4.4 Deep learning and OMICs 33

2 Aims 35

3 Materials and methods 37
3.1 Ethics declaration 37
3.2 Project workflows 37
3.2.1 Study I 37
3.2.2 Study II 38
3.2.3 Study III 38
3.2.4 Study IV 39
3.2.5 Study V 39
3.3 Participants and cohorts 39
3.3.1 PSY cohort 39
3.3.2 SKI cohort 42
3.3.3 UK Biobank cohort 43

3.3.4 DNA methylation open-access cohorts 44
3.3.5 Transcriptome open-access cohorts 49
3.4 Sample collection and DNA methylation profiling 50
3.5 DNA methylation data preprocessing 50
3.6 Transcriptome data preprocessing 52
3.7 Proteome data preprocessing 53
3.8 SNP data collection 53
3.9 Data analyses and statistical modeling 54
3.9.1 Study I 54
3.9.2 Study II 56
3.9.3 Study III 57
3.9.4 Study IV 59
3.9.5 Study V 60
3.10 Machine learning and deep learning models in Study V 61
3.10.1 Feature selection strategies 61
3.10.2 Machine learning models 61
3.10.3 Deep learning models 62
3.10.4 Model selection, training and evaluation 64
3.11 Gene enrichment analyses 65
3.12 Thesis writing 65

4 Results 66
4.1 Study I 66
4.2 Study II 68
4.3 Study III 69
4.4 Study IV 71
4.5 Study V 72

5 Discussion 75

6 Future perspectives 78

7 Acknowledgements 80

References 82

Abbreviations

Attention deficit hyperactivity disorder (ADHD)
AKT-interacting protein (AKTIP)
Binary cross-entropy (BCE)
Bipolar Disorder (BD)
Beck Depression Inventory (BDI)
Brain-derived neurotrophic factor (BDNF)
Beck Hopelessness Scale (BHS)
Body mass index (BMI)
Beck Scale for Suicide Ideation (BSSI)
Children's Depression Inventory (CDI)
Children's Depression Rating Scale (CDRS)
Center for Epidemiologic Studies Depression Scale (CES-D)
Columbia Suicide Severity Rating Scale (CSSRS)
Development and Well-Being Assessment (DAWBA)
Deep learning (DL)
Diagnostic and Statistical Manual of Mental Disorders version 5(DSM-5)
Epigenome-wide association study (EWAS)
Findable, Accessible, Interoperable, Reusable (FAIR)
Forkhead box protein P1 (FOXP1)
Generalized liner models (GLM)
Genetic risk scores (GRS)
Genome-wide association study (GWAS)
Hamilton Depression Rating Scale (HAMD)
Liquid chromatography-tandem mass spectrometry (LC-MS/MS)
Mitotic arrest deficient 1 like 1 (MAD1L1)
Montgomery–Åsberg Depression Rating Scale (MADRS)
Minor allele frequency (MAF)
Major depressive disorder (MDD)
Machine learning (ML)
Mass-spectrometry (MS)
Mean squared error (MSE)
National Center for Biotechnology Information (NCBI)
Next generation sequencing (NGS)
National Institute of Child Health and Human Development (NICHD)
National Institutes of Health (NIH)
Normalized Protein Expression (NPX)
Principal component analysis (PCA)

Proximity extension assay (PEA)
Patient Health Questionnaire (PHQ-9)
Proteome-wide association study (PWAS)
Radial basis function (RBF)
Robust multi-array average (RMA)
Rectified Linear Unit (ReLU)
Schizophrenia (SCZ)
Single-nucleotide polymorphism (SNP)
Suicide Probability Scale (SPS)
Suicide Assessment Scale (SUAS)
Support vector machine (SVM)
Transcriptome-wide association study (TWAS)
University of California, Santa Cruz (UCSC)
Variational autoencoder (VAE)
Vacuolar protein sorting-associated protein 41 (VPS41)
Whole-genome amplified (WGA)
World Health Organization (WHO)
Expression quantitative trait locus (eQTL)
Methylation quantitative trait locus (meQTL or mQTL)

1. Introduction

1.1 Mental health, depression, suicide

1.1.1 Depression and its characterization

Mental health is a significant challenge for healthcare systems worldwide. For instance, the U.S. National Institutes of Health (NIH) estimates that as many as 20% of adults are presumed to have mental health problems [1]. Based on the EU data from 2010, the most common mental health disorders included anxiety (14.0%), insomnia (7.0%), major depression (6.9%), somatoform disorders (6.3%), alcohol and drug abuse (>4%), attention deficit hyperactivity disorder (ADHD) (5%) in the young, and dementia (1-30%, depending on age) [2]. Depression (and its most prominent form Major depressive disorder (MDD)) is one of these diagnoses that profoundly impact person's life. This condition affects an estimated 2 to 21% of the global population over a lifetime as demonstrated by multiple studies conducted in multiple countries [3].

Depression is the second leading contributor to a global disease burden according to estimates from Global Burden of Disease Consortium in 2013 [4]. MDD is reported to be associated with other psychiatric and non-psychiatric diagnoses including stroke, heart-related diseases, diabetes, Alzheimer disease, obesity, and cancer [4]. Etiologically MDD has been consistently linked to both biological factors and psychological/societal factors. These factors include individual's sex (women are twice affected compared to men) [4], genetic factors [5], epigenetic factors [6], expression factors [7], lifestyle choices (smoking and alcohol intake) [3], and even cultural factors, such as religion [8].

Currently, MDD is diagnosed by an episode of depressed mood or a loss of interest or pleasure for at least 2 weeks. This episode must be accompanied by other psychiatric/physiological markers, including loss of appetite, sleep disturbances, psychomotor agitation or retardation, fatigue, lack of concentration, thoughts of death, etc. These criteria are extensively described in latest (fifth) edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [9].

To assist in the diagnosis and identification of psychiatric conditions, including MDD, multiple psychiatric questionnaires (inventories) were developed. In these questionnaires, an individual or a related person is asked to evaluate

specific aspects of their life and their corresponding mental state in a quantitative form. Each item in these inventories has possible answers that typically correspond to a numeric value, and the total psychiatric score for the inventory is typically a sum of several individual items. The list of already developed inventories for depression includes many examples, such as the Development and Well-Being Assessment (DAWBA) [10, 11], Beck Depression Inventory (BDI) [12, 13], Montgomery–Åsberg Depression Rating Scale (MADRS) [14, 15], Hamilton Depression Rating Scale (HAMD) [16], Center for Epidemiologic Studies Depression Scale (CES-D) [17], Patient Health Questionnaire (PHQ-9) [18], and other. Some instruments were developed specifically for adolescents, including Children's Depression Inventory (CDI) [19], Children's Depression Rating Scale (CDRS) [20] and other. The aforementioned questionnaires can be administered either in the clinical setting or individually during self-assessment. In the case of self-assessment, modifications to the questionnaire may be necessary. All of these instruments yield numeric scores for characterizing depression phenotypes. Typically, predefined thresholds are employed to categorize participants into distinct groups based on their initial numeric scores.

As of 2024, there is no approved lab-based/biomarker-based method for depression diagnosis despite multiple research efforts.

1.1.2 Suicidal behavior and its characterization

Suicidal behavior represents a psychiatric state with potentially devastating consequences not only for the individual but also for their family and society at large. It is crucial to emphasize that suicide is not regarded as an isolated disease; rather, it is recognized as a specific symptom that may be associated with other disorders, primarily including MDD, Bipolar Disorder (BD), substance use disorders, and schizophrenia (SCZ) [21].
Suicide is a relatively prevalent phenomenon especially among individuals with MDD. Suicide is associated with approximately 1.8 times higher mortality rates among patients with MDD, and these individuals are expected to loose 10.6 (men) and 7.2 (women) years in life expectancy compared to individuals without the diagnosis [23, 24]. From the epidemiological perspective, it is also important to distinguish different stages of suicidal behavior *"progression"*. The lifetime prevalence of suicidal ideations (a thinking process of committing suicide) is relatively common and reaches as high as 9.2% based on the data from 17 countries [25]. Meanwhile, the probability that individual with a lifetime history of suicidal ideations would ever make a suicide attempt is approximately 30% [25], and eventually, only 5% of suicide attempts result in death according to data from the WHO [25]. The prevalence of suicide is related to multiple factors, such as sex (suicide rates are higher in men)

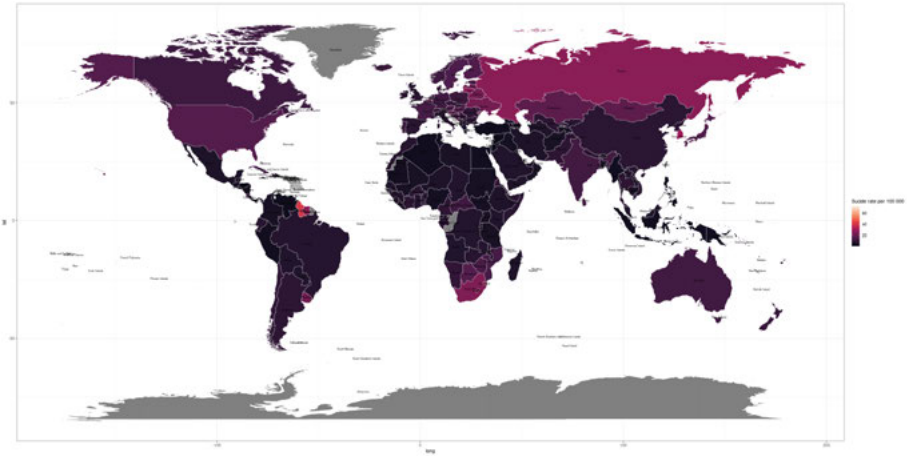

Figure 1. Suicide rate per 100 000 by country (2019, WHO). This figure has been created based on the data from [22]. The R package *ggplot2* was used to make data visualization. Lighter colors indicate higher suicide rates. Data is due 2019.

[21, 22], cultural aspects [26], lifestyle [27], country, and other factors, such as biological alterations [28].

Diagnosis of suicidal behavior is typically discussed and pursued in relation to an associated medical condition. As previously mentioned, suicide is not considered as an individual condition in DSM-5, however, it is mentioned in the associated psychiatric conditions, such as BD I/II, anxiety disorders, ADHD, SCZ, depressive disorders, and other as a diagnostic criterion. The accurate diagnosis of associated conditions should also lead to the identification of individuals *at risk* for suicide. The identification of suicidal behavior could be achieved via using screening tools including questionnaires [29, 30]. Currently, many instruments for suicide risk assessment were developed, including Suicide Assessment Scale (SUAS) [31], Beck Scale for Suicide Ideation (BSSI), Columbia Suicide Severity Rating Scale (CSSRS), Suicide Probability Scale (SPS), Screening for Depression and Thoughts of Suicide, Beck Hopelessness Scale (BHS), and many other [32].

Similar to MDD, no lab-based/biomarker-based method for suicide risk assessment has been approved so far as of 2024.

1.2 Biological factors in depression and suicide

Depression is commonly understood to be a primary result of environmental exposure. However, an increasing body of research also underscores the significance of biological components in the pathogenesis of this disorder. It

should be acknowledged that investigation of biological factors in psychiatry is a very challenging task as all of such studies are impeded by inconsistencies in the definition of phenotypes, confounding, unknown biases, as well as limited sample sizes. Thus, as one might expect, studies in depression generally produce inconsistent results, with overlaps being a rarity rather than the norm. This section provides the overview of the current state of depression research with respect to genetics, epigenetics, and transcriptomics/proteomics.

1.2.1 Overview of genetic variation

The inherent advantage in investigating genetic markers lies in the static nature of the genotype within an individual. This means that the genotype remains unchanged by environmental exposure, barring somatic mutations, and is consistent across various tissues of the participant, with the exception of anucleate cells and gametes. The study of genetic markers can be pursued through diverse methodologies, including functional studies *in vitro/in vivo* and screening approaches such as genome-wide association studies (GWAS). For research involving human subjects, GWAS is particularly common and is the primary focus of this section. GWAS is a type of cross-sectional observational study that explores the relationships between genetic variants and specific outcomes, such as depression. The primary objective of GWAS is to determine whether individuals with a certain disease or trait exhibit statistically significant variations in the frequency of specific genetic markers. Within the scope of GWAS, the range of potential marker candidates is extensive as the analysis typically encompasses the entire genome [33].

The most commonly investigated genetic markers in genomic studies are single-nucleotide polymorphisms (SNPs). A SNP represents a germline substitution of a single nucleotide at a specific locus in the genome, occurring with high relative frequency (>1% within a population). SNPs may be located within genes (intragenic) or outside of gene sequences (intergenic). The influence of SNPs on specific traits can be elucidated through various mechanisms. The most simple and direct effect could be exemplified when an intragenic SNP alters the amino acid sequence of the resultant protein, potentially modifying its structure and function, thereby affecting the phenotype [34]. However, approximately 60% of SNPs are located outside coding gene regions, and of the remaining ones, more than 90% are found in introns [35]. The impact of non-coding SNPs may involve alterations in mRNA splicing, its stability and structure, as well as protein folding [35]. In such instances, SNPs can influence gene expression, leading to their classification as expression quantitative trait loci (eQTLs). The eQTL analysis specifically aims to identify genetic variants that affect gene expression [36]. Similarly, SNPs may also be impacting epigenetic regulation, such as DNA methylation, in which case

they are referred to as methylation quantitative trait loci (meQTLs or mQTLs). Research on mQTLs represents a distinct area of study, focusing on the interaction between genetic variants and methylation patterns [37].

1.2.2 Genetic factors in depression

It is estimated that the genetic component of MDD could be as high as 35%, suggesting a genetic component in the disorder [4]. As of 2024, there are multiple studies that investigated genetic markers in relation to depression or other psychiatric condition. Since depression is a multifactorial disorder that is related to multiple *variables*, the investigation of any particular SNP is rather complicated and requires sufficient sample size. Recent GWASs shed the light on the genetic components of depression [5, 38]. However, polygenic risk scores (PRS) from the largest study with 246363 depression cases and 561190 controls showed small predictive power (1.5–3.2% of explained variance) in independent samples [39].

The overall conclusions of the studies above are mixed but lean to the neurogenic concept of depression. In the study by MDD working group of the Psychiatric GWAS consortium, the main analysis identified no loci that passed multiple comparison threshold, emphasizing potential overly high heterogeneity of MDD, divergent genetic architectures or insufficient power [5]. The second largest study by Wray *et al.*, on the other hand, identified 44 individual SNPs at the genome-wide significant level [38]. These SNPs highlighted genes that are related to CNS neuron differentiation, voltage-gated calcium channels, cytokine and immune response, synapse, and other [38]. These results somewhat match the relatively established neural plasticity theory of depression [40] and potentially the serotonin theory of depression [41]. The last largest depression GWAS (as of 2024), in turn, was a meta-analysis and identified 102 genetic markers that also highlighted related pathways [39]. Some of the genes identified in GWAS appear to be particularly interesting. For instance, mitotic arrest deficient 1 like 1 (MAD1L1), a component of the mitotic spindle assembly checkpoint, has been linked to depression [42, 43, 39, 44, 45] as well as SCZ [46, 47, 48, 49, 50, 51, 52] and BD [52, 53, 54, 55, 56] in multiple GWASs.

It should be mentioned, however, that the overall reproducibility of specific genetic markers in psychiatry is a subject of intense discussion. The recent analysis of psychiatric GWAS deposited in GWAS catalog database identified 1109 genome-wide significant SNPs (as of 2019) [57]. Interestingly enough, only 133 of these SNPs (~12%) were replicated at least in one separate publication, whereas 379 SNPs (~34%) were replicated only within the original publications [57]. Though the authors did not specify depression, it could be

reasonably assumed that depression should have somewhat similar estimates. Interestingly, the 12% estimate matches our findings on depression (see results for **Study III**).

1.2.3 Genetic factors in suicide

Overall, it is believed that suicidal behavior, same as depression, has a genetic component. Several studies suggested heritability of suicide of up to 50%, and GWASs highlighted genes, such as *TBX20*, *GNAL*, *BACE1*, *NREP* that are also linked to other neurological disorders [21]. A simple search in GWAS Catalog, a database collecting results from GWASs [58], shows that more than 340 SNPs are related to suicide behavior in more than 40 studies as of December 2023. However, neither candidate gene-based studies nor GWAS on suicide produced consistent results [21].

The study of genetic factors related to suicidal behavior faces challenges akin to those encountered in early depression GWAS. Furthermore, this field is additionally constrained by even more pronounced power limitations. For instance, the most recent and extensive meta-analytic genome-wide association study (GWAS) on suicide death and suicidal behavior included over 250000 participants, with only 8315 identified as cases [59]. By contrast, the largest GWAS on depression conducted by Howard *et al.* encompassed 246363 cases and 561190 controls, representing nearly 30 times more cases and 3 times more total samples. This large disparity between cases and controls further hampers identification of robust genetic markers.

Consistently, the aforementioned suicide GWAS identified only a single suicide-related SNP (rs73182688) that survived the correction for multiple testing in the discovery cohort, whereas no variants passed the similar threshold across five validation cohorts [59]. The enrichment analysis of suicide-related SNPs in this study indicated the involvement of immune system pathways. However, the top hit rs73182688 is related to a gene *NLGN* encoding a cell adhesion protein neuroligin facilitating synaptic formation [60]. Thus, it is further important to highlight the need of replication studies in this area.

1.2.4 Epigenetic factors

The realization that genetics alone cannot fully explain the mechanisms and properties of biological systems emerged relatively long time ago. This is exemplified by the phenotypical variations observed in monozygotic twins. Even though they could be nearly identical in appearance, such organisms frequently show inconsistencies regarding complex diseases [61]. As of today, these inconsistencies are primarily explained through gene-environmental

interactions, and the epigenetic regulation plays one of the pivotal roles in explaining these differences. The term epigenetic regulation encompasses a variety of mechanisms that influence specific characteristics of individual cells and tissues [62]. These mechanisms include histone modifications, such as methylation, demethylation or acetylation, formation of regulatory non-conding RNAs, such as microRNAs (miRNAs), as well as modifications of DNA, such as DNA methylation, and other mechanisms. In the context of this thesis, the author specifically focuses on the role of DNA methylation, though that does not imply that this specific mechanism is of greater importance than other epigenetic machinery.

1.2.5 DNA methylation biology

DNA methylation is a dynamic process involving the addition of a methyl group CH_3- at the 5-th position of cytosine within the DNA sequence by a specialized group of enzymes known as DNA methyltransferases. This process predominantly occurs not at every cytosine within the DNA, but specifically when cytosine nucleotide is followed by a guanine nucleotide (CpG site). Methylation of CpG sites is extremely common, and more than 80% of CpGs are methylated in the human genome [63]. The implications of CpG methylation are substantial, and it plays critical roles in various biological processes including development, functioning of CNS/nervous system, immune system, cell growth and diseases, such as cancer [64]. Conversely, non-CpG methylation, while less frequent, is not insignificant, and could reach even 15% of total methylation (embryonic stem cells) [65] and also has important roles in diseases, such as cancer [66] as well as in brain development [67].

The regulatory properties of DNA methylation/demethylation are related to its effects on transcription. Interestingly, the exact effect of methylation on expression of a specific gene is still frequently not completely understood. However, it is generally accepted that DNA methylation leads to gene silencing, especially if applied to CpG sites and CpG islands at promoter regions [64]. Methylation at gene body regions, on the other hand, may lead to an increase in the corresponding expression [68, 69, 70], whereas the effects of methylation at the enhancers and transcription factor binding sites is debated [68, 71].

1.2.6 DNA methylation in depression and suicide

DNA methylation plays multifaceted roles in the development and functioning of the nervous system [70]. For example, studies have shown that methylation patterns undergo temporary changes following neural activation [72]. In the realm of psychiatric disorders, in turn, differential DNA methylation has been

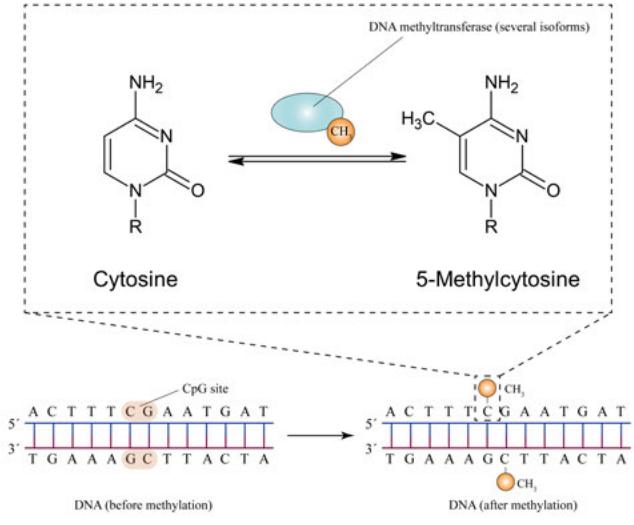

Figure 2. A schematic representation of CpG methylation

associated with a range of conditions, including MDD [6], SCZ [73], BD [74], and suicidal behavior [75]. Our research group has participated in exploring the links between DNA methylation and depression among adolescents, as well as in the general population using public datasets [76, 77, 78]. We have also investigated the relationship between DNA methylation and suicidal behavior [79]. Overall, studies on DNA methylation in depression highlighted distinct biological pathways, including neuron fate specification, stress, glial cell differentiation, development, synaptic plasticity, calcium signaling, inflammation, axon guidance, and other [80, 81, 82, 83, 84]. Multiple potential systems, such as stress response, serotonergic signaling, brain-derived neurotrophic factor (BDNF) signaling, were also found related to suicidal behavior [85, 86, 87, 88]. One group of studies, considered DNA methylation in a way similar to PRS, measuring its ability for depression prediction in isolated cohorts with a limited number of methods [80, 89, 90]. Some studies, however, did not identify any shifts in depression DNA methylation [91].

In contrast to genetic markers, DNA methylation is dynamic and vary greatly between tissues [64], thus rendering any studies of DNA methylation underpowered (due to low sample sizes) and extremely specific and applicable only within a specific tissue, for instance blood or cells in prefrontal cortex. By definition, studies on difficult to obtain tissues such as brain cells are generally rare in the context of psychiatry as such research would require obtaining post-mortem samples. This approach is also very problematic to scale and utilize for screening and diagnostic purposes. On the other hand, the blood

samples are easily available, and thus most of DNA methylation research on depression and suicide is performed, employing the blood samples. However, using blood samples also raises the question on whether the state of blood adequately reflects changes in the brain. Several studies explored this issue and found that DNA methylation in blood shows only modest correlations with the one in the brain [92, 93, 94].

Generally, it could be stated that there is currently a lack of consensus among DNA methylation studies in depression (and also suicide) attributed to several factors, including the heterogeneity of depression, population variability, confounding factors, and statistical and methodological challenges due to the low power or the choice of methodology (e.g. Hypothesis testing [95]). Typically, DNA methylation studies are conducted in a from of epigenome-wide association studies (EWAS), which involve numerous association models and require large sample sizes to yield statistically significant results after taking into the account multiple comparisons. The need for high sample sizes resulted in many studies using a "targeted" approach instead that requires less statistical power, but may limit the identification of new potential marker candidates, limit the coverage of markers [96], as well as potentially produces statistically inflated results [97]. Additionally, many nominally significant/non-significant results are likely to be not reported at all. However, all of these issues apply to all high-throughput -OMIC studies.

1.2.7 Transcriptomic and proteomic factors

Transcriptomic and proteomic analyses could offer more clear and direct insights into biological processes and their relation to phenotypic variations. Gene expression is influenced by genotype, epigenetics, and environmental conditions, reflecting the product of these factors in the relative quantities of transcripts and proteins. Transcriptome data, specifically, provides information not only regarding potential protein concentration (as transcripts are protein precursors) but also regarding regulatory factors such as non-coding RNAs. Proteomic data, on the other hand, provides precise insights into the functional proteins and reveal the actual presence and functional state of these proteins, including post-translational modifications. Additionally, both proteomic and transcriptomic analyses can better identify biomarkers and therapeutic targets for diseases of interest, as proteins are the most common targets for drugs as well as one of the most common types of soluble biomarkers.

In depression, many gene expression studies have highlighted pathways related to inflammation [98, 99], neurotransmission [100], neuroplasticity [101, 102, 99], and stress [98]. Although many studies still frequently highlight distinct genes, a significant overlap is observed in many studies, particularly

regarding inflammation. Markers such as *BDNF*, *TNF*, *IL-1*, *IL-6*, *IL-8* were reported in multiple studies [7]. Proteomic studies in depression, in turn, are less common and generally performed in small samples, especially if they involve materials coming from brain tissues. One review article provided an overview of results from 10 studies investigating proteomic differences in blood-derived samples in MDD cases compared to controls. These studies had overlapping results with several proteins reported a few times, including ceruloplasmin, alpha-2-macroglobulin, apolipoprotein B-100, apolipoprotein D, BDNF, interleukin-1 receptor antagonist protein, macrophage migration inhibitor factor, and protein S100-A12 [103]. This data suggests the impact of inflammation, signaling, and neurogenesis on depression. Interestingly, brain studies of the depression proteome, in turn, primarily highlighted signaling, ion homeostasis, or metabolic pathways rather than inflammation [104, 105].

1.3 Advances in high-throughput screening studies

The analysis of large scale -OMICs data involves sophisticated methods that were developing over decades of research within molecular biology, chemistry, statistics, computer science, and machine learning. The development of these methods allows exploring physiological conditions and disorders from multiple angles, such as genotype, epigenetics, expression, proteome, and combined. This section provides a brief introduction and overview to key theoretical aspects, methods and technologies used within the framework of the current thesis.

1.3.1 Array-based analysis of DNA methylation

DNA methylation is extremely diverse process that can encompass nearly any region within the DNA sequence. In principle, any CpG pair as well as non-CpG cytosines could be methylated, thus rendering DNA methylation profile extremely high in dimensionality. Historically, DNA methylation analyses utilized targeted approaches, predominantly employing Sanger sequencing of bisulfite-converted DNA (bisulfite sequencing). Bisulfite conversion involves treating the DNA sequence with sodium bisulfite, leading to the transformation of unmethylated cytosines into uracils. Subsequently, these uracils are amplified into thymines during PCR amplifications, resulting in T-A nucleotide pairs replacing the original G-C pairs in the double-stranded DNA. When comparing the amplified DNA with the original sequence (sequence alignment), the observed differences manifest as mismatches, enabling the inference of methylation at specific CpG sites [106]. Although Sanger sequencing is relatively accurate, it allows monitoring of only small amplified fragments at a time (800-1000 bp) [107], thus rendering this method extremely inefficient for large-scale analyses, where length of even a single gene may exceed 100000

bp [108].

With the development of next generation sequencing (NGS), performing large scale DNA analyses became possible as these methods allow analyzing entire genomes with a fraction of previous costs [109, 110]. The development of NGS also dramatically affected DNA methylation analyses, and early large-scale DNA methylation platforms (arrays), such as, Illumina Infinium HumanMethylation27 [111], featuring simultaneous analysis of 27578 CpGs, became fully available for research labs in late 2000s, with the first publications featuring these arrays appearing soon after [112]. Then, larger arrays such as Illumina Infinium HumanMethylation450 (covering more than 450000 CpGs) [113] and Illumina Infinium MethylationEPIC (covering more than 850000 CpGs) [114] became available later. Probes on these arrays target CpG sites that have multiple properties and relation to DNA/gene sequences, including promoter regions, 1-st exons of genes, gene body, 3'- and 5'-untranslated regions and other [115]. The difference between HumanMethylation450 and MethylationEPIC is that the latter array contains and additional 350000 CpGs located in potential enhancer regions as well as $\sim$90% of probes from HumanMethylation450 [116]. Illumina methylation arrays, especially MethylationEPIC after retiring of HumanMethylation450, are the mainstays of EWAS today.

The key steps of the DNA methylation array-based analysis could be outlined as follows [111]:

DNA methylation analysis with Illumina arrays

1. The DNA preparation at the beginning remained the same as in bisulfite sequencing. Genomic DNA is treated with sodium bisulfite, resulting in the conversion of unmethylated cytosines into uracils.
2. Then, bisulfite-converted DNA is whole-genome amplified (WGA) and enzymatically digested into smaller fragments
3. Subsequently, WGA-DNA is hybridized on the designed array containing locus-specific DNA oligomers linked to individual bead types. These bead type correspond to methylated and unmethylated states of investigated CpGs.
4. Allele-specific primer annealing is followed by single-base extension with fluorescent labelled ddNTPs.
5. The array is fluorescently stained and scanned after extension, thus the intensities of the unmethylated and methylated signals are measured. Obtained intensities, in turn, are used to determine DNA methylation values, defined as "β-values".
6. β-values are analyzed using statistical methods.

The obtained protocol results in the generation of methylated and unmethylated signal intensities that than are converted to β-values to express methylation quantity used to perform differential methylation analyses. In practice, before being suitable for analysis, methylation values should be background corrected, probe-bias adjusted [117], preprocessed, normalized, and filtered. The batch effect adjustment is also typically performed. The resulting β-values are ratios of the methylated probe intensity and the overall intensity as could be illustrated in the following equation:

$$\beta_i = \frac{max(y_{i,methyl}, 0)}{max(y_{i,unmethyl}, 0) + max(y_{i,methyl}, 0) + \alpha}$$

Here, $y_{i,methyl}$ depicts methylated signal at probe i, whereas $y_{i,unmethyl}$ corresponds to its unmethylated signal. The constant α (default α = 100) is an Illumina-suggested parameter to avoid potential division by 0 and regularize methylated and unmethylated intensities [118]. It should be mentioned, however, that β-values, despite being easy to interpret, follow unfavorable statistical properties, such as heteroscedasticity and non-normal distribution (β-distribution), and thus are suggested to be substituted by M-values in statistical models [118]. The M-values, in turn, are defined as $\log_2$ ratios of methylated probe intensities versus a corresponding unmethylated intensity:

$$M_i = log_2 \left(\frac{max(y_{i,methyl}, 0) + \tau}{max(y_{i,unmethyl}, 0) + \tau} \right)$$

In this equation, $y_{i,methyl}$ depicts methylated signal intensity at probe i, whereas $y_{i,unmethyl}$ corresponds to its unmethylated signal intensity. The parameter τ is also a constant offset (default τ = 1) to improve stability of M-values for small signal intensities [118]. The obtained M-values or β-values are frequently additionally adjusted for cell type heterogeneity [119] and then are used in statistical models to determine differentially methylated probes with respect to investigated phenotype.

1.3.2 High-throughput transcriptomics and proteomics

In the exploration of complex disorders, such as depression, high-throughput transcriptomics and proteomics have emerged as one of the pivotal methodologies. Similarly to DNA methylation, both of these methods have been increasingly developing over last decades and were used in the following thesis. First, let's discuss the transcriptomic approach.

Initially, transcriptomic studies were preformed leveraging hybridization-based methods, such as Northern blotting, which allowed for the detection of indi-

vidual RNA species within a sample [120]. This approach was potentially dangerous since involved working with radioactive materials and was limited in its throughput capacity. In 1990s, the microarray-based transcriptomics emerged [121], enabling a rapid paradigm shift in such studies. This technique, involving the hybridization of cDNA or cRNA to thousands of probe sequences arrayed on a chip, enabled simultaneous quantitative measurement of a wide spectrum of transcripts within a sample [121]. Several years after, many commercial arrays, such as GeneChip Human Genome U133 Plus 2.0 Array [122] or HumanHT-12 BeadChip [123] and other became the main methods for analyzing transcriptome data. Today a large compendium of transcriptome data has been accumulated in part due to large scalability of microarray technology, standardized protocols for analysis, as well as data sharing efforts. As of December 2023, almost 60000 data collections are available in the gene expression omnibus, a database for sharing -OMICs studies [124], featuring "expression profiling by array". GEO is an excellent resource for conducting multi-cohort analyses, replications, developing statistical models as well as gaining additional functional knowledge about a particular research question.

Transcriptome studies advanced even further with the introduction of NGS technologies. RNA-seq, a high-throughput sequencing method, allows for the direct sequencing and quantification of amplified transcripts (in a form of DNA) and thereby offers a comprehensive view of the transcriptome, including both coding and non-coding RNA sequences [125]. Though, beyond the scope of experiments in this thesis, it is worth mentioning that RNA-seq overcomes many array-based limitations, such as the need to know target sequences before the analysis (for some of the methods), as well as hybridizations beyond target sequences [126], generating spurious correlations. Furthermore, modern methods, even allow bypassing DNA synthesis stage via direct RNA sequencing [127].

The developments around proteomic studies generally followed a similar trajectory. Starting with a 2D gel-electrophoresis [128], this field rapidly transformed with additional milestone inventions, such as mass-spectrometry (MS) [129] and liquid chromatography-tandem mass spectrometry (LC-MS/MS) [130], enabling very precise analyses in complex samples. However, the use of all the MS methods is an extremely costly endeavor that is not affordable for most individual research groups at large scale. Thus, the area of targeted proteomics, where a research group is interested in quantifying only specific subset of proteins, was also developing in parallel. The targeted proteomic analyses today are based on either MS-based methods, such as parallel reaction monitoring [131] or via antibody-based or aptamer-based high-throughput protein quantification platforms.

Antibody-based protein detection and quantification is not new, as it has been in existence since the introduction of the enzyme-linked immunosorbent assay (ELISA) [132] and Western blotting [133] in the 70s. Despite being very accurate at protein quantification, all of these methods fall short regarding throughput capacity. Additionally, large samples amounts are required for all of the aforementioned approaches. Thus, high-throughput proteomic methods, such as aptamer-based SomaScan [134] or antibody-based Olink platform [135], emerged to address these issues. Olink-based protein quantification relies on proximity extension assay (PEA) that utilizes antibodies with covalently-linked oligonucleotide sequences containing a hybridization site. Upon interaction with a target, oligonucleotide sequences become in proximity enabling polymerase-based extension of the overlapping site, which results in a DNA template that can be detected and quantified by qPCR [135]. Obtained values from qPCR are then preprocessed, normalized based on control probes, generating NPX-expression values on log2 scale [136].

In essence, the array-based transcriptomics and multiplex Olink assays generally follow similar analysis protocols/steps without method specifics:

Table 1. *Comparison of array-based transcriptome and Olink-based proteomics workflows*

Array-based transcriptome	Olink assay
1. Sample collection.	1. Sample collection.
2. Isolation of target biological fluid. Isolation of RNA and subsequent library construction is required.	2. Isolation of target biological fluid.
3. Samples are applied on the transcriptome array. RNA or DNA fragments hybridize to array probes.	3. Samples are positionally randomized and placed on microfluidic plates. Samples are subject to PEA to generate DNA templates. Antibodies bind target proteins enabling extension reaction.
4. Transcript quantification with fluorescence intensity during array scanning.	4. Protein quantification via qPCR reaction on PEA product, involving fluorescence-based detection.
5. Data quality control. Data preparation with background correction, normalization and probe correction with methods such as robust multi-array average [137] or other.	5. Data quality control. Data correction via extension control, inter-plate control and adjustment factor. Data normalization based on median assay and plate intensities or based on reference samples.
6. Obtained values are frequently on log2 scale and ready for differential expression analysis.	6. Obtained NPX values are on log2 scale and ready for differential proteomic analysis.

The illustrated framework shows exclusively key steps, and in some cases, depending on the research question, the modifications of these steps are required. In practice, some of the analysis stages, such as PEA and qPCR in Olink, could be outsourced to specialized labs/centers that will only deliver data after all *wet lab* procedures.

1.3.3 Limitations of high-throughput biomarker discovery

Even though the advances in high-throughput methods constantly improve biomarker discovery, this area as well as the methods themselves still have limitations. First, especially in depression, the studies of biomarkers are generally biased by existing knowledge (our assumption that certain disease mechanisms must be related to a particular cell type/tissue/protein), thus limiting exploration of other less pursued aspects. For instance, one might be focused only on CNS-related pathways studying depression, ignoring other systems, such as the immune system or *vice-versa*. Second, all array-based/plate-based analyses in the high-throughput pipelines are constrained by existing platform limitations, such as the set of transcript probes on the array, methylation probes on the array, as well as assays included in Olink. Since all these methods do not cover the entire existing space of biomarkers, they produce results that are biased to certain marker instances (such as a particular set of proteins) that otherwise might be less interesting compared to other marker candidates. Third, high-throughput methods are subject to batch effects [138, 139], thus limiting comparability of samples between studies and even different batches of the same study. Lastly, nearly all high-throughput methods discussed in this chapter, including array-based DNA methylation, transcriptome, and Olink assays produce only relative measurements instead of the absolute-ones. This implies that relationships obtained from different assays may not correspond to factual biological differences. For example, if two proteins in Olink assay have the same NPX value, it does not mean that real biological concentrations of underlying proteins are identical [136].

1.4 Data analysis, modeling, machine learning and deep learning in -OMICs

In recent years, the field of biomarker discovery has witnessed remarkable advancements in data analysis, modeling, machine learning, and deep learning, significantly impacting research on complex questions such as depression pathology. These advancements have enabled deciphering of vast datasets with more accurate predictions and insights into the underlying mechanisms of diseases from different perspectives. This section provides a brief overview around common data analysis approaches that are used in biomarker studies and were applied within the present thesis.

1.4.1 Association modeling in the -OMICs field

The most common approach in any kind of observational study is to see and investigate a relationship between certain variables. Once observed, this *lead* could be further explored in the confirmatory analyses, experiments, or be used as a prediction tool. Many -OMICs studies, such as GWAS, EWAS, transcriptome-wide association study (TWAS), proteome-wide association study (PWAS) serve this purpose of exploration of potential disease-related factors.

The analysis of relationships between variables lies within the realm of statistics. Depending on a research question, data type, samples size, etc., many statistical methods could be leveraged for data analysis. The most simple approaches include correlations, such as Pearson (parametric) or Spearman correlations (non-parametric) that could be used to quickly see relationships between two continuous variables [140, 141]. Another relatively simple group of methods/approaches includes two-sample hypothesis testing, for example paired and unpaired t-tests (parametric), Mann-Whitney U test (non-parametric), Wilcoxon Signed Rank Test (non-parametric), Pearson's χ^2 test (non-parametric), as well as multi-group methods, such as ANOVA (parametric) and Kruskal Wallis Test (non-parametric) [140, 141]. Though being relatively straightforward and commonly used in life science research, these methods are not very informative for many practical questions. This not only due to well-known shortcomings in the Hypothesis testing paradigm [142, 143, 144, 145] but also due to intrinsic limitations of these approaches. Overall, none of the methods above, except for correlations, provide a reliable information regarding the effect size of associations without supplementary analyses. Moreover, none of these methods could take into the account information regarding confounding factors that are of paramount importance, especially in the psychiatric field. Thus, these approaches are generally kept only for simple experimental designs and secondary analyses.

In practice, many -OMICs studies rely on multivariate linear model-based approaches to investigate relationships between the outcome and a corresponding -OMIC factor (DNA methylation, gene expression, etc.). Multivariate models allow incorporating information regarding the confounding factors in contrast to methods specified previously. The exact model specification and assumptions could vary. For example, multiple linear regression models could be summarized as follows:

$$y_i \sim \beta_{0,i} + \beta_{1,i}X_1 + \beta_{2,i}X_2 + ... + \beta_{N,i}X_N + \varepsilon$$

or

$$y_i \sim \sum_{j=1}^{N} \beta_{i,j}X_j + \varepsilon$$

In these equations, y_i denotes an -OMIC factor (such as methylation at i-th CpG), whereas $\beta_{0,i}$ indicates model intercept, $\beta_1 ... \beta_N$ indicate coefficients of a linear model, terms $X_1 ... X_N$ correspond to model covariates (confounding factors), and N is the number of factors. Lastly, the term ε shows a random error. The biological importance of a certain variable on the outcome is interpreted via size and sign of the corresponding β-coefficient. The test for statistical significance of the underlying coefficient could be achieved via Hypothesis testing, namely estimating a t-statistic by dividing a coefficient by its standard error $\beta_{i,j}/SE(\beta_{i,j})$ and projecting it to a T-distribution with specified degrees of freedom to cut the area in the tails of this distribution. The corresponding area in the tails of the distribution is estimated, and if the area is below a pre-defined threshold (typically 0.05), the result treated as statistically significant.

Classical linear regression models assume a linear relationship between the level of an -OMIC factor and corresponding predictors. These models require many assumptions to be met (at least partially), including independence of observations (the most important), normal distribution of target variable (-OMIC factor), normal distribution of residuals, homoscedasticity, absence of multicollinearity, and other. In many cases, these assumptions may not be met or the investigated research question is formulated differently (such as binary outcome), thus other methods, such generalized liner models (GLM), including binary logistic regression, could be used instead. These methods could be also combined with more sophisticated approaches to infer significance of coefficients. For instance, the commonly used R-package *limma* uses GLM to study associations. However, the obtained T-statistics are additionally moderated using the empirical Bayes method [146]. Though, it should be mentioned that regardless of the statistical method of choice, all cross-sectional studies with the design $y_i \sim \sum_{j=1}^{N} \beta_{i,j}X_j + \varepsilon$ only allow studying associations, whereas inference about causation could not be made.

1.4.2 Machine learning in the depression -OMICs field

The developments of machine learning (ML) approaches and improvements in hardware have led to ubiquitous applications of ML in various domains. This developments also affected biomarker discovery and the -OMIC field. Many ML models are designed to work with multi-dimensional data, thus making them a good choice for bioinformatic applications. In these study designs, an investigated phenotype or outcome is frequently formulated as a product of a pre-defined list of environmental and/or biological factors, and the objective is to predict/classify/categorize the outcome of interest based on these factors [147]. ML models were used for various applications, including diagnostics, chromatin modeling, phenotypic stratification, protein structure prediction, etc [147].

ML models could serve different purposes; two main ones include regression, where the objective is to predict a continuous outcome, and classification, which implies categorization of the outcome. The number and types of models suggested is extensive and continuously developes. Some of the regression model examples include multiple linear regression, polynomial regression, decision tree regression, random forest regression, support vector regression, Lasso-regression, Ridge-regression, and artificial neural networks, etc. The classification models, in turn, could include variations of binary logistic regression (though formally it is a regression algorithm), support vector classifiers, decision trees, K-nearest neighbors, ensemble boosting methods (such as AdaBoost), ensemble bagging methods (such as Random forest), and artificial neural networks, etc. All of the mentioned methods have different underlying machinery and assumptions but could be used interchangeably in many cases. The availability of computational tools, such as Python package *scikit-learn* [148], simplifies the use of ML models even for non-specialists.

In the depression field, the use of ML models is relatively in infancy, and not so many studies were performed compared to other domains. In many existing studies, sociodemographic factors were used as input for ML models to detect depression [149, 150, 151, 152, 153, 154]. Other data inputs were also suggested, including data from social networks [155], metabolome [156], and functional MRI [157]. In a more related space, some studies also investigated the application of blood DNA methylation as a predictive tool for depression [84, 89, 90] with moderate performances. These studies, however, either used a limited number of potentially suitable ML methods or performed evaluations in small samples. The largest study, explored the predictive power of blood DNA methylation in the cohort from Scotland (1223 MDD cases and 1824 controls for training and 553 MDD cases and 1417 controls for testing) exclusively using Lasso regression models (L1 penalized regression models)

[89]. Interestingly, there was no study investigating stability of DNA methylation features and predictions in the settings of multiple cohorts.

1.4.3 Brief intro to deep learning

Deep learning (DL) is a specific branch of ML that deserves a separate mention as it is relatively different in principle compared to standard ML models. All previous models were relying on a predefined set of features that are utilized by a model to make a prediction regarding the outcome. Essentially, it means that exact operations on features being used are somewhat known and data representation (or its form) is also known. DL models, on the other hand, offer an additional layer of abstraction as these models are not only capable of optimizing model parameters to accomplish prediction tasks but also capable of learning new abstract representations (hidden features/embeddings) of the initial data to improve its performance. The calculation of these abstract representations is achieved via stacking individual layers of non-linear transformations sequentially that learn features of the data required for a particular task. To effectively extract the features, DL models must be trained (as any other ML model) to adjust the weights (parameters) within the layers of non-linear transformations [158].

Multiple types of neural networks and architectures were proposed, including convolutional neural networks [159], recurrent neural networks [160], graph neural networks [161], and transformers [162]. The simplest representation of a non-linear transformation is a fully-connected layer of a neural network. Each fully-connected layer is composed of units called neurons that could be represented in following equation:

$$y = f(\sum_{i=1}^{n} \omega_i x_i + b)$$

Here, y shows the output of a neuron. The term f represents a non-linear activation function, such as ReLU or other that is applied on the dot product of neuron weights ω that individually correspond to neuron inputs x. In addition, a bias term b, a constant value is added to the product (also called weighted sum) to further adjust the input to the activation function. The input of a neuron is either features from data (if the layer is the first in the model) or outputs from previous layers [163]. The output of the neuron is the result of non-linear transformation introduced by the activation function. The output of a layer is a vector of outputs from all neurons in this layer. And the whole model, in turn, is a composition of the layers.

During the training process, the optimizer (a learning algorithm) tries to optimize parameters within layers. This process is achieved by finding such a set

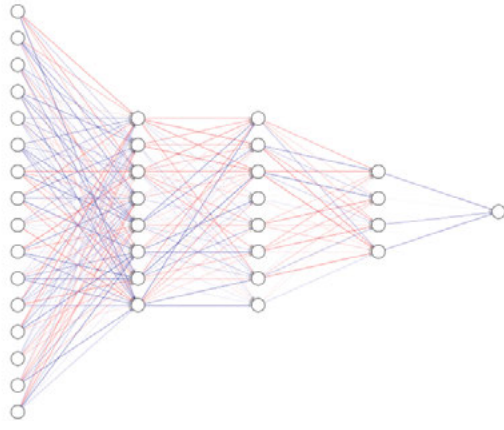

Figure 3. An example of DL model with five fully-connected layers, containing 16, 8, 8, 4, and 1 nodes, respectively

of parameters that minimize the main loss function (or objective function) for the whole model applied on a batch of data. The loss function is a function that mathematically evaluates the performance of the model on the required task, and multiple functions could be utilized for different applications. For example, such loss function could be represented by a mean squared error (MSE) for regression tasks:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Where: N is the number of samples in the dataset, y_i is the real value of the i-th sample, and $\hat{y}_i$ is the predicted value for the i-th sample.

For classification purposes, in turn, a binary cross-entropy (BCE) loss could be used for binary classification:

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$$

Where: N is the number of samples in the dataset, y_i is the real class of the i-th sample (either 0 or 1), and $\hat{y}_i$ is the predicted probability for the i-th sample, indicating the likelihood of belonging to the class labeled as 1.

To find the optimal parameter values, the algorithm needs to compute the gradient (gradient vector) of a loss function with respect to each of the trainable weights (or weight vector) in the model that indicates by what amount the loss

function would change if the weight were increased by a tiny amount [158]. Subsequently, the weight vector is adjusted in the opposite direction to the gradient vector. This process is generally scalable and could be applied to multi-layer architectures via backpropagation that utilizes the chain-rule property of derivatives to compute gradients in the direction from the output to the input [158, 163]. Thus, all layers in the model that are to be optimized should enable the *gradient flow*, or, in other words, they should be differentiable.

1.4.4 Deep learning and OMICs

The applications of DL models in the realm of -OMICs is a relatively new venue. As highly-dimensional biological data appeared quite recently, less models and architectures were applied in this domain compared to other applications, such as image classification. Depending on the model architecture, DL models in OMICs could serve various purposes, with basic ones including classification and regression. Specifically for this thesis, the author is primarily focused on DNA methylation use cases.

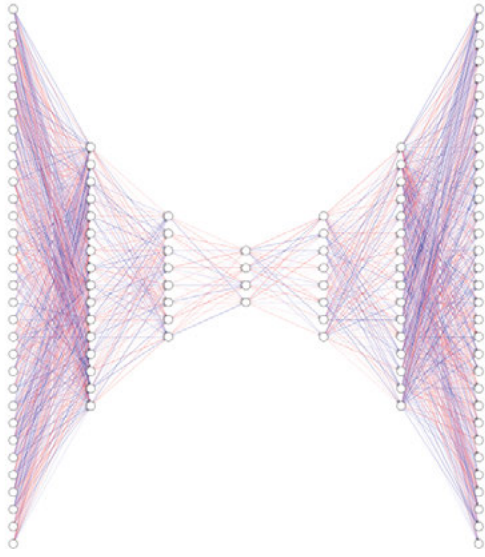

Figure 4. An example of fully-connected autoencoder model with seven fully-connected layers, shrinking input data from 32-dim space to 4-dim space and reconstructing it back.

Multiple DL models and architectures were suggested in various biological domains involving DNA methylation data. For example, flat fully-connected models were proposed for detection of gastric cancer [164], cancer cell type of origin prediction [165], survival analysis prediction [166], predictions in

the malformations of cortical development [167], metabolomics [168], heart failure [169] and other. Another architecture type that was used previously and is related to the present work is called autoencoder [170]. This is a particular type of neural network that is designed to compress the data into its low-dimensional representation (latent space) and reconstruct it back.

Essentially, this approach allows reducing the dimensionality of the data, capturing information on interaction between its features. The autoencoder approach allows capturing subtle data aspects as it is non-linear (properties of neural networks) compared to other commonly-used methods, such as principal component analysis (PCA). In practice, data reconstruction may not be interesting in many applications and only encoded representation is the main focus of the model. In certain situations, the autoencoder model could be augmented by a classification component so a model not only learns hidden data representation but does it in such a way so it optimizes its application for classification tasks. The training of such models could be either simultaneously (one-model approach) or sequentially (two-model frameworks). The autoencoder-based approach was successfully utilized for various applications, including multi-omics integration, expression prediction, cancer prognosis, Alzheimer's disease prognosis, and other [171, 172, 173, 174, 175, 176, 177, 178, 179].

2. Aims

The aims within the following thesis were to investigate, genetic, epigenetic, and expression markers for depression and suicide. We either analyzed their effects independently or juxtaposed markers from different -OMIC analyses or view an entire -OMIC layer (such as DNA methylation) as a composite marker. The aims for included studies are illustrated in the attached figure.

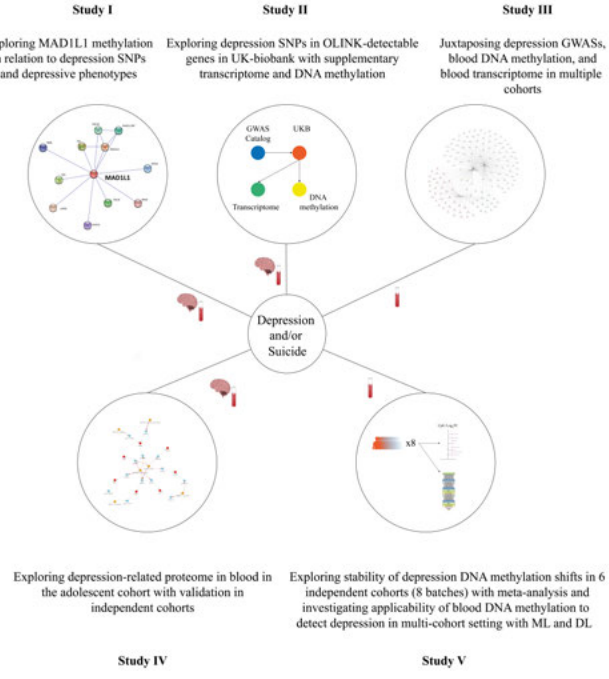

Figure 5. The main objectives of included studies.

In **Study I**, the main focus was depression-related gene *MAD1L1* that was previously reported to be associated with depression SNPs and frequently mentioned as a stress-related marker. We explored how depression SNPs in this gene are related to its methylation, and whether dependent methylation at *MAD1L1* is related to depression/suicide phenotypes. In **Study II**, we primarily focused on the genes that could be utilized as depression markers presented in the proteomic panels, such as Olink. We investigated whether GWAS catalog depression-related SNPs located in Olink genes could be replicated in UK

biobank cohort and whether these associations are supported by DNA methylation and transcriptome in the blood/brain. The main idea behind **Study III** was to compare results from different depression -OMIC layers, including genetics, DNA methylation, and transcriptome. We juxtaposed findings from depression-related GWAS in the GWAS Catalog database with differential methylation analysis in 3 cohorts and with differential expression analysis in 3 cohorts. We explored how the identified genes overlap, whether the identified changes could be related at a functional level, such as eQTL,meQTL, and eQTM, and whether these genes are related to patterns in drugs and clinical trials. **Study IV** was focused on the proteomic side of depression. We used Neuro Exploratory panel from Olink to identify related changes in blood. We took potential biases, such as antidepressant intake into the account and validated several associations in the independent transcriptome datasets. Lastly, the idea behind **Study V** was to use the entire blood DNA methylation as a depression marker. We explored the stability of DNA-methylation features in 8 independent datasets with meta-analysis. Moreover, we performed the analysis and comparison of common machine learning and deep learning strategies for the depression detection purposes, using blood DNA methylation data in a multi-cohort setting.

3. Materials and methods[1]

3.1 Ethics declaration

The studies included in this thesis used two domestic and several public-access cohorts. The domestic cohorts included the Psychiatric health in adolescents study (PSY cohort) and suicide cohort at Karolinska Institutet (SKI cohort). The PSY study was approved by the Regional Ethics Committee of Uppsala (DNR: 2011/446), and all participants gave their written informed consent regarding participation in the study. The SKI study was approved by the Ethics Review Board in Stockholm (DNRs: 00-194,2015/1454-32). All participants gave their written informed consent.

In all of the studies included in the thesis, publicly available cohorts deposited at GEO were used. Data from these studies comes from already published research that was approved by corresponding regional and national committees. All relevant information regarding initial ethical approvals of these studies is available in the corresponding source publications. The use of open-access cohort data was approved by Etikprövningsmyndigheten in Uppsala (DNR: 2023-03977-01).

3.2 Project workflows

3.2.1 Study I

The main objective of Study I was to explore the effects of *MAD1L1* previously-reported depression SNPs and their effects. We investigated whether these effects were mediated through DNA methylation, and if these methylation changes were related with depression phenotypes and gene expression. The first step was to collect depression-related SNPs from *MAD1L1* with GWAS Catalog and DisGeNet. Then, we performed SNP-CpG association analysis both in screening and recall samples of the PSY to identify CpGs that are related to several depression SNPs in both samples. The identified CpGs, in turn, were used to model associations with depression phenotypes and *MAD1L1* gene expression in several cohorts. Lastly, for getting an additional understanding regarding the biology of identified CpGs, we projected these findings

[1]Some of the text may partially or fully overlap with the included articles of studies **I-V** or Licentiate thesis "The role of genetic, epigenetic, and proteomic factors for psychiatric disease" presented by Aleksandr V. Sokolov at Uppsala University on Feb 1, 2023.

at genomic coordinates with chromatin state models to see overlap with depression transcription factors and regulatory elements. We also investigated blood-brain correlations of the interrogated CpGs to explain discrepancies in association directions.

3.2.2 Study II

In this study, we investigated potential relevancy of Olink Neuroexploraty panel genes for depression characterization. First, we identified all SNPs reported in the GWAS Catalog database for the Olink Neuroexploraty genes in the European population. Next, the identified SNPs were investigated in the UK Biobank for association with depression. The associations in the UKB cohort were studied at individual SNP levels as well as at genetic score levels. Subsequently, we performed additional functional analyses for genes overlapping with depression-related SNPs. The functional analyses included: DNA methylation analyses in PSY screening, GSE72680, and GSE125105 and transcriptome analyses in GSE53987, GSE98793 and GSE46743.

3.2.3 Study III

In **Study III**, the main focus was to identify depression-related changes that would be supported through evidence in several cohorts and at several -OMIC levels. For the genetic evidence, we collected all SNPs related to depression (and not comorbid disorders) in the GWAS Catalog database. The identified SNPs were analyzed in terms of replicability between GWAS as well as associated gene enrichment. In parallel, we performed differential DNA methylation analyses in three largest depression blood DNA methylation cohorts available for the group at a time (PSY screening, GSE72680, and GSE125105). We identified a set of differentially methylated CpGs nominally significant in several cohorts with matching direction for associations. The associated genes were analyzed for enrichment in biological processes. Subsequently, we also used three cohorts in blood (GSE98793, GSE46743, and GSE64930) to identify transcriptome changes in depression. Similarly, we identified transcripts that were differentially expressed at nominal significance in several cohorts with matching directions for associations. The transcripts were analyzed for gene enrichments. Once individual -OMIC analyses were performed, we overlaid identified changes in all three -OMIC layers to detect overlapping genes and conducted co-localization analyses of identified markers. To further strengthen understanding of overlapping results, we juxtaposed these findings with eQTL, eQTM, meQTL analyzes to see potential correlations between the obtained markers. Lastly, we also mapped identified genes to agents in drug databases and clinical trials on depression to see trends in clinical development.

3.2.4 Study IV

The workflow in **Study IV** was centered around proteomics in the PSY cohort. Initially, two big runs of Olink Neuroexploraty panels were conducted in 2021 and 2022. We analyzed differential proteomics in whole blood in PSY, using a combined non-overlapping sample from screening and recall. To validate the findings obtained in the PSY, we utilized transcriptome cohorts in the brain and blood. These cohorts included GSE53987, GSE98793, GSE4674, and GSE64930.

3.2.5 Study V

In this study we considered DNA methylation as an aggregated measure rather than focusing on individual CpGs. The main objective of the study was to investigate the applicability of blood DNA methylation for depression detection based on the data from multiple batches, using ML and DL approaches. First, we collected and performed standardized preprocessing of the data from 8 cohorts, including PSY screening and recall, GSE125105, GSE74414, GSE72680, GSE113725, GSE198904_DHRC, GSE198904_OBS. The obtained data was then merged based on overlapping CpG and either kept as is (cohort level normalization) or further harmonized. Subsequently, harmonized data was used to conduct pooled differential methylation analysis (mega-analysis), whereas cohort-level data was used to conduct differential methylation analysis per cohort. Obtained per-cohort estimates were then meta-analyzed and overlapping significant CpGs from both analyses were obtained. Lastly, both datasets were tested in depression classification scenarios, utilizing ML and DL classifiers. The combined dataset was split into cross-validation dataset that included data from PSY screening, GSE125105, GSE72680, GSE113725, GSE198904_DHRC and last hold-out set containing data from PSY recall, GSE74414_MPIP2, and GSE198904_OBS. The evaluation of classifiers was performed via 10 repetitions of 3-fold cross-validation and confirmed by the last independent test set.

3.3 Participants and cohorts

The work in the following thesis leveraged a large number of domestic and public access datasets/cohorts. In this section, the author outlined the specifics of the cohorts used, available data, and in which studies these cohorts were included.

3.3.1 PSY cohort

The first domestic cohort employed in the analyses was the Psychiatric health in adolescents study (or the PSY cohort). This study initially included 786

14–16 y.o. adolescents (more than 900 as of 2022) that were randomly selected from public schools in Uppsala County, Sweden, starting from 2012. The aim of the study was to investigate relationships between genetic, epigenetic, and proteomic shifts in the blood with relation to risk for psychiatric disorders in adolescents. The study and analyses were conducted in several waves. The waves were based on multiple consecutive visits.

1. **Visit 1**. During the first (screening) visit participants (n>900) went through standard phenotype evaluations. They reported their biological sex and body mass index (BMI), answered to several basic phenotype assessment questionnaires, and completed the DAWBA screening. Some participants managed to non-ambiguously report medications that they use. The screening visit also included the collection of blood, plasma, and serum.
2. **Visit 2**. During the second visit (recall) approximately 1 year after screening, participants (n>400) passed through similar phenotype evaluation and were assessed by additional psychiatric questionnaires that included MADRS, SUAS, SCAS, and other assessments. DAWBA scoring was repeated. The recall visit also included the collection of blood, plasma, and serum.
3. **Visits 3 & 4**. The PSY study also included two additional waves without biological material collection that were conducted at different time points. The second recall was primarily focused on social attachment questionnaires, whereas the fourth wave was distance-based and included DAWBA and PsyToolKit evaluations. Data from these waves was not included in the analyses of the presented thesis.

Psychiatric evaluation in PSY

Psychiatric evaluation of the PSY cohort was performed using computer-based DAWBA self-assessment questionnaire. The depression band of this questionnaire yields scores in the range from 1 to 5, where each score corresponds to a probability of children that have depression in the corresponding band. The scores correspond to the following probabilities: 0 (< 0.1%), 1 (~0.5%), 2 (~3%), 3 (~15%), 4 (~50%), and 5 (~70%) [11]. We dichotomized PSY participants based on the score: all participants with risk less than 50% (score less than 4) were categorized as a "low-risk" group, whereas all other participants were classified as a "high-risk" group.

Overview of the -OMIC data in PSY

The PSY cohort contained data from several -OMIC modalities that were used in all projects included in following dissertation. Due to the long-term nature of the study and natural dropouts of participants, the biological materials and corresponding data were not necessarily available for all of the participants at each time point.

- Genotyping of participants was exclusively performed at the screening stage, and this data was available for 786 participants at this time. A fraction of participants joined the study after genotyping, and thus do not contain such information. Genotype data from PSY was only used in the **Study I**.
- DNA methylation data at screening is available for 221 participants and was performed at two time points: batch one — 129 and batch two — 92. DNA methylation measurements were performed using DNA from the whole blood sample and Illumina HumanMethylation450 BeadChip. This data was used in **Study I**, **Study II**, **Study III**, **Study V**.
- DNA methylation data at recall is available for 169 individuals, and was also obtained in two batches. The recall DNA methylation data also comes from the whole blood sample but is based on Illumina Human MethylationEPIC array. Only a fraction of recall participants has DNA methylation data also at screening (78 participants), whereas 91 participants have such data only at the recall visit. DNA methylation data from the recall visit was used in **Study I** and **Study V** (91 participants).
- Proteomic data in the PSY cohort was obtained later than genotyping and methylation and was performed in several batches. The proteomic data in the PSY is currently based exclusively on the Olink Neuro Exploratory panel covering 92 proteins. Plasma samples were used to measure protein quantities. The total proteomic sample with matched DAWBA scores includes 461 non-overlapping participants (as of 2024), of which 138 come from screening and 353 from the recall. There were 56 participants measured at both screening and recall (here only included in the screening to avoid duplication). The PSY proteomic data was used only in **Study IV**.

Genetic data of the PSY

Blood samples have been collected in K2EDTA blood tubes (Greiner Bio-One, Austria). Genomic DNA has been extracted using E.Z.N.A. Blood DNA Kit (Omega Bio-Tek, USA). Then, extracted DNA was used for genotyping. Genotyping of the PSY cohort was performed at the screening stage with Illumina Infinium array (includes 700078 genetic variants) at the SNP&SEQ Technology Platform in Uppsala (www.genotyping.se). The facility is a part of the National Genomics Infrastructure supported by the Swedish Research Council for Infrastructures and Science for Life Laboratory, Sweden. After the initial quality control, data has been imputed with IMPUTE2 software as described in [78]. Analyzed imputed SNPs passed through further quality control steps: containing standard reference identifier "rs", imputation info score ≥ 0.9, and the expected first alle (A1) frequency between 0.1 and 0.9.

Proteomic data of the PSY

The protein levels in the PSY cohort were assessed using Olink Neuro Exploratory Panel that analyses 92 proteins. The proteomic data was obtained in two batches. Samples were randomized on 96-well plates before performing protein quantification with PEA. The PEA protocol is extensively described on the Olink website. Briefly, a mixture of 1 µL of EDTA-containing plasma and a 3 µL incubation mix was incubated overnight at 8 °C. Subsequently, a 96-µL extension mix with PCR reagents and PEA enzyme were added to the reaction mix and incubated at room temperature for 5 minutes. The extension reaction was performed in a thermal cycler and then was followed by 17 cycles of DNA amplification. During these steps, oligonucleotide-labelled antibodies bound to target proteins, enabling enzymatic extension of the attached oligonucleotides and amplification of the obtained DNA templates during PCR. The PCR products were then quantified with qPCR. Data was subsequently normalized and corrected using plate controls as described in the Olink manual [136]. The protein quantities are expressed as Normalized Protein Expression (NPX) on a log2 scale.

3.3.2 SKI cohort

The SKI cohort incorporates participants who were recruited to the Suicide Prevention Clinic at Karolinska University Hospital for evaluating their psychiatric state, specifically focusing on tendencies towards suicidal behavior, and for ongoing clinical supervision. The study sample excluded subjects diagnosed with intellectual disabilities, schizophrenia, intravenous substance misuse, and dementia. The main SKI cohort sample comprises 88 individuals who had experienced at least one suicide attempt before their consultation at the clinic during the period 2000 to 2005 and had DNA samples available for the analysis [79].

Psychiatric and methylation data of the SKI

In this study, participants were categorized into two distinct risk categories: severe and non-severe suicide attempters. The criteria for classifying a suicide attempt as severe included the utilization of a violent method for suicide, a high Freeman scale score, or the occurrence of suicide post-enrollment in the study, up until January 2011. The determination of the violence level in suicide attempts was assessed based on the method's aggressiveness and the degree of violence, adhering to previously established criteria [180, 181]. Concisely, methods such as consuming drugs and a solitary wrist cuts were categorized as nonviolent, in contrast to all other methods, which were deemed violent. The Freeman scale, comprising two subcomponents—reversibility and probability of interruption—serves to gauge the severity of the suicide attempt. The reversibility aspect is contingent on the suicide method and the potential harm

it could inflict. For example, ingesting a minimal quantity of low-toxicity pills is viewed as a reversible method, unlike self-inflicted shooting, which has a high likelihood of resulting in death. Conversely, the interruption probability facet assesses the likelihood of the suicide method being disrupted by external intervention, conditional on the circumstances. Each subcomponent is rated on a scale from 1 to 5, with the aggregate score ranging between 2 and 10 [182]. A threshold score greater than 6 was utilized to classify a suicide attempt as serious (severe). All participants were cross-referenced with the national Cause of Death register. Within the group of 88 participants, four individuals who ultimately committed suicide were categorized as severe suicide attempt group. Methylation profiling of the SKI cohort was performed utilizing DNA from the whole blood sample at two time points. The Illumina Human MethylationEPIC array was used to obtain methylation data for all 88 participants. The data from the SKI study was used in **Study I**.

3.3.3 UK Biobank cohort

UK Biobank resource is a substantial population-based cohort within the United Kingdom. This cohort encompasses detailed health-related information from over half a million individuals, roughly evenly distributed between male and female participants, who were between the ages of 40 and 69 at the time of their enrollment. The recruitment phase spanned from 2006 to 2010 across several centers within the UK. Initial assessments during their first visit included a range of clinical evaluations, biological specimen collection, and mental health assessments through both structured questionnaires and oral interviews. These assessments were aimed at compiling comprehensive data encompassing socio-demographic attributes, lifestyle factors, and clinical histories of the participants [183].

Phenotypic profiling of UK Biobank cohort

The phenotypic profiling of the UK Biobank cohort is documented in detail in the Supplementary Information for **Study II** manuscript. In summary, multiple sources of phenotypic data were employed to delineate participant characteristics. This approach led to the identification of five distinct depression-related traits within the dataset: "help-seeking behavior", "self-reported depression", "usage of antidepressants", "depression as per Smith's criteria" [184], and "hospital records based on ICD-10 diagnosis". Participants who aligned with at least two out of these five depressive indicators were categorized as cases of depression. Employing a minimum of two indicators to ascertain depression cases is considered an effective proxy for the condition in the absence of clinical interview data [185]. An exception was made for depression cases identified through ICD-10 coding in hospital records. Owing to the relatively high reliability of hospital-recorded ICD-10 diagnoses in indicating the disorder, individuals identified through this sole criterion were also classified as

"depression cases". Subjects who failed to satisfy any of the five delineated criteria for depression were categorized as control individuals. Moreover, respondents who answered negatively to the Mental Health Questionnaire item, "Have you been diagnosed with one or more of the following mental health problems by a professional, even if you don't have it currently?" in the context of in the Mental Health Questionnaire, were also grouped as control participants. Comprehensive details of the criteria and the specific UK Biobank codes employed for the demarcation of case and control groups are documented in the Supplementary Information of **Study II**.

In **Study II**, we also categorized participants according to coexistent comorbidities pertinent to depression. Cases and controls for three comorbid traits — namely anxiety, bipolar disorder, and schizoaffective disorders — were ascertained using a combination of self-reported data and ICD-10 diagnostic codes. A participant was classified as a case for these comorbid traits if they received either a primary or secondary diagnosis with the ICD-10 codes related to anxiety (F40-F41), bipolar disorder (F30-F31), or schizoaffective disorder (F25) based on hospital admission records. Alternatively, cases were also identified if participants reported non-cancer illnesses coded as anxiety (code:1287), bipolar disorder (code:1291), or schizophrenia (code:1289) under the UK Biobank field ID (FID): 20002. Participants who did not fulfill the criteria for anxiety, bipolar disorder, or schizophrenia were designated as control subjects for these comorbid traits.

The available phenotypic data encompassed 502717 individuals. In the context of both depression and comorbid traits, we omitted participants based on the following: those who had retracted their consent (n=174), individuals with genetic relatedness as per UK Biobank Field ID (FID): 22021 (n=150338), and participants of non-European descent, identified through UK Biobank FID: 21000 (n=59928). Additional criteria were implemented to further minimize overlap between the depression group and those with comorbid traits. Specifically, for the depression subset, any participant with a diagnosis of anxiety, bipolar disorder, or schizoaffective disorders was excluded. Conversely, for the group with comorbid traits, individuals identified as having depression were excluded.

3.3.4 DNA methylation open-access cohorts

Cohort GSE41826

Samples of the prefrontal cortex (specific Brodmann area not indicated) were acquired from the NICHD Brain Bank of Developmental Disorders. Cell nuclei were isolated and sorted into neuronal and non-neuronal categories via fluorescence-activated cell sorting, guided by NeuN expression, as elaborated

in the original study [186]. In total, this cohort comprises 29 post-mortem samples from individuals diagnosed with MDD and an equal number of control samples. This study contains DNA methylation data (HumanMethylation450) both in cases and controls. Further details are available in the **Study I** article. This cohort was only used in **Study I**. Data for this study is available at both ArrayExpress (E-GEOD-41826) and GEO (GSE41826).

Cohort GSE88890
The cohort GSE88890 involves 20 MDD individuals who died by suicide and 20 controls who died due to non-psychiatric causes. This cohort contains DNA methylation data (HumanMethylation450) from two cortical brain regions (Brodmann Area 11 (BA11) (n=40) and Brodmann Area 25 (BA25) (n=35)). Specimens were collected from the Douglas Bell Canada Brain Bank [187]. This cohort was only used in the **Study I**.

Cohort GSE72680
The GSE72680 cohort (GRADY cohort) emerged from the Grady Trauma Project, undertaken in Atlanta, GA, USA. This study involved 422 participants and was aimed at exploring the interplay between genetic factors and environmental influences in stress response. The majority of the participants were of African American descent, originating from urban environments characterized by lower socioeconomic status [83, 188]. Psychiatric assessments for this cohort were conducted using a range of questionnaires, with the BDI being the primary tool for evaluating depression [12]. Furthermore, participants were evaluated regarding the use of medications for various psychiatric conditions, encompassing depression, anxiety, and bipolar disorder, in addition to assessments of substance misuse (including tobacco, alcohol, cannabis, and heroin). Additionally, key confounding variables such as sex, BMI, and age were recorded. DNA methylation was measured from the whole blood samples, using HumanMethylation450 array. For more comprehensive information on aspects like sample collection and processing, refer to the respective publications [83, 188, 189].

Stratification of participants between cases and controls was different depending on the study. In **Study I**, we categorized individuals into groups of depressed and non-depressed based on their reported treatment status. Those receiving treatment for depression were classified under the depressed category, while the remainder were designated as control subjects. The original cohort comprised 422 participants. Subsequent to excluding individuals with incomplete data regarding gender, age, treatment for depression, BMI, and ethnicity, the final total of 377 participants were analyzed in **Study I**.

In **Study II**, an alternative stratification method was applied to the GSE72680 cohort. Depressed individuals were identified as those having a BDI score

of 21 or higher. This methodology provided a less stringent classification of participants compared to **Study I**, facilitating adjustments for depression treatment. It also encompassed individuals with depression who were not undergoing treatment.

In the studies **III & V**, we used a composite scoring procedure for depression stratification that takes both BDI and treatment information into the account. Initially, the BDI score was evaluated. A stringent threshold of 21 on the BDI was utilized to categorize participants as depressed [190], considering that scores ranging from 17 to 20 are often viewed as borderline [191]. Additionally, the status of undergoing treatment for depression was factored into the analysis. Accordingly, the classification process entailed multiple stages:

1. Individuals with a BDI score of 21 or higher were classified as depressed.
2. Participants receiving treatment for depression were also classified as depressed.
3. Those with a BDI score below 21 and lacking data on depression treatment were omitted from the study.
4. Participants not receiving depression treatment and who had absent BDI data were likewise excluded.
5. All remaining participants were categorized as non-depressed.

The refined dataset comprised 212 cases of depression and 179 control subjects.

Cohort GSE125105

The GSE125105 (MPIP cohort) cohort was assembled at the Max Planck Institute of Psychiatry in Munich, Germany. The initial composition of this dataset included 489 individuals diagnosed with depression and 210 control subjects. Standard data encompassed details on sex and age. Additional information, such as BMI and SNP-based principal components for ethnicity consideration, was obtained upon request from the Max Planck Institute of Psychiatry. BMI measurements were recorded at baseline and one week following the initiation of treatment, with the average BMI being calculated from these two values. In cases where one of the BMI values was unavailable, the available measurement was utilized as the BMI. There was an absence of phenotypic data for several vital covariates for certain participants, leading to the inclusion of only 324 patients with depression and 167 control subjects in the final analysis. The assessment of depression status was carried out in clinical settings, adhering to the criteria of the DSM-IV. Comprehensive details regarding the cohort, sample preparation, and methylation profiling methodologies are provided in the foundational publications [83, 192, 193, 194]. The methylation data from GSE125105 (HumanMethylation450) was used in studies **II, III & V**.

Cohort GSE74414

The cohort GSE74414 (GSE74414_MPIP2) was also assembled at the Max Planck Institute of Psychiatry. The focus of this study was on examining alterations in DNA methylation in response to glucocorticoid receptor activation by dexamethasone in individuals with MDD and control subjects [7]. Depression levels were assessed using the 21-item HAMD. Individuals experiencing at least a moderate depressive episode, indicated by a HAMD-21 score of 14 or higher, were categorized as suffering from depression. We analyzed only baseline data collected before any stimulation. It was noted by the study's investigator that there was one participant who was also part of the GSE125105_MPIP cohort; this individual was subsequently removed from our analysis. The final count for the sample included 32 individuals classified as depression "cases" and 49 control subjects. DNA methylation data from GSE74414_MPIP2 (HumanMethylation450) was available from GEO (GSE74414) and was used only in **Study V**.

Royal Devon and Exeter cohort

The Royal Devon and Exeter cohort (GSE113725_RDE) was acquired through the Royal Devon and Exeter (RDE) Tissue Bank, which is a component of the NIHR Exeter Clinical Research Facility, located in the UK. The primary objective of the original study was to explore the correlation between a history of depression and inflammation in the context of DNA methylation [84]. A history of depression was determined based on self-reporting, specifically if participants affirmed the question, "Has a doctor ever diagnosed you with depression requiring regular medical treatment?". In **Study V**, we focused exclusively on individuals with a reported history of depression (n=49) and control subjects (n=48). DNA methylation analysis was conducted using the Illumina Infinium HumanMethylation450 array. Data for the cohort was available from GEO (GSE113725).

Molecular Biomarkers of Antidepressant Response Cohort

The cohort GSE198904_DHRC was established through a collaborative effort between McGill University and the Douglas Mental Health University Institute in Canada [195, 196]. Classification of individuals as depressed was based on the criteria of having a current major depressive episode according to the SCID-I, accompanied by a score of 20 or higher based on HAMD-21. Control participants were enlisted via advertisements. Within the initial dataset, numerous samples originated from identical individuals, necessitating their removal prior to the analysis. Consequently, the refined dataset encompassed 32 individuals in the control group and 186 subjects classified as cases. Methylation profiling was conducted on whole blood samples, using the Illumina Human MethylationEPIC platform. Data for this cohort was available at GEO (GSE198904) and was used only in **Study V**.

Observational clinical study cohort NCT02489305
The cohort GSE198904_OBS was derived from the OBSERVEMDD0001 observational clinical study (NCT02489305) conducted by Janssen Research & Development in various locations across the U.S. [81] The study's primary aim was to identify blood biomarkers that could predict the relapse of MDD. Patients in the "post-depression" category were defined as those with a history of nonpsychotic, recurrent MDD (as per DSM-V criteria) within the preceding 24 months, a MADRS total score of 14 or lower, and evidence of a recent (within the past 3 months) positive response to oral antidepressant therapy. Control subjects were identified based on self-reported data indicating no affective disorders. As with the prior cohort, numerous samples came from identical individuals, leading to their removal before analysis. The final dataset included 115 individuals categorized as cases and 29 as controls. Methylation profiling in this cohort was similarly conducted on whole blood using the Illumina Human MethylationEPIC platform. Data for this cohort was also available at GEO (GSE198904) and was used only in **Study V**.

Cohort GSE49065
This cohort comprised transcriptomic and methylation data obtained from cultured peripheral blood mononuclear cells of 10 healthy male subjects. The original investigation focused on the relationship between DNA methylation and aging, particularly in relation to the peroxisome proliferator WY14,643 [197]. In **Study I**, we utilized only the data from cells subjected to the sham control (0.05% DMSO) to eliminate possible confounding factors. Age was the sole additional phenotypic variable considered. For methylation analysis in GSE49065, the Illumina HumanMethylation450 BeadChip was employed, while transcriptomic profiling utilized the Affymetrix Human Gene 1.1 ST Array. Detailed methodology are available in the original study [197]. This cohort was only used in **Study I** to find associations between methylation at MAD1L1 CpGs and its gene expression.

Cohort GSE56047
The initial study focused on exploring the correlations among age, DNA methylation, and gene expression patterns in $CD14^+$ monocytes, encompassing a sample of 1202 individuals [198]. For assessing DNA methylation, the HumanMethylation450 array was utilized, while gene expression was measured using the Illumina HumanHT-12 V4.0 expression array. The dataset incorporated details about the participants' age, the proportions of various cell types within the samples, and the array identifier. Additionally, a composite covariate termed "racegendersite" was included, which encompassed information about the participants' race, gender, and the location of the research site. Data from this cohort was used to find associations between methylation at MAD1L1 CpGs and its gene expression in **Study I** and to conduct cis-eQTM analysis in **Study III**.

3.3.5 Transcriptome open-access cohorts

Cohort GSE53987

The study GSE53987 encompasses data from brain samples relevant to MDD, SCZ, BD, and control groups. Detailed methodologies for sample collection and preparation are documented in the initial study [199]. The dataset includes transcriptional data from three distinct brain regions: the associative striatum, the hippocampus, and the prefrontal cortex (specifically Brodmann area 46). Transcriptomic analysis was conducted, utilizing the Affymetrix Human Genome U133 Plus 2.0 Array. Our analysis focused on contrasting the transcriptomic profiles of MDD patients (numbering between 16 to 17 individuals) with those of control subjects (ranging from 18 to 19 individuals) across each of these three brain regions separately. This cohort was used for validation analyses in the studies **II & IV**.

Cohort GSE98793

The cohort GSE98793 comprising 192 individuals forms a segment of the GlaxoSmithKline–High-Throughput Disease-specific Target Identification Program. This cohort is stratified into three groups: 64 participants diagnosed with MDD, another set of 64 individuals diagnosed with both MDD and anxiety, and a third group of 64 healthy control subjects. The dataset encompasses covariates including age, sex, and specific diagnoses (anxiety and MDD), in addition to a batch covariate. The gene expression was examined in whole blood, using the Affymetrix Human Genome U133 Plus 2.0 Array. For comprehensive details on this cohort, refer to the original publication [200]. Data from this cohort was used in studies **II, III & IV**.

Cohort GSE46743

The cohort GSE46743 created at the Max Planck Institute of Psychiatry in Munich, Germany was integral to a study focusing on the transcriptomic responses to stress within the framework of psychiatric conditions, notably depression. This study involved examining the whole blood transcriptome profiles of 160 male participants both prior to and following exposure to dexamethasone. To account for potential confounders, additional data pertaining to age and BMI were gathered. Transcriptome profiling utilized the Illumina HumanHT-12 expression beadchip. More detailed information about the study can be found in the referenced article [201]. In the current analysis, our focus was on the transcriptome profiles before the dexamethasone exposure, specifically comparing baseline expressions in patients with depression against those in control subjects, using the previously established classifications for these groups. Data from this cohort was used in studies **II, III & IV**.

Cohort GSE64930

The transcriptomic cohort GSE64930 collected at the Max Planck Institute of Psychiatry in Munich, Germany shares similarities in design with the GSE46743

study, focusing on the reaction to dexamethasone in relation to depression. This particular cohort overlaps with GSE46743, including 79 common participants. The GSE64930 cohort comprises a total of 289 individuals, encompassing both female and male participants. Initially, the phenotypic data provided was limited to the participants' sex. Additional details such as age, BMI, RNA integrity number, HAMD scores, and three surrogate variables for cell heterogeneity correction were subsequently obtained from the research team upon request. Transcriptome profiling for these cohorts was performed using the Illumina HumanHT-12 expression beadchip. Detailed cohort information can be found in the original study [202]. We used transcriptome data obtained before dexamethasone stimulation, employing the pre-established classifications for depressed and control individuals. Three participants lacking RNA integrity number data were excluded, resulting in a final count of 286 participants suitable for the analyses. This dataset was included in studies **III & IV**.

3.4 Sample collection and DNA methylation profiling

Blood samples from PSY and SKI cohorts have been collected in K2EDTA blood tubes (Greiner Bio-One, Austria). The process of genomic DNA extraction was carried out utilizing the E.Z.N.A. Blood DNA Kit from Omega Bio-Tek, USA. Subsequently, the extracted DNA was subjected to methylation profiling. The bisulfite conversion of DNA was executed using the EZ DNA Methylation kit provided by Zymo Research, USA, with all steps adhering to the guidelines provided by the manufacturer. The preparation of DNA samples, along with the array processing and scanning, was conducted at the SNP&SEQ Technology Platform located in Uppsala. A specific quantity of 250ng of the bisulfite-converted DNA was utilized for each sample. The resulting DNA methylation data in the form of IDAT files was securely transmitted via the UPPMAX cluster allocated for sensitive data (Bianca).

3.5 DNA methylation data preprocessing

Preprocessing of DNA methylation data was slightly different depending on the study objectives, whether the data was available in a form of raw IDAT files, signal intensities or β-values. All procedures related to DNA methylation data preparation were performed in the R programming language environment. Processing of DNA methylation data across PSY screening, PSY recall, GSE125105, and SKI cohorts commenced with the raw IDAT files. Data pre-processing were executed using a framework based on the *minfi* R package [203], accessible through bioconductor.org, which hosts a collection of R packages for biological research. Background noise correction in signal

intensities from raw files was achieved using the "noob" method [204]. The normalization of β-values was performed through quantile normalization, followed by a correction for bias in type I and type II probes using Beta Mixture Quantile Dilation [117] from the *wateRmelon* R package [205]. During data preprocessing, we excluded probes associated with sex chromosomes, non-CpG sites, and those lacking β-values. Additionally, we omitted probes containing an SNP with a minor allele frequency (MAF) exceeding 5% within the probe sequence, as well as those with an SNP at the CpG site or in the single-base extension area. For datasets utilizing the HumanMethylation450 arrays, cross-reactive probes as identified by Chen *et al.* [206] and Benton *et al.* [207] were removed. When working with MethylationEPIC arrays, we additionally eliminated cross-reactive and SNP-overlapping probes as determined by Pidsley *et al* [208]. To address potential batch effects arising from the use of different arrays, the "ComBat" function within the *sva* R package was employed to adjust for any biases [209, 210]. Furthermore, the *minfi* package's adaptation of the Houseman algorithm was utilized to adjust methylation data for heterogeneity in white peripheral blood cells, including $CD4^+$, $CD8^+$, natural killer cells, B-cells, monocytes, and granulocytes [211]. The data adjustment for cell type heterogeneity was performed with the regression-based approach [119]. This protocol was used for the majority of cohorts/studies presented here, with some deviations for particular cohorts.

In **Study I**, the cohort GSE41826 DNA methylation data was obtained from DNA extracted from both neuronal and non-neuronal nuclei. The dataset acquired from ArrayExpress included quantile normalized β-values, which were used directly without undergoing further normalization or adjustment processes. Additional details regarding the data processing can be found in the foundational publication [186]. Analysis of cell heterogeneity was not conducted in this instance since cell sorting had been performed prior to the methylation analysis as outlined in the original study. The data for GSE88890 was also downloaded from GEO in the form of raw β-values and passed through similar normalization and bias correction steps as in the standard pipeline. As this data comes from brain cells, standard cell-type correction was not performed. We utilized a *meffil* R package that uses reference methylation dataset from dorsolateral prefrontal cortex samples [186] to estimate the proportion of glial and neuronal cells in samples [212] that were subsequently included in the statistical models. Similarly, preprocessing of GSE72680 data also started with raw β-values and followed similar normalization and bias correction steps. Cell proportion coefficients were already provided at GEO and included in statistical models. DNA methylation data from GSE49065 and GSE56047 were also used directly from GEO. In **Study III**, a more strict protocol for GSE56047 was used and included similar CpG filtering steps as for other cohorts.

In **Study V**, many data processing routines were more standardized as the purpose of the study was assuming a combination of data from several batches. Preprocessing of IDAT-based cohorts remained same as before, and most of of the changes were applied to cohorts available from other starting file types. Preprocessing of GSE72680 started from raw signal intensities, and data was imported with *minfi*-based function "readGEORawFile". To estimate cell proportions we used R package *meffil* that allows performing such computations with β-values instead of RGSet in *minfi*. The resulting methylation β-values were adjusted for cell heterogeneity with regression-based method as described for IDAT-based cohorts. Similar processing steps were performed in GSE113725, GSE198904_DHRC,GSE198904_OBS, and GSE74414_MPIP2. Datasets in all cohorts were filtered so they include only those probes that passed filtering and QC steps in all individual batches and exists in both investigated Illumina methylation arrays. Lastly, for some of the analyses, data harmonization of merged methylation dataset was performed. It consisted of two steps: quantile normalization of probes and correction of cohort-based batch effect with "ComBat".

3.6 Transcriptome data preprocessing

Similarly to DNA methylation data, preparation of transcriptome datasets was also slightly different depending on the study objective and the dataset used. In the most of cases, however, we primarily utilized already available normalized data for analyses. Overall, all procedures were also performed within the R environment. The datasets for cohorts GSE53987, GSE98793 were originating from the Affymetrix Arrays, while the datasets GSE46743 and GSE64930 were acquired using the Illumina HumanHT-12 expression beadchip, necessitating distinct preprocessing methods. In GSE53987 and GSE98793, the initial data was accessible in the unprocessed CEL file format. The process of importing, quality control, preprocessing, and normalizing the data for these particular cohorts was executed via an *affy*-based protocol tailored for Affymetrix array data [213]. Data import was conducted using the "ReadAffy" function, followed by an examination of array intensity images, distribution of intensities to identify any possible anomalies. Following this, all preprocessing and normalization procedures were conducted within the "expresso" framework in the *affy* R package. Background correction was set to a robust multi-array average (RMA) [137]. Then the resulting log2-transformed values were quantile normalized, and a probe-wise correction based exclusively on perfect match intensities ("pmonly") was performed [214], and the summary expression values were obtained via the "medianpolish" procedure [215]. These processes were individually applied to each brain tissue sample in the GSE53987 cohort. Additionally, data for GSE98793 was corrected for batch effect (covariate)

with "ComBat".

In the cohorts GSE46743 and GSE64930, we used already normalized expression data available at GEO. Background correction and stabilization were initially performed via variance stabilization and normalization (VSN). Detailed methodologies pertaining to the initial preprocessing of these datasets are documented in the respective publications associated with these cohorts [201, 202]. In the cohorts GSE49065 and GSE56047, we also utilized normalized expression data directly from GEO.

Lastly, in **Study III**, transcriptome probes were additionally filtered based on their alignment with respected target transcripts. These procedures are extensively described in the supplementary information for **Study III** article. Briefly, all non-specific probes that target several or no transcripts presented in hg19 genome assembly were removed. All transcript target sequences were downloaded via University of California, Santa Cruz (UCSC) genome browser API [216]. When, the expected probe targets were aligned with their associated transcript target sequences (without exons) using the R package *Biostrings* [217]. All probes that matched less than 80% of the maximal potential score (we allowed 20% mismatch due to potential alternative splicing) to the reference genome were discarded.

3.7 Proteome data preprocessing

This section only applies to **Study IV**. Data from OLINK was obtained in the form of .XLSX files (excel spreadsheets) containing normalized NPX. Probes and participants were checked for QC warning, and all partcipants with such warning message were removed from the analysis. As data for the analysis was obtained in several batches, invoving several plates, the NPX values were normalized with "Combat" to remove batch effect. The batch correction was performed with covariates that included DAWBA risk group, age, sex, as well as the antidepressant intake (that was included only in analysis adjusted for antidepressant intake). The differential proteome prifiling was performed only for the assays that were detected in more than 75% of probes (43 out of 92 proteins).

3.8 SNP data collection

Studies **I, II, & III** involved analyses incorporating information on depression-related SNPs. In **Study I**, data were sourced from the latest editions of the GWAS Catalog [58] and DisGeNet [218] databases. We gathered data on eight depression-related SNPs, ultimately including six in our analyses. SNP-

related data, such as MAF, were collected from the National Center for Biotechnology Information (NCBI) and the UCSC Genome Browser. In **Studies II & III**, the initial search for SNPs and subsequent filtering were based on predefined keywords, as detailed in the supplementary materials of the respective manuscripts. The objective of this search was to identify SNPs specific to depression, deliberately excluding those associated with comorbid disorders, such as neuroticism and others. In **Study II**, the resultant list was further refined to include only SNPs identified in the European population and detectable by Olink assays. In **Study III**, we retained the initially filtered list of SNPs. Consequently, 54 SNPs were included in **Study II**, while a total of 2073 SNPs were identified in **Study III**.

3.9 Data analyses and statistical modeling

3.9.1 Study I

SNP-CpG associations

Investigation of SNP-CpG associations were performed with the R package *limma* both for screening and recall PSY samples. The association was modelled as a linear model, where methylation M-value was considered an outcome variable depending on SNP, sex, age, BMI, and analysis batch. The genetic model assumed "dominant" relatinship, i.e., minor allele is present or not present as study sample size was small and depression has high rates. The model structure is outlined below:

$$Methylation_i \sim int_i + SNP_{dom.,j} + Sex + Age + BMI + Batch + \varepsilon$$

All models were adjusted for multiple comparisons with the false discovery rate (FDR) method and additionally corrected for number of SNPs tested. An adjusted $p < 0.05$ was considered significant. Bias- and inflation-corrected t-statistics and adjusted p-values were calculated together with standard estimations from *limma* using R package *bacon*.

mQTL modeling

To investigate the architecture of SNP-CpG landscape around *MAD1L1*, we performed a cis-mQTL analysis, modelling all SNP-CpG associations within mapped *MAD1L1* coordinates $\pm$ 10000bp (chr7:1845430–2282580). SNPs passed stricter QC steps, including imputation info score ≥ 0.9 and A1 expected frequency between 0.1 and 0.9. Statistical modelling was perfromed with R package *MatrixEQTL* optimized to perform parallell computations via large matrix operations [219]. Additive genetic model was used for analysis:

$$Methylation_i \sim int_i + SNP_{add.,j} + Sex + Age + BMI + Batch + \varepsilon$$

Statistics were adjusted with the FDR method and SNIPA was used to investigate linkage disequilibrium for the Top 10 most significant variants.

CpG-Phenotype modeling

In **Study I**, we investigated several association between methylation at *MAD1L1* and corresponding psychiatric phenotypes. We used binary logistic regression to model associations. The Wald test (R "summary" function) was used to obtain two-tailed p values for coefficients. In the PSY cohort screening and recall, DAWBA depression risk was regressed against corresponding methylation at investigated CpG site and was adjusted for sex, BMI, age, and batch:

$$Depression\,risk_{DAWBA} \sim int_i + Methylation_i + Sex + Age + BMI + Batch + \varepsilon$$

In the SKI cohort, the dependent variabe represented suicide severtery, whereas predictors included sex, age, batch, binary variable for other presonality disorders, and status of alcohol addiction. BMI covariate was not included as was missing for 10 participants. The model formula was as follows:

$$Suicide\,sev. \sim int_i + Methyl._i + Sex + Age + Batch + Pers.dis. + Alc.add. + \varepsilon$$

In the validation cohorts, covarites were based on avilability of the data and biological importance and were modelled in GSE41826, GSE88890, and GSE72680, respectively:

$$Diagnosis \sim int_i + Methyl._i + Sex + Age + Ethnicity + \varepsilon$$

$$Group \sim int_i + Methyl._i + Sex + Age + Neurons + Glia + Array + \varepsilon$$

$$Depr.treatm. \sim int_i + Methyl._i + Sex + Age + Ethnicity + BMI + CD4 + CD8 + NK + Bcell + Mono + Gran + \varepsilon$$

CpG-Expression modeling

To model associations between *MAD1L1* CpGs and MAD1L1 expression, we used a linear model approach where the expression was a dependent outcome, whereas methylation and covariates were predictors. In GSE49065, we used data where cells where exposed to sham control. The model included only one covariate - age.

$$MAD1L1\,expression \sim int_i + Methyl._i + Age + \varepsilon$$

In GSE56047, models were adjusted by sex, ethnical background, age, research site, and non-targeted cell proportions. Sex, ethnical background, and research site are represented by the variable "RacegenderSite" in the original dataset. The model was formulated as follows:

$$MAD1L1\ expression \sim int_i + Methyl._i + Age + RacegenderSite + Array + Tcell + Bcell + NK + Neutro + \varepsilon$$

3.9.2 Study II

SNP-phenotype associations in UK biobank

Each single nucleotide polymorphism (SNP) was coded into three distinct genotypes: homozygous for the major allele (0), heterozygous (1), and homozygous for the minor allele (2). To augment the statistical power for analyzing co-morbid traits, phenotypes pertaining to anxiety, bipolar disorder (BD), and schizoaffective disorders were merged into a single variable termed "comorbid traits." The association analysis between individual genetic variants and depression/co-morbid traits was conducted using binary logistic regression models, which were adjusted for covariates including age, sex, and the first 10 principal genetic components. Statistical significance was determined at a p-value threshold of < 0.001 (0.05/38), taking into account the Bonferroni correction.

SNPs demonstrating an association with each phenotype, as determined by the established significance threshold (p-value < 0.001), were indetified. A standardization process was implemented for all SNPs to ensure a consistent directionality of effect. Then, two genetic risk scores were created: a) $GRS_{dep\text{-}sig}$, which aggregated the SNPs significantly correlated with depression, and b) $GRS_{dep\text{-}all}$, calculated from all SNPs, irrespective of their statistical association with depression. These genetic risk scores were then employed in binary logistic regression models as independent predictors for depression and comorbid phenotypes. Utilization of the risk scores encompassed both continuous variable or quartile-based variables. Additionally, sex-specific analysis was performed to ascertain the differential predictive efficacy of these genetic risk scores in male and female subsamples. The statistical analyses of UK Biobank data were conducted with IBM SPSS Statistics, version 26.0 (IBM SPSS, Armonk, NY, USA).

Methylation-phenotype associations

We used *limma* package to study associations with depression phenotypes. This analysis focused on probes within six genes, which were selected based on prior analyses of UK Biobank data. These genes are Dystroglycan 1 (DAG1), Fragile Histidine Triad Diadenosine Triphosphatase (FHIT), Butyrophilin Subfamily 3 Member A2 (BTN3A2), Tenascin XB (TNXB), Latent Transforming Growth Factor Beta Binding Protein 3 (LTBP3), and Neural Cell Adhesion Molecule 1 (NCAM1). In linear models, methylation M-value was treated as

an outcome variable whereas diagnosis or disease status was a binary predictor. All models included covariates depending on the available data and were formulated as follows for PSY, GSE72680, and GSE125105, respectively:

$$Methyl._i \sim int_i + Depression\,risk_{DAWBA} + Sex + Age + BMI + Batch + \varepsilon$$
$$Methyl._i \sim int_i + BD\,treatm. + AD\,treatm. + Depr.\,treatm. + Sex + Age + Ethnicity + BMI + CD4 + CD8 + NK + Bcell + Mono + Gran + \varepsilon$$
$$Methyl._i \sim int_i + Diagnosis + Sex + Age + PC1 + PC2 + \varepsilon$$

In these equations, *BD treatm.*, *AD treatm.*, and *Depr. treatm.* incicate treatment information for bipolar disorder, anxiety disorder, and depression in GSE72680. *PC1* and *PC2* indicate first two genetic principal components. Models were analyzed at nominal significance and emphasis was placed on overlapping results. FDR-adjusted statistics were also reported.

Transcriptome-phenotype associations
Analysis of differential expression was undertaken exclusively for probes associated with the specific genes harboring significant depression-linked SNPs, which met the criteria for quality control and were available for analysis. The preliminary list encompassed six genes: DAG1, FHIT, BTN3A2, TNXB, LTBP3, and NCAM1. Within the Illumina-based expression cohort GSE46743, transcripts corresponding to TNXB were absent for processed data. Limma-based models were used to indetify diffrentially expressed genes. Within the analytical models, expression levels were treated as a quantitative outcome variable, while the presence of a diagnosis or disease condition served as a dichotomous predictive factor. Models also included covarites that were included based on data avilability. The model formulas were defined as follows for cohorts GSE53987, GSE98793, and GSE46743:

$$Expression_{i,tissue} \sim int_i + Diagnosis + Sex + Age + PMI + Ethnicity + TissuepH + RIN + \varepsilon$$
$$Expression_i \sim int_i + Diagnosis + Sex + Age + Anxiety + \varepsilon$$
$$Expression_i \sim int_i + Age + BMI + \varepsilon$$

In these equations, *TissuepH*, *PMI*, and *RIN* incicate tissue pH, post-mortem interval, and RNA integrity number in GSE53987. Models were analyzed at a nominal significance and emphasis was placed on overlapping results. FDR-adjusted statistics were also reported.

3.9.3 Study III

In the differential methylation and expression analyses in **Study III**, we used nominally significant associations ($p < 0.05$), as initial cohorts have insuffi-

cient power to detect changes even at FDR. The emphasis was placed on replicated results in at least one cohort. No threshold for log_2FC in the associations was used.

Differential methylation profiling

The R package *limma* was used for differential methylation analysis. Similar to previous studies, models were adjusted for covariates depending on the avilability of data. The following models were used for PSY screening, GSE72680, and GSE125105, respectively:

$$Methyl._i \sim int_i + Depression\,risk_{DAWBA} + Sex + Age + BMI + Batch + \varepsilon$$
$$Methyl._i \sim int_i + Composite\,depr.\,score + Sex + Age + Ethnicity + BMI + CD4 + CD8 + NK + Bcell + Mono + Gran + Array + \varepsilon$$
$$Methyl._i \sim int_i + Diagnosis + Sex + Age + PC1 + PC2 + \varepsilon$$

In these equations, $Composite\,depr.\,score$ indicates composite depression score (as defined previosly), *Array* indicates array barcode from Illumina.

Differential expresion profiling

The similar principals were applied regarding differential expression analysis. Extra data was obtained for cohort GSE64930 from the study investigator. The following equations show models for GSE98793, GSE46743, and GSE64930, respectively. Models were analyzed at nominal significance and emphasis was placed on overlapping results. FDR-adjusted statistics were also reported.

$$Expression_i \sim int_i + Depression + Sex + Age + Anxiety + Batch + \varepsilon$$
$$Expression_i \sim int_i + Depression + Age + BMI + \varepsilon$$
$$Expression_i \sim int_i + Depression + Sex + Age + RIN + SV1 + SV2 + SV3 + \varepsilon$$

In these equations, *RIN* incicates RNA integrity number in GSE53987, whereas $SV1 - SV3$ correspond to surrogate variables that are used to adjust for cell heterogeneity. Models were analyzed at nominal significance and emphasis was placed on overlapping results. FDR-adjusted statistics were also reported.

cis-eQTM analysis

Cis-eQTM modeling was performed to explain links between indentified genes at differential expresion analysis and differential methylation analysis with data from GSE56046. For modeling cis-eQTM associations, we employed R-based linear regression for CpGs and genes within a 1 Mbp range, incorporating age, cell proportions, array identifier, and "racegendersite" as covariates.

Significance in eQTM associations was determined at a global FDR < 0.05 (encompassing all generated models). The model formula is presented below:

$$Expression_i \sim int_i + Methyl._i + Age + RacegenderSite + Array + Tcell + Bcell + NK + Neutro + \varepsilon$$

Chromosome map models

In **Study III**, we juxtaposed results obtained from several analyses with reference array statistics and projected them on chromosomal maps. We counted the number of depression SNP, replicated CpGs, and replicated transcripts per 1Mbp on chromosomal map intervals. In addition, we also included counts for all reference genes, number of transcripts included in the arrays, and number of CpGs included in the arrays for each chromosomal sector. We investigated fn the obtained counts could be related in a systematic fashion using R-based linear regression. Associations were modelled as follows:

$$Cross-valid\,CpGs \sim int + Depression\,SNPs + Illumina\,probes + Reference\,genes + \varepsilon$$

$$Cross-valid\,proteins_{match} \sim int + Depression\,SNPs + Transcr.\,array\,genes + Reference\,genes + \varepsilon$$

$$Cross-valid\,proteins_{match} \sim int + Cross-valid\,CpGs + Transcr.\,array\,genes + Reference\,genes + \varepsilon$$

Detailed explanation of coefficients is included in the supplementary materials to **Study III**.

3.9.4 Study IV

In the differential proteomic and expression analyses in **Study IV**, we used nominally significant associations (p < 0.05), as initial cohorts have insufficient power to detect changes at FDR. The emphasis was placed on replicated results. No threshold for log_2FC in the associations was used. The proteomic analyses were modeled in the PSY cohort. The model was based on R implementation of linear regression. Protein levels (NPX) were treated as dependent outcome variables, wherease DAWBA-based risk group with covariates were treated as predictors. The main covariates included age and sex. In the sensitivity analysis, we additionally adjusted the model for antidepressant intake. The analysis adjusted for antidepressant intake was performed as a "complete case", excluding all participants that did not clearly indicate their treatment status. The model formulas are indicated below:

$$Protein_{NPX,i} \sim int_i + Depression_{DAWBA} + Sex + Age + \varepsilon$$
$$Protein_{NPX,i} \sim int_i + Depression_{DAWBA} + Sex + Age + Antidepressant + \varepsilon$$

As nominal significant results are prone to type I error, thus we sought to replicate the findings in the independent samples. As proteomic studies on depression were not found in GEO, we used already available transcriptome cohorts that we utilized in other projects. These cohorts included GSE53987, GSE53987, GSE46743, and GSE64930. Validation analysis was performed for seven genes indicated in PSY (either with or without antidepressants) at a nominal level. A two-tailed nominal $p < 0.05$ was considered significant, and only results replicated in at least one cohort were deemed as relevant. The Bonferroni-adjusted p-values were calculated along the nominal statistics. The model formulas for all supplementary cohorts are depicted below:

$$Expression_{i,tissue} \sim int_i + Diagnosis + Sex + Age + PMI + Ethnicity + TissuepH + RIN + \varepsilon$$
$$Expression_i \sim int_i + Depression + Sex + Age + Anxiety + \varepsilon$$
$$Expression_i \sim int_i + Depression + Age + \varepsilon$$
$$Expression_i \sim int_i + Depression + Sex + Age + RIN + SV1 + SV2 + SV3 + \varepsilon$$

3.9.5 Study V

Statistical analysis in **Study V** was represented by differential methylation analysis that was both conducted as pooled (mega-analysis) and seprately in the invidual cohorts that were meta-analyzed afterwards. In the pooled analysis, data on individual participants was merged using samples from PSY screening and recall, GSE72680, GSE125105, GSE74414, GSE113725_RDE, GSE198904_DHRC, and GSE198904_OBS, forming a combined sample of 1127 cases and 815 controls. Before the analysis, data was harmonized as described above. The initial associations between DNA methylation and depression were modelled equally in both polled and meta-analysis, leveraging *limma* R package. Models treated DNA methyltion as dependent outcome, whereas depression, age, and sex were independent predictors. In the pooled analysis, models were also adjusted for cohort of origin. We adjusted models for age and sex as only these covariates were available in all cohorts.

$$Methylation_i \sim int_i + Depression + Sex + Age + Cohort + \varepsilon$$
$$Methylation_i \sim int_i + Depression + Sex + Age + \varepsilon$$

Multiple comparison correction was performed with the FDR-method.

In conducting a meta-analysis of individual cohorts, we employed pre-harmonized datasets and *limma* R package-based regression analyses applied on individual cohorts. For CpGs demonstrating nominal significance in at least one cohort, a meta-analysis of Log_2FC was subsequently conducted. We used *limma*-derived parameter values (effect sizes) and standard errors. Meta-analysis of the estimates was based on linear random-effects models implemented in the *metafor* R package. The heterogeneity of the data was claculated using the Sidik-Jonkman estimator, whereas study weights were conducted via the inverse-variance approach [220, 221]. Meta-p-values were adjusted for multiple coparisons using the FDR method. We used a χ^2 test to see if the frequency of FDR-significant CpGs is statistically increased in the list of meta-significant CpGs. The results from pooled analysis were compared with results from meta-analysis, and overlapping CpGs were used further.

3.10 Machine learning and deep learning models in Study V

3.10.1 Feature selection strategies

Since the initial DNA methylation data contains extremely large number of features for everey participant, we sought to select the most important feaures for depression detection purposes. First, we used a set of 200 CpGs, representing the Top 200 CpGs from pooled differential methylation analysis with consistent direction across all cohorts as a list for achieving *"maximal theoretical performance"* as it has positive feature selection bias. For unbiased feature selection, we implemented an automated approach within each cross-validation iteration/training sequence. The authomated approach was based on either of the following: top 200 differentially methylated CpGs from *limma*, top 10000 differentially methylated CpGs from *limma* for regularized logistic regression models, top 5%, 1%, or 0.1% of CpGs with highest variance, selection of features based on models with L1 regularization, ANOVA F-value-based selection, and ExtraTree-based selection.

3.10.2 Machine learning models

In **Study V**, a combined depression dataset was used to evaluate the performance of common ML classifiers. We used python module *scikit-learn* [148] to implement all ML models. The model list included several key configurations: binary logistic regression without regularization, binary logistic regressino with "L2" regularization ("ridge"), binary logistic regressino with "L1" regularization ("lasso"), binary logistic regressino with "L1L2" regu-

larization ("elastic net"), ecision tree, random forest, support vector machine (SVM) with linear kernel, SVM with radial basis function (RBF) kernel, AdaBoost. Classifiers were primarily used with intial default *hyper-paramters* with several exceptions. We performed grid-search optimization for SVMs as these models are very sensistive to *hyper-paramters* compared to other models and require optimization of some kind to be useful. In the logistic regression models, the solver was set to "saga", and the maximal number of iterations was set to 5000 to enable solver convergence. In the elastic net classifier, the L1 ratio was set to 0.5.

3.10.3 Deep learning models

We implemented several versions of DL models to compare their performance with standard ML classfiers in **Study V**. The rationale behind the model architectures was based on the properties of the input DNA methylation data (1D-tensor) as well as previous models that were applied for DNA methylation data in similar domain of application.

Standalone deep learning classifiers

The most direct and straightforward method involved utilizing a small, deep, fully-connected neural network for classification purposes. In these networks, the initial layer is a one-dimensional tensor (for a single sample) of dimensions (batch, N), with N equaling 200 to represent the chosen number of CpGs. The terminal layer consists of a singular node equipped with a sigmoid activation function. Between the input and output layers, various combinations of hidden layers were developed and evaluated. These combinations included a diverse number of nodes, types of layer activations, regularization methods, batch normalization, and dropout. The configuration that yielded the highest performance was found to have layers with 100, 64, and 16 nodes respectively (excluding the input and output layers), with the final two layers incorporating batch normalization. The network's initial physical layer included a 10% dropout rate. All activation functions were configured to the Rectified Linear Unit (ReLU). The loss function utilized in all simple standalone DL classifiers was set to binary cross-entropy.

Joint autoencoder-classifier frameworks

Several models involving a combined application of autoencoding and classification were proposed in other domains that utilized DNA methylation data [172, 174, 177]. In these frameworks, the DNA methylation data is "condensed" into its hidden representation, using autoencoding that enables extracting biological information. This hidden data representation then is passed directly to a classfiier. The training of both model components is performed jointly (though autoencoder could be also pre-trained). Each autoencoder includes encoder component and decoder component. In the setting of **Study V**,

we implemented autoencoder component of the model either as fully-connected autoencoder or a variational autoencoder (VAE).

A structure of fully-connected autoencoder was based on fully-connected layers without dropouts with each layer having N, 128, 64, nodes for encoder and 64, 128, N nodes for a decoder. The letter N shows the number of input CpGs that was eventually set to 200 CpGs. The bottleneck of autoencoder (hidden dimension) was set to 32. We tested different activation functions for the layers with few restrictions. The activations for the output layer of autoencoder were always set to linear for models applied on M-values or to sigmoid in case of β-values.

The VAE is an autoencoder that models hidden dimension in such a way that it follows specified distribution. The normal distribution is the most commonly used. The encoder compronent of VAE included fully-connected layers containing N, 128, and 64 nodes. The main difference is that VAE models hidden dimension as a normal distribution, so the botteleneck component has to provide these properties. To do so, the bottleneck is constructed via reparametrization and models mean μ and logarithm of variance $log(\sigma^2)$ of the underlying distribution. These parameters are then used to obtain hidden dimension, combining modelled mean value and a sample from normal distribution scaled by determined starndard deviation σ. These processes are implemented in the sampling layer and defined as:

$$\begin{aligned} \sigma &= e^{(0.5+log(\sigma^2))} \\ \varepsilon &\sim N(0,1) \\ z &= \mu + \sigma\varepsilon \end{aligned}$$

where ε is a sample from normal (Gaussian) distribution that represents noise, whereas z is a latent space. We used a 32-dimensional hidden space in the implementations of VAE. The decoder component of VAE was similar to fully-connected autoencoder models.

The classifier component of all models was represented by a small fully-connected neural network containing dropout, regularization, batch normalization, and activations that were treated as *hyper-parameters*. The last node of the classifier had either sigmoid or tanh activations depending on the classification loss.

The presence of two components in joint models warrants the use of a complex loss function that should include information not only about the data reconstruction but also classification performance. The reconstruction loss of all autoencoders was based on the methylation input type. Thus, MSE loss

was used for M-values, whereas binary cross-entropy loss was applied for β-values. The classification loss was represented by BCE (though we also tested squared hinge while searching for best model configurations). The total loss of a fully-connected autoencoder could be defined in the following equation:

$$L_{total} = aL_{recon.} + (1-a)L_{class.}$$

where L_{total}, $aL_{recon.}$, and $(1-a)L_{class.}$ indicate total, reconstruction, and classification losses, respectively. The term a represents a balancing constant $a \in [0,1]$ that was treated as *hyper-parameter*. In the case of VAE, the total loss should include information on whether hidden dimension conforms to a normal distribution. This loss is formulated through Kullback–Leibler divergence as follows:

$$KL_{loss} = -D_{KL} = -\frac{1}{2}\sum_{i=1}^{N}(1 + log(\sigma_i^2) - \mu_i^2 - \sigma_i^2)$$

where N indicates batch size, whereas $log(\sigma_i^2)$ and σ_i^2 define parameters for the sampling layer. The total loss for VAE is formulated as a sum of all losses:

$$L_{total,VAE} = L_{recon.} + KL_{loss} + aL_{class.}$$

We additionally scaled the classification loss by a parameter $a \in [0, +\infty)$ that was treated as *hyper-parameter*.

3.10.4 Model selection, training and evaluation

Model training and optimizations in **Study V** were performed at a local computer with NVIDIA RTX A5000 GPU or at HPC cluster provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). We used *Adam* [222] optimizer for training the models. The *hyper-parameter* searches were performed with a single 3-fold cross-validation on the cross-validation dataset, using averaged hold-out subset AUCs in a grid search manner. Best model configurations were tested via 10 repetitions of 3-fold cross-validation on the same dataset using averaged statistics for hold-out subsets. The last model testing was performed on independent separate hold-out set containing data from PSY recall, GSE74414_MPIP2, and GSE198904_OBS that were not used for model optimizations and unbiased feature selections. DL models were implemented in *tensorflow2*; ML models were implemented in *scikit-learn*. All analyses were performed in bash, python, and R environments.

3.11 Gene enrichment analyses

Gene enrichment analyses were conducted in Studies **I, III, and V**. In the first study, the enrichment was provided by the STRING database web tool and included only MAD1L1 interacting partners. In **Study III**, we used classical overrepresentation analysis implemented in R package *clusterProfiler* [223]. Gene universes were set based on the -OMIC layer and included reference genome (genetic level), genes from HumanMethylation450 annotation (DNA methylation), and transcriptome-detectable set by both Affymetrix Human Genome U133 Plus 2.0 and Illumina HumanHT-12 arrays (transcriptome). Enrichment statistics were corrected with FDR. Further information is available in supplementary materials to **Study III**. In **Study V**, we opted to apply enrichment analysis implemented with R package *missMethyl* as it takes into account gene length bias and overlapping genes [224]. Alternatively, a mapping of CpGs to genes was also performed based on eQTM data from the BIOS QTL browser, and the following enrichment was performed directly. Both enrichments were interpreted at nominal significance levels.

3.12 Thesis writing

The text in this thesis was written with the assistance of ChatGPT. The model was prompted to perform language correction and to improve the clarity and flow of certain sentences. Author checked and verified AI-generated text.

4. Results[1]

4.1 Study I

The analysis in **Study I** followed a classical target gene approach where we specifically focused on exploring *MAD1L1* contribution to depressive behavior. It would be reasonable to mention on how we decided investigating this gene in the first place. The product of *MAD1L1* gene is a component of mitotic spindle assembly checkpoint that is important for cell division. Functionally, this complex is inhibited via spindle and KT associated 2 (SKA2) [225]. Interestingly, SKA2 gene has been linked to suicide based on previous research [226, 227, 228, 229]. Thus, we were initially expecting that functionally-related genes may also be involved in similar phenotypes. Screening through SKA2-related genes and mitotic spindle assembly checkpoint genes, including *MAD2L1, MAD2L2, BUB1, CDC20, BUBR1, BUB3, AURKB, PRR11, SKA1-3* and other, we found that *MAD1L1* had surprisingly high number of depression-related SNPs reported previously. Thus, it was interesting to see if this gene has also associated DNA methylation depression signatures and whether they are related to previously published SNPs.

The analysis started with characterization of available data on the PSY cohort and SKI cohort. Data in the PSY was primarily skewed to unaffected individuals (low depression risk group) as expected since this cohort is primarily composed of adolescents sampled from Swedish schools. Less pronounced pattern was also observed in the SKI where ∼65% of participants were considered non-violent suicide attempters. In both cohorts, BMI was somewhat similar in "cases" and "controls". The gender distributions were a bit different, and high-risk group in the PSY was enriched with women, whereas the majority of male participants were presented in severe suicide group in the SKI.

After characterization of cohorts, the main focus was on obtaining information on *MAD1L1* SNPs in DisGeNet and GWAS Catalog. Eventually, data on six SNPs that were interesting candidates for analyses was collected (rs56072378, rs11772627, rs3823624, rs2056477, rs11514731, rs61409925). Two additional SNPs, s12668848, rs1107592 were either not fully genome-wide significant (or not close to it) or were not specifically depression-related based

[1]Some of the text may partially or fully overlap with the included articles of studies **I-V** or Licentiate thesis "The role of genetic, epigenetic, and proteomic factors for psychiatric disease" presented by Aleksandr V. Sokolov at Uppsala University on Feb 1, 2023.

on the data from the initial GWAS publication, and thus were excluded from further analysis. Once SNPs were obtained, all relevant data from the PSY genotype was extracted and used for association analyses in screening and recall samples of the study. We focused on methylation sites that are related to several depression SNPs as these could be more biologically relevant. Methylation-SNP association analyses indicated only three CpG sites (cg02825527, cg18302629, and cg19624444) were significant in both screening and recall, while depending on more than one depression SNP.

The second stage of the project was to investigate whether the identified CpGs have any kind of depression and other psychiatric associations in the depression or suicide cohorts. In the PSY cohort, none of these CpGs were related to the DAWBA risk groups. However, we observed lower methylation profiles at cg02825527 for severe suicide group in the SKI ($\exp(\beta) = 84.521$, $p \sim 0.003$ (the base level in the model was set to "severe" attempt)) and a similar statistical tendency for cg19624444 ($p \sim 0.061$) in the same cohort. Sensitivity analysis with additional BMI coefficient did not add substantial differences in the associations.

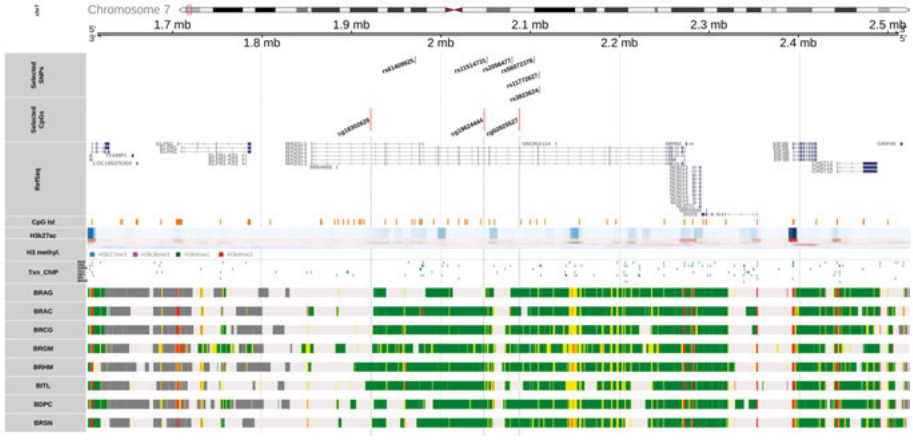

Figure 6. Genetic context of MAD1L1 depression SNPs and CpGs

Subsequently, we performed validation and functional analyses for obtained CpG sites. Methylation levels of cg02825527 were significantly increased in glial cells of depressed participants in GSE41826 ($p \sim 0.004$), whereas cg18302629 was hypomethylated ($p \sim 0.023$). However, both of these CpGs were not altered in neural cells of same cohort, whereas cg19624444 was insignificant in neither neural nor glial cells. In the second cohort, GSE88890, with two cortical brain regions, cg02825527 showed a tendency for increased methylation in the case of suicidal outcome in BA11 tissue, whereas cg19624444

and cg18302629 demonstrated association with suicide in BA25. Direction for association was matching for cg02825527 in GSE41826 and GSE88890. In the last validation cohort, cg19624444 showed strong association with depression treatment (hypomethylation), whereas cg02825527 demonstrated similar trend that failed to become statistically significant. We further investigated if the CpGs could be related with expression of MAD1L1. Though, we have not observed particularly strong associations for interrogated CpGs, cg18302629 showed nominal association with MAD1L1 transcript level in GSE56047.

4.2 Study II

The overall strategy in **Study II** was to investigate depression-related SNPs detectable by the Olink proteomic panels. We first extracted SNPs from GWAS Catalog database, including all depression-specific SNPs reported in the European population. In total, 38 SNPs were included as they passed quality control steps in the UK biobank data based on MAF, genotyping and missing rates, compliance with HWE, and linkage disequilibrium. Using binary logistic regression models, we investigated whether the included SNPs are related to composite depression scoring in the UK biobank cohort, andjusting for age, sex, and the first 10 genetic principal components. We identified eight SNPs that showed significant associations with depression, and no association for co-morbid traits were observed. Significant SNPs were spatially overlapping with sequences of a total of six genes: *DAG1*, *FHIT*, *BTN3A2*, *TNXB*, *LTBP3*, and *NCAM1*. The obtained SNPs were used to calculate a composite depression genetic risk scores (GRS). Subsequent analysis of the correlations between the calculated genetic risk scores revealed that a score derived from significant depression SNPs (GRSdep-sig) exhibited a robust association with depression (95% CI: 1.020 (1.016-1.023); p-value: 1.42×10^{-24}) but not with co-morbid traits. Furthermore, individuals in the upper quartile for GRSdep-sig demonstrated a 11.8% elevated odds of experiencing depression compared to those in the lowest quartile. Sex-stratified analyses of GRS yielded analogous outcomes, with GRSdep-sig showing a significant correlation with depression in both females and males with no association with co-morbid traits.

Subsequently, we opted to explore if identified genes with genetic evidence are related to depression at DNA methylation and expression levels in the independent cohorts. Using three publicly-available transcriptomic cohorts (GSE53987, GSE98793, and GSE46743), we were able to identify several probes that were nominally (not corrected for multiple comparisons) significantly associated with depression. No probes survived the adjustment for multiple comparison with the FDR. The direction of association for TNXB-related probes was consistent in the cohort GSE53987 across several brain tissues and included hippocampus, pre-frontal cortex (BA46), and associative striatum,

but did not match with GSE98793. Similarly, CAM1-related probes were also associated with depression in cohort GSE98793 and GSE46743 without matching association direction. All association demonstrated relatively small effect sizes.

We performed a comparable analysis using DNA methylation data from the PSY screening, GSE72680, and GSE125105 for promoter- and main gene-body-located CpGs as these could have different impacts on gene regulation [68]. Similarly to transcriptome analysis, we were able to identify several CpG sites with nominally significant associations with depression. Only cg24336152 (promoter analysis) and cg19569130 (gene body analysis) were found deregulated in more than one cohort with a consistent association direction. Both of the CpGs are are related to *TNXB*. It should be noted that none of the probes was significant in more than two cohorts.

4.3 Study III

We started multi-omic analysis of depression with identification of all listed GWAS Catalog SNPs related to depression and not co-morbid disorders. We included all SNPs identified in all available populations, resulting in 2073 unique SNPs that were linked to depression. Our first analysis was to investigate how results from different studies are replicated between GWAS and if we see some SNPs that are very confidently related to depression. We used a simple definition of replication implying that a SNP must be listed in the results in more than a single study. Interestingly, we found that only 166 out of 2073 were reported more than a single time. This result, however, is potentially biased as it included not only genome-wide significant SNPs but also other associations that authors deemed important to mention within the manuscripts or associated tables. For the genome-wide significant SNPs, the replication rate was $\sim$13% corresponding to 114 SNPs out of 1029 associations (879 unique SNPs) reported more than once. We subsequently mapped all initially identified SNPs to genes by genomic coordinates and performed enrichment analysis to see if there is any overrepresented biological process that could be related to GWAS-identified genes. In total 847 genes were analyzed, showing enrichments in synapse organization, signaling,and membrane transport processes. Reproducible SNP genes were more strongly associated with cell adhesion.

The next step in the workflow was to identify depression-related methylation changes based on the data from three available cohorts (PSY screening, GSE72680, and GSE125105). We found that none of the individual cohorts had sufficient sample sizes to identify differentially-methylated CpGs in the genome-wide setting if results are adjusted for multiple comparisons. Only

one CpG (cg20263853, gene *RERE*) passed the adjustment with the false discovery rate (FDR) of 5% in GSE72680. However, it was not replicated in two other cohorts. We focused on identification of methylation changes that should be detected at least in two out of three cohorts. We identified shared differentially-methylated CpG sites between cohorts that corresponded to 1-2% of nominally-significant CpGs per study. Only 49 CpGs were found nominally-significant in all three cohorts. In total, 2491 CpGs ($\sim$5%) were found differentially methylated in at least two cohorts, of which 1 480 had a matching direction for association. The subsequent enrichment analysis of genes overlapping with detected CpGs identified glial processes, cell adhesion, and membrane activity. The set of 49 CpGs failed to identify any biological process after the FDR correction.

We next performed an analogous analysis in the three transcriptomic cohorts (GSE98793, GSE46743, and GSE64930) to identify depression-based changes in gene expression. We used a similar strategy as in the DNA methylation analysis, and focused only on results that were replicated in at least two cohorts. Differential expression analysis highlighted 4226 nominally significant probes with matching directions in GSE98793, 1641 probes in GSE46743, and 888 probes in the cohort GSE64930. Of these probes, only 70 passed the FDR correction in GSE98793, whereas all other results were only nominally-significant. We identified 581 genes (transcripts) that showed differential expression in at least two cohorts with a matching direction for associations. The enrichment analysis of these genes highlighted strong overrepresentation of inflammatory response and innate immune pathway terms.

The most interesting step in this study was to identify if there are genes related to depression that are supported by evidence at several -OMIC levels and if depression signatures of different -OMICs show colocalization in the genome. Only three genes were supported by evidence at all three -OMIC levels and included: forkhead box protein P1 (FOXP1), vacuolar protein sorting-associated protein 41 (VPS41), and AKT-interacting protein (AKTIP). We additionally observed overlaps between two -OMIC levels: GWAS and methylation genes (n=87 genes), GWAS and expression genes (n=22 genes), and methylation with expression genes (n=41 genes). Overlap between GWAS and methylation was enriched with membrane adhesion genes. The -OMIC signatures did not demonstrate systemic colocalization based on chromosomal sectors. We additionally compared -OMIC signatures based functional relationships, including eQTL, meQTL, and eQTM. We were able to validate two SNP-expression pairs in brain tissue, 72 SNP-expression pairs in blood, two SNP-CpGs pairs in blood, and one CpG-expression pair in blood based on the data from GTEx and BIOS QTL. Some depression transcripts, such as MADD, USP4, NUP43, and other were regulated by multiple depression

eQTLs or eQTMs as discussed in the study article.

Lastly, we investigated whether identified gene sets are related to existing drugs or could be used as drug targets. The highest number of agents matched with GWAS (91/262 agents, followed by DNAm(27/121 agents) and transcriptome (30/113 agents). We subsequently explored trends in the associated clinical trials. The highest number of targets detected in depression clinical trials came from GWAS set (20 genes) compared to methylation (11 genes) and expression(12 genes). The enrichment analysis of drug classes showed that most of the identified drug sets, including GWAS and methylation, were enriched in several classes, and calcium channel blockers overlapped for several sets.

4.4 Study IV

In **Study IV**, we focused on proteomic exploration of depressive behavior. We first started with characterization of the PSY samples with available proteomic profile that included data on 461 participants. Out of these 461 participants, 61 were classified as high depression risk group based on the DAWBA score, of which the vast majority (> 91%) were women. The age was not different between high- and low-risk groups (16.67 vs 17.26). Five participants in the high-risk DAWBA group reported taking antidepressant, whereas nine participants reported taking antidepressants in the low-risk DAWBA group. Interestingly, more than 30% of all participants failed to clearly identify their antidepressant intake status.

The main analysis involved determination of depression-related proteomic signatures in the plasma of participants. The Neuro-Exploratory from Olink proteomics is designed to capture expression signatures of 92 genes. In the case of this study, only 43 proteins were detected in more than 75% of samples and were included in the analysis. We performed two types of analysis both unadjusted for antidepressants and adjusted for antidepressants. We opted to perform both analyses as antidepressants could be an important confounder regarding depression-related expression levels (though we were not aware of specific effects). On the other hand, the inclusion of this covariate leads to exclusion of many participants, thus may produce biased results and also create the possibility of over-adjustment (if antidepressant intake is not related to expression). The unadjusted analysis indicated five proteins (RBKS, CRADD, ASGR1, HMOX2, and PPP3R1) that showed different levels between DAWBA risk groups at nominal significance. The adjustment of analysis with antidepressant intake resulted in three proteins (RBKS, CRADD, and PPP3R1) that showed nominally significant differences. All of the observed associations showed very modest effect sizes.

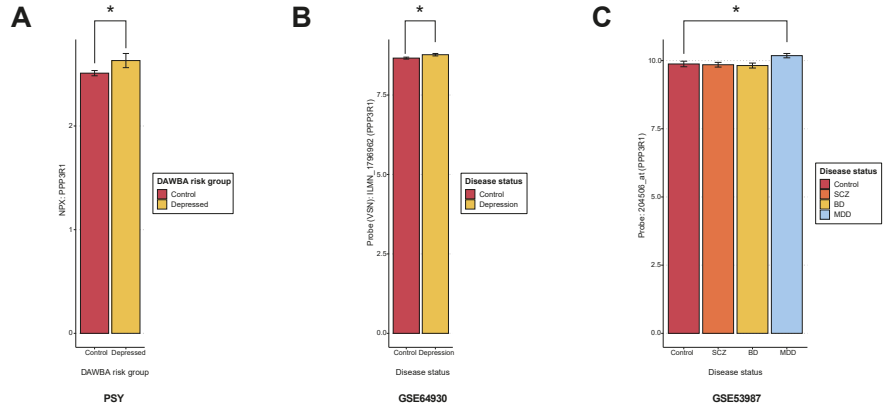

Figure 7. Observed NPX values for PPP3R1 in the PSY cohort (**A**), GSE64930(**B**), and GSE5398(**C**). In all of these cohorts PPP3R1 shows slight upregulation in depression. Asterisk indicates nominal significance.)

We wanted to validate the identified proteomic changes and used several open-access transcriptome cohorts on depression in brain and blood (GSE53987, GSE98793, GSE46743, and GSE64930) to see if we could identify similar changes as in the PSY. This analysis showed nominally significant associations for PPP3R1 in GSE53987 (prefrontal cortex) and GSE64930 (whole blood) with matching directions for associations with the PSY cohort. We additionally observed nominal significance for ASGR1 and Bonferroni-adjusted significance for CD63 in GSE98793 also with matching directions to the PSY. PMVK and RBKS were nominally significant in GSE46743 but in the opposite direction compared to the PSY cohort.

4.5 Study V

In **Study V**, we first assembled the combined DNA methylation depression dataset from several cohorts. The combined dataset included 1127 cases and 815 controls. The dataset showed class imbalance, and this trend was also noticeable in the individual cohorts. In some of the cohorts, however, such as PSY and MPIP2, the number of controls was substantially larger than the number of cases. The sex composition of the combined dataset was also not equal with the number of females being almost double than males in both cases and controls. The obtained methylation in the individual cohort-batches showed inconsistency that was stabilized after data harmonization. Harmonized dataset appeared to be somewhat homogeneous based on PCA visualization. The obtained combined datasets (harmonized and non-harmonized) were then further used for mega-/meta-analysis and the development of clas-

sification models.

The next step of the study was to analyze the stability of CpG sites with respect to depression status in different populations. To do so, we performed mega- and meta-analysis of the individual cohorts. The mega-analysis identified 20667 CpGs that were nominally associated with depression status. None of the identified CpGs survived the FDR correction for multiple comparisons. Among the nominally significant CpGs, only 723 showed consistent direction of differences between cases and controls in all cohorts. We used Top 200 of these CpGs (sorted by nominal p-values) to generate a list of positively biased CpG set for depression detection. Subsequently, we performed differential methylation analysis in the individual cohorts to meta-analyze these hits afterwards. The individual cohorts yielded 158618 unique CpGs that were associated with depression at least once. We did not identify any CpGs that were nominally significant in all included datasets. Only one CpG cg25013095 was nominally significant in 6 cohorts, showing matching directions (hypomethylation in depression) in five. Other CpGs were nominally significant in even smaller number of cohorts. The meta-analysis, in turn, identified 2451 nominally significant CpGs that were related to depression. Both mega- and meta-analysis methods produced an overlapping set of 1987 CpGs.

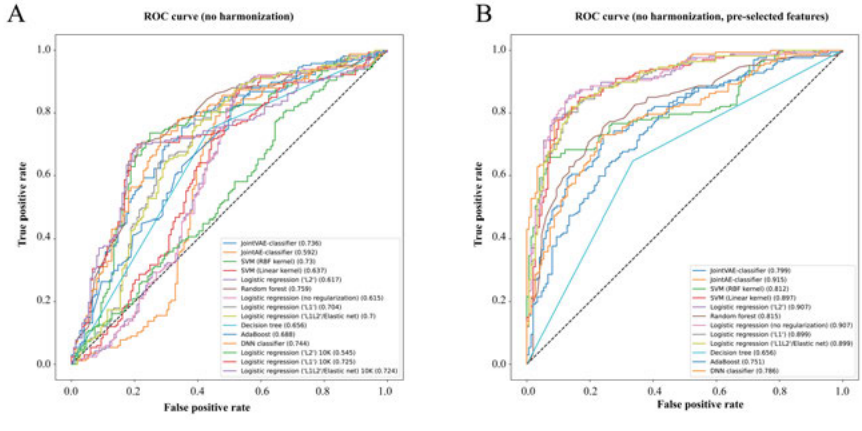

Figure 8. The performance of classification models (ROCs and AUCs) for depression detection in the final testing with Top 200 *limma* CpGs (unbiased set, section **A**) and 200 CpGs with consistent direction in all cohorts (biased set, section **B**) on data without harmonization

Identified set of 1987 CpGs did not demonstrate associations with any biological processes at FDR<5%. However, nominally, it showed associations with axon guidance, DNA damage response, membrane processes, and immune-related processes. The alternative eQTM-based enrichment analysis also iden-

tified only nominal associations, highlighting mammary gland morphogenesis and immune system-related terms. We also explored the location pattern of identified depression-related methylation. We could observe that hypermethylated depression CpGs are primarily located close to gene sequences and close to inactive regions. Interestingly, the location of hypomethylated depression CpGs was enriched with exclusively inactive chromatin states.

Lastly, we investigated the applicability of blood DNA methylation signatures for depression detection. We used multiple feature selection strategies as well as several classification algorithms. In short, our results demonstrated that simple random forest models offer the best average performance regardless of the feature selection approach with moderate average AUCs of 0.73 in CV. The application of more complex models with DL-based architectures did not provide additional benefits in the current setting. The accuracy of all models was very sensitive to data preprocessing as data harmonization eroded all predictive capacity of all classifiers with unbiased feature selection. Interestingly, all models performed "well" with biased set of initial features, highlighting the scale of potential positive bias in such studies.

5. Summary, conclusions, and discussion[1]

This thesis is a summary of several studies on depression and suicide that attempted to widen the knowledge on biological factors related to this very complicated disorder. Me and co-authors were able to identify/confirm multiple methylation and expression markers, such as multiple CpGs, MAD1L1, TNXB, FOXP1, VPS41, and other that could be related to depression pathology. Furthermore, we attempted to explain how some of the existing depression markers could be impacting cellular processes through methylation and gene expression (**Study I**).

The **Study I** was my first attempt to investigate depression biomarkers using classical "targeted" approach. MAD1L1 gene was very interesting candidate as it had many depression-related SNPs and was previously reported to be related to stress. Indeed, we were able to observe that some MAD1L1 SNPs potentially project their effect through methylation changes in the same gene. Interestingly, one of the CpGs cg19624444 was also confirmed in additional large mendelian randomization study that was published at the same time [230]. However, looking back at this analysis, the author could say that "targeted" approach has many potential issues, such as, for instance, statistical inflation. Alternatively, it could be potentially interesting to perform similar analysis utilizing all existing depression SNPs and see which methylation markers are affected, and which markers can potentially impact gene expression.

The **Study II** was somewhat different. Here the focus was to validate depression genes detected through SNPs and see if similar genes are also involved in DNA methylation and expression. We were able to see that TNXB could be potentially related to depression based on several datasets. Previous studies also suggest the role of this gene in depression [231, 232]. Though, the observed actual changes in **Study II** could be mostly interpreted as very weak or "negative" result, this study provided very interesting opportunity to test depression gene identification through multiple -OMICs in several datasets. This concept was fully implemented in the subsequent **Study III**.

[1] Some of the text may partially overlap with the included articles of studies **I-V** or Licentiate thesis "The role of genetic, epigenetic, and proteomic factors for psychiatric disease" presented by Aleksandr V. Sokolov at Uppsala University on Feb 1, 2023.

The **Study III** was designed to identify depression-related markers that are supported by evidence in several -OMICs. It is biologically reasonable that if something is truly related to disease, we should observe some evidence for it at several biological levels (-OMICs). During this study, we also investigated potentially interesting trends that could be generally related psychiatric research. It is somewhat exciting and also disappointing to confirm that depression SNPs have very low replication rate of $\sim$13%, nearly matching previously identified trend in all psychiatric GWAS [57]. It could be speculated that such low replication originate through insufficient power of majority of GWAS and intrinsic differences between studies yet it also makes sense that depression pathology should be consistent to a certain (potentially large) degree even in different populations as this disease is very prevalent regardless of population. In **Study III**, we explored the strategy of focusing on overlapping hits between the studies rather than their formal statistical significance. We observed somewhat unexpected trend that results in different independent -OMICs tend to reflect different proposed depression theories. Specifically, we observed that GWAS point to neurogenic pathology of depression, highlighting processes related to neurogenesis, signaling, etc, and this is in line with previous studies in animals and humans [233, 156, 234, 235]. DNA methylation, in turn, indicated the roles of glial and adhesion processes in depression. Interestingly, some of the research items suggest that depression could be primarily related to glial pathology [236, 237, 238]. Lastly, we observed nearly exclusive contribution of inflammation and immunity to depression based on transcriptome. And, indeed, it is a clear trend in recent research linking inflammation to many disorders, including depression [7, 239, 240, 241, 242, 243, 244, 245]. Interestingly, we identified three genes: forkhead box protein P1 (*FOXP1*), vacuolar protein sorting-associated protein 41 (*VPS41*), and AKT-interacting protein (*AKTIP*) as depression markers supported by all three -OMICs and validated some of the associations using public eQTL datasets and we also attempted to investigate some systematic dependencies between -OMICs at the genome level.

In the **Study IV**, authors were excited regarding recent developments in large-scale proteomic analyses, primarily facilitated by Olink platforms. We saw it as an opportunity to investigate proteomic factors in the domestic adolescent PSY cohort. Proteomic studies are interesting in this way as they measure the exact proteins and not their precursors such as in RNAseq or transcriptome arrays. We identified several candidate genes related to depressive phenotype based on the neuro exploratory platform. As the sample size is relatively small, we sought to replicate these findings in the independent samples. PPP3R1 was a particularly interesting finding showing consistent associations in three cohorts, one of which consisting of post-mortem prefrontal cortex samples. Previous studies strongly implicated prefrontal cortex deregulation in MDD, revealing functional, structural, and systems-level abnormalities [246]. It is

unclear why PPP3R1 appears to be deregulated in blood. We suggested a potential model on how PPP3R1 and other observed markers could be related to depression and explain the observed association. However, given small effect sizes, lack of power in all samples, and lack of protein coverage (as we used only Olink neuro exploratory panel), the proposed model is extremely speculative in nature.

Lastly, in the **Study V**, we wanted to try a different approach compared to studies **I-IV**. Here we viewed DNA methylation not as a collection of separate markers but rather as a combined measure that can characterize an individual. This is not a novel approach and is called DNA methylation-based scoring, similar to polygenic risk score. We investigated the stability of DNA methylation between different cohorts and how blood DNA methylation could identify depression status, and which classification strategy may be most suitable for this task. Interestingly, we identified a pool of DNA methylation markers that were associated with depression in multiple datasets. This panel of markers could be potentially used for depression characterization based on blood. We investigated the predictive capacity of DNA methylation in blood using various ML and DL approaches, and contrary to recent trend in complex models, simple random forest appears to be the most stable solution for this kind of applications. This notion is consistent to trends observed on tabular data [247]. In this study, we implemented different kinds of feature selection strategies, both the ones used in other ML tasks, as well *limma*-based approach that should "mimic" individual transcriptomic analyses. We could see that in the optimistic scenario (biased features), models could achieve quite high accuracy. However, in more realistic circumstances where we don't know features that "work", we can't expect accuracy significantly above 70-75% (at least based on our sample of models and available data) and could see huge variations based on data preprocessing. The accuracy of 70-75%, in my opinion, has limited applicability in diagnostic settings, but may be somewhat useful, if someone is interested in studying another phenotype and wishes to know if there is any depression confounding in the process. Though, it is also speculative, how well blood DNA methylation can distinguish between different psychiatric conditions, if at all.

6. Future perspectives

After studying depression (even though it applies to psychiatry in general) for several years, I could say that the field is facing many potential methodological challenges. Based on multiple previous studies as well as my personal results, it is very clear that depression studies are not very consistent in terms of identified hits. Personally, I don't support the argument that populations are too different to be compared and there are multiple sources of variance, such as different labs, questionnaires used, etc. that make different studies non-comparable. If the research field is studying something which is of true "causal" nature and has significant pathological importance for the disease, we should see some consistencies between the studies. I assume that these consistencies should be of higher level than we currently observe. Thus, we either studying a "null field" [248] and all of the observed effects are products of random noise, bias, and confounding or we use potentially an incorrect strategy in the first place.

At the moment, we can actually observe consistent findings on a system level, such as "inflammation", "serotonergic signaling", and similar broad terms. The difference comes if we look at the individual markers. My main assumption, which is speculative, is that the main problems arise from the statistical methods and research designs used in the field. Currently, studies like GWAS, EWAS or TWAS mostly attempt to identify markers based on some arbitrary pre-defined thresholds (such as p-values, logFC, and alike). Essentially, in every study, a set of markers with best of these or other statistics is obtained and then frequently there is a minimal/no comparison with other research, especially for "non-top" hits. It appears to me that every time we use such an approach, we potentially become victims of a type of overfitting problem. In ML, we can always build a model with perfect parameters that work with 100% accuracy in training but fails to predict anything on a new data. Similarly in -OMICs, we identify biomarkers with best statistics that fits the most our particular cohort or dataset. Then, it appears that these top markers may not generalize well, such as example of psychiatric GWAS with $\sim$12% replication rate in at least one independent study [57].

My main suggestion is that the field would benefit from comparative studies as well as meta-analyses where the focus would be made to identify consistent changes in multiple cohorts and populations. These individual changes don't have to be super significant or have very large effects in a single dataset, yet

they have to be consistent so we minimize the chance of obtaining a bias or random noise. Meta-analyses are proven very effective in answering complex multifactorial questions, and thus, for instance, we are aware that smoking is related to increased cancer risks [249, 250, 251]. I found it quite useful to use this approach in **Study V** and think that other -OMICs and psychiatric phenotypes could be studied in the same manner. At the moment, I am also working on a project involving expression markers for suicide, utilizing this approach, and quite excited to see the result. However, it is important to mention that such comparative studies and meta-analyses are only possible if the research data is openly shared, and thus open-science paradigms, such as FAIR Data Principles [252] should be facilitated.

However, what if we in fact study a null field? Then, again, first, meta-analyses would fail to identify consistencies. Second, we should include other factors (alongside biology) when we design our studies to eliminate/reduce confounding. For example, we may consider asking about participant's experience in recent time, family situation, relationship status, performance at work, and other factors. I am not a general supporter of qualitative research, but I think some elements of it alongside quantitative core (such as statistical modeling) would be beneficial for depression and psychiatry in general.

7. Acknowledgements

The PhD is a long and significant journey and I would like to thank a list of people that directly or indirectly contributed to it, supported me, and helped me during this time. The list includes, but not limited to:

Supervisor Prof. Helgi Schiöth

Thank you for providing this opportunity to do a PhD, involving me in many interesting projects, and giving many practical advices regarding research, scientific writing, and publishing papers. I have learned a lot during these four years!

Co-supervisor Prof. Jussi Jokinen

Thank you for guidance regarding the SKI cohort, suicide phenotypes and helping with my first paper!

Co-supervisor Dr. Gull Rukh

Thank you for guidance in several project that we did with UK biobank and many practical advices!

My family

Thank you for the continuous invaluable support and believing in me! I miss you and I wish I was able to see you more often in person . . .

Amelia Juslin

Thank you for being an amazing girlfriend and friend, being adorable, supportive and caring, and all nice time we spent together! Hope it will continue this way :D

Oreste Affatato

Thank you for numerous statistical discussions, being an amazing friend, movie night host, tiramisu maker :D, and helping a lot during my PhD.

Renan Maciel

Thank you for being an amazing friend, sauna and coffee discussions, as well as for numerous fun activities we had together!

Markus de Ruijter

Thank you for being an amazing friend, "Ccccc-coffee-time-time-time" as well as cool parties and activities we had!

Daria Belotcerkovtceva

Thank you for being an amazing friend, sauna discussions and numerous party stuff!

Muataz Lafta

Thank you for being an amazing friend and many projects we worked on together!

Salahuddin Mohammad
Thank you for being an amazing friend, many funny and effortless coffee breaks and discussions and wonderful BBQs!
Misty Attwood, Diana-Maria Manu, Jörgen Jonsson
Thank you for scientific discussions, giving many practical tips and research thoughts!
Pei, Elisa, Manon, Felix, Maud, Megha, Melissa, Felippe, Lieve, Satti
Thank you for being amazing colleagues and nice time we had during fikas, parties, and other moments!

References

[1] U.S. National Institutes of Health (NIH), National Institute of Mental Health https://www.nimh.nih.gov/health/statistics/mental-illness [2022].

[2] H. U. Wittchen, F. Jacobi, J. Rehm, et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. *Eur Neuropsychopharmacol*, 21(9):655–679, September 2011. Number: 9.

[3] Luis Gutiérrez-Rojas, Alejandro Porras-Segovia, Henry Dunne, Nelson Andrade-González, and Jorge A. Cervilla. Prevalence and correlates of major depressive disorder: a systematic review. *Braz J Psychiatry*, 42(6):657–672, December 2020. Number: 6.

[4] Christian Otte, Stefan M. Gold, Brenda W. Penninx, et al. Major depressive disorder. *Nat Rev Dis Primers*, 2(1):16065, December 2016. Number: 1.

[5] Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Stephan Ripke, Naomi R. Wray, et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry*, 18(4):497–511, April 2013. Number: 4.

[6] Signe Penner-Goeke and Elisabeth B. Binder. Epigenetics and depression. *Dialogues Clin Neurosci*, 21(4):397–405, December 2019. Number: 4.

[7] Nicole Mariani, Nadia Cattane, Carmine Pariante, and Annamaria Cattaneo. Gene expression studies in Depression development and treatment: an overview of the underlying molecular mechanisms and biological processes to identify biomarkers. *Transl Psychiatry*, 11(1):354, June 2021. Number: 1.

[8] Arjan W. Braam and Harold G. Koenig. Religion, spirituality and depression in prospective studies: A systematic review. *Journal of Affective Disorders*, 257:428–438, October 2019.

[9] Yiru Fang and Zhiguo Wu. Advance in Diagnosis of Depressive Disorder. *Adv Exp Med Biol*, 1180:179–191, 2019.

[10] R. Goodman, T. Ford, H. Richards, R. Gatward, and H. Meltzer. The Development and Well-Being Assessment: description and initial validation of an integrated assessment of child and adolescent psychopathology. *J Child Psychol Psychiatry*, 41(5):645–655, July 2000. Number: 5.

[11] Anna Goodman, Einar Heiervang, Stephan Collishaw, and Robert Goodman. The 'DAWBA bands' as an ordered-categorical measure of child mental health: description and validation in British and Norwegian samples. *Soc Psychiatry Psychiatr Epidemiol*, 46(6):521–532, June 2011. Number: 6.

[12] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. An inventory for measuring depression. *Arch Gen Psychiatry*, 4:561–571, June 1961.

[13] Aaron T. Beck, Robert A. Steer, and Margery G. Carbin. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8(1):77–100, January 1988. Number: 1.

[14] S. A. Montgomery and M. Asberg. A new depression scale designed to be sensitive to change. *Br J Psychiatry*, 134:382–389, April 1979.

[15] P. Svanborg and M. Asberg. A new self-rating scale for depression and anxiety states based on the Comprehensive Psychopathological Rating Scale. *Acta Psychiatr Scand*, 89(1):21–28, January 1994. Number: 1.
[16] M. Hamilton. A rating scale for depression. *J Neurol Neurosurg Psychiatry*, 23:56–62, February 1960.
[17] P. M. Lewinsohn, J. R. Seeley, R. E. Roberts, and N. B. Allen. Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychol Aging*, 12(2):277–287, June 1997.
[18] K. Kroenke, R. L. Spitzer, and J. B. Williams. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*, 16(9):606–613, September 2001.
[19] Shuyan Sun and Shanshan Wang. The Children's Depression Inventory in Worldwide Child Development Research: A Reliability Generalization Study. *J Child Fam Stud*, 24(8):2352–2363, August 2015.
[20] E. O. Poznanski, S. C. Cook, and B. J. Carroll. A depression rating scale for children. *Pediatrics*, 64(4):442–450, October 1979.
[21] Gustavo Turecki, David A. Brent, David Gunnell, et al. Suicide and suicide risk. *Nat Rev Dis Primers*, 5(1):74, October 2019. Number: 1.
[22] World Health Organization, Noncommunicable diseases and mental health https://apps.who.int/gho/data/view.sdg.3-4-data-ctry?lang=en [2022].
[23] Mary A. Whooley and Jonathan M. Wong. Depression and cardiovascular disorders. *Annu Rev Clin Psychol*, 9:327–354, 2013.
[24] Edward Chesney, Guy M. Goodwin, and Seena Fazel. Risks of all-cause and suicide mortality in mental disorders: a meta-review. *World Psychiatry*, 13(2):153–160, June 2014. Number: 2.
[25] Matthew K. Nock, Guilherme Borges, Evelyn J. Bromet, et al. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *Br J Psychiatry*, 192(2):98–105, February 2008. Number: 2.
[26] Joyce P. Chu, Peter Goldblum, Rebecca Floyd, and Bruce Bongar. The cultural theory and model of suicide. *Applied and Preventive Psychology*, 14(1-4):25–40, June 2010.
[27] Xiaozhi Li, Guijun Chi, Alyx Taylor, et al. Lifestyle Behaviors and Suicide-Related Behaviors in Adolescents: Cross-Sectional Study Using the 2019 YRBS Data. *Front. Public Health*, 9:766972, November 2021.
[28] Ghanshyam N. Pandey. Biological basis of suicide and suicidal behavior. *Bipolar Disord*, 15(5):524–541, August 2013.
[29] Robyn Thom, Charlotte Hogan, and Eric Hazen. Suicide Risk Screening in the Hospital Setting: A Review of Brief Validated Tools. *Psychosomatics*, 61(1):1–7, February 2020. Number: 1.
[30] Bjørn Odd Koldsland, Lars Mehlum, Liv Solrunn Mellesdal, Fredrik A. Walby, and Lien M. Diep. 3726067. *BMC Res Notes*, 5:417, August 2012.
[31] B. Stanley, L. Träskman-Bendz, and M. Stanley. The suicide assessment scale: a scale evaluating change in suicidal behavior. *Psychopharmacol Bull*, 22(1):200–205, 1986. Number: 1.
[32] In-Chul Baek, Soobin Jo, Eun Ji Kim, et al. A Review of Suicide Risk Assessment Tools and Their Measured Psychometric Properties in Korea.

Front. Psychiatry, 12:679779, June 2021.
[33] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, et al. Genome-wide association studies. *Nat Rev Methods Primers*, 1(1):59, December 2021. Number: 1.
[34] David F Burke, Catherine L Worth, Eva-Maria Priego, et al. Genome bioinformatic analysis of nonsynonymous SNPs. *BMC Bioinformatics*, 8(1):301, December 2007.
[35] Ryan Hunt, Zuben E. Sauna, Suresh V. Ambudkar, Michael M. Gottesman, and Chava Kimchi-Sarfaty. Silent (Synonymous) SNPs: Should We Care About Them? In Anton A. Komar, editor, *Single Nucleotide Polymorphisms*, volume 578, pages 23–39. Humana Press, Totowa, NJ, 2009. Series Title: Methods in Molecular Biology.
[36] Alexandra C. Nica and Emmanouil T. Dermitzakis. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci*, 368(1620):20120362, 2013.
[37] Alicia K Smith, Varun Kilaru, Mehmet Kocak, et al. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*, 15(1):145, 2014.
[38] Naomi R. Wray, Stephan Ripke, Manuel Mattheisen, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet*, 50(5):668–681, May 2018.
[39] David M. Howard, Mark J. Adams, Toni-Kim Clarke, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci*, 22(3):343–352, March 2019. Number: 3.
[40] Steven R. Wainwright and Liisa A. M. Galea. The neural plasticity theory of depression: assessing the roles of adult neurogenesis and PSA-NCAM within the hippocampus. *Neural Plast*, 2013:805497, 2013.
[41] Joanna Moncrieff, Ruth E. Cooper, Tom Stockmann, et al. The serotonin theory of depression: a systematic umbrella review of the evidence. *Mol Psychiatry*, 28(8):3243–3256, August 2023.
[42] David M. Howard, Mark J. Adams, Masoud Shirali, et al. Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat Commun*, 9(1):1470, April 2018. Number: 1.
[43] Daniel F. Levey, Murray B. Stein, Frank R. Wendt, et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat Neurosci*, 24(7):954–963, July 2021. Number: 7.
[44] Mats Nagel, Philip R. Jansen, Sven Stringer, et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat Genet*, 50(7):920–927, July 2018. Number: 7.
[45] Yulu Wu, Hongbao Cao, Ancha Baranova, et al. Multi-trait analysis for genome-wide association study of five psychiatric disorders. *Transl Psychiatry*, 10(1):209, June 2020. Number: 1.
[46] S. E. Bergen, C. T. O'Dushlaine, S. Ripke, et al. Genome-wide association

study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry*, 17(9):880–886, September 2012. Number: 9.

[47] Stephan Ripke, Colm O'Dushlaine, Kimberly Chambert, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet*, 45(10):1150–1159, October 2013. Number: 10.

[48] Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet*, 43(10):969–976, September 2011. Number: 10.

[49] Li Su, Tingting Shen, Guifeng Huang, et al. Genetic association of GWAS-supported MAD1L1 gene polymorphism rs12666575 with schizophrenia susceptibility in a Chinese population. *Neurosci Lett*, 610:98–103, January 2016.

[50] Zhiqiang Li, Jianhua Chen, Hao Yu, et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat Genet*, 49(11):1576–1583, November 2017. Number: 11.

[51] Antonio F. Pardiñas, Peter Holmans, Andrew J. Pocklington, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet*, 50(3):381–389, March 2018. Number: 3.

[52] Douglas M. Ruderfer, Ayman H. Fanous, Stephan Ripke, et al. Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol Psychiatry*, 19(9):1017–1024, September 2014. Number: 9.

[53] Patrick Sleiman, Dai Wang, Joseph Glessner, et al. GWAS meta analysis identifies TSNARE1 as a novel Schizophrenia / Bipolar susceptibility locus. *Sci Rep*, 3:3075, October 2013.

[54] M. Ikeda, A. Takahashi, Y. Kamatani, et al. A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder. *Mol Psychiatry*, 23(3):639–647, March 2018. Number: 3.

[55] Liping Hou, Sarah E. Bergen, Nirmala Akula, et al. Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Hum Mol Genet*, 25(15):3383–3394, August 2016. Number: 15.

[56] Eli A. Stahl, Gerome Breen, Andreas J. Forstner, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat Genet*, 51(5):793–803, May 2019. Number: 5.

[57] Tanya Horwitz, Katie Lam, Yu Chen, Yan Xia, and Chunyu Liu. A decade in psychiatric GWAS research. *Mol Psychiatry*, 24(3):378–389, March 2019. Number: 3.

[58] Annalisa Buniello, Jacqueline A. L. MacArthur, Maria Cerezo, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*, 47(D1):D1005–D1012, January 2019. Number: D1.

[59] Qingqin S. Li, Andrey A. Shabalin, Emily DiBlasi, et al. Genome-wide association study meta-analysis of suicide death and suicidal behavior. *Mol Psychiatry*, 28(2):891–900, February 2023.

[60] M.-F. Lisé and A. El-Husseini. The neuroligin and neurexin families: from structure to function at the synapse. *Cell Mol Life Sci*, 63(16):1833–1849, August 2006.
[61] Albert H. C. Wong, Irving I. Gottesman, and Arturas Petronis. Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Hum Mol Genet*, 14 Spec No 1:R11–18, April 2005.
[62] Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*, 33(S3):245–254, March 2003.
[63] Ryan Lister, Mattia Pelizzola, Robert H. Dowen, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, November 2009. Number: 7271.
[64] Maxim V. C. Greenberg and Deborah Bourc'his. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol*, 20(10):590–607, October 2019.
[65] Bernard H. Ramsahoye, Detlev Biniszkiewicz, Frank Lyko, et al. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U.S.A.*, 97(10):5237–5242, May 2000.
[66] Deepa Ramasamy, Arunagiri Kuha Deva Magendhra Rao, Thangarajan Rajkumar, and Samson Mani. Non-CpG methylation-a key epigenetic modification in cancer. *Brief Funct Genomics*, 20(5):304–311, September 2021.
[67] Junjie U Guo, Yijing Su, Joo Heon Shin, et al. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci*, 17(2):215–222, February 2014.
[68] Peter A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, 13(7):484–492, May 2012. Number: 7.
[69] Fabienne Brenet, Michelle Moh, Patricia Funk, et al. DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One*, 6(1):e14524, January 2011. Number: 1.
[70] Lisa D. Moore, Thuc Le, and Guoping Fan. DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, January 2013. Number: 1.
[71] Allegra Angeloni and Ozren Bogdanovic. Enhancer DNA methylation: implications for gene regulation. *Essays Biochem*, 63(6):707–715, December 2019. Number: 6.
[72] Junjie U. Guo, Dengke K. Ma, Huan Mo, et al. Neuronal activity modifies the DNA methylation landscape in the adult brain. *Nat Neurosci*, 14(10):1345–1351, August 2011. Number: 10.
[73] Thabo Magwai, Khanyiso Bright Shangase, Fredrick Otieno Oginga, et al. DNA Methylation and Schizophrenia: Current Literature and Future Perspective. *Cells*, 10(11):2890, October 2021. Number: 11.
[74] Gabriel R. Fries, Qiongzhen Li, Blake McAlpin, et al. The role of DNA methylation in the pathophysiology and treatment of bipolar disorder. *Neurosci Biobehav Rev*, 68:474–488, September 2016.
[75] Ana L. Romero-Pimentel, Daniel Almeida, Said Muñoz-Montero, et al. Integrative DNA Methylation and Gene Expression Analysis in the Prefrontal

Cortex of Mexicans Who Died by Suicide. *Int J Neuropsychopharmacol*, 24(12):935–947, December 2021. Number: 12.

[76] Diana M. Ciuculete, Sarah Voisin, Lara Kular, et al. meQTL and ncRNA functional analyses of 102 GWAS-SNPs associated with depression implicate HACE1 and SHANK2 genes. *Clin Epigenetics*, 12(1):99, July 2020. Number: 1.

[77] Diana M. Ciuculete, Adrian E. Boström, Anna-Kaisa Tuunainen, et al. Changes in methylation within the STK32B promoter are associated with an increased risk for generalized anxiety disorder in adolescents. *J Psychiatr Res*, 102:44–51, July 2018.

[78] Diana M. Ciuculete, Sarah Voisin, Lara Kular, et al. Longitudinal DNA methylation changes at MET may alter HGF/c-MET signalling in adolescents at risk for depression. *Epigenetics*, 15(6-7):646–663, July 2020. Number: 6-7.

[79] Jussi Jokinen, Adrian E. Boström, Ali Dadfar, et al. Epigenetic Changes in the CRH Gene are Related to Severity of Suicide Attempt and a General Psychiatric Risk Score in Adolescents. *EBioMedicine*, 27:123–133, January 2018.

[80] Weijing Wang, Weilong Li, Yili Wu, et al. Genome-wide DNA methylation and gene expression analyses in monozygotic twins identify potential biomarkers of depression. *Transl Psychiatry*, 11(1):416, August 2021. Number: 1.

[81] Qingqin S. Li, Randall L. Morrison, Gustavo Turecki, and Wayne C. Drevets. Meta-analysis of epigenome-wide association studies of major depressive disorder. *Sci Rep*, 12(1):18361, November 2022. Number: 1.

[82] Yukako Nakamura, Masahiro Nakatochi, Shohko Kunimoto, et al. Methylation analysis for postpartum depression: a case control study. *BMC Psychiatry*, 19(1):190, June 2019. Number: 1.

[83] Anthony S. Zannas, Meiwen Jia, Kathrin Hafner, et al. Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-κB-driven inflammation and cardiovascular risk. *Proc Natl Acad Sci U S A*, 116(23):11370–11379, June 2019. Number: 23.

[84] Bethany Crawford, Zoe Craig, Georgina Mansell, et al. DNA methylation and inflammation marker profiles associated with a history of depression. *Hum Mol Genet*, 27(16):2840–2850, August 2018. Number: 16.

[85] Hooriyah S. Rizavi, Omar S. Khan, Hui Zhang, et al. Methylation and expression of glucocorticoid receptor exon-1 variants and FKBP5 in teenage suicide-completers. *Transl Psychiatry*, 13(1):53, February 2023.

[86] Serina Cheung, Julia Woo, Miriam S. Maes, and Clement C. Zai. Suicide epigenetics, a review of recent progress. *J Affect Disord*, 265:423–438, March 2020.

[87] Stefania Policicchio, Sam Washer, Joana Viana, et al. Genome-wide DNA methylation meta-analysis in the brains of suicide completers. *Transl Psychiatry*, 10(1):69, February 2020. Number: 1.

[88] Bhaskar Roy, Richard C. Shelton, and Yogesh Dwivedi. DNA methylation and expression of stress related genes in PBMC of MDD patients with and without serious suicidal ideation. *J Psychiatr Res*, 89:115–124, June 2017.

[89] Miruna C. Barbu, Xueyi Shen, Rosie M. Walker, et al. Epigenetic prediction of

major depressive disorder. *Mol Psychiatry*, 26(9):5112–5123, September 2021.
[90] Shaunna L. Clark, Mohammad W. Hattab, Robin F. Chan, et al. A methylation study of long-term depression risk. *Mol Psychiatry*, 25(6):1334–1343, June 2020.
[91] P.-F. Kuan, M. A. Waszczuk, R. Kotov, et al. An epigenome-wide DNA methylation study of PTSD and depression in World Trade Center responders. *Transl Psychiatry*, 7(6):e1158, June 2017. Number: 6.
[92] Esther Walton, Johanna Hass, Jingyu Liu, et al. Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. *Schizophr Bull*, 42(2):406–414, March 2016. Number: 2.
[93] Matthew N. Davies, Manuela Volta, Ruth Pidsley, et al. Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol*, 13(6):R43, June 2012. Number: 6.
[94] Eilis Hannon, Katie Lunnon, Leonard Schalkwyk, and Jonathan Mill. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics*, 10(11):1024–1032, 2015. Number: 11.
[95] David Colquhoun. The reproducibility of research and the misinterpretation of p-values. *R Soc Open Sci*, 4(12):171085, December 2017. Number: 12.
[96] Vania Januar, Richard Saffery, and Joanne Ryan. Epigenetics and depressive disorders: a review of current progress and future directions. *Int J Epidemiol*, 44(4):1364–1387, August 2015. Number: 4.
[97] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci*, 22(11):1359–1366, November 2011.
[98] Annamaria Cattaneo, Clarissa Ferrari, Lorinda Turner, et al. Whole-blood expression of inflammasome- and glucocorticoid-related mRNAs correctly separates treatment-resistant depressed patients from drug-free and responsive patients in the BIODEP study. *Transl Psychiatry*, 10(1):232, July 2020.
[99] Annamaria Cattaneo, Massimo Gennarelli, Rudolf Uher, et al. Candidate genes expression profile associated with antidepressants response in the GENDEP study: differentiating between baseline 'predictors' and longitudinal 'targets'. *Neuropsychopharmacology*, 38(3):377–385, February 2013. Number: 3.
[100] Yasuto Kunii, Wenyu Zhang, Qing Xu, et al. CHRNA7 and CHRFAM7A mRNAs: co-localized and their expression levels altered in the postmortem dorsolateral prefrontal cortex in major psychiatric disorders. *Am J Psychiatry*, 172(11):1122–1130, November 2015.
[101] Joanna Mossakowska-Wójcik, Agata Orzechowska, Monika Talarowska, Janusz Szemraj, and Piotr Gałecki. The importance of TCF4 gene in the etiology of recurrent depressive disorders. *Prog Neuropsychopharmacol Biol Psychiatry*, 80(Pt C):304–308, January 2018.
[102] Kinga Bobińska, Joanna Mossakowska-Wójcik, Janusz Szemraj, et al. Human neuropsin gene in depression. *Psychiatr Danub*, 29(2):195–200, June 2017.
[103] Rhian Lauren Preece, Sung Yeon Sarah Han, and Sabine Bahn. Proteomic approaches to identify blood-based biomarkers for depression and bipolar

disorders. *Expert Rev Proteomics*, 15(4):325–340, April 2018.
[104] Thomas S. Wingo, Yue Liu, Ekaterina S. Gerasimov, et al. Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nat Neurosci*, 24(6):810–817, June 2021.
[105] Yue-Ting Deng, Ya-Nan Ou, Bang-Sheng Wu, et al. Identifying causal genes for depression via integration of the proteome and transcriptome from brain and blood. *Mol Psychiatry*, 27(6):2849–2857, June 2022.
[106] Felix Krueger, Benjamin Kreck, Andre Franke, and Simon R. Andrews. DNA methylome analysis using short bisulfite sequencing data. *Nat Methods*, 9(2):145–151, January 2012. Number: 2.
[107] Beate M. Crossley, Jianfa Bai, Amy Glaser, et al. Guidelines for Sanger sequencing and molecular assay monitoring. *J Vet Diagn Invest*, 32(6):767–775, November 2020.
[108] Inês Lopes, Gulam Altab, Priyanka Raina, and João Pedro de Magalhães. Gene Size Matters: An Analysis of Gene Length in the Human Genome. *Front Genet*, 12:559998, 2021.
[109] Michael L. Metzker. Sequencing technologies — the next generation. *Nat Rev Genet*, 11(1):31–46, January 2010.
[110] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6):333–351, June 2016.
[111] Illumina.com; HumanMethylation27 Manual https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_dna_methylation_analysis_infinium.pdf [2023].
[112] Simone Bork, Stefan Pfister, Hendrik Witt, et al. DNA methylation pattern changes upon long-term culture and aging of human mesenchymal stromal cells. *Aging Cell*, 9(1):54–63, February 2010.
[113] Marina Bibikova, Bret Barnes, Chan Tsan, et al. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295, October 2011. Number: 4.
[114] Sebastian Moran, Carles Arribas, and Manel Esteller. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3):389–399, March 2016. Number: 3.
[115] Illumina.com; HumanMethylation450 Data Sheet https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_humanmethylation450.pdf [2023].
[116] Illumina.com; HumanMethylationEPIC Data Sheet https://emea.support.illumina.com/content/dam/illumina-support/documents/downloads/productfiles/methylationepic/infinium-methylation-epic-ds-1070-2015-008.pdf [2023].
[117] Andrew E. Teschendorff, Francesco Marabita, Matthias Lechner, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, 29(2):189–196, January 2013. Number: 2.
[118] Pan Du, Xiao Zhang, Chiang-Ching Huang, et al. Comparison of Beta-value

and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11:587, November 2010.
[119] Meaghan J. Jones, Sumaiya A. Islam, Rachel D. Edgar, and Michael S. Kobor. Adjusting for Cell Type Composition in DNA Methylation Data Using a Regression-Based Approach. *Methods Mol Biol*, 1589:99–106, 2017.
[120] J. C. Alwine, D. J. Kemp, and G. R. Stark. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A*, 74(12):5350–5354, December 1977.
[121] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, October 1995.
[122] Mark D. Robinson and Terence P. Speed. A comparison of Affymetrix gene expression arrays. *BMC Bioinformatics*, 8:449, November 2007.
[123] Illumina, Inc, Illumina HumanHT-12 information sheet https://www.illumina.com/Documents/products/datasheets/datasheet_humanht_12.pdf [2022].
[124] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*, 41(Database issue):D991–995, January 2013. Number: Database issue.
[125] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, January 2009.
[126] Michał J. Okoniewski and Crispin J. Miller. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7:276, June 2006.
[127] Miten Jain, Robin Abu-Shumays, Hugh E. Olsen, and Mark Akeson. Advances in nanopore direct RNA sequencing. *Nat Methods*, 19(10):1160–1164, October 2022.
[128] P. H. O'Farrell. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*, 250(10):4007–4021, May 1975.
[129] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem*, 60(20):2299–2301, October 1988.
[130] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, March 2003.
[131] Amelia C. Peterson, Jason D. Russell, Derek J. Bailey, Michael S. Westphall, and Joshua J. Coon. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol Cell Proteomics*, 11(11):1475–1488, November 2012.
[132] E. Engvall and P. Perlmann. Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. *Immunochemistry*, 8(9):871–874, September 1971.
[133] H. Towbin, T. Staehelin, and J. Gordon. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc Natl Acad Sci U S A*, 76(9):4350–4354, September 1979.
[134] Larry Gold, Jeffrey J. Walker, Sheri K. Wilcox, and Stephen Williams. Advances in human proteomics at high scale with the SOMAscan proteomics

platform. *N Biotechnol*, 29(5):543–549, June 2012.
[135] Martin Lundberg, Anna Eriksson, Bonnie Tran, Erika Assarsson, and Simon Fredriksson. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids Res*, 39(15):e102, August 2011.
[136] Olink.com; Whitepaper: Data normalization and standardization https://www.olink.com/content/uploads/2022/04/white-paper-data-normalization-v2.1.pdf [2023].
[137] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, April 2003. Number: 2.
[138] Ying Yu, Naixin Zhang, Yuanbang Mai, et al. Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. *Genome Biol*, 24(1):201, September 2023.
[139] Wilson Wen Bin Goh, Wei Wang, and Limsoon Wong. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol*, 35(6):498–507, June 2017.
[140] Sorin Draghici. *Statistics and Data Analysis for Microarrays Using R and Bioconductor*. Chapman and Hall/CRC, 0 edition, April 2016.
[141] Lyman Ott and Michael Longnecker. *An introduction to statistical methods & data analysis*. Cengage Learning, Australia, seventh edition edition, 2016. OCLC: ocn898088669.
[142] Jacob Cohen. The earth is round ($p < .05$). *American Psychologist*, 49(12):997–1003, December 1994.
[143] Jeffrey A. Gliner, George A. Morgan, Nancy L. Leech, and Robert J. Harmon. Problems With Null Hypothesis Significance Testing. *Journal of the American Academy of Child & Adolescent Psychiatry*, 40(2):250–252, February 2001.
[144] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*, 31(4):337–350, April 2016.
[145] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. Moving to a World Beyond "$p < 0.05$". *The American Statistician*, 73(sup1):1–19, March 2019.
[146] Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.
[147] Rufeng Li, Lixin Li, Yungang Xu, and Juan Yang. Machine learning meets omics: applications and perspectives. *Brief Bioinform*, 23(1):bbab460, January 2022.
[148] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine Learning in Python. 2012. Publisher: arXiv Version Number: 4.
[149] Chiyoung Lee and Heewon Kim. Machine learning-based predictive modeling of depression in hypertensive populations. *PLoS One*, 17(7):e0272330, 2022.
[150] Shaowu Lin, Yafei Wu, and Ya Fang. A hybrid machine learning model of depression estimation in home-based older adults: a 7-year follow-up study. *BMC Psychiatry*, 22(1):816, December 2022.
[151] Jovana Cvetković. Breast Cancer Patients' Depression Prediction by Machine

Learning Approach. *Cancer Invest*, 35(8):569–572, September 2017.
[152] Md. Sabab Zulfiker, Nasrin Kabir, Al Amin Biswas, Tahmina Nazneen, and Mohammad Shorif Uddin. An in-depth analysis of machine learning approaches to predict depression. *Current Research in Behavioral Sciences*, 2:100044, November 2021.
[153] Sang Won Kim and Min Cheol Chang. The usefulness of machine learning analysis for predicting the presence of depression with the results of the Korea National Health and Nutrition Examination Survey. *Ann Palliat Med*, 12(4):748–756, July 2023.
[154] Yu Jin, Shicun Xu, Zhixian Shao, et al. Discovery of depression-associated factors among childhood trauma victims from a large sample size: Using machine learning and network analysis. *J Affect Disord*, 345:300–310, January 2024.
[155] Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, et al. Depression detection from social network data using machine learning techniques. *Health Inf Sci Syst*, 6(1):8, December 2018.
[156] Zerui You, Chengyu Wang, Xiaofeng Lan, et al. The contribution of polyamine pathway to determinations of diagnosis for treatment-resistant depression: A metabolomic analysis. *Prog Neuropsychopharmacol Biol Psychiatry*, 128:110849, January 2024.
[157] Yunsong Luo, Wenyu Chen, Ling Zhan, Jiang Qiu, and Tao Jia. Multi-feature concatenation and multi-classifier stacking: An interpretable and generalizable machine learning method for MDD discrimination with rsfMRI. *Neuroimage*, page 120497, December 2023.
[158] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.
[159] Yann LeCun, Bernhard Boser, John Denker, et al. Handwritten Digit Recognition with a Back-Propagation Network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.
[160] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997.
[161] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans Neural Netw*, 20(1):61–80, January 2009.
[162] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention Is All You Need. 2017. Publisher: arXiv Version Number: 7.
[163] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, 2016.
[164] Ge Zhang, Zijing Xue, Chaokun Yan, Jianlin Wang, and Huimin Luo. A Novel Biomarker Identification Approach for Gastric Cancer Using Gene Expression and DNA Methylation Dataset. *Front Genet*, 12:644378, 2021.
[165] Chunlei Zheng and Rong Xu. Predicting cancer origins with a DNA methylation-based deep neural network model. *PLoS One*, 15(5):e0226461, 2020. Number: 5.
[166] Michela Carlotta Massi, Lorenzo Dominoni, Francesca Ieva, and Giovanni

Fiorito. A Deep Survival EWAS approach estimating risk profile based on pre-diagnostic DNA methylation: An application to breast cancer time to diagnosis. *PLoS Comput Biol*, 18(9):e1009959, September 2022.
[167] Samir Jabari, Katja Kobow, Tom Pieper, et al. DNA methylation-based classification of malformations of cortical development in the human brain. *Acta Neuropathol*, 143(1):93–104, January 2022.
[168] Md Mohaiminul Islam, Ye Tian, Yan Cheng, Yang Wang, and Pingzhao Hu. A deep neural network based regression model for triglyceride concentrations prediction using epigenome-wide DNA methylation profiles. *BMC Proc*, 12(Suppl 9):21, 2018.
[169] Xuetong Zhao, Yang Sui, Xiuyan Ruan, et al. A deep learning model for early risk prediction of heart failure with preserved ejection fraction by DNA methylation profiles combined with clinical features. *Clin Epigenetics*, 14(1):11, January 2022.
[170] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.*, 37(2):233–243, February 1991.
[171] Dibyendu Bikash Seal, Vivek Das, Saptarsi Goswami, and Rajat K. De. Estimating gene expression from DNA methylation and copy number variation: A deep learning regression model for multi-omics integration. *Genomics*, 112(4):2833–2841, July 2020.
[172] Laura Macías-García, María Martínez-Ballesteros, José María Luna-Romera, et al. Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance. *Artif Intell Med*, 110:101976, November 2020.
[173] Tzong-Yi Lee, Kai-Yao Huang, Cheng-Hsiang Chuang, Cheng-Yang Lee, and Tzu-Hao Chang. Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. *Comput Biol Chem*, 87:107277, May 2020.
[174] Li Chen, Andrew J. Saykin, Bing Yao, Fengdi Zhao, and Alzheimer's Disease Neuroimaging Initiative (ADNI). Multi-task deep autoencoder to predict Alzheimer's disease progression using temporal DNA methylation data in peripheral blood. *Comput Struct Biotechnol J*, 20:5761–5774, 2022.
[175] Somayah Albaradei, Francesco Napolitano, Maha A. Thafar, et al. MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. *Comput Struct Biotechnol J*, 19:4404–4411, 2021.
[176] Zhenxing Wang and Yadong Wang. Extracting a biologically latent space of lung cancer epigenetics with variational autoencoders. *BMC Bioinformatics*, 20(Suppl 18):568, November 2019.
[177] Joshua J. Levy, Alexander J. Titus, Curtis L. Petersen, et al. MethylNet: an automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinformatics*, 21(1):108, March 2020.
[178] Li Tong, Jonathan Mitchel, Kevin Chatlin, and May D. Wang. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Med Inform Decis Mak*, 20(1):225, September 2020.
[179] Tongjun Gu and Xiwu Zhao. Integrating multi-platform genomic datasets for kidney renal clear cell carcinoma subtyping using stacked denoising

autoencoders. *Sci Rep*, 9(1):16668, November 2019.
[180] L. Träskman, M. Asberg, L. Bertilsson, and L. Sjöstrand. Monoamine metabolites in CSF and suicidal behavior. *Arch Gen Psychiatry*, 38(6):631–636, June 1981. Number: 6.
[181] M. Asberg, L. Träskman, and P. Thorén. 5-HIAA in the cerebrospinal fluid. A biochemical suicide predictor? *Arch Gen Psychiatry*, 33(10):1193–1197, October 1976. Number: 10.
[182] D. J. Freeman, K. Wilson, J. Thigpen, and R. K. McGee. Assessing intention to die in self-injury behavior. In *Psychological Assessment of Suicidal Risk*, pages 18–42. 1974.
[183] Cathie Sudlow, John Gallacher, Naomi Allen, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*, 12(3):e1001779, March 2015. Number: 3.
[184] Daniel J. Smith, Barbara I. Nicholl, Breda Cullen, et al. Prevalence and characteristics of probable major depression and bipolar disorder within UK biobank: cross-sectional study of 172,751 participants. *PLoS One*, 8(11):e75362, 2013. Number: 11.
[185] Kylie P. Glanville, Jonathan R. I. Coleman, David M. Howard, et al. Multiple measures of depression to enhance validity of major depressive disorder in the UK Biobank. *BJPsych Open*, 7(2):e44, February 2021. Number: 2.
[186] Jerry Guintivano, Martin J. Aryee, and Zachary A. Kaminsky. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*, 8(3):290–302, March 2013. Number: 3.
[187] T. M. Murphy, B. Crawford, E. L. Dempster, et al. Methylomic profiling of cortex samples from completed suicide cases implicates a role for PSORS1C3 in major depression and suicide. *Transl Psychiatry*, 7(1):e989, January 2017. Number: 1.
[188] Elisabeth B. Binder, Rebekah G. Bradley, Wei Liu, et al. Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults. *JAMA*, 299(11):1291–1305, March 2008. Number: 11.
[189] Charles F. Gillespie, Bekh Bradley, Kristie Mercer, et al. Trauma exposure and stress-related disorders in inner city primary care patients. *Gen Hosp Psychiatry*, 31(6):505–514, December 2009. Number: 6.
[190] Paul E. Holtzheimer, Jason Veitengruber, Chia C. Wang, et al. Utility of the Beck Depression Inventory to screen for and track depression in injection drug users seeking hepatitis C treatment. *Gen Hosp Psychiatry*, 32(4):426–432, August 2010. Number: 4.
[191] L. K. Kerr and L. D. Kerr. Screening tools for depression in primary care: the effects of culture, gender, and somatic symptoms on the detection of depression. *West J Med*, 175(5):349–352, November 2001. Number: 5.
[192] A. Heck, R. Lieb, A. Ellgas, et al. Investigation of 17 candidate genes for personality traits confirms effects of the HTR2A gene on novelty seeking. *Genes Brain Behav*, 8(4):464–472, June 2009.
[193] Martin A. Kohli, Susanne Lucae, Philipp G. Saemann, et al. The neuronal transporter gene SLC6A15 confers risk to major depression. *Neuron*,

70(2):252–265, April 2011.
[194] Susanne Lucae, Daria Salyakina, Nicholas Barden, et al. P2RX7, a gene coding for a purinergic ligand-gated ion channel, is associated with major depressive disorder. *Hum Mol Genet*, 15(16):2438–2445, August 2006.
[195] Yu Sun, Wayne Drevets, Gustavo Turecki, and Qingqin S. Li. The relationship between plasma serotonin and kynurenine pathway metabolite levels and the treatment response to escitalopram and desvenlafaxine. *Brain Behav Immun*, 87:404–412, July 2020.
[196] Chelsey Ju, Laura M. Fiori, Raoul Belzeaux, et al. Integrated genome-wide methylation and expression analyses reveal functional predictors of response to antidepressants. *Transl Psychiatry*, 9(1):254, October 2019.
[197] Wilma T. Steegenga, Mark V. Boekschoten, Carolien Lute, et al. Genome-wide age-related changes in DNA methylation and gene expression in human PBMCs. *Age (Dordr)*, 36(3):9648, June 2014. Number: 3.
[198] Lindsay M. Reynolds, Jackson R. Taylor, Jingzhong Ding, et al. Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nat Commun*, 5:5366, November 2014.
[199] Thomas A. Lanz, Veronica Reinhart, Mark J. Sheehan, et al. Postmortem transcriptional profiling reveals widespread increase in inflammation in schizophrenia: a comparison of prefrontal cortex, striatum, and hippocampus among matched tetrads of controls with subjects diagnosed with schizophrenia, bipolar or major depressive disorder. *Transl Psychiatry*, 9(1):151, May 2019. Number: 1.
[200] Gwenaël G. R. Leday, Petra E. Vértes, Sylvia Richardson, et al. Replicable and Coupled Changes in Innate and Adaptive Immune Gene Expression in Two Case-Control Studies of Blood Microarrays in Major Depressive Disorder. *Biol Psychiatry*, 83(1):70–80, January 2018. Number: 1.
[201] Janine Arloth, Ryan Bogdan, Peter Weber, et al. Genetic Differences in the Immediate Transcriptome Response to Stress Predict Risk-Related Brain Function and Psychiatric Disorders. *Neuron*, 86(5):1189–1202, June 2015. Number: 5.
[202] Sarah R. Moore, Thorhildur Halldorsdottir, Jade Martins, et al. Sex differences in the genetic regulation of the blood transcriptome response to glucocorticoid receptor activation. *Transl Psychiatry*, 11(1):632, December 2021. Number: 1.
[203] Martin J. Aryee, Andrew E. Jaffe, Hector Corrada-Bravo, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, May 2014. Number: 10.
[204] Timothy J. Triche, Daniel J. Weisenberger, David Van Den Berg, Peter W. Laird, and Kimberly D. Siegmund. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res*, 41(7):e90, April 2013. Number: 7.
[205] Ruth Pidsley, Chloe C. Y Wong, Manuela Volta, et al. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, 14:293, May 2013.
[206] Yi-an Chen, Mathieu Lemire, Sanaa Choufani, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium

HumanMethylation450 microarray. *Epigenetics*, 8(2):203–209, February 2013. Number: 2.
[207] Miles C. Benton, Alice Johnstone, David Eccles, et al. An analysis of DNA methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss. *Genome Biol*, 16:8, January 2015.
[208] Ruth Pidsley, Elena Zotenko, Timothy J. Peters, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*, 17(1):208, October 2016. Number: 1.
[209] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, March 2012. Number: 6.
[210] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, January 2007. Number: 1.
[211] Eugene Andres Houseman, William P. Accomando, Devin C. Koestler, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13:86, May 2012.
[212] J. L. Min, G. Hemani, G. Davey Smith, C. Relton, and M. Suderman. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics*, 34(23):3983–3989, December 2018. Number: 23.
[213] Laurent Gautier, Leslie Cope, Benjamin M. Bolstad, and Rafael A. Irizarry. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315, February 2004. Number: 3.
[214] Hans Binder and Stephan Preibisch. GeneChip microarrays—signal intensities, RNA concentrations and probe sequences. *J. Phys.: Condens. Matter*, 18(18):S537–S566, May 2006. Number: 18.
[215] Fabrice Berger and Enrico Carlon. From hybridization theory to microarray data analysis: performance evaluation. *BMC Bioinformatics*, 12:464, December 2011.
[216] UCSC, REST API https://genome.ucsc.edu/goldenPath/help/api.html [2022].
[217] Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.64.1 https://bioconductor.org/packages/Biostrings [2022].
[218] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*, 48(D1):D845–D855, January 2020. Number: D1.
[219] Andrey A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, May 2012. Number: 10.
[220] Dean Langan, Julian P. T. Higgins, Dan Jackson, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods*, 10(1):83–98, March 2019.
[221] Kurex Sidik and Jeffrey N. Jonkman. Simple Heterogeneity Variance Estimation for Meta-Analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(2):367–384, April 2005.
[222] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic

Optimization. 2014. Publisher: arXiv Version Number: 9.
[223] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*, 2(3):100141, August 2021. Number: 3.
[224] Jovana Maksimovic, Alicia Oshlack, and Belinda Phipson. Gene set enrichment analysis for genome-wide DNA methylation data. *Genome Biol*, 22(1):173, December 2021.
[225] Mengyu Xie and Youquan Bu. SKA2/FAM33A: A novel gene implicated in cell cycle, tumorigenesis, and psychiatric disorders. *Genes Dis*, 6(1):25–30, March 2019. Number: 1.
[226] Jerry Guintivano, Tori Brown, Alison Newcomer, et al. Identification and replication of a combined epigenetic and genetic biomarker predicting suicide and suicidal behaviors. *Am J Psychiatry*, 171(12):1287–1296, December 2014. Number: 12.
[227] A. B. Niculescu, D. Levey, H. Le-Niculescu, et al. Psychiatric blood biomarkers: avoiding jumping to premature negative or positive conclusions. *Mol Psychiatry*, 20(3):286–288, March 2015. Number: 3.
[228] A. B. Niculescu, D. F. Levey, P. L. Phalen, et al. Understanding and predicting suicidality using a combined genomic and clinical risk assessment approach. *Mol Psychiatry*, 20(11):1266–1285, November 2015. Number: 11.
[229] Z. Kaminsky, H. C. Wilcox, W. W. Eaton, et al. Epigenetic and genetic variation at SKA2 predict suicidal behavior and post-traumatic stress disorder. *Transl Psychiatry*, 5:e627, August 2015.
[230] Xueyi Shen, Doretta Caramaschi, Mark J. Adams, et al. DNA methylome-wide association study of genetic risk for depression implicates antigen processing and immune responses. *Genome Med*, 14(1):36, March 2022. Number: 1.
[231] Daniel L. McCartney, Rosie M. Walker, Stewart W. Morris, et al. Altered DNA methylation associated with a translocation linked to major mental illness. *NPJ Schizophr*, 4(1):5, March 2018. Number: 1.
[232] Krassimira A. Garbett, Andrea Vereczkei, Sára Kálmán, et al. Coordinated messenger RNA/microRNA changes in fibroblasts of patients with major depression. *Biol Psychiatry*, 77(3):256–265, February 2015. Number: 3.
[233] Anthony A. Grace. Dysregulation of the dopamine system in the pathophysiology of schizophrenia and depression. *Nat Rev Neurosci*, 17(8):524–532, August 2016. Number: 8.
[234] C. Ménard, G. E. Hodes, and S. J. Russo. Pathogenesis of depression: Insights from human and rodent studies. *Neuroscience*, 321:138–162, May 2016.
[235] Arvid Carlsson. Early psychopharmacology and the rise of modern brain research. *J Psychopharmacol*, 4(3):120–126, May 1990. Number: 3.
[236] Raz Yirmiya, Neta Rimmerman, and Ronen Reshef. Depression as a microglial disease. *Trends Neurosci*, 38(10):637–658, October 2015. Number: 10.
[237] Butian Zhou, Zhongqun Zhu, Bruce R. Ransom, and Xiaoping Tong. Oligodendrocyte lineage cells and depression. *Mol Psychiatry*, 26(1):103–117, January 2021. Number: 1.
[238] Haixia Wang, Yi He, Zuoli Sun, et al. Microglia in depression: an overview of microglia in the pathogenesis and treatment of depression. *J Neuroinflammation*, 19(1):132, June 2022. Number: 1.

[239] Robert Dantzer, Jason C. O'Connor, Gregory G. Freund, Rodney W. Johnson, and Keith W. Kelley. From inflammation to sickness and depression: when the immune system subjugates the brain. *Nat Rev Neurosci*, 9(1):46–56, January 2008. Number: 1.

[240] Eric S. Wohleb, Tina Franklin, Masaaki Iwata, and Ronald S. Duman. Integrating neuroimmune systems in the neurobiology of depression. *Nat Rev Neurosci*, 17(8):497–511, August 2016. Number: 8.

[241] Ole Köhler, Michael E. Benros, Merete Nordentoft, et al. Effect of anti-inflammatory treatment on depression, depressive symptoms, and adverse effects: a systematic review and meta-analysis of randomized clinical trials. *JAMA Psychiatry*, 71(12):1381–1391, December 2014. Number: 12.

[242] Harris A. Eyre, Tracy Air, Simon Proctor, Sebastian Rositano, and Bernhard T. Baune. A critical review of the efficacy of non-steroidal anti-inflammatory drugs in depression. *Prog Neuropsychopharmacol Biol Psychiatry*, 57:11–16, March 2015.

[243] Jennifer L. Warner-Schmidt, Kimberly E. Vanover, Emily Y. Chen, John J. Marshall, and Paul Greengard. Antidepressant effects of selective serotonin reuptake inhibitors (SSRIs) are attenuated by antiinflammatory drugs in mice and humans. *Proc Natl Acad Sci U S A*, 108(22):9262–9267, May 2011. Number: 22.

[244] C. H. Browning. Nonsteroidal anti-inflammatory drugs and severe psychiatric side effects. *Int J Psychiatry Med*, 26(1):25–34, 1996. Number: 1.

[245] Nils Kappelmann, Janine Arloth, Marios K. Georgakis, et al. Dissecting the Association Between Inflammation, Metabolic Dysregulation, and Specific Depressive Symptoms: A Genetic Correlation and 2-Sample Mendelian Randomization Study. *JAMA Psychiatry*, 78(2):161–170, February 2021.

[246] Diego A. Pizzagalli and Angela C. Roberts. Prefrontal cortex and depression. *Neuropsychopharmacology*, 47(1):225–246, January 2022.

[247] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data?, July 2022. arXiv:2207.08815 [cs, stat].

[248] John P. A. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, August 2005.

[249] Sara Gandini, Edoardo Botteri, Simona Iodice, et al. Tobacco smoking and cancer: a meta-analysis. *Int J Cancer*, 122(1):155–164, January 2008.

[250] Linda M. O'Keeffe, Gemma Taylor, Rachel R. Huxley, et al. Smoking as a risk factor for lung cancer in women and men: a systematic review and meta-analysis. *BMJ Open*, 8(10):e021611, October 2018.

[251] Peter N. Lee, Barbara A. Forey, and Katharine J. Coombs. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC Cancer*, 12:385, September 2012.

[252] Mark D. Wilkinson, Michel Dumontier, I. Jsbrand Jan Aalbersberg, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3:160018, March 2016.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 2093

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-540129

ACTA UNIVERSITATIS
UPSALIENSIS
2024