fakultät statistik

technische universität dortmund

Variable Selection Methods for Detecting Interactions in Large Scale Data

Dissertation

in Fulfillment of the Requirements for the Degree of

Dr. rer. nat.

Submitted to the

Department of Statistics at TU Dortmund University

by

Sven Teschke

Dortmund, April 2025

Referees:
Prof. Dr. Katja Ickstadt (TU Dortmund)
Prof. Dr. Tamara Schikowski (IUF Düsseldorf)
JProf. Dr. Christian Staerk (TU Dortmund)

Abstract

Large-scale data sets comprising millions of variables p, as is typical in the field of genetics, offer a wealth of information. However, it is a considerable challenge to extract this information from the data. From a biological perspective, it is desirable that this will lead to a better understanding of the development of diseases. Moreover, it is imperative to consider the interactions of genetic factors with each other and with the environment. Taking into account interactions further exacerbates the problem of the high dimensionality of the data. In addition to the computational challenges of processing the data at all, most statistical models are inapplicable or difficult to interpret in these scenarios. To address this research gap, a variable selection method was developed in this thesis that accounts for a multivariate structure and can be applied to arbitrarily large amounts of data. The selection of variables is executed through the utilization of cross-leverage scores (CLS). Due to their construction the CLS correspond to the variables individual leverage on the correlation with the multidimensional subspace spanned by the data with the outcome variable. Thus, they are directly linked to the importance of a variable also in the sense of an interaction effect. Further, under mild assumptions, each CLS equals its corresponding parameter in the least squares solution up to a small bounded additive error. In addition, in this thesis, methods have been developed and improved for the approximation of the CLS in large data. A notable advantage of these methods is their ability to be calculated streamwise, thereby overcoming the problem of processing on standard computers. Overall, a two-step procedure is recommended. In the first step, variables are selected using CLS. In the subsequent step, an established method is to be applied to the reduced data, which is appropriate for the research question, but limited in the number of input variables. The primary article of this dissertation introduces the methodology of these approaches and validates them by simulations as well as mathematically. In two additional articles, this method is employed to two large scale datasets, in order to answer biological questions. Once, in the framework of a two-step approach to identify SNP-environment interactions in COPD. In the second step, the recently developed logicDT model is applied to the reduced data. In the other paper, the CLS are directly incorporated into the calculation of so-called profile scores to estimate the risk of Alzheimer's disease based on DNA methylation and metabolomics data.

Acknowledgments

First, I would like to express my deep gratitude to Katja for her exemplary guidance and unwavering support throughout the course of my research. I am grateful for the constructive feedback and fruitful discussions that have contributed to the improvement of my work. I would especially like to thank her for the patience and trust she has always shown, even when I have been absent for long time periods. I would also like to express my gratitude to my co-supervisor, Tamara, who consistently demonstrated the practical applications and meaningfulness of the theoretical concepts. And in addition, I would like to express my profound gratitude to Christian for reviewing this thesis.

I would like to express my profound gratitude to my co-authors Claudia, Timur and Alex for the inspiring collaboration and to all my colleagues at TU Dortmund University and at the Research Training Group 2624 *Statistical Methods for High-Dimensional Data in Toxicology*. I am appreciative of the opportunity to conduct my doctoral research within this exceptional professional and friendly environment. Many thanks to Jörg for the great organization and warm-hearted mentoring of the RTG. The enthusiasm for statistical research has always been inspiring. Of course, many thanks to the DFG for funding this great project.

Heartfelt thanks to Vito for always being there for me and for alleviating the pain that comes along with doing a PhD. All in all, I would like to thank my friends and family, doctors and therapists, and all those who have patiently accompanied me throughout my academic journey, as well as through the struggles on the sidelines and the many afflictions that often knocked me back.
To paraphrase the words and doctor-like wisdom of a friend:

'Kopf in den Sand is' auch keine Option.'
Rest in peace Simon!

List of Publications

This cumulative thesis is based on the following manuscripts:

Published Article

- Article 1: **Teschke, S.**, Ickstadt, I. & Munteanu, A. (2024). *Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores.* Biometrical Journal. **66**(8). 10.1002/bimj.70014 [112]

Contribution of the author:
The author of this thesis elaborated and enhanced the ideas using cross leverage scores for variable selection. He gives the central idea for the stream wise calculations and implemented the data analyses and the simulation studies. Alexander Munteanu provided the major contribution to the theoretical foundations. The author wrote the first draft of the manuscript. Discussions and revising was done together with Alexander Munteanu and Katja Ickstadt.

Submitted Articles

- Article 2: **Teschke, S.**, Ickstadt, K., Schikowski, T., Wigmann, C. (2025). *Using Cross Leverage Scores for Detecting SNP-Environment Interactions Effects on COPD.* Genetic Epidemiology (submitted) [113]

Contribution of the author:
The author of this thesis implemented the data analyses and wrote the first draft of the manuscript. Discussions and revising was done together with Claudia Wigmann under the supervision of Katja Ickstadt and Tamara Schikowski.

- Article 3: Tug, T., Liang, D., **Teschke, S.**, Tan, Y., Gearing, M., Levey, A.I., Lah, J.J., Wingo, A.P., Wingo, T.S., Lau, M., Ickstadt, K., Hüls, A. (2025). *Development and Application of Brain Tissue Based Multi-Omics Risk Scores for Alzheimer's Disease.* Alzheimer's & Dementia: The Journal of the Alzheimer's Association (submitted) [118]

Contribution of the author:
The author of this thesis contributed the theoretical methods according to the calculation of Cross Leverage Scores for the calculation of profile scores for DNA methylation and metabolome data. Further, he was involved in discussions and revising the manuscript.

Contents

1 Introduction & Motivation

Recent advancements of technological capacity to store, communicate, and compute information have enabled the accumulation of an expanding volume of data, cf. [47]. The term Big Data is used to refer to massive data sets with large, varied, and complex structures that are difficult to store, analyze, and visualize [99]. In this thesis, the focus is on a large amount of variables, which is referred to here as *large scale data*. For example, in genomics, the amount of data has increased substantially, so biologists also encounter challenges regarding Big Data [74]. However, this also offers new opportunities to gain insight into how a disease is influenced by the genome. In addition to genomics, the growth of data in other omics fields such as proteomics and metabolomics over the past decade has provided an essential knowledge base, cf. [100].
Extracting information from large scale data poses a challenge from a computational perspective, but also from a statistical-methodological perspective, cf. [33]. Conventional methods mostly encounter limitations in terms of the number of variables they can process effectively. This challenge is further compounded when the number of variables, denoted by p, typically exceeds the number of observations, denoted by n, resulting in $n << p$ problems. Thus, statistical analyses become infeasible or even impossible for data sets of considerable size.
Modern machine learning (ML) methods are designed to handle millions of variables and provide good predictions in big data problems. Nevertheless, the interpretation of these models is often exceedingly difficult. Indeed, they are described as 'black box' models [69].
In this work, our primary focus is on identifying important factors rather than on achieving optimal prediction accuracy. Consequently, it is essential that a model is highly interpretable. Therefore, this thesis is based on the development of a variable selection method for regression models that is applicable to large scaled data, that is, data sets with millions of variables. In this thesis, this method is applied to regression models for binary, continuous, and ordinal outcomes.
This approach can then be used to identify significant genetic and environmental risk factors associated with health outcomes. Genetic factors are, for instance, single variations in the DNA, called Single Nucleotide Polymorphisms (SNPs).

It has been shown that it is crucial to consider interactions of genetic factors [71] as well as interactions of genetic variations with the environment for a better understanding of complex diseases [89]. As the consideration of the interactions is of particular interest, the issue of dimensionality poses an even more substantial challenge. In a scenario with p variables, the amount of possible 2-way combinations

is already quadratic $\Theta(p^2)$ and the number of choices grows exponentially in the order of interactions k. Finally the number of all possible combinations of any degree sum up to 2^p.
Existing methods either mostly do not account for interactions, or if they do, they are limited in the maximum number of variables. ML methods such as Random Forests [9] can account for interactions for prediction, but they are difficult to interpret. Moreover, Random Forests have difficulties to detect interactions when the respective variables have negligible marginal effects [131], which is often the case in genetic analyses [67]. To overcome this research gap, it is imperative that the developed variable selection method also takes interactions into account.

The challenge is therefore twofold. From a biological perspective, it is necessary to ascertain the influence of genetic, environmental, and other factors on the development of diseases. This understanding can then provide the basis for risk reduction strategies. From a statistical perspective, the interpretation of methods is essential to cope with the specific challenges of genetic data, such as the large number of variables. It is imperative to deliberate on the requisites from both domains and thereby bridge the gap between theory and application. However, the statistical methods can be applied to other genetic questions, but also to completely different research fields.

In the main article of this dissertation *Detecting interactions in high-dimensional data using cross leverage scores*, a variable selection method based on so-called cross leverage scores (CLS)[1] is developed [112]. These scores serve as a metric for assessing the importance of a variable independently of whether it is a main effect or a participant in an important interaction effect. This is determined by the mathematical construction of the scores. The application of a QR-decomposition of the transposed common matrix of variables and outcome enables the identification of each variable's leverage on the correlations they have with the multidimensional space spanned by the data with the outcome variable. A mathematical justification of the CLS is given in the paper. So, the primary benefit of this approach is that the CLS of each variable directly contains information about the importance of a possible interaction, without having to consider every possible combination.
As this score merely indicates which variables are important yet does not show which variables interact with each other, a two-step procedure is proposed. After selecting these variables with the $q << p$ most extreme CLS in a subsequent analysis, methods such as logic regression [98] or logic decision trees (logicDT) [65], which are particularly suitable for identifying interactions but limited in the number of variables they can process, are employed to the selected data.
Furthermore, this article proposes approximation approaches for the CLS. This includes two heuristic stream wise calculations, *Sliding Window* and *Random Window*, as well as a sketching approach based on subspace embeddings. The latter is a transfer of dimension reduction methods that are based on sampling and random projections from reducing observations to selecting variables. With these approximation approaches it is possible to apply the variable selection based on CLS to arbitrarily large data. In the article, SNP data is considered with binary outcomes.

[1] CLS is used for single cross-leverage score and for several scores to simplify readability. The abbreviation 'CLSs' is only used when the context is clear and the distinction is important.

The performance of the approaches is tested in complex simulation studies and on a well-studied standard data set, with the goals of determining the ability to find important variables and of understanding how variable selection affects subsequent analyses such as logicDT or Random Forests.

In the second article, *Using Cross Leverage Scores for Detecting SNP-Environment Interactions Effects on COPD* [113], the aforementioned methods are applied to the practical question of how genetic (SNPs) and environmental factors affect lung function, namely Chronic Obstructive Pulmonary Disease (COPD). This is highly relevant, as early detection of risk factors can generally prevent diseases or allow to save on expensive treatment, as this is the case with COPD [41]. For this task, the data of the SALIA cohort study [102] is considered. One objective of this study is to identify important factors associated with the risk of lung diseases. This includes data on clinical and personal data, such as smoking or the body mass index (BMI), SNP data from the whole genome, as well as environmental data on air pollution. In total there are more than 7.5 million variables for above 500 observations. The novel two-step procedure proposed in [112], with a variable selection based on CLS in the first step, facilitates the simultaneous consideration of all these data, thereby enabling the incorporation of interaction effects. Secondly, the explicit important factors and their interactions are obtained by applying the recently developed (bagged) logicDT model [65] to the reduced data. This novel procedure is intended to gain new insights into the interaction of the various factors. A comprehensive literature review was conducted for both the selected variables CLS and the factors identified by logicDT. This is done, on the one hand, to evaluate the practical applicability of the developed method, and, on the other hand, to prove the plausibility of the identified factors, in order to give a suggestion for further examinations in biological experiments.

In the third article, *Development and Application of Brain Tissue Based Multi-Omics Profile Scores for Alzheimer's Disease*, which was written in collaboration with Timur Tug et al. [118], the CLS was applied in a different way, namely to calculate so-called (multi-omics) profile scores, which are a variant of risk scores. Moreover, this method is applied here for the first time to epigenomic and metabolomic data. Even if they are often only evaluated individually, it is recommended to consider omics data of different types jointly [73]. The analysis of these multi-omics data is expected to provide new insights into the biological mechanisms underlying Alzheimer's disease. In a study on genome wide DNA methylation and metabolomics from brain tissues, various approaches for the calculation of the Profile Scores are compared. The property that the CLS in the theory approximates the solution in an ordinary least square (LS) problem [112] is utilized for the calculation of the profile scores. The CLS are used as the weights for the weighted sum needed for the calculation of profile scores. The variable selection step is directly included, since only the variables with the most extreme scores are included in the weighted sum.

In conclusion, the three papers propose the CLS for variable selection in various large scale data settings. The property of the CLS to preserve the multivariate structure of data is used to answer different biological questions. This shows that the CLS can be utilized in a flexible manner across various data types and objectives even beyond the field of biology.

The research is embedded in the Research Training Group (RTG) 2624 *Biostatistical Methods for High-Dimensional Data in Toxicology* at TU Dortmund University. The RTG is a collaboration with HHU Düsseldorf, the University of Cologne, and the two Leibnitz Institutes, IUF in Düsseldorf and IfADo in Dortmund. The aims of the research in the RTG are the development and application of biostatistical methods for the analysis of high-dimensional data for modeling and risk assessment in toxicology. This thesis is based on the project 'R1: High-dimensional regression for screening of important genetic and environmental factors' in cooperation of TU Dortmund University and the IUF.

The remainder of the thesis is structured as follows. Chapter 2 provides the biological background and describes the data. Section 2.1 gives an overview of the biological terms and the motivation for the research work with regard to the biological questions. Section 2.2 describes the data of the SALIA cohort study (considered in [113]) and of the Alzheimer study (considered in [118]). Chapter 3 introduces the methods. The present state of the art of statistics in Genetics is delineated in section 3.1 where the strengths and weaknesses of existing methods are discussed. Section 3.2-3.4 describe the central and novel methods that are used in the (articles of this) thesis. A summary of each of the three chapters is given in chapter 4. Finally, chapter 5 presents a conclusion and an outlook. The articles are contained in Appendix A.

2 Biological Background and Data

2.1 Omics Data & Diseases

From a biological perspective, this dissertation explores the impact of genetic variations in human DNA on health. In the following the most salient genetic terms are introduced. However, the focus is less on the exact biological processes than on investigating the statistical associations. For detailed descriptions, please refer to appropriate specialized literature, e.g. [44].

The essential components of the DNA (Deoxyribonucleic acid) are four different nucleotide bases, namely cytosine (C), thymine (T), adenine (A) and guanine (G).The spatial structure of the bases is expressed in the form of a double helix. This insight was postulated in 1953 by J.D. Watson and F.H.C. Crick in their article *A Structure for Deoxyribose Nucleic Acid* [124], for which they were awarded the Nobel Prize in 1962. However, it should be noted that this publication is largely based on the research of Rosalind Franklin and her PhD student Raymon Gosling and the famous "photo 51" which was published in the article *Molecular Configuration in Sodium Thymonucleate* [36]. Based on this X-ray photo and the associated calculations, Watson and Crick succeeded in proving the double helix structure. Nonetheless, they hid the fact that they knew Franklin's results and the photo, and that they could not finish their work without her research. Franklin died four years before Watson and Crick were awarded the Nobel Prize for their work. Throughout her life, however, she never experienced the fame she deserved and even in the Nobel Prize speech neither she nor her doctoral student Gosling were mentioned, see [125]. One of many examples of scientific research being credited to men when it deserves at least equal credit to women.
Two nucleotide chains linked by hydrogen bonds form the aforementioned double helix. Due to their structure it is only possible that guanine is linked to cytosine and adenine to thymine. In living organisms a long DNA chain with genetic information is called a chromosome. Humans have diploid set of 23 chromosomes, in total 46 individual chromosomes, inheriting one chromosome from the mother and one from the father. The 23 chromosomes are composed of a total of over 3 billion base pairs. Genes are sections of DNA that contain e.g. the code for a specific protein. The totality of all genetic information, including protein-coding genes and non-coding genes is called genome. The occurrence of a variation of a single base is denoted as a single-nucleotide polymorphisms (SNP), when it occurs in more than 1% of the population. Otherwise, i.e. rare variations are called mutations. A variation of

a gene sequence is called allele. The frequency of the second most common allele in a population is called minor allele frequency (MAF). In the absence of a single variation with respect to the reference, the term *homozygous reference* is employed. In the event of the variation is present on a single chromosome, the term *heterozygous variant* is utilized, and if the variation is observed on both chromosomes, the term *homozygous variant* is used. Mathematically, a SNP is coded as SNP $\in \{0, 1, 2\}$. Such a SNP occurs about every 500th to 1000th position in the genome.
It took until the beginning of the 21st century to map the entire genome. After completing a working draft of the human genome sequence in 2000, the International Human Genome Sequencing Consortium announced the successful completion of the Human Genome Project on April 14, 2003, exactly 50 years after the Watson-Crick paper. Then, on October 20, 2004, the International Human Genome Sequencing Consortium publishes its scientific description of the completed human genome sequence covering $\sim 99\%$ of the genome with error rate of 1 in 100000 [54]. Another finding is, that the human genome seems to encode only $20000-25000$ protein-coding genes.
The Human Genome Project has initiated and enabled several large and countless smaller projects in human biology. For example, the international HapMap project [115], parts of which are incorporated into this dissertation, see [112]. Moreover, it is the basis for genome-wide association studies (GWAS), which attempt to identify SNPs located on or near genes involved in the development of diseases. To analyze these studies and identify important SNPs or differences in the population, sophisticated statistical methods are required. A detailed overview of the history of achievements in genetic research is provided by [61]. It was called the 'SNP revolution' back in 1999, but research in this area is still ongoing [11]. Due to cheaper and better sequencing methods, research on statistical methods that provide reliable results in high-dimensional data is required to make meaningful use of the mass of information.
SNPs have been demonstrated to be associated with (complex) diseases [105] e.g. Crohn's disease [75] or neurofibromatosis type 1 [91]. However, not only individual SNPs, but also their interaction is causal for many diseases [19]. Disease variation in phenotypes is based on highly dynamic, interconnected and non-linear biochemical networks, which in turn are based on genetic variants. When the effect of a variant on a complex trait depends on the genotype of a second variant, it is called epistasis. It is important to distinguish this from the so-called linkage disequilibrium (LD), which is the dependent association of two alleles in a population, which of course occur together and do not necessarily have an important influence. It is crucial to investigate how naturally occurring genetic variants jointly act, instead of summing up the effects of individual variants independently, to understand and predict complex diseases [71]. Diseases that can be traced back to interactions between SNPs are e.g. breast cancer [15] and venous thrombosis [45]. Further, it is also common that individual SNP has no effect on the phenotype, but their interaction has a strong effect [67].

On the other hand, the effects of alleles are often highly sensitive to the environment to which individuals are exposed. Therefore, to gain a deeper understanding of the influence of single nucleotide polymorphisms (SNPs) on disease, it is imperative to include the interactions of genes and environmental factors in the analysis [68].

Gene-Environment interactions can be defined as 'a different effect of an environmental exposure on disease risk in humans with different genotypes' [89]. For instance, the development of chronic obstructive pulmonary disease (COPD) has been demonstrated to be influenced by both genetic and environmental factors [93], as well as their interaction [1]. For respiratory outcomes, it is recommend to include known harmful environmental exposures to identify interacting genetic loci [136].

The focus of this work is on COPD (ICD10 Code: J.44.-), but the methods developed in this work are of course generally applicable and can be used to analyze other diseases. In the following, we discuss the most important aspects of the aetiology of COPD. In 2020, COPD was the third most common cause of death worldwide [86]. It is forecasted that by 2060 there will be much more annual deaths from COPD and related conditions due to the increasing prevalence of smoking in low- and middle-income countries and the ageing population in high-income countries [40]. COPD is a common, preventable, and treatable disease, but extensive under-diagnosis and misdiagnosis leads to patients receiving no treatment or incorrect treatment. Appropriate and earlier diagnosis of COPD can have a very significant public-health impact [41]. So if someone is found to have genetic variations that increase the risk of COPD, countermeasures can be taken at an early stage.
COPD is a heterogeneous disease and includes a variety of phenotypes such as airflow obstruction, emphysema or exacerbation [21]. A reliable classifier for COPD is the Tiffeneau Index, defined by the quotient of forced expiratory volume in one second (FEV_1) and the forced vital capacity (FVC) [93]. However, it is not enough to just consider the FEV_1, because a patient can have a small lung (small FEV_1) and still have no obstruction [21]. According to the Global Initiative for Chronic Obstructive Lung Disease, COPD can be divided into four levels of severity, cf. [41].
Risk factors for developing COPD are, for example the social status [59], the age [83] and the body mass index (BMI) [108]. For BMI the relationship to COPD is not trivial and necessitates further research [25]. Alongside the association between increased BMI and COPD, a BMI that is too low favors a more severe course of the disease [80]. The leading cause for COPD is smoking [62], [70], but also years of smoking [107] as well as passive smoking and air pollution is important [133]. In particular $PM_{2.5}$ (particular matter(PM) of diameter less than 2.5μm) exposure likely increases the risk of COPD [123], [23]. The specific mechanisms of this factor still need to be investigated further and in this work this value is primarily concerned as measure of air pollution. As illustrated in fig. 2.1, the dimensions of these particles are tiny, thus enabling facile insertion into the pulmonary system and subsequent dissemination into the bloodstream, with the potential for significant physiological impairment as a consequence [132].

Specific COPD phenotypes and outcomes are associated with different transcriptomes and metabolomes indicating distinct mechanisms for different phenotypes [21]. Further, variants and genes found with GWAS do not necessarily reflect a direct biological relevance for the disease [109]. For certain traits, GWAS of common SNPs are approaching signal saturation. This indicates the need to explore also other types of genetic variations [46]. Although the developed methods in this thesis were motivated by dealing with SNP and environmental data. Nevertheless, from

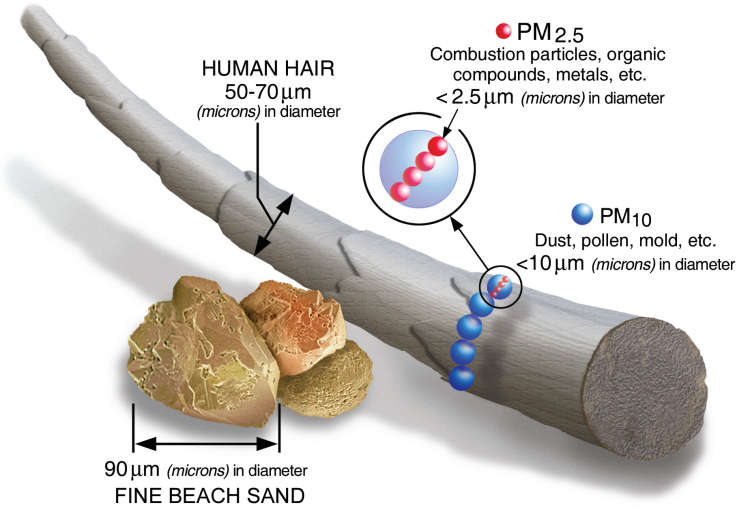

Figure 2.1: This figure illustrates the size comparisons for PM (particular matter) particles. $PM_{2.5}$ are particles with diameter of 2.5 micrometers and smaller. [119]

a statistical point of view, these methods are universally valid, and therefore also applicable to other omics data.

In the project *Development and Application of Brain Tissue Based Multi-Omics Risk Scores for Alzheimer's Disease* [118] the CLS are applied to multi-omics data. The aim is to decipher how DNA methylation (DNAm) and metabolomics interact and jointly effect Alzheimer's disease (AD) neuropathology. This pertains to the domain of epigenetics. Epigenetics encompasses changes in the regulation of gene expression without altering the DNA sequence [55]. It has been shown that genetic predisposition does not necessarily lead to the development of diseases and that epigenetic factors can contribute to phenotypic outcomes [58]. This is the reason why one twin may suffer from a disease while the other does not, even though they carry the same genetic material [24]. One important epigenetic variation is the DNA methylation. This is a chemical modification attached to individual DNA nucleotides influencing the transcription of genes. In short, we are interested in the case when a methyl group is attached to a cytosine base instead of a hydrogen atom, which usually occurs when a cytosine base is followed by a guanine base in a DNA sequence. The latter is defined as a CpG site, whereas p stands for the linking phosphate. If methylation modifies clusters of CpG sites, this is known as CpG islands which can significantly reduce or even silence the expression of the respective gene [106]. DNAm plays an important role in cell development, gene regulation and disease processes such as cancer, cf. [49]. Further, specific methylation signatures are biomarkers for diseases like rheumatoid arthritis [79], multiple sclerosis [87] and in particular Alzheimer disease [22]. DNA methylation is a remarkable field of research that can uncover the hidden mechanisms of various diseases, introduce new diagnostic and prognostic biomarkers, and propose new patient-specific therapeutic approaches for diseases [135].

Besides genomics, epigenomics, transcriptomics and proteomics there also metabolomics. The metabolome describes the totality of metabolites in a cell at a specific time. The metabolites are products of metabolic processes in organisms and the

individual metabolic signatures are unique, cf. [97]. An individual's metabolome reflects alterations in genetic, transcript, and protein profiles and is affected by the environment. Therefore, metabolomics offers a great potential for the diagnosis and prognosis of e.g. neurodegenerative diseases such as Alzheimer's diseases [128]. The interaction of metabolomics and DNAm also appears to play a role, but little is known about how these interact and jointly influence the development of Alzheimer's disease [118].

Dementia including its clinical subtypes such as Alzheimer's disease was the seventh leading cause of death among all diseases from 2000 to 2019 [126]. Alzheimer disease is the most common form of dementia (60–70% of cases) [3]. In 2019 55.2 million people worldwide are living with dementia and the annual global societal cost was estimated at US$ 1313.4 billion [129]. As age is a major risk factor, the change in demographics, along with the increasing prevalence of other risk factors such as diabetes and obesity, will lead to a massive increase in Alzheimer's dementia in the coming years [126]. Further risk factors for AD are vascular factors such as hypertension and hypercholesterolemia [20] or head injury [43].

2.2 Data

SALIA Cohort Study

As delineated in the preceding section, the interest is in what factors favor or influence the development of COPD. The investigation encompasses both environmental and genetic factors and their interactions. For a better reading, we include variables such as smoking or social status under environmental influences, even though these are not really environmental factors, as opposed to exposure to air pollution.

In the real data application in [113] the SALIA cohort study [102] is considered. An overview of the available data as well as a descriptive analysis of the data is provided here. For study design, detailed findings and results, see [113].
In total the SALIA data consists of $p_{gen} = 7643653$ SNPs and $p_{env} = 7$ clinical respective environmental data. With regard to the initial sample of 510 women, seven subjects were excluded due to partially missing values. Consequently, the total number of women under consideration is $n = 503$. For these women the medical diagnosis of COPD has been documented ('yes', 'no' or 'unknown'). The majority of subjects were not diagnosed with COPD, see table 2.1. Of the seven women who were removed from the study, none had COPD. However, given the imbalanced

$n = 503$	yes	no	unknown
COPD	19 (3.78%)	483 (96.02%)	1 (0.20%)

Table 2.1: The table shows a clinical diagnosis of chronic obstructive pulmonary disease for the 503 women in the SALIA cohort study.

distribution of cases and non-cases, the target variable in the subsequent analyses is the Global Lung Initiative (GLI) z-score of the Tiffeneau-Index (FEV_1/FVC) [94]. Furthermore, tobacco related characteristics were captured, including whether the participant is a current smoker at the time of the study (yes/no), whether they had been a smoker in the past (yes/no), whether they are currently exposed to passive smoking (yes/no), see table 2.2. In addition the variable "packyears", defined by

$n = 503$	yes	no
smoking	11 (2.19%)	492 (97.81%)
ex-smoking	83 (16.50%)	420 (83.50%)
passive smoking	193 (38.37%)	310 (61.63%)

Table 2.2: Variables related to smoking included in the SALIA cohort study.

the smoking duration in years and the current daily packs of cigarettes. For the majority of observations, namely 409 women (81.31%) it is `packyears` $= 0$. The distribution of the variable packyears is therefor characterized by the presence of mainly small values, accompanied by some outliers with large values. For better illustration the distribution of packyears > 0 is shown in fig. 2.2.
The variable 'status' measures the social status of the participants according to the maximum number of school years of them or the husband, divided into three categories: less then 10, exact 10 and more than 10 school years. About half, fall into the middle category, followed by the category with higher education (about a third), see table 2.3. The degree of influence of the environment to which the body

$n = 503$	1 ($<$ 10yr)	2($=$ 10yr)	3 ($>$ 10yr)
social status	87 (17.30%)	244 (48.51%)	172 (34.20%)

Table 2.3: The table shows the social status of the 503 women from the SALIA cohort study. The three categories are: Less then 10, exact 10 and more than 10 school years.

is exposed is, to a certain extent, dependent on the place of residence. The ratio of participants residing in urban area (Ruhr Area, ca. 51%) and those residing in rural areas (Münsterland, ca. 49%) is relatively balanced. At the time of the study, the youngest participant was 66 years old and the oldest was 79 years old. The majority of women were between 71 and 74 years old. fig. 2.3 shows the distribution of age.
Furthermore, the body mass index (BMI) of women was measured. The BMI is defined as the weight in kilograms divided by the square of the height in meters. According to the World Health Organization (WHO), an individual is categorically designated as overweight with a BMI at or above 25, and obese if their BMI exceeds 30, as well as underweight with a BMI under 18.5 [127]. The majority of the participants were classified as at least overweight according to the WHO classification criteria (65.41%). A study conducted by the RKI between 2014 and 2015 revealed that, on a national average, the majority of women over 65 in Germany belong to the overweight category. However, the percentage in this group was just

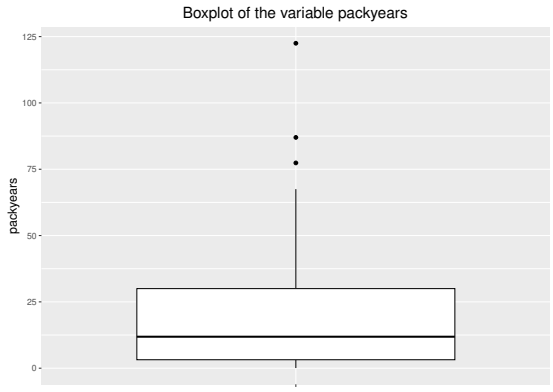

Figure 2.2: This boxplot shows the distribution of the variable packyears. However, only those variables for which packyears is greater than 0 are considered here.

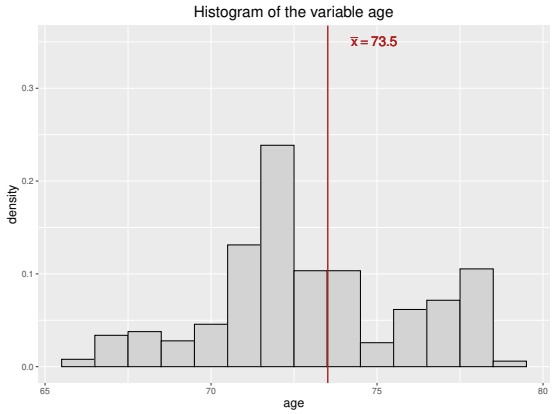

Figure 2.3: Histogram of the variable age over the 503 women in the SALIA cohort study.

58.9% with a confidence interval of $[56.5\% - 61.3\%]$ [101]. In fig. 2.4 a boxplot of the variable BMI illustrates its distribution. The median is 26.81, and there are a number of outliers with large BMI values. For air pollution the annual mean level of $PM_{2.5}$ of the year 2006 is considered. The distribution of $PM_{2.5}$ is right-skewed as can be seen in fig. 2.5.

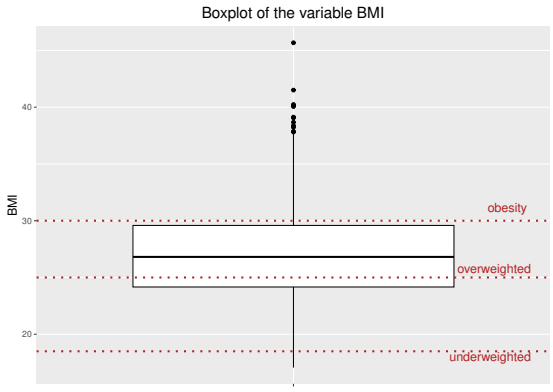

Figure 2.4: Boxplot of the variable body mass index (BMI) for the 503 women in the SALIA cohort study. With the WHO classification boundaries marked by the red lines.

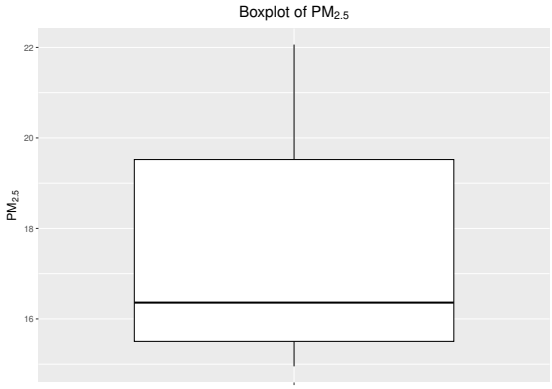

Figure 2.5: Boxplot of the variable $PM_{2.5}$ (Particular Matter with diameter less than $2.5\mu m$ for the 503 women in the SALIA cohort study.

Emory Goizueta Alzheimer's Disease Research Center Data

Genome-wide DNAm and high-resolution metabolomics data were obtained from 157 prefrontal cortex tissue samples from donor brains. The brains belong to individuals who died at different stages of AD pathology. The Emory Goizueta Alzheimer's Disease Research Center (ADRC) provided the donor brains.
The DNA methylation assessment was performed on batches of 167 prefrontal cortex samples with six replicates. For each of the CpG sites, their value is given by the ratio of the methylated signal to the total signal, yielding a continuous measure in $[0, 1]$. Preprocessing analysis and quality controls result in total to $p_{\text{DNAm}} = 789\,286$

variables (CpG sites). They were normalized for subsequent analysis. For more details on assessment and preprocessing, see [118].
For high-resolution metabolomic profiling, each sample was repeatedly analyzed using two complementary chromatographic methods (hydrophilic interaction liquid for polar metabolites and reverse-phase chromatography for less polar compounds). Then the detected signals were characterized by accurate mass-to-charge ratio (m/z), retention time (RT), and ion intensity, for details refer to [118] and [96]. This results in a total of 35348 features, averaged and log2 transformed.
Further, the following covariates were also included to adjust the model. Demographic and socioeconomic factors, personal data such as sex, race, educational attainment, socioeconomic status and age at death as well as the time between death and the removal of the biological samples, cf. [118].

The neuropathology of Alzheimer's disease can be assessed by a variety of measures that evaluate different aspects of disease progression. In [118] Braak staging [7], CERAD score [85] and ABC score [82] was considered. They can be described as follows. Higher Braak stages reflect more extensive disease progression [7]. The CERAD is divided into four levels, with higher scores indicating more advanced amyloid pathology, which is a primary indicator of Alzheimer's disease [85]. The ABC score provides a comprehensive assessment of the severity of Alzheimer's disease and helps to indicate the overall stage of the disease, categorizing it into the four levels none, low, intermediate, and high [82]. It is a combination of Braak and CERAD scores with the Thal phase. The Thal phase describes the spread of amyloid plaques in the brain in 5 phases [114]. For more details see [118] or the respective papers.

3 Statistical Methods in Genetics

3.1 Statistical Methods for GWAS

Statistics in genetics is a wide field. Its focus is on the application and development of statistical methods to derive conclusions from genetic data. Statistical methods can be utilized to quantify the risk of disease as well as identify statistically important risk factors [121]. Depending on the specific questions, various methods from different areas of statistics are available, such as Bayesian, frequentist, or machine learning methods. Further, there are methods that have been developed specifically for the analysis of genetic data, but which are of course also generally valid.
In general, when aiming to identify genetic variants that are statistically associated with a particular disease or phenotype, it is advantageous to consider the entire genome. From a statistical perspective, this presents a substantial challenge as well as an enormous potential for research into the optimal use of the information of millions upon millions of data points.
In classic genome-wide association studies (GWAS), tons of individual statistical tests are performed on all variables (e.g., SNPs) to identify those that are significantly associated with some outcome. In addition, the results of GWAS can be used for various applications, such as using the obtained effect sizes as weights for the calculation of (polygenic) risk scores, which will be discussed in more detail later in this section, cf. [28]. GWAS have revolutionized the field of complex disease genetics, providing many convincing associations for complex human traits and diseases. However, it is important to note that GWAS associations may detect variants and genes as significant that do not necessarily have direct biological relevance to disease [109]. Variants in linkage disequilibrium (LD) are often correlated, but do not have causal significance. It is therefore advisable to perform LD pruning before screening for significant variants [95].
From a statistical perspective, a further problem in GWAS is multiple testing. There exist methods for correcting the error in multiple testing, such as the Bonferroni [6] or the Benjamini-Hochberg correction [5]. However, the issue of 'p-hacking' cannot be disregarded, cf. [26].
In GWAS, e.g. linear or logistic regression models can be used to test for associations, depending on the outcome. In each of these models, clinical and external covariates are usually included in order to avoid stratification errors. Typically, these covariates are age, sex or BMI. However, for diseases with low prevalence, the inclusion of covariates known to affect the disease may increase the difficulty of identifying

associated genetic variants [92]. According to [120] the (linear) regression model for GWAS is given by

$$\mathbf{Y} = \mathbf{X}_j\beta_j + \mathbf{C}\gamma + r + e.$$

There $\mathbf{Y} \in \mathcal{R}^n$ is the outcome vector, $\mathbf{X}_j \in \mathcal{R}^n$ is the vector of SNP $j \in \{1, \ldots, p\}$ for all individuals and β_j the corresponding parameter vector of SNP j. Then $\mathbf{C} \in \mathcal{R}^{n\times(\rho+1)}$ is the matrix of ρ covariates (including an intercept term) and γ the corresponding parameter vector. r is the random effect that captures the polygenic effect of other SNPs and e a random effect of the residual errors. In the linear regression case it holds that

$$e \sim \mathcal{N}(0, \sigma_e^2\mathbf{I}) \text{ and } r \sim \mathcal{N}(0, \sigma_A^2\varphi),$$

with identity matrix $\mathbf{I}$ and the residual variance, measured by σ_e^2. The additive variation of the phenotype is reflected by σ_A^2 and φ is the standard genetic relationship matrix, which allows the regression of the level of phenotypic similarity on the level of genotypic similarity [120]. By using special link functions, the model can be adapted to other scenarios. The output of a GWAS analysis comprises a list of p-values and effect sizes for every variant that was tested. A common visualization is a so-called Manhattan plot. In a Manhattan plot, the strength of the association (e.g. $-\log_{10}$ transformed p-value) of each variant is represented as a dot on the y axis. The variants (SNPs) are arranged on the x axis according to their genomic position. In addition, a horizontal line marks a threshold of genome-wide significance level. Each dot above that line implies that the represented SNP is significant.
Although classical GWAS analyses have disadvantages as described above, it is also a well justified method that has already produced many important findings in research on various diseases such as Parkinson [35], type 2 diabetes [122], breast cancer [34] or blood pressure [32]. For a nice overview of the experimental workflow of GWAS we refer to [120].

In fact, there are many approaches that address the challenges of genetic analysis, such as high dimensionality and the imbalance between the number of features and the number of samples. In particular, the rapidly expanding field of machine learning (ML) methods has the potential to contribute to GWAS studies. For instance, a two-step algorithm that performs a SNP selection step based on a support vector machine (SVM) model, followed by statistical tests of the selected SNPs [78]. Also, ensemble methods such as Random Forests [9] have been shown to be well suited to identify the genetic variants that have relatively large effects in complex diseases [88, 12]. In [30] an overview of machine learning approaches to GWAS is given. However, a significant limitation of machine learning approaches in GWAS is that they are based on 'black box' models that make it difficult to interpret the underlying genetic effects, despite the existence of divergent methodologies for interpreting ML models, cf. [81].

The presence of interaction and non-linear effects in genetic data is a common occurrence, posing a further substantial challenge to statistical analysis. The incorporation of interactions (GWAIS) is also possible, however, in such a scenario, these would also have to be included in the model as input variables. Even for two-way interactions the number of potential combinations is already quadratic $\binom{p}{2} = \Theta(p^2)$.

This problem grows exponentially in the order of interactions and theoretically sums up to 2^p. Although common tree-based statistical learning methods, such as Random Forests, can be applied to different data structures and can account for interactions for prediction and in the tree fitting, they are still difficult to interpret. Furthermore, in the case where the interacting variables have negligible marginal effects, which very common in genetics [67], decision trees and Random Forests have difficulties in the detection of interaction effects [131]. A variety of approaches and modifications have been developed for tree-based models and Random Forests to enhance their capacity to effectively deal with interactions. Permutation based Random Forests [66] and iterative Random Forests [4] still use univariable splitting, while SNPInterForest [134] allows a combination of multiple variables as well as single variables for splitting. Interaction forests [50] additionally target interpretability and differentiate between quantitative and qualitative interaction effects. In [50] also an overview of alternative approaches is given, which takes interactions into account. A method developed specifically for analyzing SNP data is logic regression [98]. The purpose of logic regression is to identify the Boolean combinations of binary predictors that explain the variation in the outcome. It is widely used for the identification of gene-gene interaction effects, cf. [104]. However, logic regression is limited in number of variables and cannot include interactions between the binary predictors and quantitative predictors. Furthermore, if the signal is weak or if many predictors are actually predictive, single logic regression models tend to be unstable. Bagging logical regression models is one approach to address this problem [103], but similar to Random Forests, these models are no longer easily interpretable. A procedure for identifying response-associated interactions between binary predictors are logic decision trees (logicDT) [65], where also the possibility of including gene–environment interaction is given, while maintaining interpretability. A efficient test for gene–environment interaction is proposed in [64]. Since logicDT is limited in the number of variables it can handle, it requires a preselection of variables.

The application of many conventional variable selection methods is not feasible due to the high dimension. For instance, forward selection would miss variants with smaller effects that have larger effect sizes when interacting with other SNPs. One idea is to select the variables based on their cross leverage score (CLS) with the outcome [90]. The CLS measures the importance of the variable and contains information about a potential participation in an interaction effect, by its construction. Since all the information is contained in the score for each variable, it is not necessary to consider every possible combination. Further the CLS equals the solution of a LS problem with an error of $\epsilon > 0$. In the first paper of this thesis, this concept has been explored and enhanced [112]. A procedure was developed to make the CLS applicable to data of any size using methods of dimension reduction, such as sketching [27] as well as stream wise calculations. Subsequent to the variable selection step, methods such as logicDT can be applied to the reduced data set to identify the true interactions and main effects. The CLS are explained in more detail in the next section 3.2. This approach allows to consider genome wide data at once and its multivariate structure.

As previously indicated, rather than focusing on finding significant associations and interpreting features, the primary focus can be on risk assessment. The results of

a GWAS can be used for the calculation of risk to suffer a disease in a target cohort. Summary statistics named Polygenic Risk Scores (PRSs), defined as weighted sum of scores of the risk alleles, while the weights are given by the effect sizes of the variants, c.f. [14],

$$PRS = \sum_{k=1}^{K} \beta_k m_k$$

Alternative external weights from meta-analyses are used, cf. [13], but when no appropriate external weights are available it is recommended to use internal weights from the study population itself [52]. Due to overfitting it is important, that there are no shared individuals in calculation the weights and calculation the Polygenic Risk Scores, thus splitting the data in test and training data [28]. Common approaches to calculate the summary statistic are 'pruning and thresholding' where just a subset of SNPs, selected based on their p-values, is included [28] as well as methods based on shrinkage cf. [72], [53]. Latter could be classic regularization techniques like LASSO [116], ridge regression as well their combination Elastic Net [137]. Furthermore, it can also be viewed from a Bayesian perspective by shrinkage with prior distribution specification [37] or other ML models like Random Forests. To improve the prediction performance, it is recommended to incorporate the multivariate structure of the data in the calculation of the PRS, rather than use weights which result from univariate GWAS. Thus, in [60] an algorithm based on component-wise L_2-boosting is proposed that is also applicable to large scale data. As previously indicated, the multivariate structure is preserved in the case of variable selection by CLS. Further, the CLS correspond with the effect estimates of an LS problem to an error of the $\epsilon > 0$. Consequently, the CLS can be utilized directly for the calculation of risk scores.
To assess the PRS, we measure the phenotypic variance explained by the PRS. Therefore, we will consider two models: The full model, which includes the PRS and L covariates $Z \in \mathcal{R}^{n \times L}$

$$Y_{\text{full}} = \gamma_0 + \sum_{h=1}^{H} \gamma_h \text{PRS}_h + \nu Z$$

and the reduced model

$$Y_{\text{reduced}} = \gamma_0 + \nu Z$$

which includes only the covariates. The partial McFadden's R^2_{partial} [76] is then defined as

$$R^2_{\text{partial}} = 1 - \frac{\ln L_{\text{full}}}{\ln L_{\text{reduced}}}$$

with $\ln L(\cdot)$ the log- Likelihood of the full or reduced model. The larger R^2 is, the better the PRS is in predicting the risk.

In the following sections, the central methods that were used, extended, and developed in the course of this thesis are explained. Although the methods are also described in the articles of this dissertation, these are examined and explained in more detail in this section. Thus, it should be noted that this section will partly overlap with the theory parts in the articles.

The central focus of this dissertation is the development and expansion of the use of cross leverage scores (CLS) for the screening of potential main and interaction effects in large-scale data. The following sections will also address the methods of logic decision trees (logicDT) [65] and Random Forests [9] in greater detail, as these methods were applied in the subsequent analysis to the reduced data.

3.2 Cross Leverage Scores

The following approach was developed to overcome the challenge of handling the huge amount of data that occurs in genetics as well as including interaction effects in the analysis [112]. The fundamental objective is to implement a variable selection procedure, followed by the execution of more precise analyses on the reduced dataset. This is necessary because the entire dataset is too large to apply more specialized analytical methods. Using cross leverage scores (CLS) for variable selection was proposed in a manuscript from Parry et al. [90]. In the initial paper [112] (cf. Appendix A) of this dissertation, this idea was elaborated and extended, thus it can be applied to arbitrarily large amounts of data, where most statistical methods fail due to computational and theoretical reasons. Additionally, stream-wise calculations have been proposed to ensure that the calculation should not lead to memory problems even on standard computers. In addition, a theoretical justification for the CLS is provided.
The approach is motivated by dimension reduction methods which refer to the number of observations n. These were transferred to the dimension reduction of variables p, since in the context of genetics $n \ll p$ problems are typical. To be consistent with the notation of previous literature, where the case $n \gg p$ was treated in the following the transposed matrix is considered. The initial key point is to consider the common matrix of variables and outcome:

$$\tilde{X} = [X, y]^T \in \mathcal{R}^{\tilde{p} \times n}, \tag{3.1}$$

where $\tilde{p} = p + 1$ and p the number of variables and n the number of observations. $X \in \mathcal{R}^{n \times p}$ is the data matrix and $y \in \mathcal{R}^n$ denotes the response.
Then the CLS are given by the off-diagonal entries of the hat matrix H of $\tilde{X}$. The hat matrix H is given by $H = QQ^T$ [48], which can be obtained by the QR-decomposition $\tilde{X} = QR$ [42] and $Q = XR^{-1}$.
Since the focus is exclusively on the CLS $c_{i\tilde{p}}$ of variable $i \in \{1, \ldots, p\}$ to the response y, the dot product of the rows $Q_{i\cdot}$ (according to variable i) and row $Q_{\tilde{p}\cdot}$ (according to the response) has to be calculated:

$$c_{i\tilde{p}} = \langle Q_{i\cdot}, Q_{\tilde{p}\cdot} \rangle, \quad i \in \{1, \ldots, p\}. \tag{3.2}$$

Since Q forms an orthonormal basis for the column space of $\tilde{X}$ the CLS correspond to the leverage of individual variables on the correlation of the multidimensional subspace spanned by the data with the outcome variable. The argument is that a subselection based on CLS preserves information about the multivariate structure and is thus able to retain interaction effects [112]. So a CLS indicates the leverage of a variable on the outcome variable and in particular its participation in an interaction

effect rather than classic correlation. Then the q variables with the largest (absolute) CLS are selected. It is recommended to select $q = \lceil n \cdot \log n \rceil$ variables in [90] which is confirmed in experimental findings [112]. In theory, this is justified by the coupon collector's problem [31], which necessitates oversampling by a log n factor to ensure that the submatrix preserves the full rank n of the original matrix [117].
In [112] it is also proved that, under mild assumptions, each CLS equals its corresponding parameter in the least squares solution up to a small bounded additive error.

The obvious bottleneck is the QR-decomposition with running time $\Theta(pn^2)$ [42], which is prohibitively slow to compute for very large p. To overcome this problem in [112] several approximation approaches are proposed. First two stream wise calculations were introduced. In lieu of a single QR-decomposition for a large matrix including all p variables, a multitude of QR-decompositions are employed for a multitude of small matrices including $w << p$ variables until every single variable is included at least once or a maximum number of replicates is reached. The *Sliding Window* approach simply slides a window of fixed size w through the matrix and calculates the CLS for each window. In each step the outcome vector is attached to the actual considered variables. The fundamental procedure is described in algorithm 1. The representation of algorithm 1 assumes that p is divisible by

Algorithm 1 Sliding Window approach

```
1: function SW(X, y)
2:     b ← 1
3:     f ← w with w the predefined window size
4:     while p ≤ f do
5:         X^w ← X_[b:f]
6:         X̃^w ← cbind(X^w, y)^T
7:         CLS ← calculate CLS and store them
8:         b ← f + 1
9:         f ← f + w
10:    end while
11:    return all CLS
12: end function
```

w, which can lead to difficulties in practice. The algorithm is thus implemented in a manner that enables the interception of this issue. At this juncture, however, it is sufficient to point out the basic functionality. The primary benefit of this approach is that it necessitates the reading of only the part of the matrix that is currently part of the window into the computer program. The *Random Window* approach is predicated on a similar notion. However, in contrast to fixed windows, the variables are drawn randomly in each step. Thereby, the variables are drawn without replacement within the steps, yet over the steps with replacement. Their results are merged in a suitable way, akin to the Merge & Reduce technique [39].
In addition to these two heuristic approaches, theoretical dimension reduction methods are employed. These methods were originally developed for the reduction of observations and were adapted in the context of this work for $n << p$ problems. The objective is to construct a substantially smaller 'random sketch' of the input matrix

X. This reduction in size of p to $r << p$ facilitates the calculation of complex matrix operations. The columns of X are projected to a lower r-dimensional subspace by multiplying a sketching matrix $\Pi \in \mathcal{R}^{r \times \tilde{p}}$ with $\tilde{X} \in \mathcal{R}^{\tilde{p} \times n}$,

$$\Pi \tilde{X} = \tilde{X}_\star \in \mathcal{R}^{r \times n}. \tag{3.3}$$

Subsequently, the QR calculation is performed for $\tilde{X}_\star$, thereby overcoming the problem of high dimensionality. For a detailed exposition of the single steps of the approximation of CLS, see algorithm 2. This procedure allows us to approximate

Algorithm 2 Approximation of the CLS of $\tilde{X}$

Require: $\tilde{X} \in \mathbb{R}^{\tilde{p} \times n}$ $(\tilde{p} = p + 1)$
Ensure: $\hat{c}_{i\tilde{p}}, i \in \{1 \ldots, p\}$
1: Project $\tilde{X}$ to a lower r-dimensional subspace to obtain $\tilde{X}_\star = \Pi \tilde{X} \in \mathbb{R}^{r \times n}$ using, e.g., $r = \frac{n \cdot \log n}{\epsilon^2}$ [18]
2: Compute the QR-decomposition $\tilde{X}_\star = Q_\star R_\star$
3: Compute $\Omega = \tilde{X} R_\star^{-1}$ where $\Omega \in \mathbb{R}^{\tilde{p} \times n}$
4: Compute the CLS: $\hat{c}_{i\tilde{p}} = \langle \Omega_{i\cdot}, \Omega_{\tilde{p}\cdot} \rangle$

the CLS of arbitrarily large data within an arbitrary precision parameterized by $\epsilon > 0$ [27]. The sketching matrix that is used in the following step is based on a further development of a Clarkson-Woodruff embedding [17]. The sketching matrix according to Clarkson-Woodruff is a sparse matrix, where the fixed number of nonzero elements ($\{-1, 1\}$) depends on the target dimension r of the embedding. The choice of r is a trade off between the sparsity of the target dimension and approximation error ϵ. Here $r = \frac{n \cdot \log n}{\epsilon^2}$ proposed by [18] is used. For a more detailed explanation of the functionality of the CLS, refer to the first article in this thesis [112] in Appendix A. However, it should be noted that alternative sketching methods such as ε-JLT [57], the Rademacher sketch [16], and the Subsampled Randomized Hadamard Transform [2] are possible alternatives, which allow for different parameterizations in the trade-off between the time needed to multiply the sketch and the required number of rows. An overview of pssible sketching approaches as well as experimental comparisons is given in [38]. However, other approaches neither achieve a better accuracy nor a lower target dimension nor faster running time than the sparse Cohen-Sketch thus we focus on in the following.

3.3 Random Forests

Despite the disadvantages associated with Random Forests [9], as mentioned above, it is still a widely used method that can be applied to high-dimensional data due to its great flexibility. Thus, it is also utilized in two articles [112, 118] of this thesis, see Appendix A for comparisons and validation, and is therefore described in the following.

Random Forests is an ensemble method that consists of multiple classification or regression trees. The concept of a regression or decision tree is to divide the feature

space appropriately and adapt simple models in the respective areas. A decision tree (or regression tree) $\mathcal{T}$ is generally defined as a function

$$f_{\mathcal{T}} : \mathcal{X} \to \mathcal{Y},$$

given some data $\mathcal{D} = \{(x^{(i)}, y_i)\}_{i=1}^n$ with input variables $x^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)}) \in \mathcal{X}$ and an outcome $y \in \mathcal{Y}$. For $\mathcal{Y} \subseteq \mathbb{R}$ the regression case is given and for $\mathcal{Y} = \{0, 1\}$ the classification case.
The CART algorithm [10] is widely used to construct such a tree. Starting from a so-called *root node* containing all variables, the dataset is split into two *child nodes* using an appropriate (optimal) variable (*splitvariable*) and point (*splitpoint*) with the aim of reducing the prediction error. In the CART algorithm randomly $m \leq p$ potential *splitvariables* and their respective possible *splitpoints* are considered. The hyperparameter m is called *mtry*. This procedure is then repeated in the child nodes, until a predefined stopping criterion is reached, e.g. that the prediction performance cannot be improved by further splitting, or that a minimum node size is reached. Nodes that are not split further are called *end nodes*. The procedure is a greedy search, where locally optimal solutions emerge that converge to a globally optimal solution in a reasonable time. The larger the tree, i.e. the finer the subdivision of the feature space, the better the fit to the data. However, the greater the complexity, the greater the tendency of such a tree to overfit.
Using this tree for prediction, for a new observation $y_i^\star$, an assignment to one of the end nodes is made based on the input variables and the rules in the nodes of the tree. In the regression case the prediction for $y_i^\star$ is then usually made by averaging of the observations of the training data in that node. And, in the case of classification, the prediction is based on the class that predominates in that particular node.

In lieu of one single tree, Random Forests consider multiple trees with the incorporation of Bagging [8]. In the regression case, a prediction is then derived from the mean prediction of all trees in the ensemble. And, in the classification case, a new observation is assigned to the class to which it was assigned by the majority of the trees (majority vote). The idea of Bagging is that the trees in Random Forests are fitted based on Bootstrap Samples, i.e. on a sample of size n that is drawn with replacement from the data [29], to increase the tree independence. Furthermore, the trees are less correlated with each other the smaller the number of candidate *split variables* m is. A small m means that not all trees will be too similar if one or more features have a very large influence on the prediction. In the ensemble, the variance of the predictions is thus reduced. This can be proven as follows.
Let $\hat{Y^i}$ the prediction of observation i resulting from the average of the predictions of the B individual trees of the ensemble:

$$\hat{Y^i} = \frac{1}{B} \sum_{b=1}^{B} \hat{Y_b^i}. \tag{3.4}$$

Then, according to standard variance rules, the variance of the prediction is given by:

$$\mathrm{Var}(\hat{Y}^i) = \mathrm{Var}\left(\frac{1}{B} \sum_{b=1}^{B} \hat{Y}_b^i\right) = \frac{1}{B^2} \mathrm{Var}\left(\sum_{b=1}^{B} \hat{Y}_b^i\right)$$

With $\text{Var}(X_1 + \cdots + X_L) = \sum_{l=1}^{L} \text{Var}(X_l) + 2\sum_{1\leq r<s\leq L} \text{Cov}(X_r, X_s)$ and X_l for $l = 1, \ldots, L$ dependent random variables it follows:

$$\text{Var}(\hat{Y}^i) = \frac{1}{B^2}\sum_{b=1}^{B} \text{Var}\left(\hat{Y}^i_b\right) + \frac{2}{B^2}\sum_{k\neq l} \text{Cov}\left(\hat{Y}^i_{b_k}, \hat{Y}^i_{b_l}\right)$$

Assuming that the variances of the individual trees are equal with a value of σ, it follows

$$\begin{aligned}\text{Var}(\hat{Y}^i) &= \frac{1}{B^2}\sum_{b=1}^{B}\sigma + \frac{2}{B^2}\sum_{k\neq l} \text{Cov}\left(\hat{Y}^i_{b_k}, \hat{Y}^i_{b_l}\right) \\ &= \frac{\sigma}{B} + \frac{2}{B^2}\sum_{k\neq l} \text{Cov}\left(\hat{Y}^i_{b_k}, \hat{Y}^i_{b_l}\right)\end{aligned}$$

It is trivial that the first term of the sum decreases for increasing B. Further, the fundamental principle of the ensemble is that the trees are constructed in such a way that they are uncorrelated, i.e. we assume that the covariance of the estimates of two trees is (almost) zero. So also the second term decreases for a large number of trees B. Therefore, for $B \to \infty$ the variance would even tend to 0, but only if the trees are actually uncorrelated. In reality, of course, the trees are never completely uncorrelated, but this property is retained even if the trees are weakly correlated. For calculations the `R` package `ranger` is used [130].

For a tree, the observations that were not used for fitting are called Out of Bag Observations (OOB). These can be used to assess the prediction performance. Therefore, the OOB error results from the prediction for an observation i only being calculated on the basis of the trees of the ensemble for which this observation i was not used for construction, because it was not included in the bootstrap sample for this tree, i.e. 'was OOB'. It can be shown that with a sufficiently large number of trees, i.e. a large number of bootstrap samples, the OOB error is practically equivalent to the leave-one-out cross-validation error [56], but much more efficient to determine. Further, the OOB observations can be used to calculate variable importance measures.

3.4 Logic Decision Trees

To detect the explicit important factors and interactions in the reduced data set, Logic decision trees (logicDT) [65] is well suitable. It is a consistent tree-based method which is aimed to identify response-associated interaction and combines the strengths of other tree based methods while maintaining interpretability. It is possible to identify important influencing factors and their interactions between binary as well as taking continuous predictors into account.
LogicDT overcome the in section 3.1 described difficulties of other tree based methods. The direct inclusion of conjunctions as input variables and specially developed variable importance measures provide an intuitive interpretation of the

model. Thus, the important factors and their influences can be obtained directly from the model. The aim is to estimate $\mathbf{E}_{(\boldsymbol{X},Y)}[Y|\boldsymbol{X} = \cdot]$ by a function

$$\varphi : \mathcal{X} \to \mathcal{Y}$$

with $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathcal{R}^p$ the vector of binary input variables with $\boldsymbol{X} \subset \mathcal{X}$, $\mathcal{X} = \{0, 1\}^p$, $\boldsymbol{Y} \subset \mathcal{Y}$ the outcome variable and $\mathcal{D}$ a trainings data set defined as $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with independent and identically distributed observations from the joint probability distribution of $(\boldsymbol{X}, \boldsymbol{Y})$. The basic procedure of logicDT can be described as follows, for details refer to the original article [65].

Both, single input variables as well as interactions of variables are named *terms* and a set of terms is called *state* s. Then, only the terms of the state s are used as input variables for fitting a conventional decision tree. For this the training data set $\mathcal{D}$ has to be transformed in a *tree training data set* $\mathcal{D}_s$ consisting of the terms of state s and the outcome. The best local node splits identified with conventional node impurity splitting criterion and the performance of the tree is then evaluated according to the state s by calculating a *score* (e.g. the MSE for regression tasks), that measures the training data error. The ideal state is identified by simulated annealing over all appropriate states. This is done by evaluating the tree for a state s, construct a modified state s' (called *neighbor*) and fit a new tree based on s' which is then also evaluated. This is repeated until the optimal state is found. The modification of a state are slight modifications ('Exchanging or negating' and 'Adding or removing' single variables or 'Adding or removing logic terms') of the current state, avoiding tautologies and uninformative terms. The state modification is randomly drawn from a uniform distribution over all possible state modifications of the current state. If a new score $\text{Score}_{\text{new}}$ for a modified state is smaller than the current smallest score $\text{Score}_{\text{min}}$ the current state is updated with the acceptance probability defined by the simulated annealing process. The initial state consists of a single input variable that minimizes the score function, or alternatively can be chosen randomly or empty and a while loop is carried out until the optimal states is obtained (or a stopping criteria becomes true). The procedure is summarized in algorithm 3 and it is proven that the search with simulated annealing asymptotically leads with probability 1 to a globally optimal state [65].

In logicDT, continuous relationships are not included in the decision tree fitting process by including continuous variables as split variables, however they can be included by fitting regression models in the leaves of the tree. By fitting these exclusively using the binary terms that previously resulted from the splits, it is possible to uncover interactions between the binary predictors and a quantitative covariate [65].
So given the presence of continuous covariables, a likelihood-ratio test with the following test statistic is used as split criterion for fitting the tree.

$$-2\log(\Delta) := -2\log(\frac{L_{\text{reduced}}}{L_{\text{full}}}) \tag{3.5}$$

This test statistic is asymptotically χ^2 -distributed with 2 degrees of freedom under H_0, cf. [65]. Under H_0 we assume that a further splitting of the node due to a binary variable X_s leads to no different prediction models in the current tree branch.

Algorithm 3 logicDT Fitting

function LOGICDT($\mathcal{D}$ *(Training data)*)
  $s \leftarrow$ Initialize state/set of terms
  $\mathcal{D}_s \leftarrow$ Apply s to $\mathcal{D}$
  $T \leftarrow$ FITDECISIONTREE($\mathcal{D}_s$), see Algorithm 1 in [65]
  $\text{Score}_{\text{min}} \leftarrow \text{Score}(T)$
  while Global search is not finished **do**
    $s' \leftarrow$ Modify current state s
    $\mathcal{D}_{s'} \leftarrow$ Apply s' to $\mathcal{D}$
    $T' \leftarrow$ FITDECISIONTREE($\mathcal{D}_{s'}$)
    $\text{Score}_{\text{new}} \leftarrow \text{Score}(T')$
    if State s' is accepted based on $\text{Score}_{\text{min}}$ and $\text{Score}_{\text{new}}$ **then**
      $s \leftarrow s'$
      $T \leftarrow T'$
      $\text{Score}_{\text{min}} \leftarrow \text{Score}_{\text{new}}$
    end if
  end while
  return (s, T)
end function

L_{reduced} and L_{full} denotes the maximized likelihood of the reduced and full model. For a given node, this test is performed on all possible splits, thus enabling the splits to be ranked according to their p-values. The split with the smallest p-value below a predefined threshold is selected. This threshold can be seen as a stop criterion in algorithm 3. If there is no split below this threshold, no further splitting is performed.

LogicDT can be enhanced to an ensemble by using Bagging, called Bagged logicDT [65][1]. Since the combination of Bagging and simulated annealing in the respective models is very computationally intensive, it is recommended to use a greedy search. Because the main problem of greedy search approaches of getting stuck in a local optimum could be compensated by looking at different subsets of the training data set and the associated stabilization of the model. In the greedy search the same state modification used as with simulated annealing but the best modified state is chosen deterministically.
The out-of-bag observations (oob-observations) can be used for a unbiased and stabilized estimation of the generalization error and variable importance measures (VIMs). A VIM typically result from the difference between the prediction error of the full model and the prediction error of the model without the input variable of interest. Various versions of VIMs can be considered in logicDT. The *permutation VIM* [9], where the values of the input variable of interest is permuted, the *removal VIM* [77], where the respective variable is removed and *LogicVIM* [65] for binary predictors, where each possible predictor setting of the input variable of interest is considered equally to generate a prediction without knowledge about the variable of interest. In classical procedures only the importance of individual input variables is taken into account. In LogicDT, all terms, including conjunctions of variables, are

[1]In the following, logicDT is written instead of bagged logicDT to improve readability.

treated as individual input variables, so the importance of terms can be determined. If a variable itself has a strong effect only the main effect it is included in the logicDT model, regardless the variable is also involved in an interaction with a strong effect. To overcome this issue adjustments to the VIM methods are required. *Interaction VIM* [65] estimates the interaction importance by reducing the full model by multiple variables that contained in the regarded interaction at once resulting in a joint VIM. Therefore it is possible to measure the importance of specific conjunctions which are identified by logicDT. All three VIM approaches mentioned above can be used in conjunction with these adjustment for interaction effects.
For details on the methodology of logicDT and hyperparameter tuning to control the complexity or avoid overfitting refer to [65]. The method is implemented in the R package `logicDT` [63].

4 Summaries of the Articles

4.1 Article 1: Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores

The manuscript *Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores* was published in the Biometrical Journal in 2024 [112].

One aim of this dissertation is the development of methods to derive insights from high-dimensional data. Thereby, the focus is on the detection of important variables, in particular interaction effects, rather than on the optimization of the prediction. Due to that conventional statistical methods are either difficult to interpret or limited in the number of possible input variables, the goal of the first article is to overcome this research gap and to develop a variable selection method that also takes accounts for interactions, uses the entire data, and maintains interpretability.

The method was motivated by investigating the influence of single-nucleotide polymorphisms (SNPs) and their interactions on health outcomes, which is a $p \gg n$ problem. However, this approach can also be applied to other data and research questions, as can be seen in the other two papers in this thesis [113, 118].
The idea is to use Cross Leverage Scores (CLS) for variable selection, based on an article by Parry et al. [90]. The article gives a mathematical justification of the new methodology and verifies the functionality in simulation studies as well as in a small real data example.
The CLS measure each variable's leverage on the correlations with the multidimensional space spanned by the data with the outcome variable rather then ordinary correlations of single variables. Therefore, the CLS of each variable directly contains information about its importance, as well as the importance of a possible interaction involving that variable, without the need to consider every possible combination. This is true even if there is an interaction that consists only of main effects that are themselves negligibly small.
In order to extend this method to the context of high-dimensional data, it is necessary to employ approximation approaches, due to the computational complexity of the QR decomposition, see [112] or section 3.2 for details. The following three approaches are proposed for approximating the CLS: The *Sliding Window* and *Random Window* based on a window wise calculation and the *Sketching* approach

based on subspace embeddings. The utilization of these approximation approaches enables the implementation of variable selection based on CLS for arbitrarily large data sets.

The fact that important variables which are part of interaction effects can be detected by CLS has been demonstrated in several simulation studies in this article, e.g. one important 2-way interaction and nearly $2 \cdot 10^6$ 'noisy' variables and an binary outcome. In addition, it is shown that the selection by CLS has a positive effect on the prediction with Random Forests [9] as well as the detection of important factors with logicDT in the second step. The well-known HapMap dataset [115], [84] was considered for the first aspect, and simulated data was considered for the latter. In addition, the CLS is demonstrated to be more meaningful than ordinary correlations in a toy example. Moreover, a theorem demonstrates that each CLS is equal to its corresponding parameter in a least-squares solution up to a small bounded additive error.

The paper proposes a two-step procedure, as the CLS merely indicates which variables are significant, but does not specify which variables interact with each other. Subsequent to the selection of these variables according to their CLS, further analysis should be applied to the reduced data. In this context, logic decision trees, as outlined in [65], are particularly well-suited for identifying interactions but are constrained in their capacity to process a large number of variables.
The corresponding R-code is available on a GitHub repository [111] and the simulated data sets on the Zenodo platform [110].

4.2 Article 2: Using Cross Leverage Scores for Detecting SNP-Environment Interactions Effects on COPD

The manuscript *Using Cross Leverage Scores for Detecting SNP-Environment Interactions Effects on COPD* [113] was submitted to the Genetic Epidemiology Journal in the beginning of April 2025.

After the development of the variable selection method based in the first paper [112], these methods are applied to the practical question of how genetic (SNPs) and environmental factors and their interactions affect lung function, namely chronic obstructive pulmonary disease (COPD). For this task the data of the SALIA cohort study [102] is considered, which involves the examination of older women residing in the Ruhr area and adjacent Münsterland. Since COPD was the third leading cause of death worldwide in 2020 [86], a better understanding of the effects of risk factors for COPD is highly relevant. In addition, early detection of risk factors can generally prevent suffering or save on expensive treatment, as is the case with COPD [41]. COPD has been demonstrated to be influenced by both genetic and environmental factors [93], as well as their interaction [1]. In general, it is recommend to include known harmful environmental exposures to identify interacting genetic loci [136].

From the SALIA study, a total of 7643653 SNPs and 7 clinical/environmental factors such as smoking status, BMI and particulate matter ($PM_{2.5}$) are available for $n = 503$ women.
As mentioned above, it is very difficult to detect interaction effects in such large data sets. Therefore, it is recommended that the proposed two-step procedure [112] be utilized, which comprises the following steps:
Initially, a variable selection by CLS is applied using sketching to exploit the entire dataset. Subsequently, the application of bagged logicDT [65] to the reduced data is applied after some further processing steps. The important factors and their interactions are then identified by variable importance measures obtained from the logicDT model.

A comprehensive literature review is conducted in the paper itself to assess the plausibility of the identified effects. The list of important effects is intended as a starting point for biologists. In order to achieve a comprehensive understanding of the actual relationship between these factors and COPD, it is necessary to conduct a series of biological experiments. Furthermore, a comprehensive literature review was conducted on the factors selected by variable selection with CLS to evaluate the practical applicability of the developed method.

On the one hand, the literature review in the article shows that the proposed procedure, based on a variable selection step, works on real large data sets and is well suited for gene-environment analysis. On the one hand, this manuscript provides candidate genetic and environmental factors and their interactions that influence the risk of COPD, based on a whole-genome view that is only possible using the new methods. The utilization of the innovative methods is intended to provide new insights into the aetiology of COPD, which could result in significant biological benefits.

4.3 Article 3: Development and Application of Brain Tissue Based Multi-Omics Profile Scores for Alzheimer Disease

The third paper *Development and Application of Brain Tissue Based Multi-Omics Profile Scores for Alzheimer Disease* [118], was co-authored with Timur Tug, who was the first author. This paper has been submitted to Alzheimer's & Dementia: The Journal of the Alzheimer's Association end of March 2025.

After the application of CLS to the analysis of SNP data [112] and SNP-environment data in relation to COPD [113], it is applied to other omics data (DNA methylation and metabolomics) in relation to the neuropathology of Alzheimer's disease (AD) in this paper. By analyzing genome-wide DNA methylation and metabolomics data from brain tissue jointly, it is expected to provide new insights into the biological mechanisms and risk of Alzheimer's disease. Although it is

recommended to consider omics data of different types together [73], this is usually done only individually.

For this purpose novel single and multi-omics profile scores (PS) are developed. (In contrast to the standard literature, the term profile score is used here instead of risk score because the analysis is based on tissue from dead patients. However, the interpretation is equivalent.) The profile score is defined as weighted sum of selected features m_k

$$PS = \sum_{k=1}^{K} \beta_k m_k, \tag{4.1}$$

while the weights β_k for $k = 1, \ldots, K$ correspond the parameters of a regression model. The cross leverage scores were used for the calculation of the weights among various methods. This exploits the property of the CLS that they equal to its corresponding parameter in a least-squares solution up to a small bounded additive error [112]. Therefore, in this paper, the CLS is employed not only on a different type of data, but also for a different purpose. However, also the variable selection step is directly included, since only the variables with the most extreme score are included in the weighted sum. Given the high dimensionality of multi-omics data analysis, the application of CLS is particularly appropriate, since they retain the multivariate structure of the data, which is strongly recommended, cf. [60]. In the present paper, the approximation of the CLS using the sliding window approach was chosen, due to its straightforward application. For further analysis, it may be better to use sketching because there are theoretical guarantees for it.

Then, based on the profile scores, different models are considered. First, models that are based on single-omics PS of DNAm and metabolome data separately with some covariates Z:

$$Y \sim PS_{\text{DNAm}} + Z$$
$$Y \sim PS_{\text{metabolome}} + Z$$

Based on multi-omics PS of the common data:

$$Y \sim PS_{\text{DNAm+metabolome}} + Z$$

Based on jointly PS of both with and without interaction:

$$Y \sim PS_{\text{DNAm}} + PS_{\text{metabolome}} + Z$$
$$Y \sim PS_{\text{DNAm}} + PS_{\text{metabolome}} + PS_{\text{DNAm}} * PS_{\text{metabolome}} + Z$$

The paper validates the various calculations as well as the different models of the profile scores with partial McFadden's R^2 [76] underlying the recommendation of considering the Profile Scores jointly. In addition, a series of pathway analyses were conducted for a better understanding of the biological relations.
In this way, new insights into the development of Alzheimer's disease have been gained. We see that in general PS_{DNAm} shows better predictive performance than $PS_{\text{Metabolome}}$ but it is always better to consider multi omics PS to improve prediction performance.

5 Conclusion & Outlook

The objective of this dissertation was to develop variable selection methods for regression models that are capable of detecting interaction effects, are applicable to very large data sets as well as maintain interpretability. From a practical perspective, these approaches can then be applied to genome-wide analyses where several million variables occur. For example, to identify important genetic and environmental factors and their interactions associated with health outcomes. This understanding can then provide the basis for risk reduction strategies. Of course, the statistical methods can be applied to other genetic questions, but also to completely different research fields. Nevertheless, it is imperative to deliberate on the requisites from both domains and thereby bridge the gap between theory and application. This thesis results in a selection of variables on the basis of so-called Cross Leverage Scores (CLS). These scores can be estimated in various ways and applied to different research questions.

In the first article of this dissertation, *Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores*, the methodology of the cross leverage scores is introduced. The article developed a variable selection method for interactions. This method was intended for investigating the influence of single-nucleotide polymorphisms (SNPs) and their interactions on health outcomes. This is a typical $n << p$ problem with p being the number of variables and n the number of observations. Most available statistical methods fail due to dimensionality or are difficult to interpret in this setting. The issue is exacerbated further by the consideration of interactions. The total number of all possible combinations sums to 2^p. Given that p in genetic analyses is measured in the millions, the sheer volume of data alone is beyond the processing capacity of most computers. The CLS overcome these issues. Using these scores, it is not necessary to consider every possible interaction between variables individually. Due to their construction, the CLS are directly linked to the importance of a variable also in the sense of an interaction effect. The CLS of each variable correspond to the leverage of the individual variable on the correlation with the multidimensional subspace spanned by the data with the outcome variable. Thus, a subselection by the CLS carries the information on the multivariate structure in the data, and is capable of retaining interaction effects. However, the calculation of the scores using a QR-decomposition is still time-consuming for large scale data. Thus, three approximation are proposed. For the *Sliding Window* and *Random Window*, the data is divided into consecutive windows respectively random batches of variables. The calculation is then done window wise respectively batch wise. For these two approaches there are no theoretical guarantees, but the simulation studies

in the first paper show appropriation for performing the calculations efficiently. Further, the *Sketching Approach*, based on random projections and subsampling approximates, the CLS within an arbitrary precision parameterized by $\epsilon > 0$. Using the approximation methods the CLS can be applied to arbitrarily large data. In simulation studies and on a small real data example, it is shown that the CLS are appropriate to distinguish between important and unimportant SNPs and the variables that are involved in important interactions can be detected. However, the CLS only indicate which variables are important and may be involved in an interaction rather than show the explicit interactions. Therefore, a two step procedure is proposed. A more sophisticated method, such as logic regression [98] or logicDT [65], which are not readily applicable for high-dimensional data but good in finding interaction effects in scenarios with less variables p, should be applied to the reduced data.
It is, of course, also possible to apply alternative methods to the reduced data. For instance, the method Block Forest [51] (a variant of Random Forest), was not included in any of the three papers, yet it has demonstrated strong performance in ancillary analyses. In the context of gene-environment interactions, there is typically an abundance of single-nucleotide polymorphisms (SNPs), yet a paucity of clinical and environmental analyses. Nevertheless, it is advisable to incorporate clinical data into the fitting process. Since classic Random Forests does not take any existing block structure in the data into account, it is very likely that variables from a small block will not be used to build the tree, if the block sizes are very different. Block Forests modify the split point selection to incorporate such a block structure in the data by a randomized block choice in the split point selection incorporating block-specific weights. In multi-omics experiments, where such a block structure is plausible, this method ensures that important environment or clinical factors are not misplaced among the extensive genetic data. The incorporation of such block modifications into logicDT could be an interesting area for future research.

In the second article [113], the two-step procedure was conducted for this purpose of investigating the effects of gene-environment interactions on COPD within the SALIA cohort study [102]. Given the large number of genome-wide SNPs (more than 7.5 million) in this study, the variable selection step is required. In order to take into account the information of the entire genome and its multivariate structure as well as possible interaction effects, the CLS are used for selection. These are approximated by sketching. As is the case for other diseases, it is imperative to study the interactions of SNPs as well as the interactions of SNPs with environmental factors in COPD [71, 1]. However, due to the large amount of data, most methods fail due to the large number of variables. It is hypothesized that the selection with CLS will provide the information needed to identify important interactions in the second step. Of particular interest in this work are interactions between SNPs and environmental factors. Since logicDT is capable of identifying these [65], this method will be used in the second step. The procedure itself as well as its constituent elements represent a novel approach that has not yet been utilized in this particular manner on genome-wide datasets. Therefore, the resulting candidate set of important factors and their interactions could help biologists better understand the development of COPD or even uncover novel risk factors. A comprehensive review of the extant literature was conducted to ascertain the plausibility of the associations that were identified. For the majority of the identified associations, plausible explanations were found

in the literature. However, genes were identified in connection with environmental factors for which no evidence was found in the literature. Consequently, these novel associations may represent previously uncovered risk factors for COPD. However, to fully comprehend the implications of these findings, further investigation through biological experimentation are necessary.
The fact that so many of the identified associations have plausible explanations in the literature indicates that the novel approach and its steps are meaningful. Moreover, the implementation of this method is both fast and efficient, and, unlike traditional GWAS analyses, it utilizes the entire dataset simultaneously.
A further literature search for the 50 factors with the most extreme CLS also showed that plausible reasons could be found for most of them. This finding supports the appropriateness of the variable selection by CLS. The present article thus provides a justification for the statistical methods, as well as promising results for the application in biology. However, the results can probably be improved by optimizing the hyperparameters of each method. Further research is necessary to determine the impact of such to an optimization of the results.

In a further article [118], CLS was applied for the first time to metabolomic and DNA methylation data. In this study, the CLS were employed to calculate various (multi-omics) profile scores to assess the risk for Alzheimer disease. Given the high dimensionality of the data, the CLS had to be approximated in this instance as well. In this article, the *Sliding Window* approach has been used for approximation. For this purpose, the CLS itself are of interest as these are used directly as weights for the calculation of the profile scores. Since the *Sliding Window* approach has no theoretical guarantees, it may be preferable to employ an approximation with sketching. However, if the *Sliding Window* approach is used exclusively for variable selection, it provides similar results as the true scores, since only the order of the variables is of interest. As it is plausible that discrepancies from the true CLS will cancel each other out, since the weights are relative to each other, the window approach is also appropriate here.
The paper compares different profile scores based on different ways to calculate them. It has been shown that Pruning and Thresholding and respectively Random Forests are best suited for the individual scores. A joint consideration results in a slight improvement of the model. The results of CLS are compatible to widely used methods such as elastic net. However, the calculation using CLS is by far the fastest of all the methods employed and accounts for a multivariate structure in the data. Since the latter is recommended, other methods with this property, such as component-wise L_2-boosting [60], could be considered for comparison in future analyses.

In large scale scenarios, statistical methods are usually limited by the number of variables they can effectively process or are difficult to interpret. In order to overcome the research gap, a variable selection method was developed that takes into account interactions and preserves multivariate structures in the data. The methodology can be applied to data of any size due to the approximation methods for the CLS, which were also developed in the course of this dissertation. The development of these methods was driven by the aim to provide answers to a variety of questions related to genetics, with the goal of improving the understanding of the underlying

relationships between genetics and disease. The underlying functionality of this novel approach is mathematically justified and has been demonstrated through its application in three articles. The method is flexible and can be applied to different types of data as well as to different outcomes, making it suitable for answering different questions. Table 5.1 shows which method was used for which type of data and for which purpose. It also shows the article in which the analysis was carried out. It is obvious that the application of these approaches goes beyond the field of biology.

Data	**Y**	**What?**	**How?**	**Aim**	**Source**	**Where?**
SNP	binary	CLS	exact, SW, RW, sketch	Can we detect the factors involved in key interactions?	simulated	Article 1 [112]
		CLS + logicDT	exact	Can we detect the factors involved in key interactions?	simulated	Article 1 [112]
		CLS + RF	exact	Can we improve prediction performance?	HapMap [115]	Article 1 [112]
SNP+E+C	continuous	CLS + logicDT	sketch	Detect important factors and interactions for COPD	SALIA [102]	Article 2 [113]
DNAm+ metab. + C	ordinal	CLS	SW	Calculation of Profile Scores for AD	Alzheimer [118]	Article 3 [118]

Table 5.1: Overview. abbreviations: SNP (Single Nucleotide Polymorphism), CLS (Cross Leverage Score), SW (Sliding Window), RW (Random Window), logicDT (logic decision trees), RF (Random Forests), E (environmental factors), C (other covariates), COPD (chronic obstructive pulmonary disease), DNAm (DNA methylation), metab. (metabolomic data), AD (Alzheimer Disease)

However, there are still open aspects that require further research. It is imperative that further investigation be dedicated to the mathematical justification of CLS and the underlying mechanisms that explain their efficacy. It must be proven that the CLS of each variable indeed correspond to the leverage of the individual variable on the correlation with the multidimensional space spanned by the data with the outcome variable. This proposition is still quite heuristic as the meaning of the rows of the Q matrix from the QR-decomposition not yet fully proven.
So far, the focus has been on scenarios with a small number of observations n. A large n does not alter the interpretation of the CLS; nevertheless, it complicates the calculation, akin to the effect of large p. The extant literature provides solutions for the scenario in which the number of observations n is (also) large. In order to accomplish this, it is necessary to incorporate an additional sketching step into the sketching algorithm, cf. [27].
Furthermore, it is reasonable to dedicate research efforts to the newly developed approximation methods *Sliding Window* and *Random Window* approach. These have demonstrated efficacy in both simulation studies and practical applications. However, there remains a lack of mathematical validation regarding the accuracy of the approximation of the CLS. Moreover, it is crucial to determine which approximation method is most appropriate in which cases. While sketching is justified by mathematical proofs, the *Sliding Window* approach is characterized by an intuitive functionality.
In order to achieve a comprehensive impact within the application, it is necessary to

ensure that the methodology is both straightforward and readily available. Currently, the R-Code is freely available on GitHub. However, it is advantageous to make it accessible in a R-Package or Shiny App and to implement it in other programming languages, such as Python.
In addition, the effect of the variable effects on the models in the second step should be further investigated. More precisely, the question of whether certain approximation methods are particularly suitable for certain analyses, or how the choice of parameters can be optimized, remains to be investigated. Furthermore, as previously stated, there are still issues with analyzing genetic data sets. that require investigation for subsequent analysis. These require further investigation. Thus, there is an imbalance between genetic and environmental factors. Block Forest, for instance, has been demonstrated to be an effective solution to this imbalance. One potential strategy could be to adopt the idea of Block Forest, which has recently been developed specifically for the analysis of genetic and environmental variables. Finally, it is necessary to examine how the concepts developed in this work function in other contexts. It is advantageous to utilize the ability of being able to consider arbitrarily large amounts of data while still taking interactions into account. In the field of genetics, in particular, it is expected that this approach will yield novel insights if multiple omics layers are incorporated into the analysis.

Bibliography

[1] Agustí, Á., Melén, E., DeMeo, D. L., Breyer-Kohansal, R., and Faner, R. "Pathogenesis of chronic obstructive pulmonary disease: Understanding the contributions of gene-environment interactions across the lifespan". In: The Lancet Respiratory Medicine 10.5 (2022), pp. 512–524. DOI: 10.1016/S2213-2600(21)00555-5.

[2] Ailon, N. and Liberty, E. "Fast Dimension Reduction Using Rademacher Series on Dual BCH Codes". In: Discret. Comput. Geom. 42.4 (2009), pp. 615–630. DOI: 10.1007/s00454-008-9110-x.

[3] Barker, W., Luis, C., Kashuba, A., Luis, M., Harwood, D., Loewenstein, D., Waters, C., Jimison, P., Shepherd, E., Sevush, S., Graff-Radford, N., Newland, D., Todd, M., Miller, B., Gold, M., Heilman, K., Doty, L., Goodman, I., Robinson, B., and Duara, R. "Relative Frequencies of Alzheimer Disease, Lewy Body, Vascular and Frontotemporal Dementia, and Hippocampal Sclerosis in the State of Florida Brain Bank". In: Alzheimer Disease & Associated Disorders 16 (Oct. 2002), pp. 203–12. DOI: 10.1097/00002093-200210000-00001.

[4] Basu, S., Kumbier, K., Brown, J. B., and Yu, B. "Iterative Random Forests to discover predictive and stable high-order interactions". In: Proceedings of the National Academy of Sciences 115.8 (2018), pp. 1943–1948. DOI: 10.1073/pnas.1711236115.

[5] Benjamini, Y. and Hochberg, Y. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: Journal of the Royal Statistical Society: series B (Methodological) 57.1 (1995), pp. 289–300.

[6] Bonferroni, C. "Teoria statistica delle classi e calcolo delle probabilita". In: Pubbl. del R ist. superiore di scienze economiche e commericiali di Firenze 8 (1936), pp. 3–62.

[7] Braak, H. and Braak, E. "Neuropathological stageing of Alzheimer-related changes". In: Acta Neuropathologica 82.4 (1991), pp. 239–259. DOI: 10.1007/BF00308809.

[8] Breiman, L. "Bagging predictors". In: Machine Learning 24 (1996), pp. 123–140. DOI: 10.1007/BF00058655.

[9] Breiman, L. "Random Forests". In: Machine Learning 45.1 (2001), pp. 5–32.

[10] Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. Classification and Regression Trees. 1st. Chapman and Hall/CRC, 1984. DOI: 10.1201/9781315139470.

[11] Brookes, A. J. "The essence of SNPs". In: Gene 234.2 (1999), pp. 177–186. ISSN: 0378-1119. DOI: 10.1016/S0378-1119(99)00219-X.

[12] Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., and Van Eerdewegh, P. "Identifying SNPs predictive of phenotype using Random Forests". In: Genetic Epidemiology 28.2 (2005), pp. 171–182. DOI: 10.1002/gepi.20041.

[13] Che, R. and Motsinger-Reif, A. A. "Evaluation of genetic risk score models in the presence of interaction and linkage disequilibrium". In: Frontiers in Genetics 4 (2013), p. 138. DOI: 10.3389/fgene.2013.00138.

[14] Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. "Tutorial: a guide to performing polygenic risk score analyses". In: Nature Protocols 15.9 (2020), pp. 2759–2772.

[15] Chuang, L.-Y., Lin, Y.-D., Chang, H.-W., and Yang, C.-H. "SNP-SNP Interaction Using Gauss Chaotic Map Particle Swarm Optimization to Detect Susceptibility to Breast Cancer". In: 2014 47th Hawaii International Conference on System Sciences. 2014, pp. 2548–2554. DOI: 10.1109/HICSS.2014.647.

[16] Clarkson, K. L. and Woodruff, D. P. "Numerical linear algebra in the streaming model". In: Proc. of the 41st Annual ACM Symp. on Theory of Computing (STOC). ACM, 2009, pp. 205–214. DOI: 10.1145/1536414.1536445.

[17] Clarkson, K. L. and Woodruff, D. P. "Low-Rank Approximation and Regression in Input Sparsity Time". In: Journal of the ACM 63.6 (2017), pp. 1–45. ISSN: 0004-5411. DOI: 10.1145/3019134.

[18] Cohen, M. B. "Nearly Tight Oblivious Subspace Embeddings by Trace Inequalities". In: Proceedings of the 2016 Annual SODA. 2016, pp. 278–287. DOI: 10.1137/1.9781611974331.ch21.

[19] Cordell, H. J. "Detecting gene-gene interactions that underlie human diseases". In: Nature Reviews Genetics 10.6 (2009), pp. 392–404. DOI: 10.1038/nrg2579.

[20] Crisby, M., Carlson, L. A., and Winblad, B. "Statins in the prevention and treatment of Alzheimer disease". In: Alzheimer Disease and Associated Disorders 16.3 (2002), pp. 131–136. DOI: 10.1097/00002093-200207000-00001.

[21] Cruickshank-Quinn, C. I., Jacobson, S., Hughes, G., Powell, R. L., Petrache, I., Kechris, K., Bowler, R., and Reisdorph, N. "Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD". In: Scientific Reports 8.1 (2018), p. 17132.

[22] De Jager, P. L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L. C., Yu, L., Eaton, M. L., Keenan, B. T., Ernst, J., McCabe, C., Tang, A., Raj, T., Replogle, J., Brodeur, W., Gabriel, S., Chai, H. S., Younkin, C., Younkin, S. G., Zou, F., Szyf, M., and Bennett, D. A. "Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci". In: Nature Neuroscience 17.9 (2014), pp. 1156–1163. DOI: 10.1038/nn.3786.

[23] Delavar, M. A., Jahani, M. A., Sepidarkish, M., Alidoost, S., Mehdinezhad, H., and Farhadi, Z. "Relationship between fine particulate matter (PM2.5) concentration and risk of hospitalization due to chronic obstructive pulmonary disease: A systematic review and meta-analysis". In: BMC Public Health 23.1 (2023), p. 2229. DOI: 10.1186/s12889-023-17093-6.

[24] Dennis, C. "Altered states". In: Nature 421 (2003), pp. 686–688. DOI: 10.1038/421686a.

[25] Divo, M. J., Cabrera, C., Casanova, C., et al. "Co-morbidity distribution, clinical expression and survival in COPD patients with different body mass index". In: Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation 1.2 (2014), pp. 229–238. DOI: 10.15326/jcopdf.1.2.2014.0117.

[26] Dodson, T. B. "The Problem With P-Hacking". In: Journal of Oral and Maxillofacial Surgery 77.3 (2019), pp. 459–460. ISSN: 0278-2391. DOI: 10.1016/j.joms.2018.12.034.

[27] Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. "Fast approximation of matrix coherence and statistical leverage". In: J. Mach. Learn. Res. 13 (2012), pp. 3475–3506. DOI: 10.5555/2503308.2503352.

[28] Dudbridge, F. "Power and predictive accuracy of polygenic risk scores". In: PLoS Genetics 9.3 (2013), e1003348.

[29] Efron, B. and Tibshirani, R. J. An Introduction to the Bootstrap. 1st. Chapman and Hall/CRC, 1994. DOI: 10.1201/9780429246593.

[30] Enoma, D. O., Bishung, J., Abiodun, T., Ogunlana, O., and Osamor, V. C. "Machine learning approaches to genome-wide association studies". In: Journal of King Saud University - Science 34.4 (2022), p. 101847. ISSN: 1018-3647. DOI: 10.1016/j.jksus.2022.101847.

[31] Erdős, P. and Rényi, A. "On a classical problem of probability theory". In: Magyar Tud. Akad. Mat. Kutató Int. Közl. 6 (1961), pp. 215–220.

[32] Evangelou, E., Warren, H. R., Mosen-Ansorena, D., and al., et. "Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits". In: Nature Genetics 50 (2018), pp. 1412–1425. DOI: 10.1038/s41588-018-0205-x.

[33] Fan, J., Han, F., and Liu, H. "Challenges of Big Data Analysis". In: National Science Review 1 (Aug. 2013). DOI: 10.1093/nsr/nwt032.

[34] Ferreira, M. A., Gamazon, E. R., Al-Ejeh, F., and al., et. "Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer". In: Nature Communications 10 (2019), p. 1741. DOI: 10.1038/s41467-018-08053-5.

[35] Foo, J. N., Chew, E. G. Y., Chung, S. J., Peng, R., Blauwendraat, C., Nalls, M. A., Mok, K. Y., Satake, W., Toda, T., Chao, Y., Tan, L. C. S., Tandiono, M., Lian, M. M., Ng, E. Y., Prakash, K.-M., Au, W.-L., Meah, W.-Y., Mok, S. Q., Annuar, A. A., Chan, A. Y. Y., Chen, L., Chen, Y., Jeon, B. S., Jiang, L., Lim, J. L., Lin, J.-J., Liu, C., Mao, C., Mok, V., Pei, Z., Shang, H.-F., Shi, C.-H., Song, K., Tan, A. H., Wu, Y.-R., Xu, Y.-m., Xu, R., Yan, Y., Yang, J., Zhang, B., Koh, W.-P., Lim, S.-Y., Khor, C. C., Liu, J., and Tan, E.-K. "Identification of Risk Loci for Parkinson Disease in Asians and Comparison of Risk Between Asians and Europeans: A Genome-Wide Association Study". In: JAMA Neurology 77.6 (June 2020), pp. 746–754. DOI: 10.1001/jamaneurol.2020.0428.

[36] Franklin, R. E. and Gosling, R. G. "Molecular configuration in sodium thymonucleate". In: Nature 171.4356 (1953), pp. 740–741.

[37] Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. "Polygenic prediction via Bayesian regression and continuous shrinkage priors". In: Nature Communications 10.1 (2019), p. 1776.

[38] Geppert, L. N., Ickstadt, K., Munteanu, A., Quedenfeld, J., and Sohler, C. "Random projections for Bayesian regression". In: Stat. Comput. 27.1 (2017), pp. 79–101. DOI: 10.1007/s11222-015-9608-z.

[39] Geppert, L. N., Ickstadt, K., Munteanu, A., and Sohler, C. "Streaming statistical models via Merge & Reduce". In: Int. J. Data Sci. Anal. 10.4 (2020), pp. 331–347. DOI: 10.1007/s41060-020-00226-0.

[40] GOLD. "Global Strategy for Prevention, Diagnosis and Management of COPD: 2024 Report". In: (2023).

[41] GOLD. "Global Strategy for Prevention, Diagnosis and Management of COPD: 2025 Report". In: (2024).

[42] Golub, G. H. and Van Loan, C. F. Matrix Computations. 3rd. The Johns Hopkins University Press, 1996.

[43] Gottlieb, S. "Head injury doubles the risk of Alzheimer's disease". In: BMJ: British Medical Journal 321.7269 (2000), p. 1100.

[44] Graw, J. Genetik. 6th ed. Springer-Verlag Berlin Heidelberg, 2015. DOI: 10.1007/978-3-662-44817-5.

[45] Greliche, N., Germain, M., Lambert, J., Cohen, W., Bertrand, M., Dupuis, A.-M., Letenneur, L., Lathrop, M., Amouyel, P., Morange, P.-E., and Trégouët, D.-A. "A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis". In: BMC Medical Genetics 14.36 (2013). DOI: 10.1186/1471-2350-14-36.

[46] Harris, L., McDonagh, E. M., Zhang, X., Fawcett, K., Foreman, A., Daneck, P., Sergouniotis, P. I., Parkinson, H., Mazzarotto, F., Inouye, M., et al. "Genome-wide association testing beyond SNPs". In: Nature Reviews Genetics (2024), pp. 1–15.

[47] Hilbert, M. and López, P. "The World's Technological Capacity to Store, Communicate, and Compute Information". In: Science 332.6025 (2011), pp. 60–65. DOI: 10.1126/science.1200970.

[48] Hoaglin, D. C. and Welsch, R. E. "The Hat Matrix in Regression and ANOVA". In: The American Statistician 32.1 (1978), pp. 17–22. ISSN: 00031305.

[49] Hoang, P. H. and Landi, M. T. "DNA Methylation in Lung Cancer: Mechanisms and Associations with Histological Subtypes, Molecular Alterations, and Major Epidemiological Factors". In: Cancers 14.4 (2022), p. 961. DOI: 10.3390/cancers14040961.

[50] Hornung, R. and Boulesteix, A.-L. "Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects". In: Computational Statistics & Data Analysis 171 (2022), p. 107460. ISSN: 0167-9473. DOI: 10.1016/j.csda.2022.107460.

[51] Hornung, R. and Wright, M. N. "Block Forests: Random Forests for blocks of clinical and omics covariate data". In: BMC Bioinformatics 20.1 (2019), p. 358.

[52] Hüls, A., Krämer, U., Carlsten, C., Schikowski, T., Ickstadt, K., and Schwender, H. "Comparison of weighting approaches for genetic risk scores in gene-environment interaction studies". In: BMC Genetics 18.1 (2017), p. 115. DOI: 10.1186/s12863-017-0586-3.

[53] Hüls, A., Ickstadt, K., Schikowski, T., and Krämer, U. "Detection of gene-environment interactions in the presence of linkage disequilibrium and noise by using genetic risk scores with internal weights from elastic net regression". In: BMC Genetics 18.1 (2017), p. 55. DOI: 10.1186/s12863-017-0519-1.

[54] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome". In: Nature 431.7011 (2004), pp. 931–945.

[55] Jaenisch, R. and Bird, A. "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals". In: Nature Genetics 33.Suppl 3 (2003), pp. 245–254. DOI: 10.1038/ng1089.

[56] James, G., Witten, D., Hastie, T., and Tibshirani, R. An Introduction to Statistical Learning: With Applications in R. New York: Springer, 2013.

[57] Johnson, W. and Lindenstrauss, J. "Extensions of Lipschitz maps into a Hilbert space". In: Contemporary Mathematics 26 (Jan. 1984), pp. 189–206. DOI: 10.1090/conm/026/737400.

[58] Kaminsky, Z. A., Tang, T., Wang, S. C., Ptak, C., Oh, G. H., Wong, A. H. C., Feldcamp, L. A., Virtanen, C., Halfvarson, J., Tysk, C., McRae, A. F., Visscher, P. M., Montgomery, G. W., Gottesman, I. I., Martin, N. G., and Petronis, A. "DNA methylation profiles in monozygotic and dizygotic twins". In: Nature Genetics 41.2 (2009), pp. 240–245. DOI: 10.1038/ng.286.

[59] Kanervisto, M., Vasankari, T., Laitinen, T., Heliövaara, M., Jousilahti, P., and Saarelainen, S. "Low socioeconomic status is associated with chronic obstructive airway diseases". In: Respiratory Medicine 105.8 (2011), pp. 1140–1146. ISSN: 0954-6111. DOI: 10.1016/j.rmed.2011.03.008.

[60] Klinkhammer, H., Staerk, C., Maj, C., Krawitz, P. M., and Mayr, A. "A Statistical Boosting Framework for Polygenic Risk Scores Based on Large-Scale Genotype Data". In: Frontiers in Genetics 13 (2023), p. 1076440. DOI: 10.3389/fgene.2022.1076440.

[61] Knippers, R. Eine kurze Geschichte der Genetik. Springer, 2012.

[62] Laniado-Laborín, R. "Smoking and chronic obstructive pulmonary disease (COPD): Parallel epidemics of the 21st century". In: International Journal of Environmental Research and Public Health 6.1 (2009), pp. 209–224. DOI: 10.3390/ijerph6010209.

[63] Lau, M. logicDT: Identifying Interactions Between Binary Predictors. R package version 1.0.3. 2023.

[64] Lau, M., Kress, S., Schikowski, T., and Schwender, H. "Efficient gene–environment interaction testing through bootstrap aggregating". In: Scientific Reports 13.1 (2023), p. 937.

[65] Lau, M., Schikowski, T., and Schwender, H. "logicDT: A procedure for identifying response-associated interactions between binary predictors". In: Machine Learning 113.2 (2024), pp. 933–992. DOI: 10.1007/s10994-023-06488-6.

[66] Li, J., Malley, J. D., Andrew, A. S., and al., et. "Detecting gene-gene interactions using a permutation-based Random Forest method". In: BioData Mining 9 (2016), p. 14. DOI: 10.1186/s13040-016-0093-5.

[67] Li, P., Guo, M., Wang, C., Liu, X., and Zou, Q. "An overview of SNP interactions in genome-wide association studies". In: Briefings in Functional Genomics 14.2 (Sept. 2014), pp. 143–155. DOI: 10.1093/bfgp/elu036.

[68] Lin, H.-Y., Huang, P.-Y., Tseng, T.-S., and Park, J. Y. "SNPxE: SNP-environment interaction pattern identifier". In: BMC Bioinformatics 22.1 (2021), p. 425. DOI: 10.1186/s12859-021-04326-x.

[69] Lipton, Z. C. "The mythos of model interpretability: In Machine Learning, the concept of interpretability is both important and slippery." In: Queue 16.3 (2018), pp. 31–57.

[70] Lugg, S. T., Scott, A., Parekh, D., Naidu, B., and Thickett, D. R. "Cigarette smoke exposure and alveolar macrophages: mechanisms for lung disease". In: Thorax 77.1 (2022), pp. 94–101. DOI: 10.1136/thoraxjnl-2020-216296.

[71] Mackay, T. F. and Moore, J. H. "Why epistasis is important for tackling complex human disease genetics". In: Genome Medicine 6.6 (2014), p. 42.

[72] Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. "Polygenic scores via penalized regression on summary statistics". In: Genetic Epidemiology 41.6 (2017), pp. 469–480.

[73] Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., and Ferrari, R. "Genome, Transcriptome and Proteome: The Rise of Omics Data and Their Integration in Biomedical Sciences". In: Briefings in Bioinformatics 19.2 (2018), pp. 286–302. DOI: 10.1093/bib/bbw114.

[74] Marx, V. "The Big Challenges of Big Data". In: Nature 498 (2013), pp. 255–260. DOI: 10.1038/498255a.

[75] Matthews, S. M., Eshelman, M. A., Berg, A. S., Koltun, W. A., and Yochum, G. S. "The Crohn's disease associated SNP rs6651252 impacts MYC gene expression in human colonic epithelial cells". In: PLOS ONE 14.2 (2019), e0212850. DOI: 10.1371/journal.pone.0212850.

[76] McFadden, D. "Conditional Logit Analysis of Qualitative Choice Behavior". In: Frontiers in Econometrics (1974), pp. 105–142.

[77] Mentch, L. and Hooker, G. "Quantifying uncertainty in Random Forests via confidence intervals and hypothesis tests". In: J. Mach. Learn. Res. 17 (2016), Paper No. 26, 41. ISSN: 1532-4435,1533-7928.

[78] Mieth, B., Kloft, M., Rodríguez, J. A., Sonnenburg, S., Vobruba, R., Morcillo-Suárez, C., Farré, X., Marigorta, U. M., Fehr, E., Dickhaus, T., et al. "Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies". In: Scientific Reports 6.1 (2016), p. 36671.

[79] Mok, A., Rhead, B., Holingue, C., Shao, X., Quach, H. L., Quach, D., Sinclair, E., Graf, J., Imboden, J., Link, T., Harrison, R., Chernitskiy, V., Barcellos, L. F., and Criswell, L. A. "Hypomethylation of CYP2E1 and DUSP22 Promoters Associated With Disease Activity and Erosive Disease Among Rheumatoid Arthritis Patients". In: Arthritis & Rheumatology (Hoboken, N.J.) 70.4 (2018), pp. 528–536. DOI: 10.1002/art.40408.

[80] Mokaddem Mohsen, S., Chakroun, S., Chaker, A., Ayed, K., and Jameleddine, S. "Body mass index in COPD: what relationship?" In: European Respiratory Journal 56.suppl 64 (2020). DOI: 10.1183/13993003.congress-2020.2439.

[81] Molnar, C. "Interpretable machine learning: A guide for making black box models explainable". In: Leanpub (2020).

[82] Montine, T. J., Phelps, C. H., Beach, T. G., Bigio, E. H., Cairns, N. J., Dickson, D. W., Duyckaerts, C., Frosch, M. P., Masliah, E., Mirra, S. S., et al. "National Institute on Aging–Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach". In: Acta Neuropathologica 123 (2012), pp. 1–11. DOI: 10.1007/s00401-011-0910-3.

[83] Morena, D., Izquierdo, J. L., Rodríguez, J., Cuesta, J., Benavent, M., Perralejo, A., and Rodríguez, J. M. "The Clinical Profile of Patients with COPD Is Conditioned by Age". In: Journal of Clinical Medicine 12.24 (2023), p. 7595. DOI: 10.3390/jcm12247595.

[84] Moreno, V., Gonzalez, J. R., and Pelegri, D. SNPassoc: SNPs-Based Whole Genome Association Studies. R package version 2.1-0. 2022.

[85] Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., Belle, G. van, Fillenbaum, G., Mellits, E. D., and Clark, C. "The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease". In: Neurology 39.9 (1989), pp. 1159–1165. DOI: 10.1212/wnl.39.9.1159.

[86] Murray, C. J. and Lopez, A. D. "Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study". In: The Lancet 349.9064 (1997), pp. 1498–1504. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(96)07492-2.

[87] Neven, K. Y., Piola, M., Angelici, L., Bollati, V., Allegri, M., Basilico, P., Rusconi, F., Motta, T., Lunt, M., Naccarati, A., Baccarelli, A. A., Bergamaschi, R., Murgia, N., Bonavita, S., Clerico, M., Martinelli Boneschi, F., Sorokin, S., Ottaviani, S., Galli, E., Fumagalli, M., Esposito, F., Pappalardo, S., Bovis, F., Bagnardi, V., Neri, M., Barizzone, N., Leone, M., D'Alfonso, S., Harbo, H. F., Myhr, K.-M., Celius, E. G., Lie, B. A., Sospedra, M., Martin, R., Hemmer, B., Zipp, F., Kümpfel, T., Uccelli, A., D'Alessandro, R., Comi, G., Facheris, M., Hiltunen, M., Kieseppä, T., Lönnqvist, J., Färkkilä, M., Palotie, A., Peltonen, L., Heikinheimo, P., Saarela, J., Hinttala, R., Auvinen, J., Veijola, J., Jokela, M., Pudas, R., Pirttilä, T., Elovaara, I., Lehtimäki, T., Tienari, P., Pitkänen, K., Dastidar, P., Koivisto, K., Reunanen, M., Seitsonen, S., Kovanen, P., Lauerma, A., Kallela, M., Hovatta, I., Jylhä, P., Perola, M., Kristiansson, K., Kähönen, M., Nieminen, M. S., Salomaa, V., Terwilliger, J. D., Gaspari, G., Macciardi, F., Montomoli, C., Padoan, R., Rocchi, M., Amoroso, A., Matullo, G., Rosati, G., Eoli, M., Broggi, G., La Mantia, L., Radaelli, M., Martinelli, V., Martinelli Boneschi, F., Comi, G., D'Alfonso, S., Barizzone, N., and Leone, M. "Repetitive element hypermethylation in multiple sclerosis patients". In: BMC Genetics 17 (2016), p. 84. DOI: 10.1186/s12863-016-0395-0.

[88] Nguyen, T.-T., Huang, J. Z., Wu, Q., Nguyen, T. T., and Li, M. J. "Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests". In: BMC genomics. Vol. 16. Springer. 2015, pp. 1–11.

[89] Ottman, R. "Gene-environment interaction: Definitions and study designs". In: Preventive Medicine 25.6 (1996), pp. 764–770. DOI: 10.1006/pmed.1996.0117.

[90] Parry, K., Geppert, L., Munteanu, A., and Ickstadt, K. "Cross-Leverage Scores for Selecting Subsets of Explanatory Variables". In: arXiv (Sept. 2021). DOI: 10.48550/arXiv.2109.08399.

[91] Pemov, A., Sung, H., Hyland, P. L., Sloan, J. L., Ruppert, S. L., Baldwin, A. M., Boland, J. F., Bass, S. E., Lee, H. J., Jones, K. M., Zhang, X., Program, N. C. S., Mullikin, J. C., Widemann, B. C., Wilson, A. F., and Stewart, D. R. "Genetic modifiers of neurofibromatosis type 1-associated café-au-lait macule count identified using multi-platform analysis". In: PLOS Genetics 10.10 (2014), e1004575. DOI: 10.1371/journal.pgen.1004575.

[92] Pirinen, M., Donnelly, P., and Spencer, C. C. "Including known covariates can reduce power to detect genetic effects in case-control studies". In: Nature genetics 44.8 (2012), pp. 848–851.

[93] Polkey, M. I. "Chronic obstructive pulmonary disease: aetiology, pathology, physiology and outcome". In: Medicine 36.4 (2008). Respiratory disorders Part 2 of 4, pp. 213–217. ISSN: 1357-3039. DOI: 10.1016/j.mpmed.2008.01.002.

[94] Quanjer, P. H., Stanojevic, S., Cole, T. J., Baur, X., Hall, G. L., Culver, B. H., Enright, P. L., Hankinson, J. L., Ip, M. S., Zheng, J., Stocks, J., and Initiative, E. G. L. F. "Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations". In: The European Respiratory Journal 40.6 (2012), pp. 1324–1343. DOI: 10.1183/09031936.00080312.

[95] Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., and Foulkes, A. S. "A guide to genome-wide association analysis and post-analytic interrogation". In: Statistics in Medicine 34.28 (2015), pp. 3769–3792. DOI: 10.1002/sim.6605.

[96] Ribbenstedt, A., Ziarrusta, H., and Benskin, J. P. "Development, characterization and comparisons of targeted and non-targeted metabolomics methods". In: PLOS ONE 13.11 (2018), e0207082. DOI: 10.1371/journal.pone.0207082.

[97] Roessner, U. Metabolomics. Rijeka: IntechOpen, 2012. DOI: 10.5772/1237.

[98] Ruczinski, I., Kooperberg, C., and LeBlanc, M. "Logic Regression". In: Journal of Computational and Graphical Statistics 12.3 (2003), pp. 475–511. DOI: 10.1198/1061860032238. eprint: 10.1198/1061860032238.

[99] Sagiroglu, S. and Sinanc, D. "Big data: A review". In: 2013 Internat. Conference on Collaboration Technologies and Systems (CTS). 2013, pp. 42–47. DOI: 10.1109/CTS.2013.6567202.

[100] Sanches, P. H. G., Melo, N. C. de, Porcari, A. M., and Carvalho, L. M. de. "Integrating Molecular Perspectives: Strategies for Comprehensive Multi-Omics Integrative Data Analysis and Machine Learning Applications in Transcriptomics, Proteomics, and Metabolomics". In: Biology 13.11 (2024). DOI: 10.3390/biology13110848.

[101] Schienkiewitz, A., Mensink, G., Kuhnert, R., and Lange, C. "Übergewicht und Adipositas bei Erwachsenen in Deutschland". In: Journal of Health Monitoring. 2. Robert Koch-Institut, Epidemiologie und Gesundheitsberichterstattung, 2017. DOI: 10.17886/RKI-GBE-2017-025.

[102] Schikowski, T., Sugiri, D., Ranft, U., Gehring, U., Heinrich, J., Wichmann, H.-E., and Krämer, U. "Long-term air pollution exposure and living close to busy roads are associated with COPD in women". In: Respiratory Research 6 (2005), pp. 1–10. DOI: 10.1186/1465-9921-6-152.

[103] Schwender, H. and Ickstadt, K. "Identification of SNP interactions using logic regression". In: Biostatistics 9.1 (June 2007), pp. 187–198. DOI: 10.1093/biostatistics/kxm024.

[104] Schwender, H., Selinski, S., Blaszkewicz, M., Marchan, R., Ickstadt, K., Golka, K., and Hengstler, J. G. "Distinct SNP Combinations Confer Susceptibility to Urinary Bladder Cancer in Smokers and Non-Smokers". In: PLOS ONE 7 (Dec. 2012), pp. 1–12. DOI: 10.1371/journal.pone.0051880.

[105] Shastry, B. S. "SNPs in disease gene mapping, medicinal drug development and evolution". In: Journal of human genetics 52.11 (2007), pp. 871–880.

[106] Spencer, H. "Epigenetic Inheritance". In: Encyclopedia of Evolutionary Biology. Ed. by Kliman, R. M. Oxford: Academic Press, 2016, pp. 1–5. ISBN: 978-0-12-800426-5. DOI: 10.1016/B978-0-12-800049-6.00050-0.

[107] Straus, S. E., McAlister, F. A., Sackett, D. L., and Deeks, J. J. "The accuracy of patient history, wheezing, and laryngeal measurements in diagnosing obstructive airway disease". In: JAMA 283.14 (2000), pp. 1853–1857. DOI: 10.1001/jama.283.14.1853.

[108] Sun, Y., Zhang, Y., Liu, X., Liu, Y., Wu, F., and Liu, X. "Association between body mass index and respiratory symptoms in US adults: A national cross-sectional study". In: Scientific Reports 14.1 (2024), p. 940. DOI: 10.1038/s41598-024-51637-z.

[109] Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. "Benefits and limitations of genome-wide association studies". In: Nature Reviews Genetics 20.8 (2019), pp. 467–484.

[110] Teschke, S. "(simulated Data:) Detecting interactions in high-dim. data using CLS". In: Zenodo (2024). DOI: 10.5281/zenodo.12742957.

[111] Teschke, S. "special_issue_CEN". In: Github (2024). https://github.com/SvenTeschke/special_issue_CEN.

[112] Teschke, S., Ickstadt, K., and Munteanu, A. "Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores". In: Biometrical Journal 66.8 (2024), e70014. DOI: 10.1002/bimj.70014.

[113] Teschke, S., Ickstadt, K., Schikowski, T., and Wigman, C. "Using Cross Leverage Scores for Detecting SNP-Environment Interactions Effects on COPD". In: Genetic Epidemiology (submitted) (2025).

[114] Thal, D. R., Rüb, U., Orantes, M., and Braak, H. "Phases of Aβ-deposition in the human brain and its relevance for the development of AD". In: Neurology 58.12 (2002), pp. 1791–1800. DOI: 10.1212/WNL.58.12.1791.

[115] The International HapMap Consortium. "The international HapMap project". In: Nature 426 (2003), pp. 789–796. DOI: 10.1038/nature02168.

[116] Tibshirani, R. "Regression shrinkage and selection via the lasso". In: Journal of the Royal Statistical Society Series B: Statistical Methodology 58.1 (1996), pp. 267–288.

[117] Tropp, J. A. "Improved Analysis of the subsampled Randomized Hadamard Transform". In: Advances in Adaptive Data Analysis 3.1-2 (2011), pp. 115–126. DOI: 10.1142/S1793536911000787.

[118] Tug, T., Liang, D., Teschke, S., Tan, Y., Gearing, M., Levey, A. I., Lah, J. J., Wingo, A. P., Wingo, T. S., Lau, M., Ickstadt, K., and Hüls, A. "Development and Application of Brain Tissue Based Multi-Omics Profile Scores for Alzheimer's Disease". In: Alzheimer's & Dementia: The Journ. of the Alzheimer's Assoc. (submitted) (2025).

[119] U.S. Environmental Protection Agency. Particulate matter (PM) basics. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics. Accessed: 2025-02-27. 2025.

[120] Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. "Genome-wide association studies". In: Nature Reviews Methods Primers 1.1 (2021), p. 59.

[121] Van Hout, C. V. "Chapter 10 - Statistical approaches to rare disease analyses". In: Genomics of Rare Diseases. Ed. by Gonzaga-Jauregui, C. and Lupski, J. R. Translational and Applied Genomics. Academic Press, 2021, pp. 205–213. ISBN: 978-0-12-820140-4. DOI: 10.1016/B978-0-12-820140-4.00011-9.

[122] Vujkovic, M., Keaton, J. M., Lynch, J. A., and al., et. "Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis". In: Nature Genetics 52 (2020), pp. 680–691. DOI: 10.1038/s41588-020-0637-y.

[123] Wang, Q. and Liu, S. "The Effects and Pathogenesis of PM2.5 and Its Components on Chronic Obstructive Pulmonary Disease". In: International Journal of Chronic Obstructive Pulmonary Disease 18 (2023), pp. 493–506. DOI: 10.2147/COPD.S402122.

[124] Watson, J. D. and Crick, F. H. "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid". In: Nature 171.4356 (1953), pp. 737–738.

[125] Watson, J. D. "Involvement of RNA in the synthesis of proteins". In: Science 140.3562 (1963), pp. 17–26. DOI: 10.1126/science.140.3562.17.

[126] WHO. Global status report on the public health response to dementia. Geneva: World Health Organization, 2021.

[127] WHO Consultation on Obesity and WHO. Obesity: Preventing and managing the global epidemic. Geneva, Switzerland: World Health Organization, 2000.

[128] Wilkins, J. M. and Trushina, E. "Application of metabolomics in Alzheimer's disease". In: Frontiers in Neurology 8 (2018), p. 719. DOI: 10.3389/fneur.2017.00719.

[129] Wimo, A., Seeher, K., Cataldi, R., Cyhlarova, E., Dielemann, J. L., Frisell, O., Guerchet, M., Jönsson, L., Malaha, A. K., Nichols, E., Pedroza, P., Prince, M., Knapp, M., and Dua, T. "The worldwide costs of dementia in 2019". In: Alzheimer's & Dementia: The Journal of the Alzheimer's Association 19.7 (2023), pp. 2865–2873. DOI: 10.1002/alz.12901.

[130] Wright, M. N. and Ziegler, A. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: Journal of Statistical Software 77.1 (2017), pp. 1–17. DOI: 10.18637/jss.v077.i01.

[131] Wright, M. N., Ziegler, A., and König, I. R. "Do little interactions get lost in dark Random Forests?" In: BMC Bioinformatics 17.1 (2016), p. 145. DOI: 10.1186/s12859-016-0995-8.

[132] Xing, Y. F., Xu, Y. H., Shi, M. H., and Lian, Y. X. "The impact of PM2.5 on the human respiratory system". In: Journal of Thoracic Disease 8.1 (2016), E69–E74. DOI: 10.3978/j.issn.2072-1439.2016.01.19.

[133] Yang, I. A., Jenkins, C. R., and Salvi, S. S. "Chronic obstructive pulmonary disease in never-smokers: Risk factors, pathogenesis, and implications for prevention and treatment". In: The Lancet Respiratory Medicine 10.5 (2022), pp. 497–511. DOI: 10.1016/S2213-2600(21)00506-3.

[134] Yoshida, M. and Koike, A. "SNPInterForest: A new method for detecting epistatic interactions". In: BMC Bioinformatics 12 (2011), p. 469. DOI: 10.1186/1471-2105-12-469.

[135] Younesian, S., Mohammadi, M. H., Younesian, O., Momeny, M., Ghaffari, S. H., and Bashash, D. "DNA methylation in human diseases". In: Heliyon 10.11 (2024), e32366. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2024.e32366.

[136] Zeng, X., Vonk, J. M., Jong, K. de, Xu, X., Huo, X., and Boezen, H. M. "No convincing association between genetic markers and respiratory symptoms: results of a GWA study". In: Respiratory Research 18.1 (2017), p. 11. DOI: 10.1186/s12931-016-0495-4.

[137] Zou, H. and Hastie, T. "Regularization and variable selection via the elastic net". In: Journal of the Royal Statistical Society Series B: Statistical Methodology 67.2 (2005), pp. 301–320.

A Articles

- A1: Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores
- A2: Using Cross Leverage Scores for Detecting SNP-Environment Interactions Effects on COPD
- A3: Development and Application of Brain Tissue Based Multi-Omics Profile Scores for Alzheimer Disease

Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores

Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores

Sven Teschke[1] | Katja Ickstadt[1,2] | Alexander Munteanu[1]

[1]Faculty of Statistics, TU Dortmund University, Dortmund, Germany | [2]Lamarr-Institute for Machine Learning and Artificial Intelligence, Dortmund, Germany

Correspondence: Sven Teschke (teschke@statistik.tu-dortmund.de)

Received: 13 December 2023 | **Revised:** 23 July 2024 | **Accepted:** 16 October 2024

Funding: Sven Teschke and Katja Icksatdt were supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, project R1) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Project Number 427806116). Katja Ickstadt acknowledges the support of BMBF and MKW.NRW within the Lamarr-Institute for Machine Learning and Artificial Intelligence. Alexander Munteanu was supported by the TU Dortmund - Center for Data Science and Simulation (DoDaS) and by the German Research Foundation (DFG), grant MU 4662/2-1 (535889065).

Keywords: cross leverage scores | genetics | high-dimensional data | interaction effects | sketching | variable selection

ABSTRACT

We develop a variable selection method for interactions in regression models on large data in the context of genetics. The method is intended for investigating the influence of single-nucleotide polymorphisms (SNPs) and their interactions on health outcomes, which is a $p \gg n$ problem. We introduce cross leverage scores (CLSs) to detect interactions of variables while maintaining interpretability. Using this method, it is not necessary to consider every possible interaction between variables individually, which would be very time-consuming even for moderate amounts of variables. Instead, we calculate the CLS for each variable and obtain a measure of importance for this variable. Calculating the scores remains time-consuming for large data sets. The key idea for scaling to large data is to divide the data into smaller random batches or consecutive windows of variables. This avoids complex and time-consuming computations on high-dimensional matrices by performing the computations only for small subsets of the data, which is less costly. We compare these methods to provable approximations of CLS based on sketching, which aims at summarizing data succinctly. In a simulation study, we show that the CLSs are directly linked to the importance of a variable in the sense of an interaction effect. We further show that the approximation approaches are appropriate for performing the calculations efficiently on arbitrarily large data while preserving the interaction detection effect of the CLS. This underlines their scalability to genome wide data. In addition, we evaluate the methods on real data from the HapMap project.

1 | Introduction

In this paper, we present a method to quickly and efficiently detect and select interaction effects in large data sets. In addition to the main effects, interaction effects are often of great relevance, for instance, when detecting associations of interacting genetic variations with certain diseases or physical conditions (Li et al. 2014). We consider single-nucleotide polymorphisms (SNPs), which are individual variations of nucleotides in the (human) DNA with a prevalence of more than 1% (in contrast, when the prevalence is less than 1%, it is called a mutation). Not only individual SNPs, but also interactions between SNPs can be associated with a certain disease such as breast cancer (Chuang et al. 2014) or venous thrombosis (Greliche et al. 2013). In general, it is difficult to search for important interaction effects, because all possible combinations of p variables have to be considered individually and have to be included in the model. In a model with p variables, the amount of possible two-way combinations is

already quadratic $\binom{p}{2} = \Theta(p^2)$ and the number of choices grows exponentially in the order of interactions k, as $\left(\frac{p}{k}\right)^k \leq \binom{p}{k} \leq \left(\frac{ep}{k}\right)^k$. Finally, the number of all possible combinations of any degree sum up to 2^p. This poses a crucial limitation, as it is common in genetics to consider higher order interactions (Schwender et al. 2012). Even for relatively small p, it is prohibitive to consider all possible interactions and it becomes increasingly difficult in higher dimensional data sets, for instance, in genetics where data sets with several millions of variables but comparably small numbers of observations $n \ll p$ are common. Nevertheless, research on this topic is very difficult because most available statistical methods fail in these extremely high-dimensional settings. Prominent examples include logic regression (Ruczinski, Kooperberg, and LeBlanc 2003) or the maximum entropy conditional probability model (MECPM) approach (Miller et al. 2009), which are limited to a small number of variables.

In this paper, we introduce a method for calculating the so-called cross leverage scores (CLSs) that indicate for each variable their leverage on the outcome variable and in particular their participation in an interaction effect. In particular, we avoid considering every possible combination of variables. By considering the original matrix of variables as a whole, we only need to calculate p individual scores in total. These indicate whether a variable is important or not in terms of an interaction effect. The method is therefore characterized by the fact that we consider the full multivariate model instead of p univariate models in which ones checks whether a single variable has an effect or not, which has so far been common (Uffelmann et al. 2021) but is neither desirable nor sustainable from a statistical point of view.

Indeed, any multivariate information on interaction effects is lost in single-variable analyses. In particular, it can occur that the variables participating in important SNP interaction effects have only moderate or no influence at all if their individual main effect is considered (Li et al. 2014). Another issue is that several works claim genome-wide studies but consider only a few thousand variables (e.g., Terada et al. 2016; Yang et al. 2008). Such methods often do not scale to arbitrarily large data without further computational improvements.

Methods that perform a hierarchical search for interactions are common, for example, the genome-wide association analysis using LASSO-penalized logistic regression (Wu et al. 2009). An example for a two-step hierarchical search first selects only a small subset of variables with large main effects and in a second step searches for interacting variables only within this subset. A variable that interacts with others, but is missing an individual main effect would thus be removed from the consideration before the actual interaction analysis. Here, we propose a new hierarchical approach: in a first step, a quick variable selection is performed to reduce the number of variables based on CLS, such that variables participating in an interaction are retained. This enables in a second step to apply a more sophisticated method such as logic regression (Ruczinski, Kooperberg, and LeBlanc 2003), which is not readily applicable for high-dimensional data sets, but good in finding interaction effects in scenarios with smaller p.

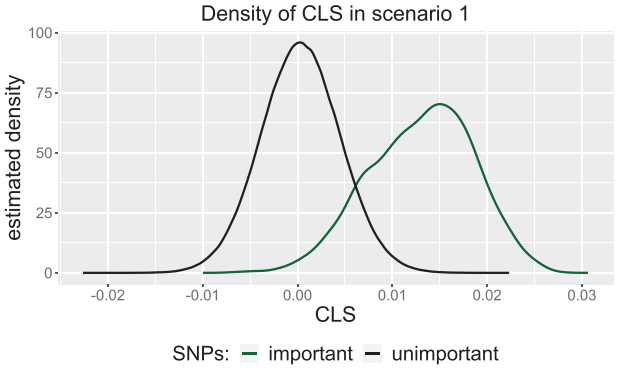

FIGURE 1 | Kernel density estimates of the cross leverage scores of the important (green) and unimportant (black) variables.

The idea of using CLS for this purpose was first mentioned in previous work of our group (Ding, Ickstadt, and Munteanu 2022; Parry et al. 2021). We build on this idea and further develop the approach of using CLS as a tool for detecting variable interactions. Here, we make this approach available for genome-wide studies by developing scalable approximations with small bounded error guarantees, and applicable to arbitrarily large data sets. The approximation methods can also be extended to a pure data stream algorithm that reads the data successively while using only very limited memory of the computing environment such as `R` (for example, in the context of genetics, chromosome by chromosome).

Our paper is structured as follows. First, we will motivate the use of CLS for detecting interactions and selecting the participating variables. Then, we will show how they can be suitably approximated so that the method becomes applicable to very large data sets. In a simulation study, we will analyze how well the method is able to detect important variables participating in interactions of different order, and due to certain similarities, we compare their performance with correlations and uniform sampling as a baseline method. In particular, we will also give an example, where it is impossible to detect important interactions using correlations, but our CLSs reveal the participating variables. In addition, we will evaluate the effect of variable preselection by our and other methods in a simulated and a real data scenario.

2 | Motivation

To show that the CLSs are a useful measure for distinguishing between important and unimportant variables, we consider a scenario in which we simulate $n = 120$ observations with one single important two-way interaction and 1998 noisy variables. Note that $n \ll p$. In total, we simulate 1000 data sets independently in the same manner. The simulations are described in detail in Section 4.1. We calculate the CLS between the individual variables and the binary target variable y as formally described in Section 3. In Figure 1, we plot the kernel density estimates of the calculated

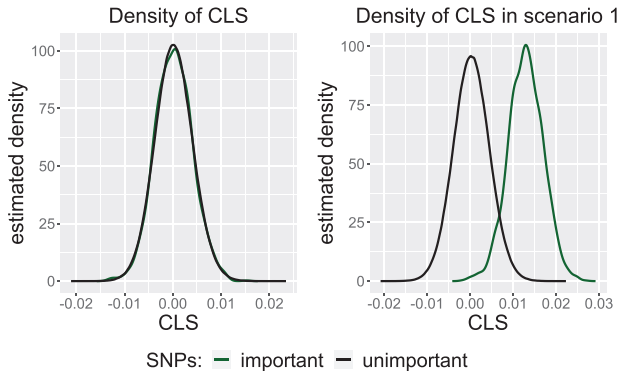

FIGURE 2 | Kernel density estimates of the cross leverage scores of the important (green) and unimportant (black) variables. On the left, we consider two main effects with marginal influence, and in the right plot, we add a high interaction effect while the main effects persist to be low.

scores over all data sets, distinguishing between unimportant and important variables by color.

We see that the kernel density estimates for the unimportant and important variables differ strongly. The CLSs of the unimportant noise variables are concentrated around zero, while the CLSs of the important variables are larger. This suggests that we can use the CLS as a measure for distinguishing between unimportant and important variables and that we should select the variables with the largest CLS.

We will see later in Proposition 3.1 that each CLS equals their corresponding parameter in a least squares solution up to a small bounded additive error. This might suggest that it is just a different way of detecting main effects. We thus show that we actually measured the interaction effect and not solely main effects. To this end, we simulate a scenario in which the main effects of the first two variables are chosen to be negligibly small and no interaction effect is present. For a direct comparison, we consider the same scenario but with an added high interaction effect for the same two variables.

In the left plot in Figure 2, we see that the CLS of the first two variables cannot be distinguished from the other noise variables due to the absence of main or interaction effects. In contrast, we see in the right plot that we can distinguish between the important and unimportant variables using CLS because they indicate that the two variables participate in the added interaction. Although main effects are not increased in the generating model, they show up in the fitted model.

Using a toy example (Parry et al. 2021), we show that a variable selection with CLS outperforms the selection with correlation. We construct a data set with $n = 16$ observations and $p = 60$ variables with $x_{ij} \in \{0, 1\}$ for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$, and a binary response y. We construct the data set in such a way that the response takes the value $y = 1$ whenever $x_1 = x_2 = 1$ or $x_3 = x_4 = 1$, and $y = 0$ otherwise. By construction, there are two two-way interactions. The remaining values for x_5 to x_{60} are chosen from $\{0, 1\}$ uniformly at random:

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & x_{1,5} & \cdots & x_{1,60} \\ 1 & 1 & 1 & 0 & x_{2,5} & \cdots & x_{2,60} \\ 0 & 0 & 1 & 1 & x_{3,5} & \cdots & x_{3,60} \\ 0 & 0 & 1 & 1 & x_{4,5} & \cdots & x_{4,60} \\ 1 & 0 & 0 & 1 & x_{5,5} & \cdots & x_{5,60} \\ 1 & 0 & 1 & 0 & x_{6,5} & \cdots & x_{6,60} \\ 0 & 1 & 0 & 1 & x_{7,5} & \cdots & x_{7,60} \\ 0 & 1 & 1 & 0 & x_{8,5} & \cdots & x_{8,60} \\ 1 & 1 & 0 & 1 & x_{9,5} & \cdots & x_{9,60} \\ 1 & 1 & 1 & 0 & x_{10,5} & \cdots & x_{10,60} \\ 0 & 0 & 1 & 1 & x_{11,5} & \cdots & x_{11,60} \\ 0 & 0 & 1 & 1 & x_{12,5} & \cdots & x_{12,60} \\ 1 & 0 & 0 & 1 & x_{13,5} & \cdots & x_{13,60} \\ 1 & 0 & 1 & 0 & x_{14,5} & \cdots & x_{14,60} \\ 0 & 1 & 0 & 1 & x_{15,5} & \cdots & x_{15,60} \\ 0 & 1 & 1 & 0 & x_{16,5} & \cdots & x_{16,60} \end{bmatrix}, y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad (1)$$

The data, especially the first two variables, are constructed in such a way that $\text{cor}(x_1, y) = \text{cor}(x_2, y) = 0$. So, when selecting the variables according to their correlation with the response, we would never select the first two variables x_1 and x_2. In contrast, the CLSs of the first two variables are nonzero. It is therefore conceivable that we detect them being part of the two-way interaction by selection via CLS. Of course, this is no guarantee, since the CLS may also depend on the other (noisy) variables.

We thus conduct a small simulation study in which we simulate 1000 data sets according to the scheme described above in Equation (1). We count how often how many of the first two important variables are detected when selecting the variables with the $q = \lceil n \log n \rceil = 45$ largest CLS or correlation. In the case of CLS, we find two out of two important variables in the median average, and zero out of two in the case of correlation. Consider Figure A1 in Appendix A.2 to see that even for small q where the selection with CLS deteriorates, it still outperforms the selection using correlations. This example shows that in some cases, we can detect interaction effects with CLS, where correlations necessarily miss the relevant information and thus miss the important variables.

3 | Methods

As mentioned before, the idea to consider the CLSs is based on a paper of Parry et al. (2021). Since we deal with a $p \gg n$ problem, we exchange the role of observations and variables: to make our notation consistent with previous literature, where the case $n \gg p$ was treated (Drineas et al. 2012), we consider the matrix

$$\tilde{X} = [X, y]^T \in \mathbb{R}^{\tilde{p} \times n} \quad (2)$$

with $\tilde{p} = p + 1$ and p the number of variables and n the number of observations. $X \in \mathbb{R}^{n \times p}$ is the data matrix and $y \in \mathbb{R}^n$ denotes the response. We obtain the CLS from the off-diagonal entries of the hat matrix H of $\tilde{X}$. The hat matrix H is given by $H = QQ^T$

(Hoaglin and Welsch 1978) where Q forms an orthonormal basis for the column space of $\tilde{X}$, which can be obtained by its QR-decomposition $\tilde{X} = QR$ (Golub and Van Loan 1996). Since we are only interested in the CLS $c_{i\tilde{p}}$ between the variables $i \in \{1, \dots, p\}$ and the response y, we can avoid the time-consuming matrix multiplication of $Q \in \mathbb{R}^{\tilde{p}\times n}$ and $Q^T \in \mathbb{R}^{n\times\tilde{p}}$ by instead calculating only the dot products of rows $Q_{i\cdot}$ with row $Q_{\tilde{p}\cdot}$:

$$c_{j\tilde{p}} = \langle Q_{i\cdot}, Q_{\tilde{p}\cdot}\rangle, \quad i \in \{1, \dots, p\}. \tag{3}$$

Equation (3) is computed from the orthogonal basis Q. This means that the CLSs correspond to the leverage of individual variables on the correlation of the multidimensional subspace spanned by the data with the outcome variable, rather than correlations of single variables. Our intuition is that an according subselection thus carries more information on the multivariate structure, and is capable of retaining interaction effects as shown in our previous example in Figure 2.

Under the mild assumption that the subspace is well aligned with the response vector, we show next that each CLS $c_{i,p+1}$ for $i \in [p]$ equals their corresponding parameter in the least squares solution up to a small additive error. We note that variables with large parameters are thus recovered by CLS, and are more likely to participate in an interaction, since parameters near zero would cancel a possible interaction effect.

Proposition 3.1. *Let $\tilde{X} = [X, y] \in \mathbb{R}^{n\times(p+1)}$, for $p \gg n$, where X and thus also $\tilde{X}$ have full rank n. Let $\tilde{X} = U\Sigma V^T$ be its SVD. Consider the smallest norm solution to the ℓ_2 regression problem: $\beta^{OLS} \in \arg\min_{\beta\in\mathbb{R}^p} \|X\beta - y\|_2^2$. Assume for $\frac{1}{2} > \eta > 0$ that $\sum_{i=1}^{p} \langle V_{i\cdot}, V_{y\cdot}\rangle^2 \geq \frac{1}{\eta}\|V_y\|_2^4$. Then, it holds that*

$$\max_{i\in[p]} |\beta_i^{OLS} - c_{i,p+1}| < \frac{3\eta}{2}.$$

Proof. See Appendix A.1. □

3.1 | CLS Calculation

The obvious bottleneck of the approach described above is the QR-decomposition with running time $\Theta(pn^2)$ (Golub and Van Loan 1996), which is prohibitively slow to compute for p as large as several millions. The data might not even fit into main memory, which aggravates the situation. However, without the Q matrix, we are not able to calculate the CLS. Massively parallel QR-decomposition algorithms are applicable in such cases but this requires a large compute cluster, which is not always available and this is no remedy for the total workload that remains $\Theta(pn^2)$ (Demmel et al. 2012).

In this paper, we describe and compare three possible solutions to the problem that easily run on standard commodity hardware: specifically, we introduce two new heuristics, the *Sliding Window approach* and the *Random Window approach*. In both cases, instead of one QR-decomposition for a large matrix, a lot of QR-decompositions for many small submatrices are performed and their results are merged in a suitable way, akin to the Merge & Reduce technique (Geppert et al. 2020). The third solution is the *Sketching approach* based on the ideas of Drineas et al. (2012). The QR-decomposition is calculated after reducing the large dimension $\tilde{p}$ by sketching $\tilde{X} \in \mathbb{R}^{\tilde{p}\times n}$ to obtain $\tilde{X}_\star \in \mathbb{R}^{r\times n}$ with significantly smaller dimension $r \ll \tilde{p}$.

In what follows, the main goal is to approximate the CLS in an efficient way, based on which we make a preselection retaining the important variables. This enables more refined methods that are not suitable for large data sets—such as logic regression (Ruczinski, Kooperberg, and LeBlanc 2003)—to operate on the reduced data subsequently. We select the variables that have the most extreme CLS. It was recommended to select $q = \lceil n \log n \rceil$ variables in (Parry et al. 2021), which is confirmed in our experimental findings. Theoretically, this is supported by the coupon collector's problem (Erdős and Rényi 1961) that requires oversampling by a $\log n$ factor so that the selection contains at least as many variables to ensure that the submatrix preserves the full rank n of the original matrix (Tropp 2011).

3.2 | New Window-Based Approaches

We first introduce two new approaches, which share the concept of a *moving* window. For both, we consider the matrix X consisting of all variables (e.g., SNPs). In the *Sliding Window approach*, as illustrated in Figure 3, we iterate through this matrix with a window consisting of w consecutive variables at a time and add the response y, respectively. We consider the submatrix starting at some index j for each window $\tilde{X}^{j,j+w-1} = [X_{\cdot,j:(j+w-1)}, y]^T$. If the window size is small enough, we can efficiently calculate the QR-decomposition and determine the CLS in the standard way described above. We start with the first window $X^{1,w}$ comprising the variables 1 up to w plus the response y and store their corresponding scores. Then, we move the window w variables forward and repeat this until we reach the end of the matrix X, where we calculate the final set of scores. The window size should be kept as constant as possible as is common in other data stream algorithms (e.g., Geppert et al. 2020; Vitter 1985). Therefore, w should be (roughly) a divisor of p. In a simulation study, we found that the size of w does not have a large impact on the result (see Figure A7 in Appendix A.2). The results indicate that w should be at least of order $\Omega(n \log n)$ to avoid rank-deficient submatrices, which is in line with our previous discussion. Above this value, the gain of using larger w declines and becomes negligible. It should thus be chosen as large as possible but small enough to ensure that we can perform the calculations on submatrices efficiently. However, we found a more critical trade-off in computation time between computing many small matrices or few large matrices, which will be discussed in detail in Section 7.

In the *Random Window approach*, the selection of variables in each window is chosen uniformly at random. For each window, we sample w out of p variables and add the response y, respectively. We repeat this a number of R times, but the algorithm also stops when every variable was chosen at least once before R steps. To ensure that every variable is chosen at least once, R should chose sufficiently large. If only one variable is chosen in each step, this can be quantified as $R = \Theta(p \log p)$ by the coupon collector's theorem (Erdős and Rényi 1961). If the window size w is increased, we take a sample of w variables without replacement at a time that only increases the inclusion probability of each variable. Therefore, if this is repeated independently $R = O(\frac{p}{w} \log p)$ times,

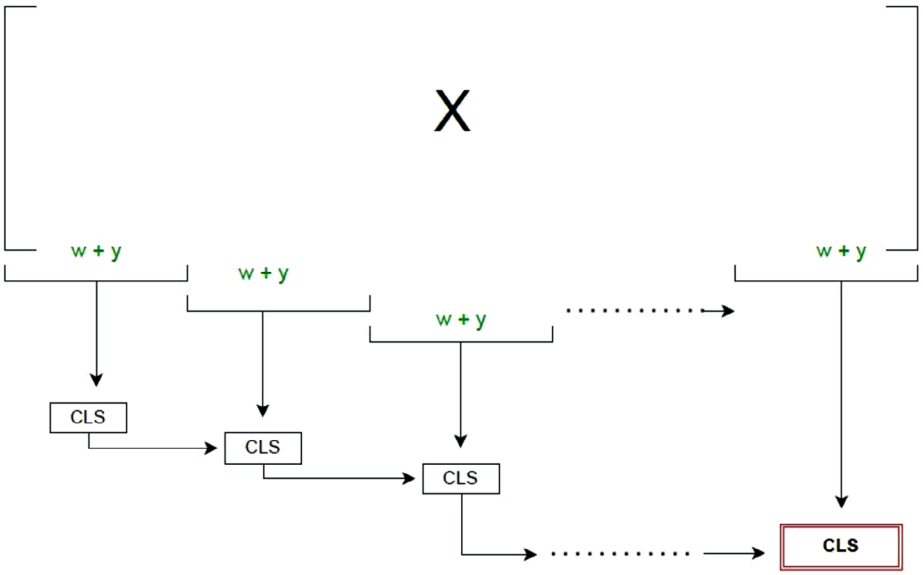

FIGURE 3 | The Sliding Window approach for $X \in \mathbb{R}^{n \times p}$. For one window, we consider a submatrix comprising w variables out of p and attach the response y to it. We slide through the whole matrix and calculate the cross leverage scores, respectively. At the end, we have a final set of cross leverage scores for all variables.

the resulting sample includes all variables with good probability. This will be discussed in more detail in Section 7. Analogously to previous considerations, w should be at least in $\Omega(n \log n)$ to avoid rank-deficient submatrices. In Section 7, we discuss the parameter choice for the different approaches.

Each individual window is sampled from the set of variables without replacement, but variables can be sampled repeatedly in different windows. We always store the corresponding scores to the variables for every window. When one variable is chosen repeatedly, we compare the new score to the previous and store only the largest. Hereby, we try to avoid a possibly bad influence of the Random Window selection on the estimates of the CLS. The Random Window approach is illustrated in Figure A2 in Appendix A.2.

The advantage of these approaches is that they enable calculating CLS in scenarios where the direct calculations fail because there are too many variables to compute the QR-decomposition as a whole. In genetic application, our data are often so large that we cannot even read it into the main memory of the computing environment such as R. Since our two approaches consider only small windows of the data at once, we only need to store and process the subset of data that we need for the current window. So, we mainly save a lot of memory by calculating the scores in that way, but indirectly also running time, since the two window approaches avoid swapping between the fast internal and slow external memory. However, one limitation of both approaches is that there are no theoretical guarantees available on the quality of the estimates they yield for the CLS. Therefore, we additionally consider a third approximation approach that provides rigorous theoretical guarantees.

3.3 | Sketching Approach

We would like to calculate the CLS as in Equation (3), but we need to avoid the costly QR-decomposition of the huge matrix $\tilde{X}$, which might even become intractable beyond some amount of data. The idea is to construct a significantly smaller "random sketch" of the input matrix of which we can calculate the QR-decomposition. We therefore project the columns of $\tilde{X}$ to a lower dimensional subspace by multiplying a properly chosen sketching matrix $\Pi \in \mathbb{R}^{r \times \tilde{p}}$ with $\tilde{X} \in \mathbb{R}^{\tilde{p} \times n}$:

$$\tilde{X}_\star = \Pi \tilde{X} \in \mathbb{R}^{r \times n}. \tag{4}$$

The *Sketching approach* is a method to approximate the CLS by means of sketching (Drineas et al. 2012), which is a common data reduction tool for the design of algorithms for large data, distributed data, and data streams (Munteanu 2023). This technique allows us to approximate the CLS of arbitrarily large data within an arbitrary precision parameterized by $\varepsilon > 0$. Our idea is transferring the idea of data reduction based on random projections and subsampling from reducing observations to selecting variables. To this end, we consider the transposed data matrix $\tilde{X} = [X, y]^T \in \mathbb{R}^{\tilde{p} \times n}$ again with $p \gg n$ and $\tilde{p} = p + 1$. We use a sketching matrix Π that is a further development of a Clarkson–Woodruff embedding (Clarkson and Woodruff 2017). Π is a sparse matrix, whose nonzero entries are $\{-1, 1\}$. The number of nonzero entries per column is fixed, but depends on the target dimension of the embedding. The original Clarkson–Woodruff embedding attains the sparsest possible structure, involving only one single nonzero entry per column. We give a simple example of such an embedding in Equation (5) for reducing from five to three dimensions.[1]

$$\begin{pmatrix} 0 & 0 & 1 & -1 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} x_3 - x_4 \\ -x_1 \\ x_2 + x_5 \end{pmatrix}. \tag{5}$$

[1] The example is taken from Nayebi, Munteanu, and Poloczek (2019).

ALGORITHM 1 | Approximation of the CLS of $\tilde{X}$.

Input: $\tilde{X} \in \mathbb{R}^{\tilde{p} \times n}$ $(\tilde{p} = p + 1)$

Output: $\hat{c}_{i\tilde{p}}, i \in \{1 \dots, p\}$

1: Project $\tilde{X}$ to a lower r-dimensional subspace to obtain $\tilde{X}_\star = \Pi\tilde{X} \in \mathbb{R}^{r \times n}$ using, e.g., $r = \frac{n \cdot \log n}{\varepsilon^2}$ (Cohen 2016)

2: Compute the QR-decomposition $\tilde{X}_\star = Q_\star R_\star$

3: Compute $\Omega = \tilde{X} R_\star^{-1}$ where $\Omega \in \mathbb{R}^{\tilde{p} \times n}$

4: Compute the CLS: $\hat{c}_{i\tilde{p}} = \langle \Omega_{i\cdot}, \Omega_{\tilde{p}\cdot} \rangle$

In general, there is a trade-off between the sparsity and the target dimension. The target dimension (r in Equation (4)) of the original Clarkson–Woodruff embedding is $r = \Theta(n^2)$. In our experiments, using oblivious subspace embeddings by Cohen (2016), we can sketch $\tilde{X}$ from $\tilde{p}$ down to only $r = O(\frac{n \log n}{\varepsilon^2})$ dimensions at the cost of a less sparse embedding with $O(\log n)$ nonzeros per column. We note that the most recent variant by Chenakkod et al. (2024) achieves even optimal size $r = O(\frac{n}{\varepsilon^2})$ with slightly worse sparsity of $O(\log^4 n)$ nonzeros per column. Despite these differences between sketching techniques, the main property, we are interested in the scope of this paper, is the approximation error bound depending on ε, which remains the same for all variants. We thus refer the interested reader to the following related literature for further comparison between different sketching techniques.

Other dense sketching approaches such as ε-JLT (Johnson and Lindenstrauss 1984), the Rademacher sketch Clarkson and Woodruff (2009), and the Subsampled Randomized Hadamard Transform (Ailon and Liberty 2009) are possible alternatives, which allow for different parameterizations in the trade-off between the time needed to multiply the sketch and the required number of rows. We refer to Geppert et al. (2017) for experimental comparisons and overview. However, as pointed out previously, these approaches neither achieve a better accuracy nor a lower target dimension nor faster running time than the sparse Cohen–Sketch. We thus focus on the Cohen–Sketch (Cohen 2016) in our presentation and experiments.

From $\tilde{X}_\star \in \mathbb{R}^{r \times n}$ in Equation (4), we can easily determine the QR-decomposition $\tilde{X}_\star = Q_\star R_\star$ and calculate $R_\star^{-1}$ in time and space independent of $\tilde{p}$. Now we can determine a matrix Ω as an approximation of the matrix Q from the original QR-decomposition $\tilde{X} = QR$:

$$\Omega = \tilde{X} R_\star^{-1} \in \mathbb{R}^{\tilde{p} \times n}.$$

Finally, we can compute the approximation of the CLS of every single variable with the $\tilde{p}$th variable (the attached response) by calculating the dot product of the respective rows of Ω:

$$\hat{c}_{i\tilde{p}} = \langle \Omega_{i\cdot}, \Omega_{\tilde{p}\cdot} \rangle, \quad i \in \{1, \dots, p\}.$$

The whole procedure is summarized in Algorithm 1. The output of the algorithm satisfies the following guarantee (cf. Drineas et al. 2012, Lemma 5) with respect to the original CLS $c_{i\tilde{p}}$:

$$\forall i \in \{1, \dots, p\} : |c_{i\tilde{p}} - \hat{c}_{i\tilde{p}}| \leq \varepsilon \|Q_{i\cdot}\|_2 \|Q_{\tilde{p}\cdot}\|_2 \leq \varepsilon. \quad (6)$$

See Drineas et al. (2012) for details. In particular, the guarantee given by Equation (6) implies that the large CLSs are well preserved. We remark that the largest entries of the CLS vector can be approximated within strictly lower dimensions than the original least squares problem Mai et al. (2023), which implies further computational benefits.

4 | Data

4.1 | Simulated Data

We simulate two different scenarios of SNP data, where we distinguish between three different genotypes: homozygous referent, homozygous variant, and heterozygous variant. These are usually encoded using the values $\{0, 1, 2\}$ and the last two indicate the presence of an SNP. In the first scenario S_1, we consider one two-way interaction and $p - 2$ noisy variables.

$$S_1 = (\text{SNP}_1 \wedge \text{SNP}_2).$$

For the number of variables, we choose $p \in \{2\,000, 20\,000, 200\,000, 2\,000\,000\}$. We always fix $n = 120$ observations and a binary outcome $y \in \{0, 1\}$. We simulate 1000 independent data sets for $p \in \{2\,000, 20\,000\}$ and 100 data sets for $p \in \{200000, 2000000\}$. We use the function `simulateSNPglm` from the `R` package `scrime` (Schwender 2018) to simulate the SNP data. All data are generated from the model

$$Y \sim \text{Bernoulli}(\texttt{pred})$$

$$\texttt{pred} = \frac{1}{1 + \exp(-\texttt{lin.pred})},$$

where $\texttt{lin.pred} = \beta_0 + \beta_1 M_1$, and $M_1 = (\text{SNP}_1 \wedge \text{SNP}_2)$. Further, we choose $\beta_0 = \log(\frac{0.3}{1-0.3})$ and $\beta_1 = \log(50)$. These values for β are chosen such that the specified interactions have a relatively large effect on the target variable. So, the probability that $y = 1$ is 0.3 if M_1 is `FALSE` and roughly 0.95 if M_1 is `TRUE`. For all SNPs, we draw the minor allele frequency from a uniform distribution on the interval $[0.15, 0.45]$. To validate how the approaches can deal with more complex interactions, we design a second scenario

$$S_2 = (\text{SNP}_1 \wedge \text{SNP}_2) \vee (\text{SNP}_3 \wedge \text{SNP}_4).$$

For S_2, we choose $\texttt{lin.pred} = \beta_0 + \beta_1 M_1 + \beta_2 M_2$ with $M_1 = (\text{SNP}_1 \wedge \text{SNP}_2)$, $M_2 = (\text{SNP}_3 \wedge \text{SNP}_4)$, and $\beta_0 = \log(\frac{0.3}{1-0.3})$ and $\beta_1 = \beta_2 = \log(50)$.

All simulated data are available at Zenodo (Teschke 2024).[2] The code is available on `GitHub`.[3]

4.2 | Real Data

We also consider a small real data application, for which we use the HapMap data set taken from the `R` package `SNPassoc` (Moreno, Gonzalez, and Pelegri 2022). This data set consists of

[2] https://doi.org/10.5281/zenodo.12742957
[3] https://github.com/SvenTeschke/special_issue_CEN

TABLE 1 | This table shows the values we have chosen for the parameters of the different methods. The choice of values is discussed in Section 7. Unsuitable parameters are highlighted in bold.

	Random window		Sliding window	Sketching
p	**w**	**R**	**w**	ε
2 000	200	200	200	$\varepsilon \in \{\mathbf{0.5}, \mathbf{0.2}, \mathbf{0.1}\}$
20 000	2 000	500	2 000	$\varepsilon \in \{\mathbf{0.5}, \mathbf{0.2}, \mathbf{0.1}\}$
200 000	2 000	1 000	2 000	$\varepsilon \in \{0.5, 0.2, 0.1\}$
2 000 000	2 000	2 000	2 000	$\varepsilon \in \{0.5, 0.2, 0.1\}$

9307 SNPs that belong to 22 chromosomes from the HapMap project (Thorisson et al. 2005) of the National Human Genome Research Institute. Additionally, the data contain information on the individuals' origin. We distinguish between European (CEU) and Yoruba (YRI) to encode the binary response. We note that the HapMap data set is rather small. Nevertheless, we decided to use it because we are primarily interested in a comparison between the approximation and the original, nonapproximated CLS method. For conducting the latter analysis, the data must necessarily have a small and tractable dimension p. Furthermore, our choice of the well-researched HapMap data set is a baseline reference and thus allows comparisons to other related work.

5 | Results of Simulation Study

All calculations were performed using `R` version `R 4.3.0` (R Core Team 2023). In the following, we focus on the two simulated scenarios comprising a single two-way interaction and two two-way interactions, respectively. Hereby, we address the problem that arises when there is so much data to process, that the computations become too expensive and we can no longer calculate the CLS in the conventional way. To evaluate how well the approaches work, we count how many of the important variables we find on average when we choose the q variables with the largest CLS out of different dimensions p. As discussed above, we set $q = \lceil n \log n \rceil = 575$ with $n = 120$. For the different approaches, we choose the parameter settings summarized in Table 1. The specific choices are evaluated and discussed later in Section 7.

In the sketching approach, we reduce the dimension from p to $r = \lceil \frac{n \log n}{\varepsilon^2} \rceil$. It holds that $r \in \{2\,300, 14375, 57500\}$ for $n = 120$ and $\varepsilon \in \{\mathbf{0.5}, \mathbf{0.2}, \mathbf{0.1}\}$, respectively. Note that some combinations of ε and p are not meaningful, since for very small ε and small values of p, we would increase the dimension p, in which case it is advisable to preserve the initial dimension. Nevertheless, we will use all ε for all p in our evaluation. For saving running time in the Random Window approach, we use optimized values $R \lesssim \frac{p}{w} \log p$, that is, lower than discussed in Sections 3.2 and 7. For comparing the performance of our new approaches to standard methods, we calculate the correlations of each variable with the response and select the q variables attaining the largest correlations. Additionally, we compare to uniform sampling as a standard baseline.

In Scenario 1 (see Table A1 in Appendix A.2), all methods find on median average both important variables for $p = 2\,000$ and $p = 20\,000$. For $p = 200\,000$ and $p = 2\,000\,000$, we still find one out of two important variables on median average for all methods, whereby the CLS can only be calculated with approximation methods for $p = 2\,000\,000$. The choice of the median is reasonable because the median is a robust measure of location and the number of variables found is dependent on the data set. To get a better comparison of the methods, we also consider the nonrobust mean. In Figure 4, we plot the mean number of important variables (out of two) found for different values of q and p across the various approaches.

We see that also for smaller values of q, we find on average almost two out of two important variables for all approaches for $p = 2\,000$ and $p = 20\,000$. For an increasing q, the number of variables found on average increases rapidly in the beginning and flattens out later, but we see that a smaller value of q would also suffice in practice. All methods are working well and the approximations differ only slightly. The gray line shows what we would expect if we select q variables uniformly at random. We see that this is consistently outperformed. For higher dimensions $p = 200\,000$ and $p = 2\,000\,000$, we see a similar pattern, but it should be noted that the plot includes considerably larger values of q on the horizontal axis here. Again, all approaches perform similarly well. However, on average, we do not find both important variables this time for the recommended q, but again, we clearly outperform the expected value of variables chosen uniformly at random (gray line). It seems that correlations work best in this case, but as we have seen in the toy example in Section 2, it cannot be relied upon in general. The plot indicates that we should choose a larger q for very large p.

In the more complex Scenario 2, we first investigate how many out of the four important variables we find on median average (see Table A2 in Appendix A.2). We still find all important variables when selecting the variables with the largest CLS. This holds for both, the regular calculation of the CLS, and for the sketching methods, but also for the baseline using correlations. For larger p, we do not find all important variables any more, but still 3/4 for $p = 20\,000$, and at least 1/4 for $p = 200\,000$ and $p = 2\,000\,000$ for all methods. In this case, it becomes interesting how often we find the respective complete interactions. Note that if we find only two important variables in total, it makes a difference whether we find a complete interaction or whether we find only one variable out of each interaction. For this purpose, we show how many variables we find on average for different values of q and different approaches for each of the two interactions separately. In Figure 5, we plot the curves for $p = 2\,000$.

The case $p = 20\,000$ can be found in Figure A3 in Appendix A.2. As can be seen from Table A2 in Appendix A.2, both interactions for $p = 2\,000$ are very likely to be found for $q = 575$. However, it is remarkable how well the first interaction is found even for a smaller q. Even for $p = 20\,000$, we found on average nearly all the important variables from the first interaction. For $p = 2\,000$ as well as for $p = 20\,000$, the first interaction is detected slightly more often for smaller values of q. The plots for larger values of p can be found in Figure A4 and Figure A5 in Appendix A.2. Also for $p = 200\,000$, we see that the first interaction was found in more cases. For $p = 2\,000\,000$, the two interactions are found

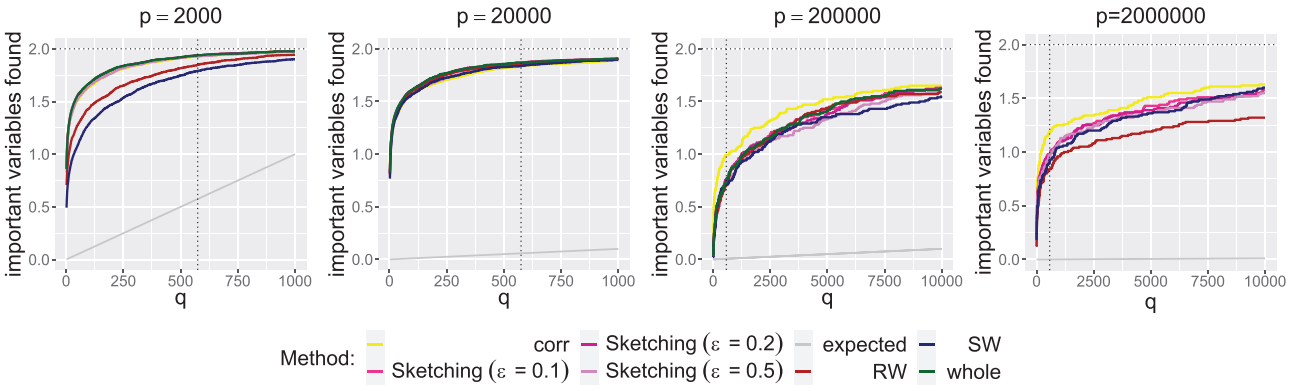

FIGURE 4 | Important variables out of two important interacting variables (horizontal dotted line) found on average if we select the variables with the q largest CLS. At $q = 575$, there is a vertical dotted line. We distinguish between the approaches by color. The gray line shows what we would expect if we select q variables uniformly at random.

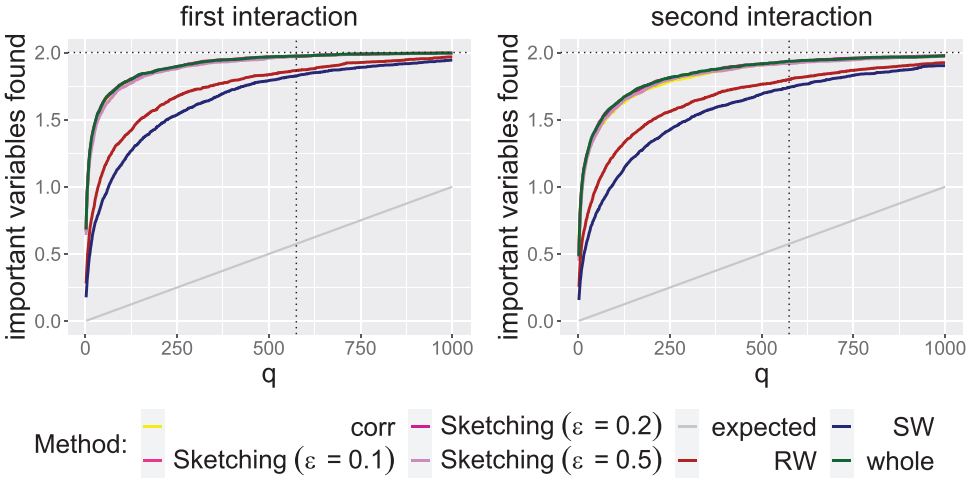

FIGURE 5 | Important variables out of two (horizontal dotted line) found on average if we select the variables with the q largest CLS for the single two-way interactions in the second scenario with $p = 2\,000$. At $q = 575$, there is a vertical dotted line. We distinguish between the approaches by color. Left: results for the first interaction. Right: results for the second interaction.

roughly equally often. Again, it should be noted that the plot includes considerably larger values of q on the horizontal axis. It can be seen, however, that for smaller q, significantly fewer of the important variables are found here. For increasing p, q should also be increased, if necessary. Nevertheless, a selection with CLS outperforms a uniform random selection. Again, it can be stated that all methods, including the approximation methods, achieve similar results.

6 | Real Data Application and Predictive Performance

In this section, we investigate how the variable selection using CLS affects predictive performance. To this end, we consider Random Forests (Breiman 2001) for the HapMap data, and logic regression (Ruczinski, Kooperberg, and LeBlanc 2003) for another simulated data set. The simulated data set is generated analogously to scenario S_1 from the previous section, but we consider only small values of p, so that we can apply logic regression to the full data for the sake of comparison. Our aim is to investigate whether we can improve the predictive performance by applying the variable selection methods instead of considering the whole data set.

6.1 | Random Forests on HapMap Data

The data consists of $p = 7\,648$ SNPs with a binary response and $n = 120$ observations in total. We use Random Forests here to handle the full data with such a "large" value of p, which would not be possible using logic regression. We use the `R`-package `ranger` (Wright and Ziegler 2017), with default parameter settings. To ensure that the measured effect on the performance comes from the specific choice of variable selection methods, and is not only a consequence of the mere variable reduction, we also consider the performance gain using a uniform random sample of variables as a baseline we would like to improve upon.

We perform the variable selection on a training data set on which we also train the Random Forest (using only the selected variables) and then test the predictive performance on the remaining test data. The fraction of training data is chosen to be 2/3, while the remaining 1/3 fraction represents the test data. Since we have

$n = 120$ observations, this means that 80 observations are used for training and 40 for testing.

Note that for the whole data set without preselection, we already have a prediction error of 0%. Thus, we choose a small value of $q = 10$. We want to see what happens when we select $q = 10$ random variables compared to a selection of $q = 10$ variables by the largest CLS or correlations, respectively. We see (Table A3 in Appendix A.2) that using a uniform random sample, the prediction error increases to a nonzero error rate of 12.5%. On the other hand, the prediction error after variable selection according to the CLS remains at 0%, This supports that *meaningful* variables were selected by CLS (and correlations).

6.2 | Logic Regression on Simulated Data

To consider the predictive performance with logic regression, we need a data set with considerably smaller p. We simulate again 1000 independent data sets with one two-way interaction in the same way as for simulation scenario S_1, but only for $p = 500$. Analogously to the HapMap data, we consider the predictive performance comparing no variable selection, variable selection with CLS or correlations, and a uniform random sample of variables. We split the data into test and training set and select $q = 25$ variables based on the training data. For the calculations, we use the `R`-package `logicDT` (Lau 2023). In Table A4 in Appendix A.2, we summarize the median prediction errors for all methods.

We see that the variable selection with CLS and correlations yield a prediction error of 35%. According to the t-test, both outperform the model without variable selection significantly, which have a prediction error of 37.5%. Again, a selection by uniform sample has a much worse prediction error of 47.5%. In all cases, the prediction quality with logic regression is quite low,[4] but we have shown that it improves significantly by using variable selection with CLS. And even in better performing settings such as Random Forests on the HapMap data, variable selection with CLS is useful.

Finally, we are interested in the variable importance measures (VIMs) of the logic terms identified by the logic regression model (Schwender and Ickstadt 2008). We obtain VIMs from the model for various main effects as well as for interaction effects of different orders. In the `logicDT` package, we use the parameter setting `vim_type = ''logic''` and `ave = ''before''`, which compares the average performance over fixing one Boolean variable to $\{0, 1\}$, respectively. We refer to Lau (2023) for details on available parameter settings. We would like to investigate whether we can obtain better results for the VIMs after performing a variable selection using CLS. To this end, we select the $q = 25$ variables with the largest CLS on the training data, and calculate the VIMs on the test data. We also consider the VIMs using a uniform random sample of $q = 25$ variables for the sake of comparison. Additionally, we calculate the VIMs in the same way without any preselection. We count how often the important interaction between X_1 and X_2 is among the top v most important variables, respectively, logic terms over 1000 data sets in terms of the calculated VIM. In Figure 6, it is shown that applying a variable selection method beforehand results in a considerable improvement, and we find the relevant interaction more often in the top v of logical terms measured by the variable importance.

[4] Note that its main purpose is detecting interacting variables rather than making preditions.

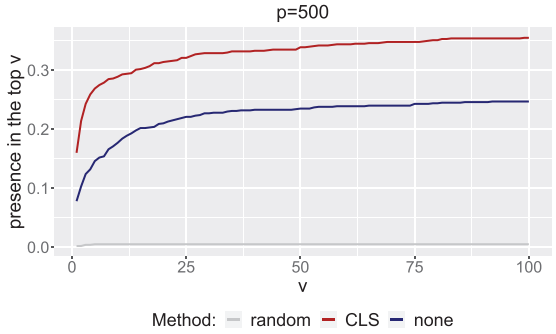

FIGURE 6 | Presence of interaction between X_1 and X_2 in the top v important variables in the logic regression model: without variable selection (blue line), selection by the q largest CLS (red line), and uniform random selection of q variables (gray line).

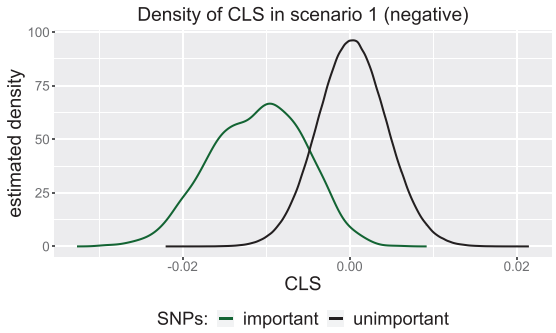

FIGURE 7 | Kernel density estimates of the CLS of the important (green) and unimportant (black) variables in the scenario where an existing two-way interaction has an negative influence on the response ($p = 2\,000$).

7 | Discussion and Conclusion

When we consider more complex scenarios, such as, for instancethree- or four-way interactions, identifying all variables of the respective interaction becomes more difficult. However, if p is not too large, our methods still work. In case of $p = 2\,000$, we still find all important variables for the three-way interaction and at least two out four important variables for the four-way interaction on median average, when selecting these variables with the $q = 575$ largest CLS. For larger p, this becomes more difficult and the results deteriorate. Known findings in high-dimensional statistics suggest that a $\log(p)$ dependence might be necessary (Amini and Wainwright 2009; Mai et al. 2023).

So far, we have only considered scenarios where an interaction of SNPs has a positive effect on the response. In this case, it makes sense to select the variables attaining the largest CLS. However, if we consider scenarios in which the absence of an SNP interaction favors a disease ($y = 1$), then the corresponding important variables have a largely *negative* CLS. This is illustrated in Figure 7 (in contrast to Figure 1).

TABLE 2 | This table shows values for $R \approx \frac{p}{w} \log p$ in the Random Window approach.

p	w	R
2 000	200	76
20 000	2 000	99
200 000	2 000	1221
2 000 000	2 000	14,509

So, if we do not know anything about the data, it is recommended to select the variables attaining the largest absolute values of their CLS. Nevertheless, also in scenario S_1, the absolute scores could still be used to distinguish the important from the unimportant variables, see Figure A6 in Appendix A.2.

In our paper, we select exactly the q variables with the largest (absolute) CLS. But, in the presence of outliers and to preserve the whole space spanned by the data, it was recommended to sample q variables weighted by their CLS including a small fraction of standard leverage scores (Parry et al. 2021) instead of greedily selecting the largest scores. However, in this paper, we used the CLS as a measure of importance of individual variables that justifies selecting the variables without sampling.

Another point to discuss is the choice of parameters for the two window-based approaches. For the following analyses, we considered the scenario S_1 with one two-way interaction and $p = 20\,000$. How should we choose the window width w in the Sliding Window and Random Window approaches? First, the smaller we choose the window size w, the faster we can calculate the QR-decomposition, but this also increases the number of QR-decompositions to be computed. Our analyses have shown that w should not be chosen too small, in particular, $w \gtrsim n \log n$, but beyond some threshold, its choice does not make a big impact on the performance. Figure A7 in Appendix A.2 illustrates this. However, the intuitive reasoning holds that the larger w is, the better interactions can be detected on average. This applies to both window-based approaches. In the Random Window approach, a larger value of w implies that it takes less iterations until each variable is drawn at least once, but the individual computations for the QR-decompositions in each iteration take longer. Similar considerations apply to the parameter R, which determines how many windows we consider in the Random Window approach. A large number provides an increased running time, but no performance improvement, as long as R is chosen large enough so that each variable is selected at least once (with high probability), which is the case for $R \approx \frac{p}{w} \log p$. In Table 2, we show the respective values for R. Nevertheless, for the specification of R and w, it is recommended to consider the trade-off between how fast we want to calculate the scores and how large we (have to) choose the individual submatrices. In the sketching approach, we can vary the parameter ε, which determines the accuracy guarantee of the approximation of the CLS and provides a trade-off to the size of the sketched matrix that again has a crucial impact on the time and space required to perform the QR-decomposition.

In conclusion, we showed that the CLSs are appropriate for distinguishing important from unimportant variables, even if they only have an indirect influence on the response through variable interactions. The advantage of CLS is that they can be calculated directly from the data matrix and that it is not necessary to consider the vast space of every possible variable interaction. Especially for higher order interactions and more complex scenarios, this brings massive advantages compared to standard methods. Our approximations to CLS work reliably and interaction effects can be identified even for very large p in the order of millions, where previous methods face severe computational limitations. With the selected variables via the CLS, conventional downstream regression and classification analyses can be applied, which would fail on the original high-dimensional input. The previous selection using CLS has a positive influence on their predictive performance and calculation of VIMs. Not to mention that many methods would simply not be applicable without a prior variable selection respectively reduction in the presence of massive amounts of data commonly considered in genome-wide analyses. Even though we have limited this paper to SNP data and binary response, the methods presented here can also be applied to arbitrary real-valued data types.

Acknowledgments

The authors would like to thank Leo N. Geppert, Katharina Parry, and Rieke Deborah Moeller-Ehmcke for valuable discussions and for providing parts of their code.

Open access funding enabled and organized by Projekt DEAL.

Conflicts of Interest

The authors have declared no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

References

Ailon, N., and E. Liberty. 2009. "Fast Dimension Reduction Using Rademacher Series on Dual BCH Codes." *Discretete Computational Geometry* 42, no. 4: 615–630.

Amini, A. A., and M. J. Wainwright. 2009. "High-Dimensional Analysis of Semidefinite Relaxations for Sparse Principal Components." *The Annals of Statistics* 37, no. 5B: 2877–2921.

Ben-Israel, A., and T. N. Greville. 2003. *Generalized Inverses: Theory and Applications*. Vol 15. New York: Springer Science & Business Media.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45, no. 1: 5–32.

Chenakkod, S., M. Derezinski, X. Dong, and M. Rudelson. 2024. "Optimal Embedding Dimension for Sparse Subspace Embeddings." In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, 1106–1117.

Chuang, L.-Y., Y.-D. Lin, H.-W. Chang, and C.-H. Yang. 2014. "SNP-SNP Interaction Using Gauss Chaotic Map Particle Swarm Optimization to Detect Susceptibility to Breast Cancer." In *2014 47th Hawaii International Conference on System Sciences*, 2548–2554.

Clarkson, K. L., and D. P. Woodruff. 2009. "Numerical Linear Algebra in the Streaming Model." In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, 205–214.

Clarkson, K. L., and D. P. Woodruff. 2017. "Low-Rank Approximation and Regression in Input Sparsity Time." *Journal of the ACM* 63, no. 6: 1–45.

Cohen, M. B. 2016. "Nearly Tight Oblivious Subspace Embeddings by Trace Inequalities." In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 278–287.

Demmel, J., L. Grigori, M. Hoemmen, and J. Langou. 2012. "Communication-optimal Parallel and Sequential QR and LU Factorizations." *SIAM Journal on Scientific Computing* 34, no. 1: A206–A239.

Ding, Z., K. Ickstadt, and A. Munteanu. 2022. "Bayesian Analysis for Dimensionality and Complexity Reduction." In *Machine Learning under Resource Constraints - Volume 3: Applications*, 58–70. Berlin: De Gruyter.

Drineas, P., M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. 2012. "Fast Approximation of Matrix Coherence and Statistical Leverage." *Journal of Machine Learning Research* 13: 3475–3506.

Erdős, P., and A. Rényi. 1961. "On a Classical Problem of Probability Theory." *"A Magyar Tudományos Akadémia Matematikai. Kutató Intézetének Közleményei* 6: 215–220.

Geppert, L. N., K. Ickstadt, A. Munteanu, J. Quedenfeld, and C. Sohler. 2017. "Random Projections for Bayesian Regression." *Statistical Computation* 27, no. 1: 79–101.

Geppert, L. N., K. Ickstadt, A. Munteanu, and C. Sohler. 2020. "Streaming Statistical Models via Merge & Reduce." *International Journal of Data Science and Analytics* 10, no. 4: 331–347.

Golub, G. H., and C. F. Van Loan. 1996. *Matrix Computations*. Baltimore: The Johns Hopkins University Press.

Greliche, N., M. Germain, J. Lambert, et al. 2013. "A Genome-Wide Search for Common SNP x SNP Interactions on the Risk of Venous Thrombosis." *BMC Medical Genetics* 14, no. 36: 1.

Hoaglin, D. C., and R. E. Welsch. 1978. "The Hat Matrix in Regression and ANOVA." *The American Statistician* 32, no. 1: 17–22.

Johnson, W., and J. Lindenstrauss. 1984. "Extensions of Lipschitz Maps into a Hilbert Space." *Contemporary Mathematics* 26: 189–206.

Lau, M. 2023. *logicDT: Identifying Interactions Between Binary Predictors*. R package version 1.0.3.

Li, P., M. Guo, C. Wang, X. Liu, and Q. Zou. 2014. "An Overview of SNP Interactions in Genome-Wide Association Studies." *Briefings in Functional Genomics* 14, no. 2: 143–155.

Mai, T., A. Munteanu, C. Musco, A. Rao, C. Schwiegelshohn, and D. P. Woodruff. 2023. "Optimal Sketching Bounds for Sparse Linear Regression." In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 11288–11316.

Miller, D. J., Y. Zhang, G. Yu, et al. 2009. "An Algorithm for Learning Maximum Entropy Probability Models of Disease Risk That Efficiently Searches and Sparingly encodes Multilocus Genomic Interactions." *Bioinformatics* 25, no. 19: 2478–2485.

Moreno, V., J. R. Gonzalez, and D. Pelegri. 2022. *SNPassoc: SNPs-Based Whole Genome Association Studies*. R package version 2.1-0.

Munteanu, A. 2023. "Coresets and Sketches for Regression Problems on Data Streams and Distributed Data." In *Machine Learning under Resource Constraints, Volume 1 - Fundamentals*, 85–98. Berlin, Boston: De Gruyter.

Nayebi, A., A. Munteanu, and M. Poloczek. 2019. "A Framework for Bayesian Optimization in Embedded Subspaces." In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 4752–4761.

Parry, K., L. N. Geppert, A. Munteanu, and K. Ickstadt. 2021. "Cross-Leverage Scores for Selecting Subsets of Explanatory Variables." *arXiv preprint 2109.08399*.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ruczinski, I., C. Kooperberg, and M. LeBlanc. 2003. "Logic Regression." *Journal of Computational and Graphical Statistics* 12, no. 3: 475–511.

Schwender, H. 2018. *Scrime: Analysis of High-Dimensional Categorical Data Such as SNP Data*. R package version 1.3.5. https://CRAN.R-project.org/package=scrime.

Schwender, H., and K. Ickstadt. 2008. "Identification of SNP interactions using logic regression." *Biostatistics* 9: 187–198.

Schwender, H., S. Selinski, M. Blaszkewicz, et al. 2012. "Distinct SNP Combinations Confer Susceptibility to Urinary Bladder Cancer in Smokers and Non-Smokers." *PLoS One* 7: 1–12.

Terada, A., R. Yamada, K. Tsuda, and J. Sese. 2016. "LAMPLINK: Detection of Statistically Significant SNP Combinations From GWAS Data." *Bioinformatics* 32, no. 22: 3513–3515.

Teschke, S. 2024. "(simulated Data:) Detecting Interactions in High Dimensional Data Using Cross Leverage Scores." Zenodo. https://doi.org/10.5281/zenodo.12742957.

Thorisson, G. A., A. V. Smith, L. Krishnan, and L. D. Stein. 2005. "The International HapMap Project Web site." *Genome Research* 15, no. 11: 1592–1593.

Tropp, J. A. 2011. "Improved Analysis of the Subsampled Randomized Hadamard Transform." *Advances in Adaptive Data Analysis* 3, no. 1–2: 115–126.

Uffelmann, E., Q. Huang, N. Munung, et al. 2021. "Genome-Wide Association Studies." *Nature Reviews Methods Primers* 1, no. 59.

Vitter, J. S. 1985. "Random Sampling With a Reservoir." *ACM Transactions on Mathematical Software* 11, no. 1: 37–57.

Wright, M. N., and A. Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77, no. 1: 1–17.

Wu, T. T., Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. 2009. "Genome-Wide Association Analysis by LASSO Penalized Logistic Regression." *Bioinformatics* 25, no. 6: 714–721.

Yang, C., Z. He, X. Wan, Q. Yang, H. Xue, and W. Yu. 2008. "SNPHarvester: A Filtering-Based Approach for Detecting Epistatic Interactions in Genome-Wide Association Studies." *Bioinformatics* 25, no. 4: 504–511.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix A

A.1 | Poof of Proposition 3.1

Proof. Let $\tilde{X} = [X, y] \in \mathbb{R}^{n\times p+1}$, for $p \gg n$, where X and thus also $\tilde{X}$ have full rank n. Consider the ℓ_2 regression problem

$$\min_{\beta\in\mathbb{R}^p} \|X\beta - y\|_2^2 . \tag{A1}$$

Since $p \gg n$ the problem is underconstrained and therefore there exists a subspace $L \subset \mathbb{R}^p$ such that any $\beta \in L$ satisfies $X\beta = y$. Consequently, any $\beta \in L$ constitutes an optimal solution where Equation (A1) equals 0.

Now, consider the (thin) singular value decomposition (Golub and Van Loan 1996) $\tilde{X} = U\Sigma V^T$, where $U \in \mathbb{R}^{n\times n}$ is orthonormal, $\Sigma \in \mathbb{R}^{n\times n}$ is diagonal whose entries are the singular values $\sigma_i(\tilde{X}) > 0$, for $i \in [n]$, and $V^T \in \mathbb{R}^{n\times p+1}$ has orthonormal rows. We further let $V^T = [V_X^T, V_y^T]$,

where V_X^T carries the first p columns corresponding to the columns of X, while V_y^T is the $(p+1)$st column that corresponds to y. With these decompositions in place, Equation (A1) can be rewritten as the minimization of

$$\|X\beta - y\|_2^2 = \|U\Sigma(V_X^T\beta - V_y^T)\|_2^2.$$

In particular, note that it implies

$$\|X\beta - y\|_2^2 = 0 \quad \Longleftrightarrow \quad \|V_X^T\beta - V_y^T\|_2^2 = 0.$$

Next, we make use of the fact that a zero error solution can be obtained in closed form by means of the Moore–Penrose pseudoinverse (Ben-Israel and Greville 2003), which for the RHS yields the (unique) least ℓ_2-norm solution

$$\beta^{OLS} = V_X(V_X^TV_X)^{-1}V_y^T.$$

If V_X^T were orthonormal, this would simplify even further to the vector comprising the CLS as its coordinates, since in that case $(V_X^TV_X)^{-1} = I^{-1} = I$ and we have $c_{\cdot,p+1} = V_XV_y^T$ by definition. However, V_X is not quite orthonormal since it was obtained by removing one row vector V_y from the orthonormal basis V. But since we removed only one row of small, bounded norm, we can argue that it is close to being orthonormal in what follows. Interpreting the matrix inner product as a sum of rank-1 tensors, we have that

$$\begin{aligned}\|V^TV - V_X^TV_X\|_2 &= \left\|\sum_{i=1}^{p+1} V_i^TV_i - \sum_{i=1}^{p} V_i^TV_i\right\|_2 \\ &= \|V_y^TV_y\|_2 = \|V_y\|_2^2. \end{aligned} \tag{A2}$$

We now assume that after normalization of the space, the columns are well aligned with the normalized response vector. Note that this is natural to assume since otherwise the regressors would be (almost) orthogonal to the response, in which case a regression model would not make any sense at all. Formally, we require that for some data dependent constant $\eta \in (0, 1/2]$ the following holds

$$\sum_{i=1}^{p}\left(V_i\frac{V_y^T}{\|V_y\|_2}\right)^2 \geq \frac{1}{\eta}\|V_y\|_2^2. \tag{A3}$$

Note that the threshold is independent of the data dimensions and is *not* required to grow with the numbers of observations or variables. We only require some *constant* amount of the projected information to be aligned with the column that belongs to the response variable.

Now consider the standard leverage score of the response variable, which we define and bound from above as

$$\ell_y = \sup_{u\in\mathbb{R}^n\setminus\{0\}}\frac{|V_y^Tu|^2}{\|Vu\|_2^2} \overset{CSI}{\leq} \frac{\|V_y^T\|_2^2\|u\|_2^2}{\|Vu\|_2^2} = \|V_y\|_2^2,$$

where we used the Cauchy–Schwarz inequality (CSI). We further note that strict equality holds for the unit vector $u = V_y^T/\|V_y^T\|_2$. Consequently, it follows that

$$\begin{aligned}\|V_y\|_2^2 &= \|V_y^T\|_2^2\left\|\frac{V_y^T}{\|V_y^T\|_2}\right\|_2^2 \Big/ \left\|V\frac{V_y^T}{\|V_y^T\|_2}\right\|_2^2 \\ &= \frac{\|V_y^T\|_2^2}{\left(V_y\frac{V_y^T}{\|V_y^T\|_2}\right)^2 + \sum_{i=1}^{p}\left(V_i\frac{V_y^T}{\|V_y^T\|_2}\right)^2} \\ &\overset{(A3)}{\leq} \frac{\|V_y\|_2^2}{\|V_y\|_2^2\left(1+\frac{1}{\eta}\right)} \leq \eta.\end{aligned}$$

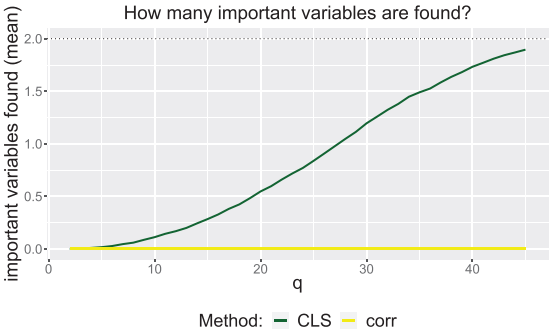

FIGURE A1 | Toy example: On the y-axis, we plot how many of the first two important variables we found on average when we choose the variables with the q largest cross leverage scores (green), respectively, largest correlation (yellow).

Combining the previous calculations (A2) with our assumption (A3) thus implies that

$$\|I - V_X^TV_X\|_2 \leq \eta. \tag{A4}$$

To get a bound on the β^{OLS} estimator above, we need a similar bound for the inverse $(V_X^TV_X)^{-1}$ quantifying the amount by which it deviates from the identity $(V^TV)^{-1} = I$.

To this end, consider the eigenvalue decomposition $V_X^TV_X = Q\Lambda Q^T$, where $Q \in \mathbb{R}^{n\times n}$ is an orthonormal basis satisfying $Q = Q^T = Q^{-1}$ and thus $QQ^T = Q^TQ = I$, and $\Lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose entries are the eigenvalues $\lambda_i = \sigma_i^2(V_X) > 0$, for $i \in [n]$.

We have

$$\begin{aligned}\|(V^TV)^{-1} - (V_X^TV_X)^{-1}\|_2 &= \|I - (V_X^TV_X)^{-1}\|_2 \\ &= \|QQ^T - (Q\Lambda Q^T)^{-1}\|_2 \\ &= \|QQ^T - Q\Lambda^{-1}Q^T\|_2 \\ &= \|I - \Lambda^{-1}\|_2 \\ &= \max_{i\in[n]}\left|1 - \frac{1}{\lambda_i}\right| = \max_{i\in[n]}\left|\frac{1-\lambda_i}{\lambda_i}\right| \\ &\leq \frac{\max_{i\in[n]}\left|1-\sigma_i^2(V_X)\right|}{\min_{i\in[n]}\sigma_i^2(V_X)} \overset{(A4)}{\leq} \frac{\eta}{1-\eta} \leq 2\eta.\end{aligned}$$

Then, it follows that

$$\begin{aligned}\max_{i\in[p]}|\beta_i^{OLS} - c_{i,p+1}| &= \|V_X(V_X^TV_X)^{-1}V_y^T - V_XV_y^T\|_\infty \\ &= \|V_X(V_X^TV_X)^{-1}V_y^T - V_X(V^TV)^{-1}V_y^T\|_\infty \\ &\leq \|V_X((V_X^TV_X)^{-1} - (V^TV)^{-1})V_y^T\|_2 \\ &\leq \|V_X\|_2\|(V_X^TV_X)^{-1} - (V^TV)^{-1}\|_2\|V_y^T\|_2 \\ &\leq 1\cdot 2\eta\cdot\sqrt{\eta} < \frac{3\eta}{2},\end{aligned}$$

which implies that each CLS$c_{i,p+1}$ for $i \in [p]$ equals their corresponding parameter in the least ℓ_2-norm solution up to a small additive error bounded in terms of η. □

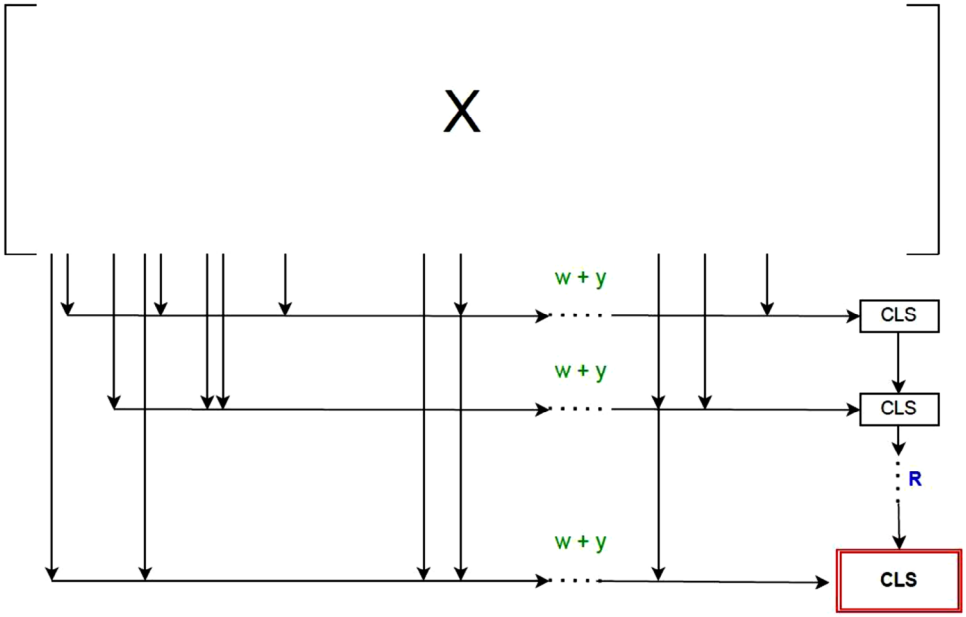

FIGURE A2 | The Random Window approach for $X \in \mathbb{R}^{n \times p}$. For each window, we consider a random submatrix comprising w out of p variables and attach the response y. We then calculate the cross leverage scores for the submatrix. We repeat this R times to get a final set cross leverage score to the sampled variables.

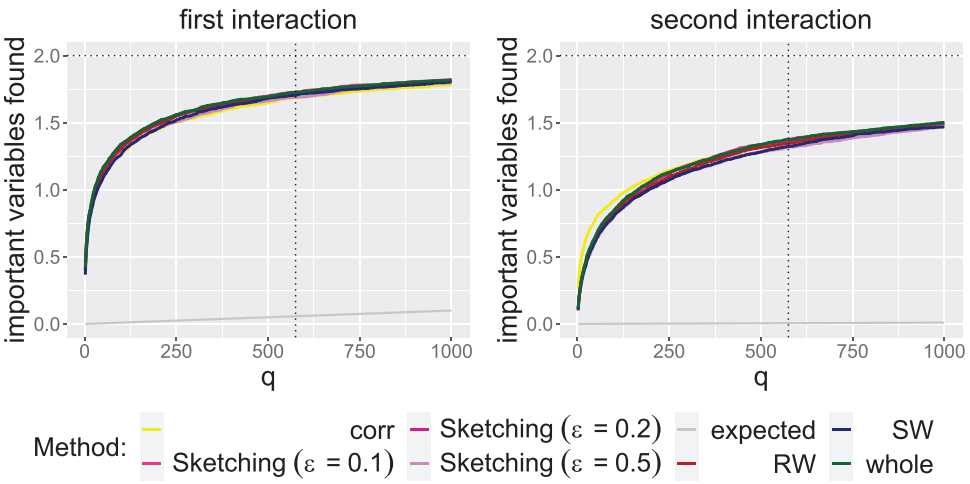

FIGURE A3 | Important variables out of two (horizontal dotted line) found on average if we choose the variables with the q largest CLS for the single two-way interactions in the second scenario with $p = 20\,000$. At $q = 575$, there is a vertical dotted line. We distinguish between the approaches by color. On the left side, we plotted the results for the first interaction and on the right side for the second.

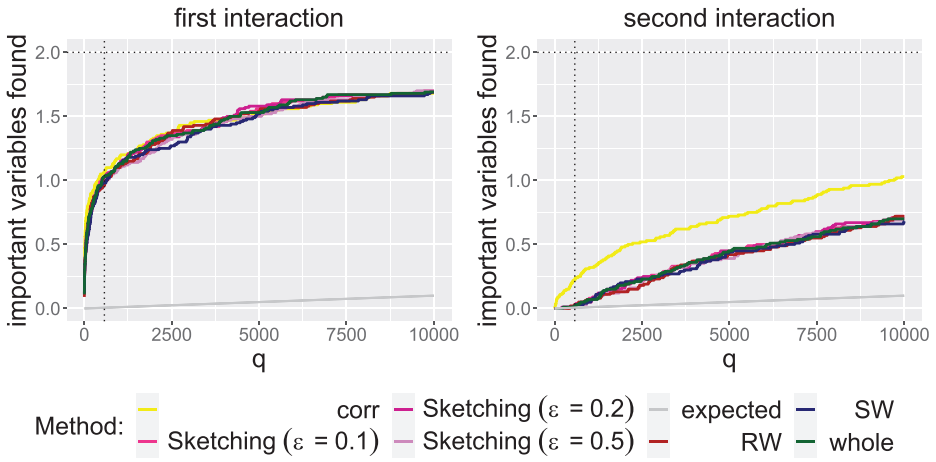

FIGURE A4 | Important variables out of two (horizontal dotted line) found on average if we choose the variables with the q largest CLS for the single two-way interactions in the second scenario with $p = 200\,000$. At $q = 575$, there is a vertical dotted line. We distinguish between the approaches by color. Left: results for the first interaction. Right: results for the second interaction.

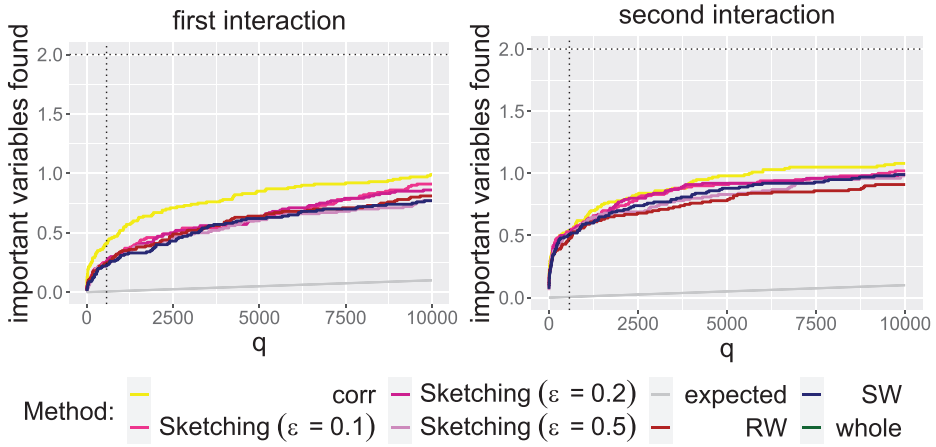

FIGURE A5 | Important variables out of two (the vertical dotted line) found on average if we choose the variables with the q largest CLS for the single two-way interactions in the second scenario with $p = 2\,000\,000$. At $q = 575$, there is a horizontal dotted line. We distinguish between the approaches by color. On the left side, we plotted the results for the first interaction and on the right side for the second.

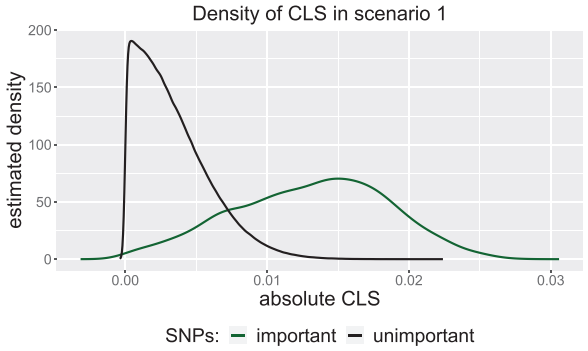

FIGURE A6 | Kernel density estimates of the absolute CLS of the important (green) and unimportant (black) variables in the scenario S_1 and $p = 2000$.

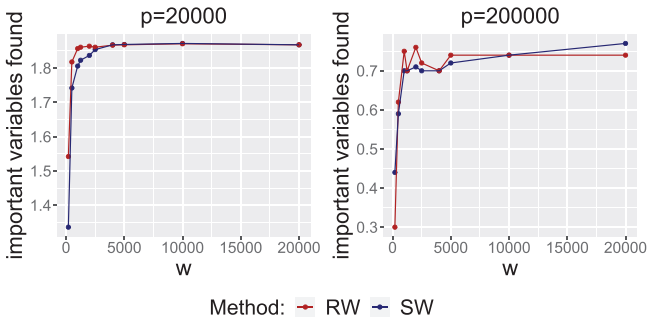

FIGURE A7 | Number of important variables of two found on average if we choose the variables with the $q = n\log(n)$ largest CLS for the single two-way interactions in the second scenario with $p = 20\,000$ and $p = 200\,000$ for different window sizes w. Different approaches are indicated by color.

TABLE A1 | This table shows how many important variables out of two we find with the different methods in median average in scenario 1 (one two-way interaction) if $q = 575$. "CLS" means to calculate the CLS for the whole matrix, and RW and SW stand for the Random Window or Sliding Window Approach. We consider different values for ε $(0.5, 0.2, 0.1)$ in the Sketching approach and the correlation as baseline method. Note that the case $p = 2\,000\,000$ is only feasible for the approximation methods, not for the exact CLS calculations, indicated by X.

Median		**Window**		**Sketching ε =**			
p	**CLS**	**RW**	**SW**	**0.5**	**0.2**	**0.1**	**correlation**
2 000	2	2	2	2	2	2	2
20 000	2	2	2	2	2	2	2
200 000	1	1	1	1	1	1	1
2 000 000	X	1	1	1	1	1	1

TABLE A2 | This table shows how many important variables out of two we find with the different methods in median average in scenario 2 (two 2-way interactions) if $q = 575$. "CLS" means to calculate the CLS for the whole matrix, and RW and SW stand for the Random Window or Sliding Window Approach. We consider different values for ε $(0.5, 0.2, 0.1)$ in the sketching approach and the correlation as baseline method. Note that the case $p = 2,000,000$ is only feasible for the approximation methods, not for the exact CLS calculations, indicated by X.

median		**Window**		**Sketching ε =**			
p	**CLS**	**RW**	**SW**	**0.5**	**0.2**	**0.1**	**correlation**
2 000	4	4	4	4	4	4	4
20 000	3	3	3	3	3	3	3
200 000	1	1	1	1	1	1	1
2 000 000	X	1	1	1	1	1	1

TABLE A3 | Prediction error on HapMap data with and without variable and random selection beforehand.

Method:	**Without**	**Random**	**CLS**	**corr**
Prediction error:	0 %	12.5 %	0 %	0 %

TABLE A4 | Prediction error on scenario S_1 with $p = 500$ with and without variable and random selection beforehand.

method:	**Without**	**Random**	**CLS**	**corr**
Prediction error:	37.5%	47.5%	35.0%	35.0%

Using Cross Leverage Scores for Detecting SNP-Environment Interactions Effects on COPD

USING CROSS LEVERAGE SCORES FOR DETECTING SNP-ENVIRONMENT INTERACTION EFFECTS ON COPD

Sven Teschke
TU Dortmund University
IUF Düsseldorf

Katja Ickstadt
TU Dortmund University
Lamarr-Institute for Machine Learning and Artificial Intelligence

Tamara Schikowski
IUF Düsseldorf
tamara.schikowski@iuf-duesseldorf.de

Claudia Wigmann
IUF Düsseldorf

ABSTRACT

Like other diseases, chronic obstructive pulmonary disease (COPD) has a multifactorial aetiology, meaning that it is caused by both genetic and environmental factors, as well as the interaction between these factors. In this paper, we use the SALIA cohort study to investigate which factors, and which interactions of factors, are particularly important for the risk of COPD. Due to millions of genetic factors, this is a highly dimensional problem. Opposed to genome-wide association (interaction) studies, which model each factor or interaction individually, we aim to use all information simultaneously. We propose a two-step procedure: First we use novel cross leverage scores (CLS) to select a subset of important factors. Second we apply recently proposed logic decision trees (logicDT) to find the explicit interactions. LogicDT is specifically designed to account for interactions but is limited in the number of possible input variables. However, using sketching methods for the calculation of the CLS in the variable selection step allows us to use information from the entire genome at once including all possible combinations. With this sequence of novel methods we gain new insights into the development of COPD and thus contribute to the early initiation of measures to prevent severe disease progression.

Keywords Gene-Gene and Gene-Environment Interaction · COPD · Cross Leverage Scores · logic decision trees

1 Introduction

Chronic obstructive pulmonary disease (COPD) is widespread in the population, has major negative effects on daily living of affected patients and is one of the leading causes of death worldwide [12], [92]. COPD is caused by both genetic and environmental factors [103]. Numerous studies have investigated the causes of COPD individually. However, the interaction of genetic and environmental factors is important in assessing the risk of disease in general [77] and of COPD in particular [2]. The precise consideration of possible interactions is important, as they otherwise obscure significant genetic or environmental effects [95]. For genetic factors, we look specifically at single nucleotide polymorphisms (SNPs), which are variations in individual base pairs in DNA. However, it is very difficult to detect these interaction effects from a statistical point of view, due to the huge amount of SNPs identified throughout the genome. Typically, millions of variables p are present, yet the number of observations n is often (just) in the hundreds, which leads to a $n \ll p$ problem. This issue poses a considerable challenge for Genome Wide Association Studies (GWAS). In GWAS every single variant is tested against the outcome individually. The incorporation of interactions (GWAIS) is also possible, however, in such a scenario, these would also have to be included in the model as input variables. Even for two-way interactions the number of potential combinations is already quadratic $\binom{p}{2} = \Theta(p^2)$. This problem grows exponentially in the order of interactions and theoretically sums up to 2^p. Besides the fact that standard computers are incapable of handling this amount of data, most statistical methods also fail due to the large dimensionality. Additionally, GWAS frequently encounter challenges in identifying minor yet significant effects. This limitation also makes GWAS

inappropriate for searching for interactions. Further, there is still need to optimize approaches for multiple testing correction while screening the whole genome [111].
Popular tree-based statistical learning methods such as Random Forests [15] which can deal with a huge amount of data are applicable to different data structures and mainly characterized by a good prediction performance, rather than by their interpretability. Yet interpretability is crucial in this work in order to detect important associations. Random Forests can also include interactions in the fitting process. However, when the interacting variables have negligible marginal effects, Random Forests have difficulties to detect these interaction effects [151]. There are also various solutions for interaction-focused modifications for Random Forests such as interaction forests [52], but they are again limited in the number of variables.
Another tree based method developed specifically for analyzing SNP data is logic regression [113]. Logic regression tries to identify Boolean combinations of binary predictors that explain the variation in the outcome. This method is widely used for the identification of SNP-SNP interaction effects, cf. [123]. To address that single logic regression models tend to be unstable, if the signal is weak or if many predictors are actually predictive, bagging can be included [122]. However, similar to Random Forests, these models are no longer easily interpretable. Furthermore, logic regression is limited in the number of variables and cannot include interactions between binary predictors and quantitative predictors, yet interactions between binary (genetic) factors and quantitative (environmental) factors are crucial. A methodology that addresses the aforementioned issues while maintaining interpretability is bagged logic regression trees (bagged logicDT) [70]. In the following, we will use the just term logicDT instead of bagged logicDT for better readability. However, since this method is limited in the number of variables it can handle, it requires a preselection of variables.
To address this issue, we have developed a variable selection method that takes interactions into account by providing an interpretable score for each variable [136]. The so-called Cross Leverage Score (CLS) of a variable indicates how important a variable is, also in the sense of a possible interaction effect. This method works in arbitrarily large amounts of data due to a calculation with sketching. Sketching means that we cleverly reduce the dimension of the data matrix in order to make the CLS calculation possible, while obtaining approximately the same result as for the whole data matrix with theoretical guarantee. The distinguishing feature of this approach is its capacity to exploit the entirety of a genomic sequence, including potential interactions, without the necessity of examining every possible combination separately. In this work we transfer this method to the detection of gene-environment interactions. However, the score does not indicate which interactions are explicitly present. Therefore, we propose a two-step procedure, where variable selection with CLS should be followed by a more in-depth analysis with e.g. logicDT.

The aim of this work is twofold. For the first time, we apply this two-step procedure in the context of gene-environment interaction analysis to gain new insights into the relationship between COPD and genetic and environmental factors. Second, we evaluate the recently developed methods on a real data set.
If we can identify gene-environment interactions associated with COPD, preventive measures can be taken at an early stage. Early detection of genetic risk factors and identification of harmful environmental exposures should improve the course of the disease and, in the best case, reduce mortality. We look at the well characterized SALIA cohort study (**S**tudy on the influence of **A**ir pollution on **L**ung function, **I**nflammation and **A**ging) [120]. Starting in the 1980s, this study recruited women from the formerly highly polluted Ruhr area and adjacent rural areas.

In section 2, an overview of the data is provided, along with an explanation of the basic vocabulary. This is followed by a detailed description of the statistical methodology in section 3. In section 4, a descriptive analysis of the SALIA cohort study is presented, followed by an application of the variable selection based on the Cross Leverage Scores. This selection is then evaluated through extensive literature research to assess its association with COPD. The subsequent analysis is performed with logicDT to identify the explicit important factors and their interactions. We conclude with a discussion of the results in section 5 and summarize the main points as well as provide an outlook (section 6).

2 Background & Data

One of the leading external causes for COPD is smoking [68], [80]. Nevertheless, COPD also is also associated to other environmental exposures like air pollution or passive smoking [155]. In particular $PM_{2.5}$ (particular matter(PM)) of diameter less than 2.5μm) exposure increases the risk of COPD, although specific mechanisms still need to be investigated further [142], [25]. In addition, there are many other clinical or personal factors that are related to COPD such as the social status [60], the age [89], years of smoking [134] and the body mass index (BMI) [135]. Nevertheless, the relationship between BMI and COPD is not trivial and necessitates further research [30]. But it also shows that COPD cannot be explained by external factors alone and that a substantial proportion of the risk of developing COPD is caused by genetic variations [21].
Single nucleotide polymorphisms are a well known variation of single base pairs in the human DNA and are related to complex and common diseases [126]. For example, a variation in the *C1orf87* gene is associated with COPD among many other genes [115]. Further, not only individual SNPs, but also interactions between SNPs can be associated with

diseases, e.g. breast cancer [22] or venous thrombosis [45]. In addition, it is also imperative to consider the interaction between single SNPs and environmental factors, as this interaction is of great importance [2].

This paper analyzes data from SALIA cohort study [120], which involves the examination of older women residing in the Ruhr area and adjacent Münsterland. This study is of particular interest, since it pertains on the one hand the Ruhr area as a region that has historically experienced substantial environmental impact from its coal and steel industry. And even in the 21st century, the region continues to experience high levels of pollution, a consequence of its dense population. And on the other hand women from rural areas of Southern Münsterland are included.
In 1985-1994 4874 middle aged women in industrialized and adjacent rural areas in North-Rhine Westphalia, Germany, were recruited and took part in the baseline examination. After a follow-up questionnaire in 2006 [65] a second follow-up including clinical examinations was performed in 2007–2010 in a randomly selected subgroup who had a lung function measurement at baseline [119]. Genetic data, more precisely SNP data of the whole genome, for $n = 510$ women was gathered, cf. [58]. In total various investigations have been carried out on the women, for a detailed description of the study design, the questionnaires and the assessment of environmental factors and other confounders, see [119] or [139].
To investigate the influence of air pollution we consider $PM_{2.5}$. The $PM_{2.5}$ levels were estimated by the method of optimal interpolation at a spatial resolution of 2×2 km and provided by the German Environment Agency (Umwelt Bundesamt, UBA) [39]. The daily mean levels were rescaled to a spatial resolution of 1×1 km and assigned to the participants' home addresses. For the current analysis we calculated the annual mean levels of the year 2006 measured in μg per m^3. Further, we consider data on the living area (urban/rural), the smoking status, packyears and if the participant was exposed to passive smoking at the time of second follow-up. In addition, we include age and BMI as well as data on the social status. COPD was determined using pre-bronchodilator lung function measurements from the second follow-up. The quotient of the lung function parameters forced expiratory volume in 1 second (FEV1) and forced vital capacity (FVC), Tiffeneau index, was used as outcome instead of a binary COPD variable (yes/no). The Tiffeneau index was z-standardized using the Global Lung Initiative reference values [106]. An overview of the available data can be found in table 1.
The study was conducted in accordance with the Declaration of Helsinki, and approved by the Medical Ethics Committee of the University of Bochum (approval number 2732, date of approval: 4 April 2006) and all participants provided written informal consent.

The following fig. 1 outlines the basic procedure in this analysis, the individual work steps are then discussed in more detail in the individual chapters. We start with the full data $\mathcal{D}$ and screen it for missing values, obvious measurement errors etc. We then make adjustments to the data for further processing, which we describe in more detail in section 4. In order to apply methods such as logic regression to the data, a variable selection is carried out using CLS in the next step. We then perform LD pruning (cf. [110]) on the reduced data and prepare the remaining variables for the application of logic decision trees. In the last step we identify the set of important factors and their interactions with variable importance measures obtained from the logicDT model.

3 Statistical Methods

Cross Leverage Scores

This following approach was specifically developed to overcome the challenge of handling the huge amount of data that occurs in Genetics. The fact that interaction effects are of particular interest complicates the challenge substantially. We further developed the idea of using cross leverage scores (CLS) for variable selection [96] and made it applicable to any amount of data where most statistical methods fail due to computational and theoretical reasons or are difficult to interpret [136]. The CLS indicate for each variable their leverage on the outcome including all possible interaction effects. Using sketching approaches the CLS can be approximated with small bounded error guarantees so that they can be applied to arbitrary large data sets. Moreover, also a stream-wise calculation is possible [136]. Since $n \ll p$ problems are typical in the context of genetics, the aim is to reduce the number of variables p by a variable selection based on the CLS. In general the CLS are defined as follows. Let

$$\tilde{X} = [X, y]^T \in \mathbb{R}^{\tilde{p} \times n} \tag{1}$$

with $\tilde{p} = p + 1$ and p the number of variables and n the number of observations. $X \in \mathbb{R}^{n \times p}$ is the data matrix and $y \in \mathbb{R}^n$ denotes the outcome. Then the CLS are given by the off-diagonal entries of the hat matrix H of $\tilde{X}$. The hat matrix H is given by $H = QQ^T$ [51] where Q forms an orthonormal basis for the column space of $\tilde{X}$, which can be obtained by its QR-decomposition $\tilde{X} = QR$ [43]. Since we are only interested in the CLS $c_{i\tilde{p}}$ between the variables $i \in \{1, \dots, p\}$ and the response y, the CLS are given by the dot products of the respective rows:

$$c_{j\tilde{p}} = \langle Q_{i\cdot}, Q_{\tilde{p}\cdot} \rangle, \quad i \in \{1, \dots, p\}. \tag{2}$$

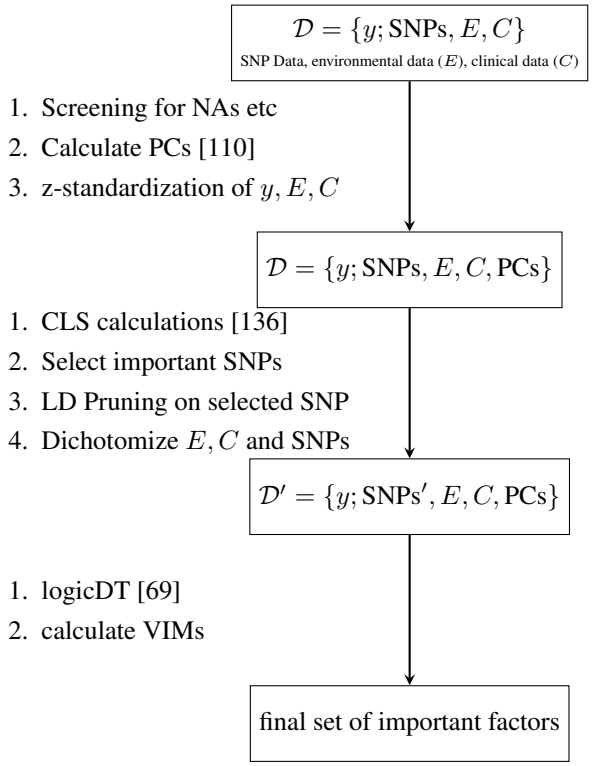

Figure 1: This pipeline describes the individual (intermediate) steps of the two-step procedure consisting of variable selection and subsequent analysis. Beginning with the full data set $\mathcal{D}$ consisting of the Single Nucleotide Polymorphisms (SNPs), Exposure data E and clinical data C, resulting in the final set of important factors and their interactions. Abbreviations: PCs: Prinicipal Components, LD: Linkage Disequilibrium logicDT: logic decision trees, VIMs: variable importance measures

The obvious bottleneck is the QR-decomposition with running time $\Theta(pn^2)$ [43], which is prohibitively slow to compute for p as large as several millions. The sketching algorithm is based on [32] using a further development of Clarkson-Woodruff embedding [23] as sketching matrix. This technique allows us to approximate the CLS of arbitrarily large data within an arbitrary precision parameterized by $\epsilon > 0$. We then select the variables that have the most extreme CLS. It was recommended to select $q = \lceil n \cdot \log n \rceil$ variables in [96] which has been validated in experimental findings [136].

Logic decision trees

Logic decision trees (logicDT) have been developed for the explicit identification of important factors and response-associated interactions and can be extended to an ensemble method with bagging [70]. LogicDT is a consistent tree-based method based on the idea of logic regression [113]. It combines the strengths of other tree based methods such as Random Forests [15] or logic bagging [122] while maintaining interpretability. Although, logicDT is specifically tailored to finding interactions between binary predictors, it is able to take continuous covariables into account by fitting regression models in the decision tree branches. The main idea is to use sets of Boolean conjunctions as input variables for fitting decision trees and searching for the best performing model. Using bagging, the out-of-bag observations (oob-observations) can be used for unbiased and stabilized estimation of the generalization error and variable importance measures (VIMs). By directly including conjunctions as input variables and the integrated variable importance measures, the important factors and their effects can be obtained directly from the model. The algorithm for fitting a logicDT model is given in algorithm 1 in addition to further explanations, cf. section D. The code for this method is available in the R-package `logicDT` [69]. We can therefore use this method to find interactions in the gene-environment analysis. For this, we consider the Tiffeneau Index as outcome and the SNPs and environmental factors as input variables. However, since this method is limited in the number of input variables, we must first perform a variable selection, for which we use the CLS described above.

4 Results

4.1 Descriptive Analyses

After cleaning the data according missing values we have in total data on $p_{gen} = 7643653$ SNPs and $p_{env} = 7$ clinical and environmental variables for 503 women. The data is summarized in table 1. Based on self-report, only 3.78% of the

outcome	Categories (%)	mean, $[\min, \max]$
COPD	yes (3.78%) no (96.02%) unknown (0.20%)	
Tiffeneau Index (FEV_1/FVC)		$-0.223 \quad [-4.02, 2.75]$
variables		
social status	low (17.30%) medium (34.20%) high (48.51%)	
living area	rural (49.10%) urban (50.90%)	
age		$73.52 \quad [66, 79]$
BMI		$27.29 \quad [17.07, 45.67]$
smoking status	no (97.81%) yes (2.19%)	
passive smoking	no (61.63%) yes (38.37%)	
ex-smoking	no (83.50%) yes (16.50%)	
packyears		$3.79 \quad [0, 122.5]$
$PM_{2.5}$, $\mu g/m^3$		$17.33 \quad [14.95, 22.06]$
SNPs	$\{0, 1, 2\}$	

Table 1: Overview of available variables. For the categorical variables the categories and respective frequencies are shown. For the remaining variables the mean, maximum and minimum are shown.
Abbreviations: COPD: chronic obstructive pulmonary disease, FEV_1: Forced expiratory volume in 1 second, FVC: Forced vital capacity, BMI: body mass index, $PM_{2.5}$: particular matter with diameter smaller than $2.5\mu m$, SNP: single nucleotide polymorphism.

women have a clinical diagnosis of COPD. Given the imbalanced distribution of cases and non-cases, the target variable in the subsequent analyses will be the Global Lung Initiative (GLI) z-score of the Tiffeneau-Index (FEV_1/FVC) [106] as an important continuous parameter regarding the diagnosis of COPD. The variable 'status' measuring the social status of the participant according to the maximum number of school years attended by her or her husband (low = less, medium = exact and high = more than 10 years). In addition, the BMI of the women was measured, see fig. 2. The BMI is defined as the weight in kilograms divided by the square of the height in meters. With BMI, an individual can be categorized in underweight, normalweight and overweight or obese according to the World Health Organization (WHO). The median BMI is 26.81 and the majority of the participants were classified as at least overweight according to the WHO classification criteria (65.41%). Similar frequencies have been found for women over 65 in whole Germany, as demonstrated by an RKI study conducted between 2014 and 2015, however the percentage in this group was just 58.9% (95%-CI:$[56.5\% - 61.3\%]$) [118]. Besides the association of large BMI and COPD, it is postulated that a BMI below 18.5 is associated with a more severe course of COPD [88]. In order to reflect the environmental influences in the form of particulate matter, we consider the annual mean level of $PM_{2.5}$ of the year 2006. The distribution of $PM_{2.5}$ is right-tailed, therefore the majority is less exposed than the rest, see fig. 3. However the median exposure is very large and even the minimum exposure ($14.95\mu g/m^3$) is substantially larger than the air quality guideline level of $10\mu g/m^3$ given by the WHO for the year 2005. This level even has been lowered to $5\mu g/m^3$ for 2021 due to new evidence of effects on mortality occurring at concentrations below $10\mu g/m^3$ [150].

4.2 Detecting important variables

First, we'll apply variable selection with CLS to the full dataset, incorporating SNPs and environmental/clinical variables. Migration can lead to genetic diversity within homogeneous populations, known as 'substructure'. We can capture this substructure by calculating principal components of the genotype data and including these as additional variables [110]. In order to ensure that the influence of continuous variables with evidently larger values is not overestimated in comparison to the categorical variables, it is recommended that they are standardized [31]. Thus, we z-standardize the continuous variables, which is common practice in gene-environment analyses, cf. [133].
Given the possibility to access all data at once on a high-performance computer, the sketching approach was used instead of streamwise calculation. The hyperparameter epsilon determines how much the dimension of the original analysis is reduced. The smaller ϵ, the smaller the approximation error, but the less we reduce the matrix and thus the less we reduce the computational time. Here we set $\epsilon = 0.2$. We then select the $q = n \cdot \log n \approx 3129$ variables with the absolute largest CLS for further analysis. Before proceeding with the subsequent analysis, we perform pruning of the selected SNPs with regard to high Linkage Disequilibrium (LD), as correlated variables cause problems with logicDT [70]. LD pruning is a common practice in genome-wide analyses, see [110].

To evaluate the novel method of variable selection we conduct a literature research to determine if SNPs or the respective genes were associated with COPD / Tiffeneau index or related diseases before. Thereby, we focus on the genes according to the 50 SNPs with largest CLS scores, listed in table 4. The objective is not to ascertain the precise correlations between the SNPs and COPD, but rather to ascertain that the SNPs selected are reasonably plausible and not arbitrary. Consequently, there is no claim of completeness in the search for associations between the genes and COPD. The detailed analysis of the associations is shown in the appendix section A. For the majority of the top 50 SNPs, we find associations of the respective genes with COPD.

4.3 Subsequent Analysis

To identify the explicitly important factors and their interactions, we use logicDT and the reduced data set selected by CLS. Although logicDT can theoretically consider interactions between binary and continuous variables for the fitting process, we first have to dichotomize the continuous variables to calculate importance values for their interactions. We set variables to 0 for values smaller than the mean and 1 for larger than the mean. We divided the BMI into three classes: overweight, normalweight and underweight. Then, the ordinal variables and the SNPs have to be converted into binary dummy variables using the R-package `logicFS` [124]. For example for an arbitrary SNP $S_i \in \{0, 1, 2\}$ we obtain

$$\begin{aligned} S = 0 \rightarrow \;& S_1 = 0 \wedge S_2 = 0 \\ S = 1 \rightarrow \;& S_1 = 1 \wedge S_2 = 0 \\ S = 2 \rightarrow \;& S_1 = 1 \wedge S_2 = 1 \end{aligned}$$

In the following, however, we are solely interested in whether a variable has been identified as important or is involved in an important interaction, rather than the specific expression.
Since we are specifically interested in the interactions of SNPs and environmental variables (in a wider sense: also comprising social status, exposure to tobacco smoke etc.), we first consider separately those variable combinations in which at least one environmental variable is involved, this also includes main effects (table 2). It should be noted that the importance of an individual variable also contains the importance of the interactions in which this variable is involved. Therefore, no statement can be made about the order if variable A has a higher importance value than, for example, the combination B $\wedge$ C. If we are interested in the ranking of importance, we therefore have to considered the variables separately for each order of interaction [70].

The three important environmental variables are the smoking status, packyears and $PM_{2.5}$. The smoking status has the largest variable importance score among the environmental variables, but also second largest score of all variables (see table 3), followed by packyears. This is plausible due to numerous studies that have shown a relationship between COPD and smoking, cf. [80]. Also, $PM_{2.5}$ has been reported to be a very important risk factor for COPD, e.g. [142],[25].

For the two-way interactions we obtain the following top combinations according to their variable importance score. In appendix section A, we have already described the connection between the *EDEM1* gene and COPD [42]. In another context, investigating the pathophysiological effects of particulate matter pollution on the development of metabolic disorders such as type II diabetes, up-regulation of *EDEM1* in $PM_{2.5}$ exposed mice and $PM_{2.5}$ induced stress in the endoplasmic reticulum were demonstrated [86]. An interaction of $PM_{2.5}$ and *EDEM1* in other disease pathways is therefore conceivable.
For variations in or next to *TMEM174* we didn't find any direct association to COPD. However, the area next to that gene was identified as top significant restored differentially methylated region associated with wheeze in smokers [128].

factors	respective gene	literature source
smoking status		[80]
packyears		[80]
$PM_{2.5}$		[142], [25]
rs1836577 $\wedge$ $PM_{2.5}$	*EDEM1* $\wedge$ $PM_{2.5}$	[42], [86]
packyears $\wedge$ rs4560527	packyears $\wedge$ *TMEM174*	
social status $\wedge$ rs1335625	social status $\wedge$ *NRG3*	[74], [75]
packyears $\wedge$ rs16850204	packyears $\wedge$ *SYT2*	[13], [66], [85]
smoking $\wedge$ rs3892395	smoking $\wedge$ *CPNE4*	[29]
packyears $\wedge$ rs736020	packyears $\wedge$ *DHRS9*	[7]
packyears $\wedge$ rs10446705	packyears $\wedge$ *PDHA2*	[20]
smoking $\wedge$ rs2211907	smoking $\wedge$ *TMPRSS15*	[94]
rs7478081 $\wedge$ social status $\wedge$ rs1335625	*MPP7* $\wedge$ social status $\wedge$ *NRG3*	[59], [75], [74]
packyears $\wedge$ rs4752656 $\wedge$ rs7074695	packyears $\wedge$ *TACC2* $\wedge$ *PLXDC2*	[81], [130]
(packyears $\wedge$ rs16850204) $\wedge$ rs6909929	(packyears $\wedge$ *SYT2*) $\wedge$ *DTNBP1*	[137], [13], [66]
rs13253753 $\wedge$ status $\wedge$ rs1335625)	*NSMAF* $\wedge$ social status $\wedge$ *NRG3*	[131], [75], [74]
$PM_{2.5}$ $\wedge$ rs1388261 $\wedge$ rs3917932	$PM_{2.5}$ $\wedge$ *LINC00923* $\wedge$ *CSF3R*	[19], [62], [105]
smoking $\wedge$ rs10036629 $\wedge$ rs3892395	smoking $\wedge$ *HEXB* $\wedge$ *CPNE4*	[156] [29]
packyears $\wedge$ rs16850204 $\wedge$ rs56814306	packyears $\wedge$ *SYT2* $\wedge$ *AMZ1*	[115], [13], [66], [85]
smoking $\wedge$ rs34159233 $\wedge$ rs2467458	smoking $\wedge$ *PEBP4* $\wedge$ *ETNK1*	[24]
packyears $\wedge$ rs4853546 $\wedge$ rs7074695	packyears $\wedge$ *STAT4* $\wedge$ *PLXDC2*	[27], [130]
smoking $\wedge$ rs3892395 $\wedge$ rs56038426	smoking $\wedge$ *CPNE4* $\wedge$ *CTLA4*	[55],[29]
(rs1455767 $\wedge$ rs736020) $\wedge$ packyears	(*MCTP2* $\wedge$ *DHRS9*) $\wedge$ packyears	[7]
smoking $\wedge$ rs12895105 $\wedge$ rs4853546	smoking $\wedge$ *EGLN3* $\wedge$ *STAT4*	[38],[27]
smoking $\wedge$ rs7074695 $\wedge$ rs944254	smoking $\wedge$ *PLXDC2* $\wedge$ *CDH4*	[130]

Table 2: The most important gene-environment interactions as well as the most important environmental factors identified by variable importance measures in a logicDT model. Although $\wedge$ is usually used for conjunction of Boolean expressions, in this context we use it to denote any interaction between variables or genes.

Next, *NRG3* is related to the susceptibility towards schizophrenia, but it also may be closely associated with cognitive deficit, especially attention performance in chronic schizophrenia [74]. It could therefore also have an effect on the school education which is used to measure the social status variable in the SALIA study. Further *NRG3* was shown to be associated with Tiffeneau index [75].
The expression of SYT2 protein encoded by *SYT2* is related to Nicotine-derived nitrosamine ketone and tobacco-specific nitrosamines derived from smoking [13]. Further, recent research has shown that SYT2 protein can be used to combat mucus hypersecretion, which is a major cause of airway obstruction in the pathophysiology of COPD [66]. Latter gives hope for a drug which targets SYT2 as an effective therapy for COPD, as the MD Anderson Cancer Center announced 2022 in a press release [85]. It is therefore very plausible that an interaction between a locus that influences *SYT2* expression together with a smoking variable is related to COPD.
In a study on chronic mucus hypersecretion, a variation close to the *CPNE4* gene was identified in heavy smoking males [29], which makes the detected interaction with smoking reasonable.
In the study showing the already described association between COPD and *DHRS9* (see appendix section A), only patients diagnosed with COPD and with more than 10 packyears of smoking were enrolled, cf. [7]. This suggests an actual interaction of packyears and change in the *DHRS9* gene.
Further, *PDHA2* expression was down-regulated by cigarette smoke exposure, however just in a study on male infertility in smoking persons [20].
Lastly, *TMPRSS15* was found as important gene for prediction of inhaled corticosteroids response in patients with asthma [94]. Because inhaled corticosteroids are also used in the treatment of COPD, the interaction with smoking might be interesting.

For the most important three-way interactions, the following findings can be derived from the literature. Some of these overlap with the two-way interactions in that a third variable is added to the interaction. The relationship between *NRG3* and social status in relation to COPD has already been described in the previous section. In addition, the factors appear to be related to the *MPP7* gene, which has already been associated with asthma [59]. Similar has been observed for the interaction of *NRG3* and social status. The third factor, *NSMAF*, is part of the biochemical pathway of the interaction of the endothelial nitric oxide synthase and angiotensin converting enzyme genes and cigarette smoking in COPD [131]. With logicDT we identify an interaction of the variable packyears with SNPs on the genes *TACC2*

and *PLXDC2*. An association between *TACC2* and smoking-induced COPD has already been shown above, cf. [81]. Further, there is an association between *PLXDC2* and pulmonary hypertension [130] which is a common complication of COPD [17]. Packyears and *PLXDC2* are also part of a threefold interaction with *STAT4*. And a *STAT4* activation in smokers and patients with chronic obstructive pulmonary disease has been reported [27]. Another important threefold interaction where *STAT4* and smoking participate is with *EGLN3*. Studies have shown that variation in the *EGLN3* gene increased the risk for COPD. and in a study within Chinese participants, it was also shown that this risk for COPD is significantly increased when adjusting for smoking status [38]. Next, in addition to the already described interaction of smoking (in terms of packyears) and *SYT*2, an interaction of these two factors with *DTNBP1* is also plausible, as *DTNBP1* was identified as a candidate gene underlying the causal link between cigarette smoke exposure and the observed increased risk for interstitial lung abnormalities and idiopathic pulmonary fibrosis [137]. The same holds for *AMZ1* which has been shown to be an associated locus for COPD and lung function [115] and may interact with packyears and *SYT2*. The next identified interaction is $(\mathrm{PM}_{2.5} \wedge \textit{LINC00923} \wedge \textit{CSF3R})$. *CSF3R* on the one hand was found to be significantly up-regulated in transcriptome-wide analyses of the effects of ambient $\mathrm{PM}_{2.5}$ [19] and on the other hand differentially regulated in the blood of COPD patients [62]. To what extent gene *LINC00923* is related to particular matter and *CSF3R* in the context of COPD requires further analysis. At least both genes were listed as two of several downregulated genes in the context of type 1 diabetes [105], so that an interaction also in COPD may be possible. For the interaction $(\text{smoking} \wedge \textit{HEXB} \wedge \textit{CPNE4})$ the according protein to *HEXB* was found to be decreased in female COPD patients compared to female smokers [156] and the possible relation of smoking and *CPNE4* was shown above. The next three-way interaction involves *ETNK1* whose association to COPD (cf. section A) was shown in a cohort comprised of current and former smokers [24]. An interaction with smoking is therefore not surprising. The biological role of *PEBP4*, which is also involved in that interaction, is unclear. For all factors of the combination $(\text{smoking} \wedge \textit{CPNE4} \wedge \textit{CTLA4})$ we showed associations to COPD as well as for the association between *CPNE4* and smoking. Moreover, for *CLTA4* there also exists a significant interaction with exposure to tobacco smoke in the context of activated tumor-resident regulatory T cells in non-small cell lung cancer [55]. We have already shown for *DHRS9* both the direct association with COPD and the association in interaction with packyears with COPD. But the role of *MCTP2* in that interaction is unclear. Last but not least, for the combination of smoking and *PLXDC2* with *CDH4* we already explained the relation of the first two, but did not find a common relation to *CDH4* yet.

Next we consider the important gene-gene interactions according their VIMs from the logicDT model. As aforetime, the VIMs are considered separately according to the number of variables involved in the interaction (table 3).

The SNP with the largest VIM is located on the *DYTN* gene. *DYTN* is associated with Chronic Hypersensitivity Pneumonitis [40]. The smoking status is the second most important variable across all variables. The relation of smoking and COPD has already been described several times in the work, cf. [80]. Next, *PPM1A* encodes the PPM1A protein whose activation seems promising as a therapeutic strategy for pulmonary fibrosis [164]. The next gene *DNAJC15* encodes the MCJ protein whose levels rise in human lungs affected by COPD [117]. For *RGS7*, *TRPS1* as well as *PDHA2* we did not find a direct relation to COPD. However *PDHA2* is related to smoking [20]. In research on the molecular interactions between SARS-CoV-2 and lung cells in patients with COPD, the gene *FAM98A* showed to be differentially expressed in patients with COPD compared with non-smoking control subjects [1]. *CACNA2D1* was reported as candidate biomarker when comparing frequent exacerbation to infrequent exacerbation COPD patients [153] and *TMEM151A* is a biomarker for asthma [18].

The two-way interaction of SNPs from the loci *SORCS2* $\wedge$ *DYTN* have the second largest variable importance across all two fold interactions. While we described the possible association of *DYTN* above, it is reported that *SORCS2* is interacting with MMP-12 [35], which in turn is associated to the development of COPD in smokers [26]. *DYTN* is involved in four of the five most important two-way interactions. This explains why the main effect also has a large VIM. For the interacting features of *DYTN*: *CRACDL*, *LHX2* as well as *TRPS1* we did not find a relation to COPD. The first interaction without that gene is $(\textit{KIF26B} \wedge \textit{TMEM151A})$. The number of mutations in the*KIF26B* gene was increased in COPD patients with myocardial infarction [37]. The interacting gene *TMEM151A* is not just a biomarker for asthma, but was also highly expressed in brain tissues in studies among metabolic syndrome, which represents a collection of cardiovascular risk factors associated with e.g. myocardial infarction [76]. This could be the possible biological relationship between these two genes. A direct link to COPD can be established for both genes involved in the interaction of $(\textit{TRPV4} \wedge \textit{EPAH5})$, cf. [109] and [159]. Analyzing the RNA-binding protein AU-rich-element factor-1, *SETX* was identified as one of the target genes that is differentially expressed in COPD patients compared to smoking controls [116]. But for the interacting *TRAM2* we again could not find an association to COPD in the literature. It was reported that *RAB28* may regulate NRF2 proteolysis. The transcription factor NRF2 is a critical player in the battle against oxidative stress, mostly caused by smoking, which is major driving force of COPD [121]. This gene interacts with the above described *DNAJC15*. SNPs in the *GABRB2* have so far mainly been associated with schizophrenia [158] as well as neurological disorders like epilepsy [84]. Also *GNA14* is associated with epilepsy [5]. It

factors	respective gene	literature source
rs16838595	*DYTN*	[40]
raucher	–	[80]
rs3014557	*RGS7*†	
rs7154604	*PPM1A*†	[164]
rs35272958	*DNAJC15*†	[117]
rs1180626	*TRPS1*	
rs10446705	*PDHA2*†	
rs112455680	*FAM98A*†	[1]
rs4335074	*CACNA2D1*†	[153]
rs4930351	*TMEM151A*†	[18]
rs2048949 ∧ rs16838595	*CRACDL* ∧ *DYTN*	[40]
rs4689067 ∧ rs16838595	*SORCS2* ∧ *DYTN*	[35], [26], [40]
rs16838595 ∧ rs7049157	*DYTN* ∧ *LHX2*	[40]
rs1093959 ∧ rs4930351	*KIF26B* ∧ *TMEM151A*	[37], [18], [76]
rs16838595 ∧ rs1180626	*DYTN* ∧ *TRPS1*	[40]
rs2338568 ∧ rs13109328	*TRPV4* ∧ *EPHA5*	[109] [159]
rs3014557 ∧ rs35272958	*RGS7* ∧ *DNAJC15*	[117]
rs4962060 ∧ rs761441	*SETX* ∧ *TRAM2*	[116]
rs2673423 ∧ rs35272958	*RAB28* ∧ *DNAJC15*	[121], [117]
rs7019154 ∧ rs1899928	*GNA14* ∧ *GABRB2*	[5],[158]
rs16838595 ∧ rs2131464 ∧ rs773832	*DYTN* ∧ *RP11-180C1.1* ∧ *FAM98A*	[40], [1]
(rs2479047 ∧ rs4762523) ∧ rs16838595	(*TCF7L2* ∧ *ANKS1B*) ∧ *DYTN*	[46], [157], [34]
rs247527 ∧ rs7019154 ∧ rs1899928	*STARD4* ∧ *GNA14* ∧ *GABRB2*	[104], [5], [158]
(rs6470471 ∧ rs9401874) ∧ rs7154604	(*FAM84B* ∧ *CENPW*) ∧ *PPM1A*	[154], [129], [164]
rs12480562 ∧ rs3014557 ∧ rs35272958	*PTPRT* ∧ *RGS7* ∧ *DNAJC15*)	[28], [112], [117]
(rs12117683 ∧ rs12132738) ∧ rs16838595	(*H3F3A* ∧ *GBP5*) ∧ *DYTN*	[9], [127], [40]
rs35272958 ∧ rs2673423 ∧ rs952503	*DNAJC15* ∧ *RAB28* ∧ *PCSK5*	[117], [121], [61]
rs1180626 ∧ rs2496322 ∧ rs8051080	*TRPS1* ∧ *EFHD2* ∧ *RBFOX1*	[41], [67]
(rs2515793 ∧ rs7478081) ∧ rs7153483	(*TMPRSS13* ∧ *MPP7*) ∧ *SPTLC2*	[11], [59], [63]
(rs17381675 ∧ rs7049157) ∧ rs16838595	(*MN1* ∧ *LHX2*) ∧ *DYTN*	[4], [40]

Table 3: The most important gene-gene interactions as well as the most important main effects identified by variable importance measures in a logicDT model. Although ∧ is usually used for conjunction of Boolean expressions, in this context we use it to denote any interaction between variables or genes.

is unclear to what extent the two genes, or variations in these genes, interact biologically. It is also unclear how they affect COPD.

For the first three-way interaction (*DYTN* ∧ *RP11-180C1.1* ∧ *FAM98A*) we have already shown the association of *DYTN* and *FAM98A* with COPD, but we could not find anything comparable for *RP11-180C1.1* in the literature. The as transcription factor acting protein TCF4 encoded by *TCF7L2* is associated with COPD and showed to be correlated with packyears in a study of smokers with or without COPD [46]. And *ANKS1B* is related to asthma [157] and smoking-related clear cell renal cell carcinoma [34]. So the respective association with the smoking could be the reason for the interaction. The tree-fold interaction is completed by *DYTN*, which we explained in relation to COPD above. In addition to the two-way interaction of (*GNA14* ∧ *GABRB2*) described above, there also seems to be a three-way interaction with *STARD4*. Like the other two, *STARD4* is also being linked to epilepsy [104]. *FAM84B* is related to asthma [154] and *CENPW* implicated by protein quantitative trait loci signal to be associated with forced vital capacity (FVC) [129]. The interaction of these two genes interacts with *PPM1A*, for which we showed above a link to pulmonary fibrosis. The next three fold interaction is (*PTPRT* ∧ *RGS7* ∧ *DNAJC15*). *PTPRT* is related to COPD [28], [112] as well as that is the case for *DNAJC15*. As for the twofold interaction with *DNAJC15* above we did not find any relation to COPD for *RGS7*. An increase in *H3F3A* mRNA patients with advanced COPD compared to a control group composed of ex-smokers who had normal lung function is reported [9]. In appendix section A we show the potential contribution of *GBP5* to COPD [127]. So for all genes of (*H3F3A* ∧ *GBP5*) ∧ *DYTN* there is somehow an association to COPD. The interaction (*DNAJC15* ∧ *RAB28*) is described above, however they also interact together with *PCSK5*. For *PCSK5* a significant correlation with FEV_1 changes after long-acting β-2 agonist treatment was found in patients with stable COPD [61]. The next three-way interaction is (*TRPS1* ∧ *EFHD2* ∧ *RBFOX1*). *EFHD2* encodes the EFHD2 protein which was found as one of 90 markers for the study of the molecular pathogenesis of COPD combined with type II

respiratory failure [41], for *RBFOX1* see appendix section A. For *TRPS1* we could not find a published association. For $((TMPRSS13 \wedge MPP7) \wedge SPTLC2)$ we found that *TMPRSS13* is associated with SARS-CoV-2 [11], *MPP7* associated with asthma and *SPTLC2* with chronic inflammatory lung diseases like COPD [63]. *DYTN* is also included in the last interaction, again in conjunction with *LHX2*. The third gene in this combination, *MN1*, is associated with COPD [4].

5 Discussion

The aim of this study was to apply a newly designed analysis pipeline for variable selection and subsequent modelling of genetic and environmental factors and their interactions in the context of COPD. The genetic data consisted of very high dimensional genome-wide SNP data.
In the initial step, we performed a variable selection using cross leverage scores. Employing a sketching method facilitated the incorporation of all available variables into the selection step simultaneously. We selected the variables with the largest absolute CLS. Then, in an intermediate step, we applied LD pruning to these variables to avoid problems with multicollinearity for the subsequent analysis. This process yielded a total of 514 variables. For the top 50 of these, an extensive literature search was conducted to investigate whether the genes in which the SNPs are located have been associated with COPD in the past. For the majority of these genes, an association with COPD, COPD phenotypes, or at least similar lung diseases was identified. This finding serves to substantiate the efficacy of the recently developed method of variable selection with CLS [136], affirming its ability to fulfill its intended function.
For the subsequent analysis we applied the logicDT model, where we included all environmental and clinical data regardless of whether they were selected in the variable selection or not. For the fitting process logicDT is able to include interactions between binary and continuous features. But since we are interested in calculating VIMs, we had to dichotomize the clinical and environmental variables. Although this means that information is being lost, we believe the dichotomized variables are sufficient as this is screening approach to identify possible biomarkers and interactions between biomarkers.
For both the most important gene-environment and the most important gene-gene interactions, as well as main effects, we then again conducted a literature search to explain possible relationships. For most interactions, studies and results can be found from which relationships can be derived. In some cases these are clear from direct associations with COPD. In other cases the factors occur together with regard to different diseases. However, it is important to note that these are statistical associations and even though the literature search seems to provide a proof of concept for the applied analysis pipeline, biological and mechanistic insights into the relationships with COPD are still lacking for the possible biomarkers. In order to obtain a biological proof, experiments and pathway analyses have to be carried out. Our results provide a basis for focusing on specific genes and environmental variables. Many of the literature sources are from this year or last year. Therefore, it is quite possible that the candidate interactions include genes or SNPs that have been little studied and may provide new insights.
It is also pleasing to note that with the consistent logicDT method, the most important variables also include those that were among the top variables in terms of their CLS. This confirms that our newly developed selection method also works in practice.
In the variable selection step we use sketching for the calculation of CLS. As mentioned above, the approximation to the CLS can be improved by adjusting the hyperparameters or the calculations can be simplified. In addition, there are other methods to approximate the CLS or compute it streamwise. We have developed two window-wise approaches. The sliding window approach and the random window approach. For more details, see [136].
The main focus of this work was the identification of candidate biomarkers and the evaluation of the methodology based on literature review. However, we can also perform prediction calculations on the reduced data. Again, logicDT is well suited for this, but other methods such as Block forests [53], Interaction forests [52] or any other method are possible. Furthermore, we have seen that the calculation of risk scores based on the variables selected by the CLS is also a useful tool for estimating the risk of diseases.

6 Conclusion & Outlook

We have transferred the novel variable selection technique based on Cross Leverage Scores and sketching to a genome-wide dataset from the SALIA cohort study. Using the CLS allows us to use the information from the entire genome at once. Further, the CLS also exploit possible interaction effects of variables that have only small main effects. An extensive literature search shows that this selection is indeed a reasonable one, as published associations with COPD can be found for most genes. We obtain a reduced dataset on which novel as well as state of the art statistical methods can be subsequently applied. In the subsequent analysis, we apply the logicDT model to the reduced data and are able to identify the truly important main and interaction effects by using variable importance measures. Again, a literature review shows that the effects found are plausible. However, the obtained list of statistical important genetic and environmental candidate factors should be validated by biological experiments e.g. pathway analysis in a next step.

Our results can therefore be taken as an indication to take a closer look at these genes in further biological experiments and risk analysis. All in all we introduced a new two-step procedure, including recently developed methods, which is applicable to arbitrary large data as well as data from other fields. Concurrently, this method does not pose a significant challenge to conventional computers.

Funding

Sven Teschke, Katja Ickstadt, Tamara Schikowski and Claudia Wigmann were supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, project R1) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Project Number 427806116). Katja Ickstadt acknowledges the support of BMBF and MKW.NRW within the Lamarr-Institute for Machine Learning and Artificial Intelligence. The SALIA cohort study was supported by the Ministry of the Environment of the state North Rhine-Westphalia (Düsseldorf, Germany), the Federal Ministry of the Environment (Berlin, Germany), the German Federal Ministry of Education and Research (BMBF) as well as by grants HE-4510/2-1, KR 1938/3-1, LU 691/4-1 and SCHI 1358/3-1 from the Deutsche Forschungsgemeinschaft (DFG), 617.0-FP266 from the German Statutory Accident Insurance (DGUV), and grant agreement number 211250 from the European Community's Seventh Framework Program (FP7/2007-2011). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of manuscript.

Acknowledgments

We thank all study members and staff involved in data collection and also the respective funding bodies for SALIA.
Study directorate: R. Dolgner; U. Krämer, U. Ranft, T. Schikowski, A. Vierkötter
Scientific Team Baseline: A.W. Schlipköter, M.S. Islam; A. Brockhaus, H. Idel, R. Stiller-Winkler, W. Hadnagy, T. Eikmann
Scientific Team Follow-up: D. Sugiri, A. Hüls, B. Pesch, A. Hartwig, H. Käfferlein, V. Harth, T.Brüning, T. Weiss
Study Nurses: G. Seitner-Sorge, V. Jäger, G. Petczelies, I. Podolski, T. Hering, M. Goseberg
Administrative Team: B. Schulten, S. Stolz
During the last decades many scientists, study nurses and laboratories were involved in conducting the study. We are most grateful for all the women from the Ruhr area and Borken who participated in the study over decades and the local health Departments for organizing the study. We thank Stefan Feigenspan from the German Federal Environmental Agency for providing data on air pollution, and information about the used model, including the explanation of some results in response to our questions. We also wish to thank the Deutsche Wetterdienst for providing data from the COSMO-REA6 model.

Conflict of Interest

The authors have declared no conflict of interest.

Data Availability Statement

Due to privacy laws in Germany the data of the cohort study is not available. The R code can be obtained by contacting the authors of the paper.

References

[1] A. Agusti, O. Sibila, S. Casas-Recasens, N. Mendoza, L. Perea, A. Lopez-Giraldo, and R. Faner. "Molecular interactions of SARS-CoV-2 in lung tissue of patients with chronic obstructive pulmonary disease". In: Annals of the American Thoracic Society 18.11 (2021), pp. 1922–1924. DOI: 10.1513/AnnalsATS.202006-619RL.

[2] Á. Agustí, E. Melén, D. L. DeMeo, R. Breyer-Kohansal, and R. Faner. "Pathogenesis of chronic obstructive pulmonary disease: Understanding the contributions of gene-environment interactions across the lifespan". In: The Lancet Respiratory Medicine 10.5 (2022), pp. 512–524. DOI: 10.1016/S2213-2600(21)00555-5. URL: https://doi.org/10.1016/S2213-2600(21)00555-5.

[3] M. S. Artigas, D. W. Loth, L. V. Wain, and et al. "Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function". In: Nature Genetics 43 (2011), pp. 1082–1090. DOI: 10.1038/ng.941.

[4] M. S. Artigas et al. "Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation". In: Nature communications 6.1 (2015), p. 8658. DOI: 10.1038/ncomms9658.

[5] A. Al-Asmi et al. "Familial temporal lobe epilepsy as a presenting feature of choreoacanthocytosis". In: Epilepsia 46.8 (2005), pp. 1256–1263. DOI: 10.1111/j.1528-1167.2005.65804.x.

[6] S. M. Bache and H. Wickham. magrittr: A Forward-Pipe Operator for R. R package version 2.0.3. 2022. URL: https://CRAN.R-project.org/package=magrittr.

[7] S. Bai et al. "Identification of Proteomic Signatures in Chronic Obstructive Pulmonary Disease Emphysematous Phenotype". In: Frontiers in Molecular Biosciences 8 (2021), p. 650604. DOI: 10.3389/fmolb.2021.650604.

[8] J. N. Baraniuk, B. Casado, L. K. Pannell, P. B. McGarvey, P. Boschetto, M. Luisetti, and P. Iadarola. "Protein networks in induced sputum from smokers and COPD patients". In: International Journal of Chronic Obstructive Pulmonary Disease 10 (2015), pp. 1957–1975. DOI: 10.2147/COPD.S75978.

[9] C. A. Barrero et al. "Histone 3.3 participates in a self-sustaining cascade of apoptosis that contributes to the progression of chronic obstructive pulmonary disease". In: American journal of respiratory and critical care medicine 188.6 (2013), pp. 673–683. DOI: 10.1164/rccm.201302-0342OC.

[10] K. Bazemore, J. Joo, W. T. Hwang, and B. E. Himes. "Clarifying Chronic Obstructive Pulmonary Disease Genetic Associations Observed in Biobanks via Mediation Analysis of Smoking". In: AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science (2024), pp. 499–508.

[11] M. Benlarbi et al. "Identification and differential usage of a host metalloproteinase entry pathway by SARS-CoV-2 Delta and Omicron". In: IScience 25.11 (2022). DOI: 10.1016/j.isci.2022.105316.

[12] C. E. Berry and R. A. Wise. "Mortality in COPD: Causes, Risk Factors, and Prevention". In: COPD 7.5 (2010), pp. 375–382. DOI: 10.3109/15412555.2010.510160. URL: https://doi.org/10.3109/15412555.2010.510160.

[13] H.-J. Boo et al. "The tobacco-specific carcinogen-operated calcium channel promotes lung tumorigenesis via IGF2 exocytosis in lung epithelial cells". In: Nature communications 7.1 (2016), p. 12961. DOI: 10.1038/ncomms12961.

[14] L. Breiman. "Bagging predictors". In: Machine Learning 24 (1996), pp. 123–140. DOI: 10.1007/BF00058655.

[15] L. Breiman. "Random forests". In: Machine learning 45 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[16] J. Callegari, F. S. Magnet, S. Taubner, M. Berger, S. B. Schwarz, W. Windisch, and J. H. Storre. "Interfaces and Ventilator Settings for Long-Term Noninvasive Ventilation in COPD Patients". In: International Journal of Chronic Obstructive Pulmonary Disease 12 (2017). PMID: 28721033, pp. 1883–1889. DOI: 10.2147/COPD.S132170.

[17] A. Chaouat, R. Naeije, and E. Weitzenblum. "Pulmonary hypertension in COPD". In: The European Respiratory Journal 32.5 (2008), pp. 1371–1385. DOI: 10.1183/09031936.00015608.

[18] S.-T. Chen and N. Yang. "Constructing ferroptosis-related competing endogenous RNA networks and exploring potential biomarkers correlated with immune infiltration cells in asthma using combinative bioinformatics strategy". In: BMC genomics 24.1 (2023), p. 294. DOI: 10.1186/s12864-023-09400-7.

[19] W. Chen, Y. Han, Y. Yao, J. Liu, Y. Wu, K. Frank, and T. Zhu. "Transcriptome-wide analyses of the effects of ambient PM2.5 and carbonaceous constituents: results of the AIRLESS project". In: ISEE Conference Abstracts 2020.1 (2020). DOI: 10.1289/isee.2020.virtual.P-0011.

[20] X. Chen et al. "Alteration of sperm protein profile induced by cigarette smoking". In: Acta Biochimica et Biophysica Sinica 47.7 (June 2015), pp. 504–515. DOI: 10.1093/abbs/gmv045. eprint: https://academic.oup.com/abbs/article-pdf/47/7/504/7690615/gmv045.pdf.

[21] M. H. Cho, B. D. Hobbs, and E. K. Silverman. "Genetics of chronic obstructive pulmonary disease: Understanding the pathobiology and heterogeneity of a complex disorder". In: The Lancet Respiratory Medicine 10.5 (2022), pp. 485–496. DOI: 10.1016/S2213-2600(21)00510-5.

[22] L.-Y. Chuang, Y.-D. Lin, H.-W. Chang, and C.-H. Yang. "SNP-SNP Interaction Using Gauss Chaotic Map Particle Swarm Optimization to Detect Susceptibility to Breast Cancer". In: 2014 47th Hawaii International Conference on System Sciences. 2014, pp. 2548–2554. DOI: 10.1109/HICSS.2014.647.

[23] K. L. Clarkson and D. P. Woodruff. "Low-Rank Approximation and Regression in Input Sparsity Time". In: Journal of the ACM 63.6 (2017), pp. 1–45. ISSN: 0004-5411. DOI: 10.1145/3019134.

[24] C. I. Cruickshank-Quinn, S. Jacobson, G. Hughes, R. L. Powell, I. Petrache, K. Kechris, R. Bowler, and N. Reisdorph. "Metabolomics and Transcriptomics Pathway Approach Reveals Outcome-Specific Perturbations in COPD". In: Scientific Reports 8.1 (Nov. 20, 2018), p. 17132. DOI: 10.1038/s41598-018-35372-w.

[25] M. A. Delavar, M. A. Jahani, M. Sepidarkish, S. Alidoost, H. Mehdinezhad, and Z. Farhadi. "Relationship between fine particulate matter (PM2.5) concentration and risk of hospitalization due to chronic obstructive pulmonary disease: A systematic review and meta-analysis". In: BMC Public Health 23.1 (2023), p. 2229. DOI: 10.1186/s12889-023-17093-6.

[26] I. K. Demedts, A. Morel-Montero, S. Lebecque, Y. Pacheco, D. Cataldo, G. Joos, R. Pauwels, and G. Brusselle. "Elevated MMP-12 protein levels in induced sputum from patients with COPD". In: Thorax 61.3 (2006), pp. 196–201. DOI: 10.1136/thx.2005.042432.

[27] A. Di Stefano et al. "STAT4 activation in smokers and patients with chronic obstructive pulmonary disease". In: European Respiratory Journal 24.1 (2004), pp. 78–85. DOI: doi.org/10.1183/09031936.04.00080303.

[28] I. Diamant, D. Clarke, J. E. Evangelista, N. Lingam, and A. Ma'ayan. "Harmonizome 3.0: integrated knowledge about genes and proteins from diverse multi-omics resources". In: Nucleic Acids Research 53.D1 (Nov. 2024), pp. D1016–D1028. ISSN: 1362-4962. DOI: 10.1093/nar/gkae1080.

[29] A. E. Dijkstra et al. "Dissecting the genetics of chronic mucus hypersecretion in smokers with and without COPD". In: European Respiratory Journal 45.1 (2014), pp. 60–75. DOI: doi.org/10.1183/09031936.00093314.

[30] M. J. Divo, C. Cabrera, C. Casanova, et al. "Comorbidity distribution, clinical expression and survival in COPD patients with different body mass index". In: Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation 1.2 (2014), pp. 229–238. DOI: 10.15326/jcopdf.1.2.2014.0117.

[31] P. Doebler, A. Doebler, P. Buczak, and A. Groll. "Interactions of scores derived from two groups of variables: Alternating lasso regularization avoids overfitting and finds interpretable scores". In: Psychological Methods 28.2 (2023), pp. 422–437. DOI: 10.1037/met0000461.

[32] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. "Fast approximation of matrix coherence and statistical leverage". In: J. Mach. Learn. Res. 13 (2012), pp. 3475–3506. DOI: 10.5555/2503308.2503352. URL: https://dl.acm.org/doi/10.5555/2503308.2503352.

[33] W. Du, J. A. Stiber, P. B. Rosenberg, G. Meissner, and J. P. Eu. "Ryanodine Receptors in Muscarinic Receptor-mediated Bronchoconstriction". In: Journal of Biological Chemistry 280.28 (2005), pp. 26287–26294. ISSN: 0021-9258. DOI: 10.1074/jbc.M502905200. URL: https://www.sciencedirect.com/science/article/pii/S002192582056821X.

[34] J. E. Eckel-Passow, D. J. Serie, B. M. Bot, R. W. Joseph, J. C. Cheville, and A. S. Parker. "ANKS1B is a smoking-related molecular alteration in clear cell renal cell carcinoma". In: BMC Urology 14 (2014), p. 14. DOI: 10.1186/1471-2490-14-14.

[35] J. Eriksson Ström, S. Kebede Merid, R. Linder, J. Pourazar, A. Lindberg, E. Melén, and A. F. Behndig. "Airway MMP-12 and DNA methylation in COPD: an integrative approach". In: Respiratory Research 26.1 (2025), pp. 1–12. DOI: 10.1186/s12931-024-03088-3.

[36] M. E. Ezzie et al. "Gene expression networks in COPD: microRNA and mRNA regulation". In: Thorax 67.2 (2012), pp. 122–131. ISSN: 0040-6376. DOI: 10.1136/thoraxjnl-2011-200089. URL: https://thorax.bmj.com/content/67/2/122.

[37] A. Faiz et al. "COPD Patients Display Increased Peripheral Blood Somatic Mutations Which Associate With the Prevalence of Co-morbidities". In: Archivos de bronconeumologia 60.2 (2024), pp. 119–121. DOI: 10.1016/j.arbres.2023.12.011.

[38] S. Fang, J. Qiu, H. Yu, H. Fan, X. Wu, Z. Fang, Q. Shen, and S. Chen. "Association of EGLN1 and EGLN3 single-nucleotide polymorphisms with chronic obstructive pulmonary disease risk in a Chinese population". In: Int J Clin Exp Med 10.7 (2017), pp. 10866–10873.

[39] J. Flemming and R. Stern. "Datenassimilation auf der Basis der Optimalen Interpolation für die Kartierung von Immissionsbelastungen: Beschreibung der Methodik und praktische Anwendung für 2002". In: Freie Univ., Inst. für Meteorologie, Tropospärische Umweltforschung (2004).

[40] H. Furusawa et al. “Chronic hypersensitivity pneumonitis, an interstitial lung disease with distinct molecular signatures”. In: American journal of respiratory and critical care medicine 202.10 (2020), pp. 1430–1444. DOI: 10.1164/rccm.202001-0134OC.

[41] Q. Gan et al. “Differential expression study of lysine crotonylation and proteome for chronic obstructive pulmonary disease combined with type II respiratory failure”. In: Canadian Respiratory Journal 2021.1 (2021), p. 6652297. DOI: 10.1155/2021/6652297.

[42] P. Geraghty, N. Baumlin, M. A. Salathe, R. F. Foronjy, and J. M. D’Armiento. “Glutathione Peroxidase-1 Suppresses the Unfolded Protein Response upon Cigarette Smoke Exposure”. In: Mediators of Inflammation 2016 (2016), p. 9461289. DOI: 10.1155/2016/9461289.

[43] G. Golub and C. Van Loan. Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013. ISBN: 9781421407944. URL: https://books.google.de/books?id=X5YfsuCWpxMC.

[44] W. Gou, Z. Zhang, C. Yang, and Y. Li. “MiR-223/Pknox1 axis protects mice from CVB3-induced viral myocarditis by modulating macrophage polarization”. In: Experimental Cell Research 366.1 (2018), pp. 41–48. ISSN: 0014-4827. DOI: 10.1016/j.yexcr.2018.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0014482718301265.

[45] N. Greliche et al. “A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis”. In: BMC Medical Genetics 14.36 (2013). DOI: https://doi.org/10.1186/1471-2350-14-36.

[46] L. Guo et al. “WNT/β-catenin signaling regulates cigarette smoke-induced airway inflammation via the PPARδ/p38 pathway”. In: Laboratory Investigation 96.2 (2016), pp. 218–229. ISSN: 0023-6837. DOI: https://doi.org/10.1038/labinvest.2015.101. URL: https://www.sciencedirect.com/science/article/pii/S0023683722023108.

[47] M. Hardin et al. “A genome-wide analysis of the response to inhaled β_2-agonists in chronic obstructive pulmonary disease”. In: The Pharmacogenomics Journal 16.4 (2016), pp. 326–335. DOI: 10.1038/tpj.2015.65.

[48] U. Hedström et al. “Impaired Differentiation of Chronic Obstructive Pulmonary Disease Bronchial Epithelial Cells Grown on Bronchial Scaffolds”. In: American Journal of Respiratory Cell and Molecular Biology 65.2 (2021), pp. 201–213. DOI: 10.1165/rcmb.2019-0395OC.

[49] R. J. Hijmans. terra: Spatial Data Analysis. R package version 1.8-29. 2025. URL: https://CRAN.R-project.org/package=terra.

[50] C. Hilzendeger et al. “Reduced sputum expression of interferon-stimulated genes in severe COPD”. In: International Journal of Chronic Obstructive Pulmonary Disease 11 (2016), pp. 1485–1494. DOI: 10.2147/COPD.S105948.

[51] D. C. Hoaglin and R. E. W. and. “The hat matrix in regression and ANOVA”. In: The American Statistician 32.1 (1978), pp. 17–22. DOI: 10.1080/00031305.1978.10479237.

[52] R. Hornung and A.-L. Boulesteix. “Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects”. In: Computational Statistics & Data Analysis 171 (2022), p. 107460. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2022.107460.

[53] R. Hornung and M. N. Wright. “Block forests: random forests for blocks of clinical and omics covariate data”. In: BMC bioinformatics 20 (2019), pp. 1–17. DOI: 10.1186/s12859-019-2942-y.

[54] X. Hu, Z. Wei, Y. Wu, M. Zhao, L. Zhou, and Q. Lin. “Pathogenesis and Therapy of Hermansky-Pudlak Syndrome (HPS)-Associated Pulmonary Fibrosis”. In: International Journal of Molecular Sciences 25.20 (2024), p. 11270. DOI: 10.3390/ijms252011270.

[55] Y. Hu et al. “Exposure to tobacco smoking induces a subset of activated tumor-resident tregs in non-small cell lung cancer”. In: Translational Oncology 15.1 (2022), p. 101261. DOI: 10.1016/j.tranon.2021.101261.

[56] H. Huang, H. Feng, and D. Zhuge. “M1 Macrophage Activated by Notch Signal Pathway Contributed to Ventilator-Induced Lung Injury in Chronic Obstructive Pulmonary Disease Model”. In: Journal of Surgical Research 244 (2019), pp. 358–367. ISSN: 0022-4804. DOI: 10.1016/j.jss.2019.06.060. URL: https://www.sciencedirect.com/science/article/pii/S0022480419304561.

[57] HugeAmp. HugeAmp: A Comprehensive Resource for Molecular Data. March, 20, 2025. 2023. URL: https://md.hugeamp.org/.

[58] A. Hüls, U. Krämer, C. Herder, K. Fehsel, C. Luckhaus, S. Stolz, A. Vierkötter, and T. Schikowski. “Genetic susceptibility for air pollution-induced airway inflammation in the SALIA study”. In: Environmental Research 152 (2017), pp. 43–50. ISSN: 0013-9351. DOI: 10.1016/j.envres.2016.09.028. URL: https://www.sciencedirect.com/science/article/pii/S0013935116307174.

[59] M. Imboden et al. “Genome-wide association study of lung function decline in adults with and without asthma”. In: Journal of allergy and clinical immunology 129.5 (2012), pp. 1218–1228. DOI: doi.org/10.1016/j.jaci.2012.01.074.

[60] M. Kanervisto, T. Vasankari, T. Laitinen, M. Heliövaara, P. Jousilahti, and S. Saarelainen. “Low socioeconomic status is associated with chronic obstructive airway diseases”. In: Respiratory Medicine 105.8 (2011), pp. 1140–1146. ISSN: 0954-6111. DOI: 10.1016/j.rmed.2011.03.008. URL: https://www.sciencedirect.com/science/article/pii/S0954611111000965.

[61] J. Kang et al. “Predicting treatable traits for long-acting bronchodilators in patients with stable COPD”. In: International journal of chronic obstructive pulmonary disease (2017), pp. 3557–3565. DOI: 10.2147/COPD.S151909.

[62] T. S. Kapellos et al. “Systemic alterations in neutrophils and their precursors in early-stage chronic obstructive pulmonary disease”. In: Cell reports 42.6 (2023). DOI: 10.1016/j.celrep.2023.112525.

[63] S. Karandashova, A. B. Kummarapurugu, S. Zheng, C. E. Chalfant, and J. A. Voynow. “Neutrophil elastase increases airway ceramide levels via upregulation of serine palmitoyltransferase”. In: American Journal of Physiology-Lung Cellular and Molecular Physiology 314.1 (2018), pp. L206–L214. DOI: 10.1152/ajplung.00322.2017.

[64] A. R. Koczulla et al. “Krüppel-like Zinc Finger Proteins in End-Stage COPD Lungs with and without Severe Alpha1-Antitrypsin Deficiency”. In: Orphanet Journal of Rare Diseases 7 (2012), p. 29. DOI: 10.1186/1750-1172-7-29.

[65] U. Krämer, C. Herder, D. Sugiri, K. Strassburger, T. Schikowski, U. Ranft, and W. Rathmann. “Traffic-related air pollution and incident type 2 diabetes: results from the SALIA cohort study.” In: Environmental health perspectives 118(9) (2010), pp. 1273–1279. DOI: 10.1289/ehp.0901689.

[66] Y. Lai et al. “Screening of hydrocarbon-stapled peptides for inhibition of calcium-triggered exocytosis”. In: Frontiers in pharmacology 13 (2022), p. 891041. DOI: 10.3389/fphar.2022.891041.

[67] P. Lakshman Kumar et al. “Genetic variation in genes regulating skeletal muscle regeneration and tissue remodelling associated with weight loss in chronic obstructive pulmonary disease”. In: Journal of Cachexia, Sarcopenia and Muscle 12.6 (2021). DOI: doi.org/10.1002/jcsm.12782.

[68] R. Laniado-Laborín. “Smoking and chronic obstructive pulmonary disease (COPD): Parallel epidemics of the 21st century”. In: International Journal of Environmental Research and Public Health 6.1 (2009), pp. 209–224. DOI: 10.3390/ijerph6010209.

[69] M. Lau. logicDT: Identifying interactions between binary predictors. 3. 2023. URL: https://CRAN.R-project.org/package=logicDT.

[70] M. Lau, T. Schikowski, and H. Schwender. “logicDT: A procedure for identifying response-associated interactions between binary predictors”. In: Machine Learning 113.2 (2024), pp. 933–992. DOI: 10.1007/s10994-023-06488-6.

[71] J. Li, A. Dai, R. Hu, L. Zhu, and S. Tan. “Positive correlation between PPARgamma/PGC-1alpha and gamma-GCS in lungs of rats and patients with chronic obstructive pulmonary disease”. In: Acta Biochimica et Biophysica Sinica 42.9 (2010), pp. 603–614. DOI: 10.1093/abbs/gmq071.

[72] R. Li et al. “Differential Expression of Serum Proteins in Chronic Obstructive Pulmonary Disease Assessed Using Label-Free Proteomics and Bioinformatics Analyses”. In: International Journal of Chronic Obstructive Pulmonary Disease 17 (2022), pp. 2871–2891. DOI: 10.2147/COPD.S383976.

[73] X. Li, J. Liu, Z. Jing, and S. Li. “SLC27A3 downregulation restores Th17/Treg balance and alleviates COPD via JAK2/STAT3 pathway inhibition”. In: Allergologia et Immunopathologia 53.1 (2025), pp. 91–98. DOI: 10.15586/aei.v53i1.1215.

[74] Z. Li et al. “NRG3 contributes to cognitive deficits in chronic patients with schizophrenia”. In: Schizophrenia research 215 (2020), pp. 134–139. DOI: doi.org/10.1016/j.schres.2019.10.060.

[75] S.-Y. Liao, X. Lin, and D. C. Christiani. “Genome-wide association and network analysis of lung function in the Framingham Heart Study”. In: Genetic epidemiology 38.6 (2014), pp. 572–578. DOI: doi.org/10.1002/gepi.21841.

[76] A. M. W. Lim, E. U. Lim, P.-L. Chen, and C. S. J. Fann. “Unsupervised clustering identified clinically relevant metabolic syndrome endotypes in UK and Taiwan Biobanks”. In: Iscience 27.7 (2024). DOI: https://doi.org/10.1016/j.isci.2024.109815.

[77] H.-Y. Lin, P.-Y. Huang, T.-S. Tseng, and J. Y. Park. “SNPxE: SNP-environment interaction pattern identifier”. In: BMC Bioinformatics 22.1 (2021), p. 425. DOI: 10.1186/s12859-021-04326-x.

[78] G. Liu, J. Hu, J. Yang, and J. Song. "Predicting early-onset COPD risk in adults aged 20–50 using electronic health records and machine learning". In: PeerJ 12 (2024), e16950. DOI: doi.org/10.7717/peerj.16950.

[79] R. Lu et al. "Lung transcriptomics of radiologic emphysema reveal barrier function impairment and macrophage M1-M2 imbalance". In: medRxiv (2022). DOI: 10.1101/2022.10.21.22281369.

[80] S. T. Lugg, A. Scott, D. Parekh, B. Naidu, and D. R. Thickett. "Cigarette smoke exposure and alveolar macrophages: mechanisms for lung disease". In: Thorax 77.1 (2022), pp. 94–101. DOI: 10.1136/thoraxjnl-2020-216296. URL: https://thorax.bmj.com/content/77/1/94.

[81] R. K. Mallampalli et al. "Cigarette smoke exposure enhances transforming acidic coiled-coil–containing protein 2 turnover and thereby promotes emphysema". In: JCI insight 5.2 (2020), e125895. DOI: 10.1172/jci.insight.125895.

[82] A. Manichaikul et al. "Genome-wide association study of subclinical interstitial lung disease in MESA". In: Respiratory Research 18.1 (2017), p. 97. DOI: 10.1186/s12931-017-0581-2.

[83] N. Matveeva, I. Kiselev, N. Baulina, E. Semina, V. Kakotkin, M. Agapov, O. Kulakova, and O. Favorova. "Shared genetic architecture of COVID-19 and Alzheimer's disease". In: Frontiers in Aging Neuroscience 15 (Oct. 2023). DOI: 10.3389/fnagi.2023.1287322.

[84] P. May et al. "Rare coding variants in genes encoding GABAA receptors in genetic generalised epilepsies: an exome-based case-control study". In: The Lancet Neurology 17.8 (2018), pp. 699–708. DOI: 10.1016/S1474-4422(18)30215-1.

[85] MD Anderson Cancer Center. "Novel therapy could help people with asthma, COPD, cystic fibrosis and cancer-related lung disease". In: MD Anderson News Release (Mar. 23, 2022). Accessed: 28 Feb 2025 https://www.mdanderson.org/newsroom/novel-therapy-could-help-people-with-asthma-copd-cystic-fibrosis-and-cancer-related-lung-disease.h00-159538167.html.

[86] R. Mendez, Z. Zheng, Z. Fan, S. Rajagopalan, Q. Sun, and K. Zhang. "Exposure to fine airborne particulate matter induces macrophage infiltration, unfolded protein response, and lipid deposition in white adipose tissue". In: American journal of translational research 5.2 (2013), p. 224.

[87] L. Mentch and G. Hooker. "Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests". In: Journal of Machine Learning Research 17.26 (2016), pp. 1–41. URL: http://jmlr.org/papers/v17/14-168.html.

[88] S. Mokaddem Mohsen, S. Chakroun, A. Chaker, K. Ayed, and S. Jameleddine. "Body mass index in COPD: what relationship?" In: European Respiratory Journal 56.suppl 64 (2020). DOI: 10.1183/13993003.congress-2020.2439.

[89] D. Morena, J. L. Izquierdo, J. Rodríguez, J. Cuesta, M. Benavent, A. Perralejo, and J. M. Rodríguez. "The Clinical Profile of Patients with COPD Is Conditioned by Age". In: Journal of Clinical Medicine 12.24 (2023), p. 7595. DOI: 10.3390/jcm12247595.

[90] J. D. Morrow et al. "DNA methylation profiling in human lung tissue identifies genes associated with COPD". In: Epigenetics 11.10 (2016), pp. 730–739. DOI: 10.1080/15592294.2016.1226451.

[91] N. Mukherjee et al. "DNA methylation at birth is associated with lung function development until age 26 years". In: European Respiratory Journal 57.4 (2021). DOI: 10.1183/13993003.03505-2020. URL: https://publications.ersnet.org//content/erj/57/4/2003505.

[92] C. J. Murray and A. D. Lopez. "Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study". In: The Lancet 349.9064 (1997), pp. 1498–1504. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(96)07492-2.

[93] E. K. Nyren-Erickson, J. M. Jones, D. Srivastava, and S. Mallik. "A disintegrin and metalloproteinase-12 (ADAM12): Function, roles in disease progression, and clinical implications". In: Biochimica et Biophysica Acta (BBA) - General Subjects 1830.10 (2013), pp. 4445–4455. ISSN: 0304-4165. DOI: 10.1016/j.bbagen.2013.05.011. URL: https://www.sciencedirect.com/science/article/pii/S0304416513002079.

[94] M.-S. Ong et al. "Machine learning prediction of treatment response to inhaled corticosteroids in asthma". In: Journal of Personalized Medicine 14.3 (2024), p. 246. DOI: doi.org/10.3390/jpm14030246.

[95] R. Ottman. "Gene–Environment Interaction: Definitions and Study Design". In: Preventive Medicine 25.6 (1996), pp. 764–770. ISSN: 0091-7435. DOI: https://doi.org/10.1006/pmed.1996.0117. URL: https://www.sciencedirect.com/science/article/pii/S0091743596901176.

[96] K. Parry, L. N. Geppert, A. Munteanu, and K. Ickstadt. "Cross-Leverage Scores for Selecting Subsets of Explanatory Variables". In: arXiv preprint 2109.08399 (2021). DOI: 10.48550/arXiv.2109.08399.

[97] E. Pebesma. "Simple Features for R: Standardized Support for Spatial Vector Data". In: The R Journal 10.1 (2018), pp. 439–446. DOI: 10.32614/RJ-2018-009.

[98] E. Pebesma and R. Bivand. Spatial Data Science: With applications in R. Chapman and Hall/CRC, 2023. DOI: 10.1201/9780429459016. URL: https://r-spatial.org/book/.

[99] L. Phan, H. Zhang, Q. Wang, R. Villamarin, T. Hefferon, A. Ramanathan, and B. Kattman. "The evolution of dbSNP: 25 years of impact in genomic research". In: Nucleic Acids Research 53.D1 (2025), pp. D925–D931. DOI: 10.1093/nar/gkae977.

[100] D. E. Phelan et al. "Hypercapnia Alters Mitochondrial Gene Expression and Acylcarnitine Production in Monocytes". In: Immunology and Cell Biology 101.6 (2023), pp. 556–577. DOI: 10.1111/imcb.12642.

[101] Y. Piao, S. Y. Yun, Z. Fu, J. M. Jang, M. J. Back, H. H. Kim, and D. K. Kim. "Recombinant Human HAPLN1 Mitigates Pulmonary Emphysema by Increasing TGF-β Receptor I and Sirtuins Levels in Human Alveolar Epithelial Cells". In: Molecules and Cells 46.9 (2023), pp. 558–572. DOI: 10.14348/molcells.2023.0097.

[102] D. Pierce. ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files. R package version 1.24. 2025. URL: https://CRAN.R-project.org/package=ncdf4.

[103] M. I. Polkey. "Chronic obstructive pulmonary disease: aetiology, pathology, physiology and outcome". In: Medicine 36.4 (2008). Respiratory disorders Part 2 of 4, pp. 213–217. ISSN: 1357-3039. DOI: https://doi.org/10.1016/j.mpmed.2008.01.002. URL: https://www.sciencedirect.com/science/article/pii/S1357303908000157.

[104] E. Y. Popova, Y. I. Kawasawa, M. Leung, and C. J. Barnstable. "Temporal changes in mouse hippocampus transcriptome after pilocarpine-induced seizures". In: Frontiers in Neuroscience 18 (2024), p. 1384805. DOI: 10.3389/fnins.2024.1384805.

[105] G. Prashanth, B. Vastrad, A. Tengli, C. Vastrad, and I. Kotturshetti. "Identification of hub genes related to the progression of type 1 diabetes by computational analysis". In: BMC Endocrine Disorders 21 (2021), pp. 1–65. DOI: 10.1186/s12902-021-00709-6.

[106] P. H. Quanjer et al. "Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations". In: The European Respiratory Journal 40.6 (2012), pp. 1324–1343. DOI: 10.1183/09031936.00080312.

[107] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

[108] K. Rakkar, D. Thakker, M. A. Portelli, I. Hall, H. Schlüter, and I. Sayers. "Transcriptomics Using Lung Resection Material to Advance Our Understanding of COPD and Idiopathic Pulmonary Fibrosis Pathogenesis". In: ERJ Open Research 10.4 (2024), pp. 00061–2024. DOI: 10.1183/23120541.00061-2024.

[109] Y. Rao, X. Gai, J. Xiong, Y. Le, and Y. Sun. "Transient Receptor Potential Cation Channel Subfamily V Member 4 Mediates Pyroptosis in Chronic Obstructive Pulmonary Disease". In: Frontiers in Physiology 12 (2021), p. 783891. DOI: 10.3389/fphys.2021.783891.

[110] E. Reed, S. Nunez, D. Kulp, J. Qian, M. P. Reilly, and A. S. Foulkes. "A guide to genome-wide association analysis and post-analytic interrogation". In: Statistics in medicine 34.28 (2015), pp. 3769–3792. DOI: 10.1002/sim.6605.

[111] M. D. Ritchie and K. Van Steen. "The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation". In: Annals of translational medicine 6.8 (2018), p. 157. DOI: 10.21037/atm.2018.04.05.

[112] A. D. Rouillard, G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott, and A. Ma'ayan. "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins". In: Database 2016 (July 2016), baw100. ISSN: 1758-0463. DOI: 10.1093/database/baw100.

[113] I. Ruczinski, C. Kooperberg, and M. LeBlanc. "Logic Regression". In: Journal of Computational and Graphical Statistics 12.3 (2003), pp. 475–511. DOI: 10.1198/1061860032238. eprint: https://doi.org/10.1198/1061860032238.

[114] P. Sakornsakolpat, M. McCormack, P. Bakke, A. Gulsvik, B. J. Make, J. D. Crapo, M. H. Cho, and E. K. Silverman. "Genome-Wide Association Analysis of Single-Breath DlCO". In: American Journal of Respiratory Cell and Molecular Biology 60.5 (2019), pp. 523–531. DOI: 10.1165/rcmb.2018-0384OC.

[115] P. Sakornsakolpat et al. "Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations". In: Nature genetics 51.3 (2019), pp. 494–505. DOI: 10.1038/s41588-018-0342-2.

[116] I. Salvato et al. "Expression of targets of the RNA-binding protein AUF-1 in human airway epithelium indicates its role in cellular senescence and inflammation". In: Frontiers in Immunology 14 (2023), p. 1192028. DOI: 10.3389/fimmu.2023.1192028.

[117] A. M. Santamans et al. “MCJ: A mitochondrial target for cardiac intervention in pulmonary hypertension”. In: Science advances 10.3 (2024), eadk6524. DOI: 10.1126/sciadv.adk6524.

[118] A. Schienkiewitz, G. Mensink, R. Kuhnert, and C. Lange. “Übergewicht und Adipositas bei Erwachsenen in Deutschland”. In: Journal of Health Monitoring. 2. Robert Koch-Institut, Epidemiologie und Gesundheitsberichterstattung, 2017. DOI: 10.17886/RKI-GBE-2017-025.

[119] T. Schikowski, U. Ranft, D. Sugiri, A. Vierkötter, T. Brünong, V. Harth, and U. Krämer. “Decline in air pollution and change in prevalence in respiratory symptoms and chronic obstructive pulmonary disease in elderly women.” In: Respiratory research 11(1) (2010). DOI: 10.1186/1465-9921-11-113.

[120] T. Schikowski, D. Sugiri, U. Ranft, U. Gehring, J. Heinrich, H.-E. Wichmann, and U. Krämer. “Long-term air pollution exposure and living close to busy roads are associated with COPD in women”. In: Respiratory research 6 (2005), pp. 1–10. DOI: 10.1186/1465-9921-6-152.

[121] F.-R. Schumacher, S. Schubert, M. Hannus, B. Sönnichsen, C. Ittrich, S. Kreideweiss, T. Kurz, and J. F. Rippmann. “RNAi screen for NRF2 inducers identifies targets that rescue primary lung epithelial cells from cigarette smoke induced radical stress”. In: Plos one 11.11 (2016), e0166352. DOI: 10.1371/journal.pone.0166352.

[122] H. Schwender and K. Ickstadt. “Identification of SNP interactions using logic regression”. In: Biostatistics 9.1 (June 2007), pp. 187–198. DOI: 10.1093/biostatistics/kxm024.

[123] H. Schwender, S. Selinski, M. Blaszkewicz, R. Marchan, K. Ickstadt, K. Golka, and J. G. Hengstler. “Distinct SNP Combinations Confer Susceptibility to Urinary Bladder Cancer in Smokers and Non-Smokers”. In: PLOS ONE 7 (Dec. 2012), pp. 1–12. DOI: 10.1371/journal.pone.0051880. URL: https://doi.org/10.1371/journal.pone.0051880.

[124] H. Schwender and T. Tietz. logicFS: Identification of SNP Interactions. R package version 2.20.0. 2023. DOI: 10.18129/B9.bioc.logicFS. URL: https://bioconductor.org/packages/logicFS.

[125] Z. Shang, J. Sun, J. Hui, Y. Yu, X. Bian, B. Yang, K. Deng, and L. Lin. “Construction of a Support Vector Machine-Based Classifier for Pulmonary Arterial Hypertension Patients”. In: Frontiers in Genetics 12 (2021), p. 781011. DOI: 10.3389/fgene.2021.781011.

[126] B. Shastry. “SNPs: impact on gene function and phenotype.” In: Methods in molecular biology (Clifton, N.J.) (2009), pp. 3–222. DOI: 10.1007/978-1-60327-411-1_1.

[127] W. Shen, W. Wei, S. Wang, X. Yang, R. Wang, and H. Tian. “RNA-binding protein AZGP1 inhibits epithelial cell proliferation by regulating the genes of alternative splicing in COPD”. In: Gene 927 (2024), p. 148736. ISSN: 0378-1119. DOI: 10.1016/j.gene.2024.148736. URL: https://www.sciencedirect.com/science/article/pii/S0378111924006176.

[128] L. Shorey-Kendrick et al. “Impact of vitamin C supplementation on placental DNA methylation changes related to maternal smoking: association with gene expression and respiratory outcomes”. In: Clinical epigenetics 13 (2021), pp. 1–17. DOI: doi.org/10.1186/s13148-021-01161-y.

[129] N. Shrine et al. “New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries”. In: Nature genetics 51.3 (2019), pp. 481–493. DOI: 10.1038/s41588-018-0321-7.

[130] A. J. Smits et al. “Distinct Platelet Ribonucleic Acid Signatures in Patients with Pulmonary Hypertension”. In: Annals of the American Thoracic Society 19.10 (2022), pp. 1650–1660. DOI: 10.1513/AnnalsATS.202201-085OC.

[131] M. Stanković, V. Dordević, A. Tomović, L. Nagorni-Obradović, N. Petrović-Stanojević, M. Kovač, and D. Radojković. “Interactions of the eNOS and ACE genes and cigarette smoking in chronic obstructive pulmonary disease”. In: Journal of Medical Biochemistry 42.1 (2023), p. 94. DOI: 10.5937/jomb0-34017.

[132] G. Stelzer et al. “The GeneCards suite: from gene data mining to disease genome sequence analyses”. In: Current protocols in bioinformatics 54.1 (2016), pp. 1–30. DOI: doi.org/10.1002/cpbi.5.

[133] K. Stienstra, A. Knigge, and I. Maas. “Gene-environment interaction analysis of school quality and educational inequality”. In: NPJ Science of Learning 9.1 (2024), p. 14. DOI: 10.1038/s41539-024-00225-x.

[134] S. E. Straus, F. A. McAlister, D. L. Sackett, and J. J. Deeks. “The accuracy of patient history, wheezing, and laryngeal measurements in diagnosing obstructive airway disease”. In: JAMA 283.14 (2000), pp. 1853–1857. DOI: 10.1001/jama.283.14.1853.

[135] Y. Sun, Y. Zhang, X. Liu, Y. Liu, F. Wu, and X. Liu. “Association between body mass index and respiratory symptoms in US adults: A national cross-sectional study”. In: Scientific Reports 14.1 (2024), p. 940. DOI: 10.1038/s41598-024-51637-z.

[136] S. Teschke, K. Ickstadt, and A. Munteanu. “Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores”. In: Biometrical Journal 66.8 (2024), e70014. DOI: 10.1002/bimj.70014.

[137] A. E. Tilley et al. “Cigarette smoking induces changes in airway epithelial expression of genes associated with monogenic lung disorders”. In: American Journal of Respiratory and Critical Care Medicine 193.2 (2016), pp. 215–217. DOI: 10.1164/rccm.201412-2290LE.

[138] I. Ungvári et al. “Evaluation of a Partial Genome Screening of Two Asthma Susceptibility Regions Using Bayesian Network Based Bayesian Multilevel Analysis of Relevance”. In: PLOS ONE 7.3 (2012), e33573. DOI: 10.1371/journal.pone.0033573.

[139] M. Vossoughi et al. “Air pollution and subclinical airway inflammation in the SALIA cohort study”. In: Immunity & ageing 11.1 (2014). DOI: 10.1186/1742-4933-11-5.

[140] M. de Vries, J. Vonk, and M. Boezen. “Genetic variants and lung function decline in the LifeLines cohort study”. In: European Respiratory Journal 54.suppl 63 (2019). DOI: 10.1183/13993003.congress-2019.PA5398. URL: http://erj.ersjournals.com/content/erj/54/suppl_63/PA5398.

[141] L. Wang, H. Zhao, I. Raman, M. Yan, Q. Chen, and Q.-Z. Li. “Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease: miRNA and mRNA regulation”. In: Journal of Inflammation Research (2022), pp. 2167–2180. DOI: doi.org/10.2147/JIR.S337894.

[142] Q. Wang and S. Liu. “The Effects and Pathogenesis of PM2.5 and Its Components on Chronic Obstructive Pulmonary Disease”. In: International Journal of Chronic Obstructive Pulmonary Disease 18 (2023), pp. 493–506. DOI: 10.2147/COPD.S402122.

[143] T. Wang, W. Wang, C. Xu, X. Tian, and D. Zhang. “Genome-wide analysis in northern Chinese twins identifies twelve new susceptibility loci for pulmonary function”. In: BMC Genomics 25.1 (2024), p. 1255. DOI: 10.1186/s12864-024-11165-6.

[144] Y. Wang et al. “DRD1 downregulation contributes to mechanical stretch-induced lung endothelial barrier dysfunction”. In: Theranostics 11.6 (2021), pp. 2505–2521. DOI: 10.7150/thno.46192.

[145] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[146] H. Wickham. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.1. 2023. URL: https://CRAN.R-project.org/package=stringr.

[147] H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan. dplyr: A Grammar of Data Manipulation. R package version 1.1.4. 2023. URL: https://CRAN.R-project.org/package=dplyr.

[148] H. Wickham, E. Miller, and D. Smith. haven: Import and Export ’SPSS’, ’Stata’ and ’SAS’ Files. R package version 2.5.4. 2023. URL: https://CRAN.R-project.org/package=haven.

[149] A. Wood, P. Newby, S. Gough, and R. Stockley. “CTLA4 polymorphisms and COPD”. In: European Respiratory Journal 35.2 (2010), pp. 457–458. DOI: 10.1183/09031936.00134709.

[150] World Health Organization. “WHO global air quality guidelines: Particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide”. In: (2021). Licence: CC BY-NC-SA 3.0 IGO. URL: https://apps.who.int/iris/handle/10665/345329.

[151] M. N. Wright, A. Ziegler, and I. R. König. “Do little interactions get lost in dark random forests?” In: BMC Bioinformatics 17.1 (2016), p. 145. DOI: 10.1186/s12859-016-0995-8.

[152] S. Xiong, Q. Liu, S. Zhou, and Y. Xiao. “Identification of key genes and regulatory networks involved in the comorbidity of atrial fibrillation and chronic obstructive pulmonary disease”. In: Heliyon 9.11 (2023), e22430. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2023.e22430. URL: https://www.sciencedirect.com/science/article/pii/S240584402309638X.

[153] C. Yang et al. “Plasma Proteomics Study Between the Frequent Exacerbation and Infrequent Exacerbation Phenotypes of Chronic Obstructive Pulmonary Disease”. In: International Journal of Chronic Obstructive Pulmonary Disease (2023), pp. 1713–1728. DOI: 10.2147/COPD.S408361.

[154] I. V. Yang et al. “DNA Methylation and Childhood Asthma in the Inner City”. In: The Journal of Allergy and Clinical Immunology 136.1 (2015), pp. 69–80. DOI: 10.1016/j.jaci.2015.01.025.

[155] I. A. Yang, C. R. Jenkins, and S. S. Salvi. “Chronic obstructive pulmonary disease in never-smokers: Risk factors, pathogenesis, and implications for prevention and treatment”. In: The Lancet Respiratory Medicine 10.5 (2022), pp. 497–511. DOI: 10.1016/S2213-2600(21)00506-3.

[156] M. Yang et al. “Proteomic Profiling of Lung Immune Cells Reveals Dysregulation of Phagocytotic Pathways in Female-Dominated Molecular COPD Phenotype”. In: Respiratory Research 19.1 (2018), p. 39. DOI: 10.1186/s12931-017-0699-2.

[157] R. M. Younis, R. M. Taylor, P. M. Beardsley, and J. L. McClay. "The ANKS1B gene and its associated phenotypes: focus on CNS drug response". In: Pharmacogenomics 20.9 (2019), pp. 669–684. DOI: 10.2217/pgs-2019-0015.

[158] Z. Yu, J. Chen, H. Shi, G. Stoeber, S.-Y. Tsang, and H. Xue. "Analysis of GABRB2 association with schizophrenia in German population with DNA sequencing and one-label extension method for SNP genotyping". In: Clinical Biochemistry 39.3 (2006), pp. 210–218. ISSN: 0009-9120. DOI: 10.1016/j.clinbiochem.2006.01.009.

[159] L. Yuan et al. "Clinical characteristics and gene mutation profiles of chronic obstructive pulmonary disease in non-small cell lung cancer". In: Frontiers in Oncology 12 (2022), p. 946881. DOI: 10.3389/fonc.2022.946881.

[160] D. Zhang, X. Pu, M. Zheng, G. Li, and J. Chen. "Employing a synergistic bioinformatics and machine learning framework to elucidate biomarkers associating asthma with pyrimidine metabolism genes". In: Respiratory Research 25.1 (2024), p. 327. DOI: 10.1186/s12931-024-02954-4.

[161] S. Zhang, K. Pang, X. Feng, and et al. "Transcriptomic data exploration of consensus genes and molecular mechanisms between chronic obstructive pulmonary disease and lung adenocarcinoma". In: Scientific Reports 12 (2022), p. 13214. DOI: 10.1038/s41598-022-17552-x.

[162] Y. Zhang. "From gene identifications to therapeutic targets for asthma: Focus on great potentials of TSLP, ORMDL3, and GSDMB". In: Chinese Medical Journal Pulmonary and Critical Care Medicine 1.3 (2023), pp. 139–147. ISSN: 2772-5588. DOI: 10.1016/j.pccm.2023.08.001. URL: https://www.sciencedirect.com/science/article/pii/S2772558823000403.

[163] R. Zhao, Y. Guo, L. Zhang, and et al. "CBX4 plays a bidirectional role in transcriptional regulation and lung adenocarcinoma progression". In: Cell Death & Disease 15 (2024), p. 378. DOI: 10.1038/s41419-024-06745-z.

[164] T. Zhao et al. "Otilonium bromide ameliorates pulmonary fibrosis in mice through activating phosphatase PPM1A". In: Acta Pharmacologica Sinica (2024), pp. 1–15. DOI: 10.1038/s41401-024-01368-8.

[165] X. Zheng, D. Levine, J. Shen, S. Gogarten, C. Laurie, and B. Weir. "A high-performance computing toolset for relatedness and principal component analysis of SNP data". In: Bioinformatics 28.24 (2012), pp. 3326–3328. DOI: 10.1093/bioinformatics/bts606.

Appendix A Plausibility check of variable selection

After LD pruning of the selected SNPs the SNP with the largest absolute CLS score, thus the largest importance, is rs736020. This SNP is located next to the protein coding gene *DHRS9*. The eponymous protein encoded by this gene can be used as phenotype-specific proteomic signature in patients with COPD emphysematous phenotype [7].
Below we discuss the associations with COPD or similar lung outcomes found in the literature for the top 50 selected SNPs and the related genes. For better readability, only the genes that contain or lie close to the selected SNPs are mentioned. See table 4 for the corresponding SNPs. The mapping of SNPs and genes is achieved through the utilization of diverse platforms [132], [99] and [57].
The HEXB protein encoded by the *HEXB* gene is reported as a key protein that is decreased in female COPD patients compared to female smokers [156]. Moreover, the *HEXB* gene has been linked to hypercapnia, a condition that increases the risk of developing COPD [100]. Within the next important gene, *COG2*, a significant exon was identified corresponding to emphysema, an important COPD phenotype [79]. One treatment method for COPD patients with chronic hypercapnic respiratory failure is long-term non-invasive ventilation [16]. However, this type of ventilation leads to ventilator-induced lung injury, which is a significant problem in COPD patients [56]. Recent studies have suggested a possible link between DRD1, encoded by the *DRD1* gene, and ventilator-induced lung injury. Specifically, the findings indicate a negative correlation between pulmonary DRD1 expression levels and the duration of mechanical ventilation[144]. Another association with COPD can be derived from the microRNA *miR-223*. *miR-223* has been identified as one of the most impacted in subjects with COPD compared to smokers without obstruction [36]. Furthermore, studies have demonstrated that *miR-223* directly targets *PKNOX1*, thereby inhibiting its expression [44]. Consequently, a plausible mechanism underlying the development of COPD involves a genetic variation in the *PKNOX1* gene. One study showed that a variation in *RYR3* can affect ryanodine receptor function, which regulates calcium balance in smooth muscle cells of the airways and it can lead to impaired regulation and increased bronchoconstriction, potentially causing respiratory diseases such as asthma or COPD [33]. The next gene, *GBP5*, is a target gene of the AZGP1 protein, which regulates alternative splicing events in COPD, so *GBP5* may play a role in contributing to COPD [127]. Then, *ETS1* encodes the protein of the same name, which is part of a highly interconnected protein network forming the COPD proteome [8]. The chromobox protein CBX4, encoded by *CBX4*, has been reported to be upregulated in lung adenocarcinoma, which shares similar molecular mechanisms in the pathogenesis of COPD [161]. In addition,

the inhibition of ZEB2 transcription involves CBX4 [163] and a Bayes factor colocalization analysis, restricted to loci meeting genomewide significant criteria, found support for shared variants at ZEB2 loci for spirometry-defined COPD and heavy smoking [10]. Furthermore, mutations in *EPHA5* were reported in patients (of European descent) with non-small cell lung cancer coexisting COPD compared to patients without COPD [159]. In addition, there is an association auf *EPHA5* and asthma [162]. Next, *KIFC3* was identified as a gene downregulated in both COPD and idiopathic pulmonary fibrosis patients compared to controls [108] and *COL21A1* in COPD and atrial fibrillation [152]. The SNP `rs7738312` is located next to the *SLC22A23* gene and another variant in this gene has been shown to be associated with bronchodilator responsiveness in COPD [47]. Similarly, a variant in the *RBFOX1* gene was found to be associated with weight loss in non-Hispanic white COPD participants in the COPDGene study [67]. In addition, *RBFOX1* is one of two genes for which SNPs were found to be genome-wide significantly associated with both COVID-19 and Alzheimer's disease [67], [83]. *PAPLN* was detected to be down-regulated expression in patients with acute exacerbation of COPD compared to healthy control participants [72]. Next it is reported that the inhibition of the JAK2/STAT3 pathway has been shown to alleviate COPD [73]. So, a variation in the *JAK2* gene may be a risk factor for COPD or influence its progression. In a GWAS of subclinical interstitial lung disease, defined as high attenuation areas on CT, variants in *UBE2U* showed significant [82]. The matrisome protein corresponding to *EFEMP1* shows an altered abundance in COPD scaffolds compared to normal scaffolds. Since EFEMP1 is known to weakly bind tropoelastin and is directly involved in the regulation of elastic fiber integrity, its increased abundance may be a result of dysregulated elastic fiber homeostasis in COPD lungs [48]. Additionally, *BLOC1S4* encoded protein is a subunit of BLOC-1, where mutations may be a risk for pulmonary fibrosis [54]. Other genes, according to the top 50 SNPs, which are some how directly associated with COPD are *CTLA4* [149], *ETNK1* [24], *PPARGC1A* [71], *GLT1D1* [90], *EDEM1* [42], *TRPV4* [109], *HAPLN1* [101], *OLFM3* [78] and *ARHGAP28* [114]. Genes associated with the Tiffeneau Index (FEV_1/FVC), which we use as parameter for COPD, are *MX1* [50] and *RARB* [3]. Other genes are associated with the pathophysiological mechanisms of COPD, such as *TRAT1* [141] and *KLF10* [64] or pulmonary function like FEV_1: *COBL* [91], *TAFA5* [143] and *ARHGAP31* [140]. A number of genes have not yet been directly linked to COPD, but there is an association with asthma (*ADAM12* [93], *POLR1C* [160], *FRMD6* [138], and *FAM84B* [154]) or Pulmonary arterial hypertension (*SIK1* [125]). So far we have not found any associations for the remaining genes (*PEX2*, *CAPN15*, *CHD6*, *TMEM171*, *HCN4*, *RIN2*, *C14orf180*, *BLID* and *TRPS1*) but this does not mean that they do not exist. Conducting such an analysis for the remaining selected SNPs would exceed the scope of the present work, due to the fact that the primary objective of this section is just to validate the selection based on the CLS. In the main part of this study, we will apply logicDT to all selected variables, including the environmental and clinical variables. For the important variables found there that do not overlap with those described in this section, we then perform another literature search.

Appendix B Figures & Tables

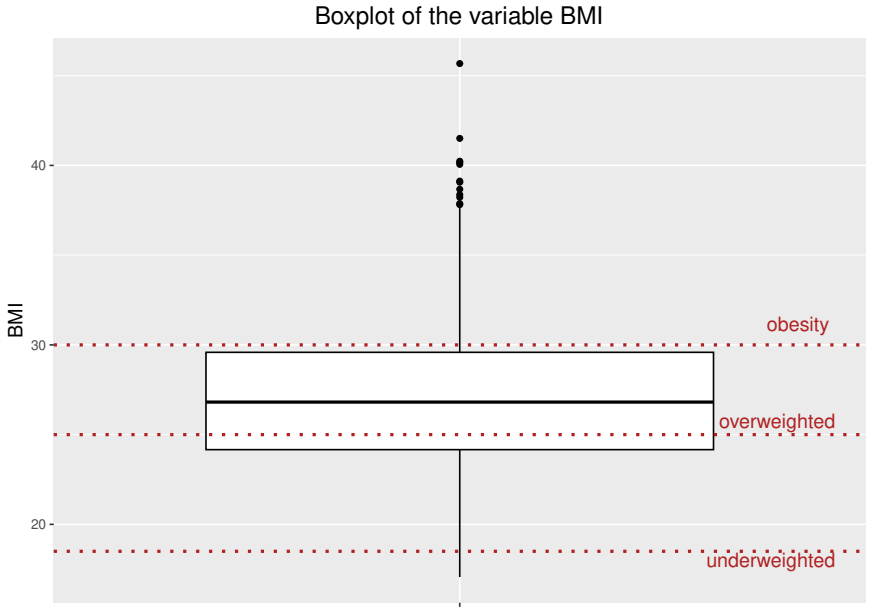

Figure 2: Boxplot of the variable body mass index (BMI) across 503 women in the SALIA cohort study

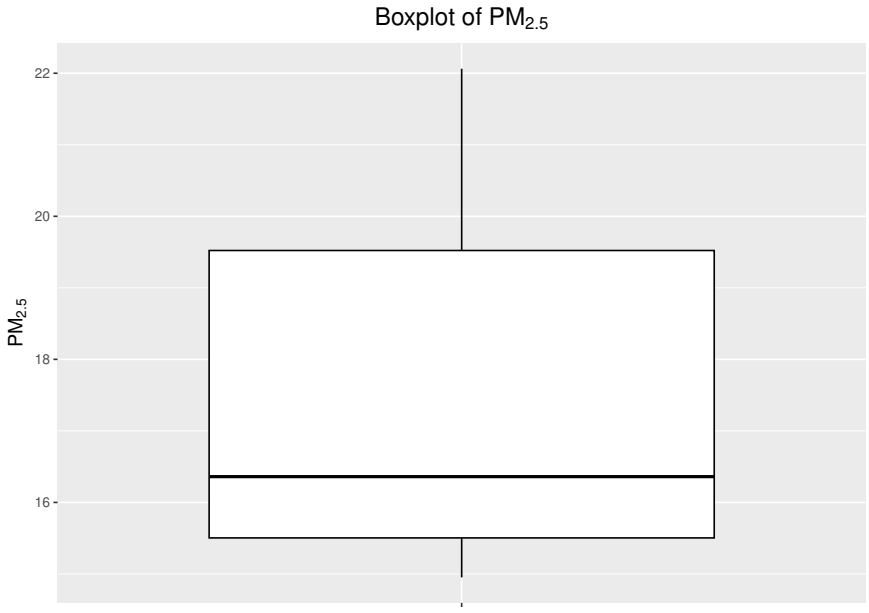

Figure 3: Boxplot of the variable $PM_{2.5}$ (Particular Matter with diameter less than $2.5\mu m$)

SNP	CLS	(nearest) protein coding gene	Literature source
rs736020	-2.039×10^{-7}	*DHRS9*†	[7]
rs10036629	1.993×10^{-7}	*HEXB*†	[156],[100]
rs1752178	1.970×10^{-7}	*COG2*†	[79]
rs11750933	-1.928×10^{-7}	*DRD1*†	[144]
rs9637290	1.918×10^{-7}	*PKNOX1*	[36], [44]
rs729411	1.884×10^{-7}	*ARHGAP31*	[140]
rs4236857	1.866×10^{-7}	*PEX2*†	
rs7182398	-1.864×10^{-7}	*RYR3*	[33]
rs12132738	-1.851×10^{-7}	*GBP5*†	[127]
rs462903	-1.827×10^{-7}	*MX1*	[50]
rs7110282	-1.818×10^{-7}	*ETS1*†	[8]
rs894862	1.818×10^{-7}	*CBX4*†	[161], [163], [10]
rs1674907	-1.812×10^{-7}	*ADAM12*	[93]
rs56038426	1.789×10^{-7}	*CTLA4*†	[149]
rs6550930	1.784×10^{-7}	*RARB*	[3]
rs2467458	1.770×10^{-7}	*ETNK1*†	[24]
rs34161784	-1.765×10^{-7}	*TRAT1*	[141]
rs13236400	1.761×10^{-7}	*COBL*†	[91]
rs1858113	1.759×10^{-7}	*PPARGC1A*†	[71]
rs2066432	1.746×10^{-7}	*POLR1C*	[160]
rs60924083	1.742×10^{-7}	*GLT1D1*	[90]
rs13109328	-1.742×10^{-7}	*EPHA5*	[162], [159]
rs2436863	-1.736×10^{-7}	*KLF10*†	[64]
rs2967169	1.733×10^{-7}	*KIFC3*	[108]
rs6470471	1.731×10^{-7}	*FAM84B*†	[154]
rs1836577	-1.72×10^{-7}	*EDEM1*†	[42]
rs8050046	1.723×10^{-7}	*CAPN15*, *MIR3176*$^{+}$	
rs6065392	-1.722×10^{-7}	*CHD6*†	
rs682798	1.720×10^{-7}	*TMEM171*, *LOC124901001*	
rs4776632	1.718×10^{-7}	*HCN4*	
rs133663	-1.715×10^{-7}	*TAFA5*	[143]
rs2841001	-1.709×10^{-7}	*COL21A1*†	[152]
rs2424251	-1.706×10^{-7}	*RIN2*	
rs7738312	1.706×10^{-7}	*SLC22A23*†	[47]
rs3923816	-1.704×10^{-7}	*FRMD6*†	[138]
rs9671900	-1.702×10^{-7}	*C14orf180*†	
rs177391	1.700×10^{-7}	*PAPLN*	[72]
rs13288137	-1.699×10^{-7}	*JAK2*†	[73]
rs7520865	1.694×10^{-7}	*UBE2U*†	[82]
rs2338568	1.689×10^{-7}	*TRPV4*	[109]
rs1237355	-1.683×10^{-7}	*BLID*	
rs6452548	-1.679×10^{-7}	*HAPLN1*	[101]
rs4786833	-1.675×10^{-7}	*RBFOX1*	[67], [83]
rs17488808	1.673×10^{-7}	*OLFM3*	[78]
rs4560527	-1.672×10^{-7}	*TMEM174*†	
rs3791676	-1.672×10^{-7}	*EFEMP1*	[48]
rs4818726	1.671×10^{-7}	*SIK1*†	[125]
rs12954151	1.670×10^{-7}	*ARHGAP28*	[114]
rs7681156	-1.700×10^{-7}	*BLOC1S4*	[54]
rs1180626	1.666×10^{-7}	*TRPS1*	

Table 4: The 50 SNPs (single nucleotide polymorphisms) with the largest Cross Leverage Scores (CLS) and the respective protein coding genes where the SNPs are located. $+$ characterizes RNA genes and † when the SNP is not in but next to the gene.

Appendix C Software

All calculations are carried out in the R 4.3.2 software [107]. Different R packages were used for the different steps of the analysis. We use packages `terra` [49], `ncdf4` [102], `sf` [98], [97] and `haven` [148] for the preprocessing of the data. Packages `stringr` [146] and `SNPRelate` [165] were used for LD pruning. The Code we need for the variable selection is required at Github[1] and for the subsequent analysis with logicDT we need `logicFS` [124] and `logicDT` [69]. Other packages used are `ggplot2` [145] for graphics and `dplyr` [147] and `magrittr` [6] for certain coding passages.

Appendix D Theory

LogicDT

Let $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathcal{R}^p$ the vector of input variables and $\boldsymbol{Y}$ the random outcome variable. A *training data set* $\mathcal{D}$ is defined as $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with independent and identically distributed observations from the joint probability distribution of $(\boldsymbol{X}, \boldsymbol{Y})$. *terms* are the possible single input variables and their interactions and a set of *terms* is called *state*. Then, the *terms* of a *state* s are exclusively used as input variables for building a decision tree by transforming the training data set $\mathcal{D}$ into a *tree training data set* $\mathcal{D}_s$ consisting exclusively of the *terms* of *state* s as well as the outcome. To find the ideal *state* s a global search with *simulated annealing* over all appropriate *state* is performed. The procedure is to evaluate the tree for a *state* s with a Score function, construct a slightly modified *state* s' called *neighbor* and fit a new tree based on s' which is then also evaluated with the Score function. If this new score $\text{Score}_{\text{new}}$ for the modified *state* s' is smaller than the current smallest score $\text{Score}_{\text{min}}$, the current *state* is updated with a predefined acceptance probability. This while loop is carried out until a stopping criteria becomes true or the optimal states is obtained. The initial *state* can consist of a single input variable that minimizes the score function, a random input variable, or it can be empty. For details according hyperparameter tuning to control the complexity of logicDT models see [70].
To cover interactions between binary predictors and a quantitative covariable, regression models are fitted in the leaves resulting from splits using only the binary terms. In this case, a likelihood-ratio test is used as split criterion for fitting the tree. It is tested if a further splitting of the node due to a binary variable X_s does not lead to different prediction models in the current tree branch.

It is also possible to consider logicDT in a ensembles model using bagging [14], called bagged logicDT. Since the combination of bagging and simulated annealing in the respective models is very computationally intensive, it is recommended to use a greedy search, where the best modified state is chosen deterministically. The respective algorithm given in the following, see algorithm 1.
The out-of-bag observations (oob-observations) can be used for unbiased and stabilized estimation of the generalization error and variable importance measures (VIMs). The VIMs typically result from the difference between the prediction error of the full model and the model without the input variable of interest. Various versions of the VIM can be considered. The *permutation VIM* [15], the *removal VIM* [87] and *Logic VIM* [70] for binary predictors. To measure the importance of interactions the *Interaction VIM* is proposed which uses a joint VIM and can be combined with the three mentioned VIM procedures [70]. To measure the importance of specific conjunctions which are identified by logicDT, each possible conjunction of the identified interaction is considered and the conjunction that leads to the most severe deviation in the outcome is selected. For more details refer to [70].

[1] `https://github.com/SvenTeschke/special_issue_CEN`

Algorithm 1 logicDT fitting

function LOGICDT($\mathcal{D}$ *(Training data)*)
    $s \leftarrow$ Initialize state/set of terms
    $\mathcal{D}_s \leftarrow$ Apply s to $\mathcal{D}$
    $T \leftarrow$ FITDECISIONTREE($\mathcal{D}_s$), see Algorithm 1 in [70]
    $\text{Score}_{\text{min}} \leftarrow \text{Score}(T)$
    while Global search is not finished **do**
        $s' \leftarrow$ Modify current state s
        $\mathcal{D}_{s'} \leftarrow$ Apply s' to $\mathcal{D}$
        $T' \leftarrow$ FITDECISIONTREE($\mathcal{D}_{s'}$)
        $\text{Score}_{\text{new}} \leftarrow \text{Score}(T')$
        if State s' is accepted based on $\text{Score}_{\text{min}}$ and $\text{Score}_{\text{new}}$ **then**
            $s \leftarrow s'$
            $T \leftarrow T'$
            $\text{Score}_{\text{min}} \leftarrow \text{Score}_{\text{new}}$
        end if
    end while
    return (s, T)
end function

Development and Application of Brain Tissue Based Multi-Omics Profile Scores for Alzheimer Disease

DEVELOPMENT AND APPLICATION OF BRAIN TISSUE BASED MULTI-OMICS PROFILE SCORES FOR ALZHEIMER'S DISEASE

Timur Tug[1,2*], Donghai Liang[2,3], Sven Teschke[1], Youran Tan[3], Marla Gearing[4,5], Allan I. Levey[5], James J. Lah[5], Aliza P. Wingo[6], Thomas S. Wingo[7,8], Michael Lau[9,10], Katja Ickstadt[1], Anke Hüls[2,3*]

[1]Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany,
[2]Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA,[3]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA,
[4]Department of Pathology and Laboratory Medicine, Emory University, 100 Woodruff Circle, Atlanta, GA 30322 USA,
[5]Department of Neurology, Emory University School of Medicine, 100 Woodruff Circle, Atlanta, GA 30322 USA,
[6]Department of Psychiatry, University of California, Davis, 2230 Stockton Blvd, Sacramento, CA 95817, USA,[7]Department of Neurology, University of California, Davis, 4860 Y Street, Sacramento, CA 95817, USA,[8]Alzheimer's Disease Research Center, University of California, Davis, 1651 Alhambra Blvd, Suite 200A, Sacramento, CA 95816,USA
[9]Mathematical Institute, Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
[10]eBay Inc., 2025 Hamilton Avenue, San Jose, CA 95125, USA*: corresponding authors: timur.tug@tu-dortmund.de, anke.huels@emory.edu

Corresponding authors:

Timur Tug, Vogelpothsweg 87, 44227 Dortmund, Germany

Email: timur.tug@tu-dortmund.de

Anke Huels, PhD, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA.

Email: anke.huels@emory.edu

Abstract (148/150 words)

INTRODUCTION

Advances in omics technologies, such as epigenomics and metabolomics, provide novel insights into the biological mechanisms underlying Alzheimer's disease (AD). However, little is known how different omics layers interact and jointly relate to AD neuropathology.

METHODS

We performed a comprehensive single- and multi-omics analysis integrating genome-wide DNA methylation and high-resolution metabolomics data from 157 frontal cortex samples. We developed novel single and multi-omics profile scores (PS) for AD pathology, using a combination of machine learning, regression, and pathway analysis.

RESULTS

The best multi-omics PS showed an R^2 of 0.15 for the ABC score (Amyloid, Braak, CERAD), independent of age, sex, race and socioeconomic factors. The pathway analyses identified lipid metabolism and signal transduction as key biological pathways among the included metabolites and CpG sites.

DISCUSSION

The integration of DNA methylation and metabolomics provides deeper insights into AD pathophysiology and identifies promising molecular targets for potential disease-modifying approaches.

Research in Context

1. **Systematic Review:** Alzheimer's disease (AD) is a multifactorial disorder with complex molecular underpinnings. Advances in omics technologies, particularly DNA methylation (DNAm) and metabolomics, have provided insights into AD pathophysiology. Prior studies identified associations between DNAm and AD-related neuropathology, while metabolomics studies highlighted alterations in lipid metabolism and oxidative stress. However, most research examines these omics layers separately, limiting insights into their interplay.

2. **Interpretation:** Our study integrates DNAm and metabolomics data using novel PS approaches to improve AD prediction. The best DNAm-based profile score (PS) achieved $R^2 = 0.11$, outperforming metabolomics-based PS ($R^2 = 0.04$). Combining both omics layers improved predictive accuracy ($R^2 = 0.15$). Key pathways enriched in both layers include lipid metabolism and signal transduction, reinforcing their role in AD pathology.

3. **Future Direction:** Future studies should expand multi-omics approaches, conduct longitudinal analyses, validate findings in diverse cohorts, and explore translational applications for early diagnosis and therapy.

BACKGROUND

Alzheimer's disease (AD) is a progressive neurological disorder that affects millions of Americans, with approximately 6.7 million people aged 65 and older currently living with the condition [1]. AD is the fifth-leading cause of death among older adults in the United States and poses a major public health challenge [2]. The financial burden is equally staggering, with the annual cost of care for AD patients projected to reach $580 billion in 2025, a figure expected to rise substantially in the coming decades [1,3]. As the prevalence of AD continues to rise globally due to aging populations, developing effective strategies to reduce its burden has become a critical priority. Substantial efforts are underway to create disease-modifying therapies targeting the molecular mechanisms of AD [4,5], but the complex and multifactorial nature of the disease has posed significant challenges.

Advances in high-throughput omics technologies, such as genomics, epigenomics, proteomics, and metabolomics, have provided new insights into the biological pathways and signatures underlying AD, offering promising avenues for innovative therapeutic strategies [6,7]. Among these, DNA methylation (DNAm) and metabolomics analyses have emerged as powerful tools for exploring AD etiology. These approaches offer complementary insights into AD's complex pathophysiology by capturing information on epigenetic changes and metabolic dysfunction—both recognized as core features of the disease [8]. DNAm, a key epigenetic mechanism, is closely linked to AD, with global hypomethylation, gene-specific methylation changes, and interactions with neuroinflammation, aging, oxidative stress, and environmental factors contributing to disease risk and progression. As the downstream product of the gene transcription and gene-environment interaction, metabolomics involves studying small molecules (metabolites) in biological systems, which can reveal changes in metabolic pathways associated with AD. Metabolic pathways that have been reported in association with AD include dysregulation in energy metabolism, lipid profiles, amino acid pathways, oxidative stress, and gut-brain axis interactions [9–12]. However, little is known how differences in DNAm and metabolomics interact and jointly influence the development of AD. Integrating DNAm and metabolomics through a multi-omics approach could illuminate shared biological pathways that may be particularly important for understanding AD etiology.

To address this knowledge gap, we conducted a comprehensive multi-omics analysis integrating genome-wide DNAm and high-resolution metabolomics data from 157 prefrontal cortex tissue samples of brain donors with varying stages of AD pathology, assessed with Braak staging, CERAD (**C**onsortium to **E**stablish a **R**egistry for **A**lzheimer's **D**isease) scoring, and the comprehensive ABC score (**A**myloid, Braak, **C**ERAD) [13,14]. Specifically, we calculated multi-omics profile scores (PS) integrating DNAm and metabolomic data, providing a holistic understanding of AD neuropathology by capturing both epigenetic and metabolic contributions. Our analysis framework not only identifies CpG sites and/or metabolomic features predictive of AD neuropathology levels but also elucidates mechanistic underpinnings by evaluating and comparing biological pathways enriched among the selected CpG sites and/or metabolomic features. The multi-omics framework developed in this study highlights the potential of combining epigenetic and metabolomic data to deepen our understanding of AD pathophysiology. This study extends the current state of research through several innovative approaches. We proposed novel brain

tissue-based multi-omics profile scores for AD, which integrate DNA methylation and metabolomics to better predict the neuropathological changes of the disease. Methodologically, advanced machine learning techniques such as Random Forest, Elastic Net and Boosting are used to efficiently analyze high-dimensional data and generate robust profile scores. These novel analysis strategies and methods and a comprehensive pathway analysis allow a more precise identification of relevant biological mechanisms.

METHODS

Study design

The Emory Goizueta Alzheimer's Disease Research Center (ADRC) established a brain bank to support Alzheimer's research, primarily enrolling research participants and patients clinically diagnosed with AD by Emory physicians. By the third quarter of 2020, the brain bank included 1,011 donors. Genome-wide DNAm and metabolomics profiling were conducted on 161 samples from donors deceased after 2007, with 159 samples passing quality control. All 161 donors had complete data for key covariates (e.g., age of death, race, sex and educational attainment) and outcome variables (e.g., ABC score, Braak Stage, CERAD). Informed consent and Institutional Review Board-approved protocols governed the research.

Assessment of AD neuropathology

The ADRC conducted comprehensive neuropathologic evaluations on all donor brains using established research protocols and diagnostic criteria [13]. These assessments, performed by experienced neuropathologists, involved a variety of stains and immunohistochemical techniques, along with semi-quantitative scoring to evaluate AD and related neuropathologies in various brain regions [15]. AD neuropathology in this project was measured using the Braak staging, CERAD score, and ABC score - each assessing different aspects of disease progression. Braak Stage classifies the spread of neurofibrillary tangles (NFTs) tau-containing protein deposits, a hallmark of AD - in the brain across six stages [16]. Early stages (I and II) indicate NFTs in transentorhinal regions, while intermediate stages (III and IV) involve limbic regions, and later stages (V and VI) show NFTs spread throughout cortical areas. Higher stages reflect more extensive disease progression and broader NFT distribution in the brain [16]. The CERAD score evaluates the density of neuritic plaques, mainly composed of beta-amyloid, and categorizes them into four levels: none, sparse, moderate, and frequent [17]. These plaques are a primary indicator of AD, with higher CERAD scores signifying a greater accumulation of plaques, reflecting more advanced amyloid pathology [17]. The ABC score combines data from the Braak and CERAD scores with the Thal phase, which describes the spread of amyloid plaques in the brain. Thal staging ranges from phase 1 (amyloid in subcortical areas) to phase 5 (widespread distribution across the brain) [18]. The

ABC score synthesizes information on NFT spread and amyloid plaque density into a single assessment of AD pathology, categorizing it into four levels: none, low, intermediate, or high [14]. This score provides a more comprehensive evaluation of AD severity and helps to indicate the overall stage of the disease.

Genome-wide DNA methylation

DNA was extracted from fresh-frozen prefrontal cortex tissues in 161 samples using the QIAGEN GenePure kit. DNAm was assessed using the Illumina Infinium MethylationEPIC BeadChips, processed in batches of 167 prefrontal cortex samples, which included six replicates. The raw intensity files were converted into a dataset containing beta values for each CpG site. These beta values were calculated as the ratio of the methylated signal to the total signal (methylated plus unmethylated), ranging from 0 to 1 on a continuous scale.

Preprocessing and quality control were conducted in R (v4.2.0) [19] using a validated quality control and normalization pipeline as previously described [20]. Out of the initial samples, 159 passed the quality control checks. After excluding single nucleotide polymorphism (SNP) probes, XY chromosome probes, and other low-quality probes, 789,286 CpG sites were retained for further analysis.

The final DNAm beta values were normalized to minimize probe-type differences and adjusted using ComBat to account for batch effects prior to downstream analyses [21]. Cell-type proportions (neuronal vs. non-neuronal cells) for each sample were estimated using the latest prefrontal cortex reference database and the R package *minfi* [22–29].

High-resolution metabolomics

High-resolution metabolomic profiling of prefrontal cortex tissue was conducted using liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS) following established protocols [30–33]. Each sample was analyzed in triplicate with two complimentary chromatographic methods: hydrophilic interaction liquid chromatography (HILIC) for polar metabolites and reverse-phase chromatography (C18) for less polar compounds to enhance the coverage of feature extraction. Detected signals were characterized by accurate mass-to-charge ratio (m/z), retention time (RT), and ion intensity [34].

Raw data were converted to .mzML format and processed with apLCMS and xMSanalyzer for peak detection, alignment, feature quantification, batch correction, and quality filtering [35,36]. Metabolic features were further screened before being included in the final analysis based on strict criteria: detected in >15% of samples, median CV <30%, and Pearson correlation $\rho > 0.7$ among technical replicates. Data

were averaged across replicates with non-zero intensities and log2-transformed for statistical analyses. Finally, we included 20,051 features at the HILIC and 15,297 features at the C18, a total of 35,348 features.

Covariates assessment

All models were adjusted for *a priori* selected covariates based on the literature, which include demographic and socioeconomic factors. Individual-level characteristics included sex, race (Black vs. White), educational attainment (high school degree or less, college degree, graduate degree) and age at death. The Area Deprivation Index (ADI) served as a proxy for neighborhood socioeconomic status, based on the 2015 Census Block Group data [37], indicating socioeconomic disadvantages in income, education, employment, and housing. The postmortem interval (PMI) refers to the time elapsed between a person's death and the collection of biological samples, such as tissue or blood. PMI is crucial in research involving post-mortem samples because it can impact on the quality and stability of molecular markers, including DNAm and metabolomics.

Statistical analysis

Following a similar analysis pipeline as established for polygenic risk scores [38]), we developed single- and multi-omics PS based on metabolomics ($PS_{Metabolome}$) and DNAm (PS_{DNAm}) data, to predict AD neuropathology. A PS represents a weighted sum of selected variables:

$$PS = \sum_{k=1}^{K} \beta_k m_k \,, \tag{1}$$

where K is the number of selected variables (e.g., metabolites and/or CpG sites), β_k the weight assigned to the k-th feature and m_k is the actual value of the k-th feature.

The analytic pipeline is outlined in **Figure 1** and includes the following steps that are described in the following section: Stage 0 – DNAm and/or Metabolomic dataset linking and cleaning; Stage 1 – Split data randomly into training and test datasets; Stage 2 – Estimate the weights in the training dataset using different regression and machine learning methods (PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach); Stage 3 - Calculate the PS in the data dataset; Stage 4 - Validate the single- and multi-omics PS in the test data; Stage 5 - Conduct pathway analyses on identified CpG sites and/or metabolites.

Stages 0 and 1 – Data preparation and splitting into training and test data

Once the data sets are cleaned and merged (Stage 0), they are randomly split into training and test sets to prevent overfitting by ensuring that the models are tested on independent data. The training set is used for model building and feature selection, while the test set is reserved for evaluating model performance. In our analysis we split the data evenly into training and test data. However, with larger sample sizes, allocating a greater proportion to the training set can enhance the model's predictive accuracy (see [38] for related recommendations for polygenic scores). We repeated this process 10 times with 10 different seeds (10 iterations) to provide a robust evaluation of our model performance. Since some of the methods described in Step 2 do not allow for missing values, we used Random Forest as implemented in the R package missRanger to impute missing values in the metabolome data set [39]. For the multi-omics PS, the DNAm and metabolites beta values were transformed into z-scores to ensure that both omics datasets are on the same scale before calculating one PS for the combined dataset.

Stage 2 - Feature selection and weight calculation in the training data

In the training data, variable selection and weight calculation are performed using five different regression and machine learning approaches: Pruning & Thresholding (PT), Elastic Net (EN), Boosting (BO), Random Forest (RF) and Windows Approach (WA) using cross leverage scores. This step identifies the most informative CpG sites and/or metabolites related to neuropathological outcomes. A detailed description of the methods can be found in the Supplementary Materials.

Pruning and thresholding (PT) are complementary techniques to simplify high-dimensional datasets by reducing redundancy and focusing on the most relevant variables. Pruning removes highly correlated or redundant variables, decreasing multicollinearity and computational demands while retaining predictive accuracy. For the pruning step, we applied the agglomerative (bottom-up) clustering approach [40] using the R packages ClassDiscovery [41], flashClust [42] and cluster [43]. Complete-linkage clustering was used to define the distance between two clusters as the maximum distance between any pair of points, and the Pearson correlation was used as the distance metric. The analyses generate different numbers of clusters with different numbers of CpG sites/metabolomic variables. A single representative is then drawn from each cluster and used for pruning. Thresholding applies to a predefined cutoff (e.g., p-value threshold) to filter out weak associations and retain only significant variables. We used ordinal logistics regression models (adjusted for covariates) to estimate associations between each and the neuropathology markers. Only CpG site and/or variables with a p-value < 0.05 were included in the PS.

Elastic Net (EN) is a regularized regression technique that combines the penalties of Lasso (L1) and Ridge (L2) to select variables and assign weights. To perform the analysis, we used the R package ordinalNet [44].

Boosting (BO) combines several weakly predictive models to improve the prediction accuracy [45]. We used an extended version of the Boosting method proposed by [46] to allow for ordinal outcomes. The approach considers the response variable as an ordered factor and computes thresholds to distinguish between categories. The boosting process iteratively updates coefficients, computes gradients, and adjusts predictions, resulting in a final model with optimized parameters and scaling corrections.

The Random Forest (RF) algorithm operates by creating multiple decision trees using bootstrap sampling and feature subset selection. The final prediction is then aggregated across all trees, using majority voting for classification or averaging for regression, ensuring robust and accurate predictions. We obtained [47]variable importance measures (VIMs) within the RF fitting process using the R package ranger [47] [48]. This method performs an automatic selection of variables and the VIMs were used as weights to calculate the PS.

The sliding windows approach (WA) involves analyzing variables within specific "windows" or genomic regions. For each window, cross leverage scores are calculated for each variable based on matrix decomposition techniques. These scores quantify the association of the variables, their interactions and the outcome. We select those *q* variables with the absolute highest cross leverage score. This method is described in more detail in [49] with an R code in the corresponding in Supplementary Materials.

Stage 3 - Profile score calculation and prediction of AD neuropathology in the test data

With the selected variables from Stage 2, PS are calculated in the test dataset.

Three different PS were calculated (after equation (1)): Single-omics PS (PS_{DNAm}; $PS_{Metabolome}$) that only contain information from one omics layer (either DNAm or metabolomics), multi-omics PS that contain both the DNAm and metabolomics data ($PS_{DNAm+Metabolome}$), and joint PS models that contain the individual single-omics PS (PS_{DNAm}; $PS_{Metabolome}$) in the same prediction model with and without an interaction term between the single-omics PS (PS_{DNAm} * $PS_{Metabolome}$). In our analysis, the outcome is ordinal, and all three models are ordinal logistic regression models [50,51] with the following equations:

Single-omics PS: $$Neurop.\ outcome \sim PS_{DNAm}\ (or\ PS_{Metabolome}) + Covariates \quad (2)$$

Multi-omics PS: $$Neurop.\ outcome \sim PS_{DNAm+Metabolome} + Covariates \quad (3)$$

Joint PS model: $$Neurop.\ outcome \sim PS_{DNAm} + PS_{Metabolome}\ (+PS_{DNAm} * PS_{Metabolome}) + Covariates \quad (4)$$

Stage 4 - Validation of PS models in the test data

The performance of the single-, multi-omics and joint PS models was evaluated in the test data based on a partial McFadden's R^2, also known as the partial pseudo-R^2, which is a measure of goodness of fit for logistic regression models, including ordinal logistic regression (can be found in the Supplementary Materials). Calculating partial R^2 allowed us to demonstrate the prediction R^2 for the PS, independent of the influence of the other covariates (sex, race, educational attainment, age at death, ADI, PMI). We further evaluated whether the PS was significantly associated with the neuropathology outcomes in the independent test data using a likelihood ratio test (p-value<0.05) in the ordinal logistic regression models from Step 3.

Stage 5 - Pathway analysis of selected metabolites and/or CpG sites

The final step involves a pathway analysis of the identified omics variables (CpG site and/or metabolomics features). Pathway analysis maps selected CpG sites and/or metabolomics features to known biological pathways, elucidating the biological mechanisms potentially underlying observed associations with neuropathology. Only the best performing PS were used for the pathway analyses.

For the DNAm PS, we conducted gene set enrichment analyses using the R package missMethyl [24,52–54] and the KEGG database [55–57].

We included the features/CpG sites that were selected in at least one, two, or three of the 10 iterations in the pathway enrichment analysis.

For the metabolomics PS, datasets for positive and negative ion modes were merged with experimental results to match mass-to-charge ratios (m/z) and retention times. Next, we used the R package metapone [58] to identify pathways from the KEGG database associated with the detected metabolites by leveraging adduct information and permutation-based statistical thresholds.

The top pathways were ranked based on p-values, and their significance was visualized with scatter plots showing the strength and size of CpG sites or metabolomic features contributing to each pathway.

Sensitivity Analyses

While the main analyses were conducted for the largest sample possible (N=154 for PS_{DNAm}, N=141 for $PS_{Metabolome}$, N=138 for the multi-omics PS and the joint PS

model), we conducted a sensitivity analysis, in which we restricted the PS_{DNAm} and $PS_{Metabolome}$ to the donors with data on DNAm and metabolomics (N=138), to validate that differences between the different PS models are not due to differences in sample size.

Results

Study population

After excluding 4 brain donors with missing covariate information, a total of 157 samples were included in the current analysis (**Table 1**). 154 of them had DNAm data available, 141 of them had metabolomics data available and 138 of them had both DNAm and metabolomics data available.

The mean age at death was 76.4 years (standard deviation [SD]: 10.0). Most participants were White (89.2%) and 10.8% self-identified as Black or African American. The study population consisted of 54.8% males and 45.2% females. The study population was predominantly well-educated with 49.7% holding a college degree, and 28.0% a graduate degree. The mean ADI score was 36.1, with a standard deviation of 24.0, indicating a wide range in socioeconomic deprivation among the participants. The prevalence of the APOE ε4 allele (56.1% with at least one APOE ε4 allele) was much higher than that in the general population in the United States, which is estimated to be around 25-30% (Huang et al., 2017). Most donors (59.2%) had high levels of AD neuropathologic changes (ABC score of "high"), 47.1% of donors were classified as Braak Stage 6 and 70.1% of donors had frequent neuritic plaques on the CERAD score, indicating a high prevalence of AD neuropathology in this study population.

Single-omics PS

First, we calculated single-omics PS (equation (1)) based on DNAm and metabolomics data (Figure 1). PS_{DNAm} and $PS_{Metabolome}$ were calculated for each of the three neuropathological outcomes (ABC score, Braak stage and CERAD score) using six different approaches (PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach). The results for the ABC score are shown in Figure 2 and the results for the other two outcomes are included in the appendix (Figures S1 and S2). The results from all three outcomes were similar.

For the PS_{DNAm} (Figure 2A), PT reaches the highest median R^2 (0.11) over all other methods and found a significant association between the DNAm PS and the ABC score in all 10 iterations. Compared to the PT approach, the RF approach performed worse in terms of both the median R^2 value of 0.02 and the number of significant associations (4 out of 10). Other noteworthy approaches for the DNAm PS are PT+EN and BO. The median R^2 was similar to the RF approach (PT+EN: 0.04; BO: 0.01) and PT+EN identified a significant association between DNAm PS and ABC

score in 6 out of 10 iterations and the BO approach in 3 out of 10 iterations. WA and WA+EN resulted in a smaller median R^2 and no significant associations with the ABC score.

For the $PS_{Metabolome}$ (Figure 2B), the RF approach performed best, leading to a median R^2 value of 0.04 across the 10 iterations, identifying a significant association between the metabolomics PS and the ABC score in 8 out of 10 iterations. While the PT approach detected a significant association in 7 out of 10 iterations, its median R^2 value was similar to the RF approach (R^2=0.04). The other approaches resulted in fewer significant associations and lower R^2 values.

The results were similar in a sensitivity analysis, in which we restricted the PS_{DNAm} and $PS_{Metabolome}$ to the donors with data on DNAm and metabolomics (N=138; Supplementary Figure S6).

Multi-omics PS

When combining the DNAm and metabolomics data in one dataset to derive a multi-omics PS ($PS_{DNAm+Metabolome}$; equation (3)), the PT approach (R^2=0.15, 10 out of 10 significant associations) and the RF approach (R^2=0.01, 2 out of 10 significant associations) performed the best (Figure 2C). The R^2 values were greater than in the single-omics PS (PT for the DNAm PS (Figure 2A)).

Joint PS models

Next, we included the individual DNAm PS and metabolomics PS in the joint PS model (with and without an interaction term between the two PS; equation (4)) to evaluate whether this results in a higher model prediction performance (Figure 3).

In general, there was only a small improvement in R^2 when combining both PS in comparison to the single-omics scores. The models with an interaction term between the two PS always performed a little better than the models without the interaction term. The best model R^2 (0.15) was reached when combining the DNAm PS based on the PT approach with the metabolomics PS based on the RF approach, which were also identified as the best PS in the individual omics data (Figure 1). R^2 for the model with the interaction term (R^2=0.15) was slightly higher than R^2 for the model without the interaction term between the two PS (R^2=0.13). For most methods (including RF and PT), the multi-omics PS R^2 were higher than the joint PS R^2.

Secondary analyses of the best-performing PS

The two best-performing single-omics PS (PT for DNAm and RF for metabolomics) showed a Pearson correlation of $\rho = 0.25$) (Figure 4). Of note, absolute correlations between the other single-omics PS were lower ranging from 0.00 to 0.19.

For the two best-performing single-omics PS, 100 iterations were carried out to receive a more precise estimate for the prediction R^2 (R^2=0.13 PT for DNAm and R^2 = 0.01 RF for metabolomics) (Supplementary Figure S7). The results reflect the previous results (R^2=0.11 PT for DNAm and R^2 = 0.04 RF for metabolomics) and

most important statements with a relatively large variability. A similar behavior is also observed for the captured variables: the values found from the 10 iterations are in the range of the 100 iterations found (e.g. 1061 (10 iterations) to 1032 (100 iterations) captured variables with RF for metabolomics).

To map CpG sites and/or metabolites that were selected by the best-performing PS (PT for DNAm and RF for metabolomics) to known biological pathways, we conducted a KEGG pathway enrichment analysis for all CpG sites/metabolites that were selected in one, two or three of the ten iterations (Figure 5). The results for all significantly enriched pathways (p-value < 0.05) in at least one of the ten iterations are summarized in Figure S5 and Figure 5 presents the enriched pathway classes which had at least one pathway with significant p-values in more than one iteration. For PS_{DNAm}, 20 KEGG pathways that can be summarized in 13 classes, were significantly enriched (p-value < 0.05) in at least one of the ten iterations (Figure 5A, Figure S5A). Among these, the most prevalent pathway classes were lipid metabolism and digestive system. For $PS_{Metabolome}$, 21 KEGG pathways that can be summarized in 9 classes, were significantly enriched (p-value < 0.05) in at least one of the ten iterations (Figure 5B, Figure S5B). Among these, the most prevalent pathway classes were lipid metabolism and signal transduction. Of note, both of these pathway classes were also identified in the PS_{DNAm} analyses.

Discussion

We performed a comprehensive single- and multi-omics analysis integrating genome-wide DNAm and high-resolution metabolomics data derived from 157 frontal cortex samples, aiming to gain a better understanding of AD neuropathology. We developed single- and multi-omics PS based on DNAm and metabolomics data to predict the neuropathological features of AD independent of age, sex, race and socioeconomic factors, using various machine learning and regression-based approaches. The best-performing PS_{DNAm}, which was calculated using the PT approach, predicted AD neuropathology levels with a median partial R^2 of 0.11 and the best-performing $PS_{Metabolome}$, which was calculated using RF, reached a median R^2 of 0.04. Combining the DNAm and metabolomics data in the same PS model only led to a small improvement in the prediction accuracy (R^2 = 0.15 for the best-performing joint PS model). Interestingly, PS_{DNAm} and $PS_{Metabolome}$ were moderately correlated with a Pearson correlation of 0.25 and the biological pathways lipid metabolism and signal transduction were enriched among the identified CpG sites as well as the identified metabolites, emphasizing the importance of these two AD-related pathways across various omics layers.

Our analysis showed that DNAm-based PS had a better predictive performance for AD neuropathology than metabolomics-based PS. The DNAm PS achieved a median R^2 value of 0.11, while the best metabolomics PS only achieved a median R^2 value of 0.04. These results suggest that DNAm may be a stronger indicator of

neuropathologic changes in AD than metabolomics. Current evidence highlights the distinct yet complementary roles of brain DNAm and metabolomics in understanding neuropathology markers associated with AD. Alterations in DNAm patterns are closely linked to neuroinflammation, oxidative stress, and other pathological processes in AD. For instance, global hypomethylation and gene-specific methylation changes have been identified as significant contributors to disease progression, impacting genes involved in synaptic function and neurodegeneration [8]. On the other hand, brain metabolomics has recently emerged as a powerful tool for identifying metabolic dysfunctions associated with AD. Research indicates that specific metabolic pathways—such as those involving lipid metabolism, energy production, and amino acid metabolism—are significantly altered in the brains of individuals with AD [10,11].

We identified two overlapping KEGG pathway classes between the CpGs identified in the PS_{DNAm} and the metabolites identified in the $PS_{Metabolome}$ that are related to AD. Notably, both analyses identified key pathway classes such as "lipid metabolism" and "signal transduction," which play significant roles in AD pathology. For instance, the alpha linolenic acid and metabolism and linoleic acid metabolism have been previously linked with AD [59–61]. Alpha-linolenic acid is an essential omega-3 fatty acid known for its anti-inflammatory properties and its role in maintaining neuronal health. Studies have shown that an imbalance in fatty acid metabolism, including alpha-linolenic acid, is linked to neurodegenerative diseases such as AD [59,60]. Linoleic acid, on the other hand, is an essential omega-6 fatty acid whose dysregulation has also been associated with AD. It influences inflammatory processes and can lead to the formation of bioactive lipids that are important for neuronal health [61].

In addition to the two shared pathways, CpG sites selected by the PS_{DNAm} were mapped to eight AD-related pathways that were unique for the DNAm data, including the pentose phosphate pathway [62], glycerophospholipid metabolism [63], ether lipids [64], arachidonic acid [65], linoleic acid [66], Ras signaling [67], the complement system [68] and impaired thermogenesis [69]. For the metabolomics data, ten unique pathways were identified with a known link to AD, including arginine metabolism and proline [70], sphingolipids [61], steroid hormones [71], cholesterol [72], C21 steroids [73], insulin signaling [74], adenosine A2A receptors [75], drug metabolism [76] and cytochrome P450 enzymes [77]. These unique pathways emphasize that different biological signatures of AD are detected by different omics layers.

Among the statistical methods used, there were clear differences in the performance for different omics data. For the DNAm data, PT was the best-performing method based on the R^2. For the metabolomics data, RF produced the highest R^2. While extensive simulation studies are needed to explain the observed differences in performance for different omics data, it could be due to the different number of variables in DNAm (789,286 CpG sites) versus metabolomics (35,348 variables) data or differences in the distribution, correlation structures, or underlying interactions

between variables. In comparison, EN and WA led to a substantially lower R^2 for both omics layers and EN also clearly underperformed in terms of computational time. BO showed a similar performance as EN, but future studies need to determine whether an improved hyperparameter optimization could increase the performance of this method.

In our study, we developed novel brain tissue–based multi-omics profile scores for AD neuropathology that integrate genome-wide DNAm and high-resolution metabolomics data, revealing that combining these omics layers modestly improves predictive accuracy compared to using either layer alone. This finding aligns with a growing body of literature exploring multi-modal prediction models for AD-related outcomes. Wang et al. [78] used unsupervised machine learning to develop an AD risk score that integrates cerebrospinal fluid, MRI, while Cary et al. [79]further advanced the field by integrating genetic, transcriptomic, and proteomic data to map AD risk onto core biological domains such as synapse function, immune response, and lipid metabolism. In a study by Liu et al. [80], comprehensive genetic prediction models were developed to compute risk scores for blood metabolites, which were subsequently employed in association analyses with Alzheimer's disease risk. Together, these studies emphasize that several omics layers are associated with AD-related outcomes, and they provide a valuable context for our results which show that the integration of these omics layers can offer deeper insights into the molecular underpinnings of AD.

There are several strengths of our study to be noted. The unique dataset includes both well-characterized DNAm and metabolomics data from 138 brain donors, allowing a detailed examination of brain-based multi-omics profiles related to AD neuropathology. Our study is characterized by methodological advances that enable a comprehensive analysis of DNAm and metabolomics. While previous studies often analyzed isolated omics data (e.g., only DNA methylation or only metabolomics), this analysis combines both types of data to obtain a holistic picture of biological processes. This opens new perspectives for understanding the interactions between epigenetic changes and metabolic dysfunctions. The calculation of PS from both datasets enables a differentiated view of the influence of various factors on AD neuropathology. Our PS are based on sophisticated machine learning techniques, such as RF, EN, and BO methods, which allow for robust feature selection even in high-dimensional datasets. By employing these innovative methodologies, our analysis not only identified CpG sites and/or metabolomic features predictive of AD neuropathology levels but also elucidated mechanistic underpinnings by evaluating enriched biological pathways among selected variables. Overall, these methodological improvements facilitate a deeper understanding of the intricate relationships between epigenetic modifications and metabolic changes in AD pathology.

In addition to its strengths, our study has some limitations that should be considered. One notable aspect is that the ADRC brain bank is enriched with AD patients and

other dementias, which makes the brain bank a convenience sample rather than a population-based one. This concentration of AD cases may reduce variability in neuropathology markers within the sample. Another consideration is the relatively small sample size when splitting our data into training and test sets to prevent overfitting. However, it is important to highlight that few studies have access to such a large autopsy sample, which is crucial for accurately measuring neuropathology markers as well as brain tissue-based DNAm and metabolomics data—this represents a significant strength of our work. While integrating DNAm and metabolomics has improved predictive power, further exploration is needed to understand how these epigenetic and metabolic changes interact and jointly influence AD pathology. The use of postmortem tissue samples also introduces some limitations; they may not fully capture dynamic metabolic and epigenetic changes throughout disease progression, potentially overlooking important temporal variations in biomarkers. Additionally, applying machine learning methods such as Pruning & Thresholding, Random Forests, Elastic Net, Boosting, and Sliding Windows can present challenges related to hyperparameter selection and optimization. While a general optimization of the hyperparameter was performed and used in each model, future work should evaluate whether these factors can be further optimized to improve the model performance and the robustness of analyses.

Future research directions should focus on expanding the sample size and diversity to validate the results and ensure their generalizability. Integrating additional levels of omics such as genomics, proteomics, transcriptomics, lipidomics, and microbiomics could provide a more comprehensive view of the pathophysiology of AD. Exploring interactions between these different layers could help to identify novel biomarkers and signaling pathways. In addition, functional studies are needed to confirm the biological relevance of the identified signaling pathways.

Conclusion

The present research highlights the potential of integrating DNAm and metabolomics data to deepen our understanding of the pathophysiology of AD. Future research should focus on expanding and diversifying study populations and longitudinal designs to translate the multi-omics insights gained into clinical applications.

References

[1] Alzheimer`s Association (2022). 2022 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*. https://doi.org/https://doi.org/10.1002/alz.12638.

[2] Centers for Disease Control and Prevention (2023). .

[3] Nandi, A. *et al.* (2024). Cost of care for Alzheimer's disease and related dementias in the United States: 2016 to 2060. *npj Aging*. https://doi.org/10.1038/s41514-024-00136-6.

[4] Long, J.M. and Holtzman, D.M. (2019). Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. *Cell*. https://doi.org/10.1016/j.cell.2019.09.001.

[5] Selkoe, D.J. (2002). Alzheimer's Disease Is a Synaptic Failure. *Science*. https://doi.org/10.1126/science.1074069.

[6] Mathys, H. *et al.* (2019). Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*. https://doi.org/10.1038/s41586-019-1195-2.

[7] Dai, X. and Shen, L. (2022). Advances and Trends in Omics Technology Development. *Frontiers in Medicine*. https://doi.org/10.3389/fmed.2022.911861.

[8] Wilkins, J.M. and Trushina, E. (2018). Application of metabolomics in Alzheimer's disease. *Frontiers in Neurology*. https://doi.org/10.3389/fneur.2017.00719.

[9] Yuan, Y. *et al.* (2025). Dysregulation of energy metabolism in Alzheimer's disease. *Journal of Neurology*. https://doi.org/10.1007/s00415-024-12800-8.

[10] Cleland, N.R.W. *et al.* (2021). Altered substrate metabolism in neurodegenerative disease: new insights from metabolic imaging. *Journal of Neuroinflammation*. https://doi.org/10.1186/s12974-021-02305-w.

[11] Korczowska-Łącka, I. *et al.* (2023). Selected Biomarkers of Oxidative Stress and Energy Metabolism Disorders in Neurological Diseases. *Molecular Neurobiology*. https://doi.org/10.1007/s12035-023-03329-4.

[12] Zeng, Y. *et al.* (2023). Identification of key lipid metabolism-related genes in Alzheimer's disease. *Lipids in Health and Disease*. https://doi.org/10.1186/s12944-023-01918-9.

[13] Besser, L.M. *et al.* (2018). The revised national Alzheimer's coordinating center's neuropathology form-available data and new analyses. *Journal of Neuropathology and Experimental Neurology*. https://doi.org/10.1093/jnen/nly049.

[14] Montine, T.J. *et al.* (2012). National Institute on aging-Alzheimer's association guidelines for the neuropathologic assessment of Alzheimer's disease: A practical approach. *Acta Neuropathologica*. https://doi.org/10.1007/s00401-011-0910-3.

[15] Deture, M.A. and Dickson, D.W. (2019). The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration*. https://doi.org/10.1186/s13024-019-0333-5.

[16] Braak, H. and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol*.

[17] Mirra, S.S. *et al.* (1991). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). *Neurology*. https://doi.org/10.1212/WNL.41.4.479.

[18] Thal, D.R. *et al.* (2002). Phases of Aβ-deposition in the human brain and its relevance for the development of AD. *Neurology*. https://doi.org/10.1212/WNL.58.12.1791.

[19] R Core Team (2021). R: A Language and Environment for Statistical Computing.

[20] Pett, L. *et al.* (2024). The association between neighborhood deprivation and DNA methylation in an autopsy cohort. *Aging*. https://doi.org/10.18632/aging.205764.

[21] Johnson, W.E. *et al.* (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. https://doi.org/10.1093/biostatistics/kxj037.

[22] Aryee, M.J. *et al.* (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu049.

[23] Guintivano, J. *et al.* (2013). A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*. https://doi.org/10.4161/epi.23924.

[24] Maksimovic, J. *et al.* (2012). SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biology*.

[25] Fortin, J.-P. and Hansen, K.D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*. https://doi.org/10.1186/s13059-015-0741-y.

[26] Fortin, J.-P. *et al.* (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw691.

[27] Fortin, J.-P. *et al.* (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*. https://doi.org/10.1186/s13059-014-0503-2.

[28] Triche, T.J. *et al.* (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt090.

[29] Andrews, S. V. *et al.* (2016). “Gap hunting” to characterize clustered probe signals in Illumina methylation array data. *Epigenetics & Chromatin*. https://doi.org/10.1186/s13072-016-0107-z.

[30] Go, Y.M. *et al.* (2015). Reference Standardization for Mass Spectrometry and High-resolution Metabolomics Applications to Exposome Research. *Toxicological Sciences*. https://doi.org/10.1093/toxsci/kfv198.

[31] Ladva, C.N. *et al.* (2018). Particulate metal exposures induce plasma metabolome changes in a commuter panel study. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0203468.

[32] Liang, D. *et al.* (2019). Perturbations of the arginine metabolome following exposures to traffic-related air pollution in a panel of commuters with and without asthma. *Environment International*. https://doi.org/10.1016/j.envint.2019.04.003.

[33] Liang, D. *et al.* (2018). Use of high-resolution metabolomics for the identification of metabolic signals associated with traffic-related air pollution. *Environment International*. https://doi.org/10.1016/j.envint.2018.07.044.

[34] Ribbenstedt, A. *et al.* (2018). Development, characterization and comparisons of targeted and non-targeted metabolomics methods. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0207082.

[35] Uppal, K. *et al.* (2013). xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-14-15.

[36] Yu, T. *et al.* (2009). apLCMS—adaptive processing of high-resolution LC/MS data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp291.

[37] Kind, A.J.H. and Buckingham, W.R. (2018). Making Neighborhood-Disadvantage Metrics Accessible — The Neighborhood Atlas. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMp1802313.

[38] Choi, S.W. *et al.* (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*. https://doi.org/10.1038/s41596-020-0353-1.

[39] Mayer, M. (2024). missRanger: Fast Imputation of Missing Values.

[40] Everitt, B.S. *et al.* (2011). Cluster Analysis 5th Edition Cluster Analysis 5th Edition WILEY SERIES IN PROBABILITY AND STATISTICS Cluster Analysis 5th Edition.

[41] Coombes, K.R. (2024). ClassDiscovery: Classes and Methods for “Class Discovery” with Microarrays or Proteomics.

[42] Langfelder, P. and Horvath, S. (2012). Fast *R* Functions for Robust Correlations and Hierarchical Clustering. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v046.i11.

[43] Maechler, M. *et al.* (2023). cluster: Cluster Analysis Basics and Extensions.

[44] Wurm, M.J. *et al.* (2021). Regularized Ordinal Regression and the ordinalNet R Package. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v099.i06.

[45] Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*. https://doi.org/10.1214/07-STS242.

[46] Lau, M. *et al.* (2024). logicDT: a procedure for identifying response-associated interactions between binary predictors. *Machine Learning*. https://doi.org/10.1007/s10994-023-06488-6.

[47] Wright, M.N. and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v077.i01.

[48] Janitza, S. *et al.* (2015). A computationally fast variable importance test for random forests for high-dimensional data.

[49] Teschke, S. *et al.* (2024). Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores. *Biometrical Journal*. https://doi.org/10.1002/bimj.70014.

[50] McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*.

[51] Simon, N. *et al.* (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2012.681250.

[52] Phipson, B. *et al.* (2015). missMethyl: an R package for analysing methylation data from Illuminas HumanMethylation450 platform. *Bioinformatics*.

[53] Maksimovic, J. *et al.* (2015). Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. *Nucleic acids research*.

[54] Phipson, B. and Oshlack, A. (2014). DiffVar: A new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biology*. https://doi.org/https://doi.org/10.1186/s13059-014-0465-4.

[55] Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/28.1.27.

[56] Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. *Protein Science*. https://doi.org/10.1002/pro.3715.

[57] Kanehisa, M. *et al.* (2025). KEGG: biological systems database as a model of the real world. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkae909.

[58] Tian, L. and Yu, T. (2024). metapone: Conducts pathway test of metabolomics data using a weighted permutation test. https://doi.org/10.18129/B9.bioc.metapone.

[59] Grimm, M.O.W. *et al.* (2017). Omega-3 fatty acids, lipids, and apoE lipidation in Alzheimer's disease: a rationale for multi-nutrient dementia prevention. *Journal of Lipid Research*. https://doi.org/10.1194/jlr.R076331.

[60] Penke, B. *et al.* (2018). The Role of Lipids and Membranes in the Pathogenesis of Alzheimer's Disease: A Comprehensive View. *Current Alzheimer Research*. https://doi.org/10.2174/1567205015666180911151716.

[61] He, X. *et al.* (2010). Deregulation of sphingolipid metabolism in Alzheimer's disease. *Neurobiology of Aging*. https://doi.org/10.1016/j.neurobiolaging.2008.05.010.

[62] Butterfield, D.A. *et al.* (2012). Redox Proteomics in Selected Neurodegenerative Disorders: From its Infancy to Future Applications. *Antioxidants & Redox Signaling*. https://doi.org/10.1089/ars.2011.4109.

[63] Yin, F. (2023). Lipid metabolism and Alzheimer's disease: clinical evidence, mechanistic link and therapeutic promise. *The FEBS Journal*. https://doi.org/10.1111/febs.16344.

[64] Fabelo, N. *et al.* (2014). Altered lipid composition in cortical lipid rafts occurs at early stages of sporadic Alzheimer's disease and facilitates APP/BACE1 interactions. *Neurobiology of Aging*. https://doi.org/10.1016/j.neurobiolaging.2014.02.005.

[65] Olivier, J.-L. (2016). Arachidonic acid in Alzheimer's disease. *Journal of Neurology and Neuromedicine*. https://doi.org/10.29245/2572.942X/2016/9.1086.

[66] Grimm, M. *et al.* (2013). Effect of Different Phospholipids on α-Secretase Activity in the Non-Amyloidogenic Pathway of Alzheimer's Disease. *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms14035879.

[67] Kumari, S. *et al.* (2023). Apoptosis in Alzheimer's disease: insight into the signaling pathways and therapeutic avenues. *Apoptosis*. https://doi.org/10.1007/s10495-023-01848-y.

[68] Tenner, A.J. (2020). Complement-Mediated Events in Alzheimer's Disease: Mechanisms and Potential Therapeutic Targets. *The Journal of Immunology*. https://doi.org/10.4049/jimmunol.1901068.

[69] Clarke, J.R. *et al.* (2018). Metabolic Dysfunction in Alzheimer's Disease: From Basic Neurobiology to Clinical Approaches. *Journal of Alzheimer's Disease*. https://doi.org/10.3233/JAD-179911.

[70] Kan, M.J. *et al.* (2015). Arginine Deprivation and Immune Suppression in a Mouse Model of Alzheimer's Disease. *The Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.4668-14.2015.

[71] Pike, C.J. *et al.* (2009). Protective actions of sex steroid hormones in Alzheimer's disease. *Frontiers in Neuroendocrinology*. https://doi.org/10.1016/j.yfrne.2009.04.015.

[72] Chang, T.-Y. *et al.* (2017). Cellular cholesterol homeostasis and Alzheimer's disease. *Journal of Lipid Research*. https://doi.org/10.1194/jlr.R075630.

[73] Vaňková, M. *et al.* (2023). The Role of Steroidomics in the Diagnosis of Alzheimer's Disease and Type 2 Diabetes Mellitus. *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms24108575.

[74] De la Monte, S.M. (2012). Brain Insulin Resistance and Deficiency as Therapeutic Targets in Alzheimers Disease. *Current Alzheimer Research*. https://doi.org/10.2174/156720512799015037.

[75] Cunha, R.A. (2008). Different cellular sources and different roles of adenosine: A1 receptor-mediated inhibition through astrocytic-driven volume transmission and synapse-restricted A2A receptor-mediated facilitation of plasticity. *Neurochemistry International*. https://doi.org/10.1016/j.neuint.2007.06.026.

[76] Mangialasche, F. *et al.* (2010). Alzheimer's disease: clinical trials and drug development. *The Lancet Neurology*. https://doi.org/10.1016/S1474-4422(10)70119-8.

[77] Bahado-Singh, R.O. *et al.* (2023). Alzheimer's Precision Neurology: Epigenetics of Cytochrome P450 Genes in Circulating Cell-Free DNA for Disease Prediction and Mechanism. *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms24032876.

[78] Wang, Z. *et al.* (2020). AD risk score for the early phases of disease based on unsupervised machine learning. *Alzheimer's and Dementia*. https://doi.org/10.1002/alz.12140.

[79] Cary, G.A. *et al.* (2024). Genetic and multi-omic risk assessment of Alzheimer's disease implicates core associated biological domains. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*. https://doi.org/10.1002/trc2.12461.

[80] Liu, S. *et al*. (2024). Identification of blood metabolites associated with risk of Alzheimer's disease by integrating genomics and metabolomics data. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-023-02400-9.

ACKNOWLEDGMENTS

T.T. was supported by the Research Training Group “Biostatistical Methods for High-Dimensional Data in Toxicology” (RTG 2624, Project I1) funded by the German Research Foundation (DFG, Project Number 427806116). This work was supported by the HERCULES Pilot Project via NIEHS P30ES019776 (Huels), the Goizueta Alzheimer’s Disease Research Center: Pilot Grant via NIA P30 AG055611 (Huels/Liang), the Rollins School of Public Health Dean’s Pilot and Innovation Grant (Huels), NIEHS R21ES032117 (Liang), R01ES035738 (Liang), NIA R01AG079170 (Huels/Wingo), U01AG088425 (Huels/Liang/Wingo), and R01AG087250 (Huels/Liang).

CONFLICT OF INTEREST STATEMENT

The authors report no competing interests.

CONSENT STATEMENT

All relevant ethical guidelines have been followed, and any necessary IRB and/or ethics committee approvals have been obtained. Written informed consent was obtained from all participants before inclusion in the study.

KEYWORDS

Alzheimer’s disease, multi-omics, profile scores, DNA methylation, metabolomics, neuropathology, machine learning

Figures and tables

Table 1. Characteristics of the study population.

		Total (N=157)	DNAm (N=154)	Metabolomics (N=141)	Multi-omics analyses (N=138)
Age at Death					
Mean (SD)		76.4 (10.0)	76.4 (10.0)	76.7 (10.2)	76.7 (10.2)
Range		57.0 - 105.0	57.0 - 105.0	57.0 - 105.0	57.0 - 105.0
Race					
White		140 (89.2%)	137 (89.0%)	124 (87.9%)	121 (87.7%)
Black		17 (10.8%)	17 (11.0%)	17 (12.1%)	17 (12.3%)
Sex					
Female		71 (45.2%)	70 (45.5%)	61 (43.3%)	60 (43.5%)
Male		86 (54.8%)	84 (54.5%)	80 (56.7%)	78 (56.5%)
Post Mortal Index (PMI) (in hours)					
Mean (SD)		11.6 (9.6)	11.6 (9.6)	11.5 (9.7)	11.5 (9.7)
Range		1.5 - 64.0	1.5 - 64.0	1.5 - 64.0	1.5 - 64.0
Education (cat.)					
High school or less		35 (22.3%)	35 (22.7%)	29 (20.6%)	29 (21.0%)
College degree		78 (49.7%)	76 (49.4%)	70 (49.6%)	68 (49.3%)
Graduate degree		44 (28.0%)	43 (27.9%)	42 (29.8%)	41 (29.7%)
Area Deprivation Index (ADI)					
Mean (SD)		36.1 (24.0)	36.3 (24.1)	34.4 (23.7)	34.6 (23.9)
Range		1.0 - 94.0	1.0 - 94.0	1.0 - 94.0	1.0 - 94.0
APOE (yes/no)					
E4 absent		69 (43.9%)	69 (44.8%)	60 (42.6%)	60 (43.5%)
E4 present		88 (56.1%)	85 (55.2%)	81 (57.4%)	78 (56.5%)
ABC score					
Not		15 (9.6%)	15 (9.7%)	13 (9.2%)	13 (9.4%)
Low		28 (17.8%)	28 (18.2%)	24 (17.0%)	24 (17.4%)

Intermediate		21 (13.4%)	20 (13.0%)	20 (14.2%)	19 (13.8%)
High		93 (59.2%)	91 (59.1%)	84 (59.6%)	82 (59.4%)
Braak Stage					
Stage 0		0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Stage 1		16 (10.2%)	16 (10.4%)	15 (10.6%)	15 (10.9%)
Stage 2		11 (7.0%)	11 (7.1%)	10 (7.1%)	10 (7.2%)
Stage 3		17 (10.8%)	17 (11.0%)	13 (9.2%)	13 (9.4%)
Stage 4		18 (11.5%)	17 (11.0%)	17 (12.1%)	16 (11.6%)
Stage 5		21 (13.4%)	21 (13.6%)	19 (13.5%)	19 (13.8%)
Stage 6		74 (47.1%)	72 (46.8%)	67 (47.5%)	65 (47.1%)
CERAD score					
No		34 (21.7%)	34 (22.1%)	30 (21.3%)	30 (21.7%)
Sparse		3 (1.9%)	3 (1.9%)	3 (2.1%)	3 (2.2%)
Moderate		10 (6.4%)	10 (6.5%)	9 (6.4%)	9 (6.5%)
Frequent		110 (70.1%)	107 (69.5%)	99 (70.2%)	96 (69.6%)

Donors included in the column “total” had information on all covariates, neuropathology outcomes and either DNAm or metabolomics data available. Of these, 154 donors (listed in the column “DNAm” had DNAm available and 141 donors (listed in the column “metabolomics”) had metabolomics data available. Donors included in the column “multi-omics analyses” had data on DNAm and metabolomics and were included in the multi-omics analyses.

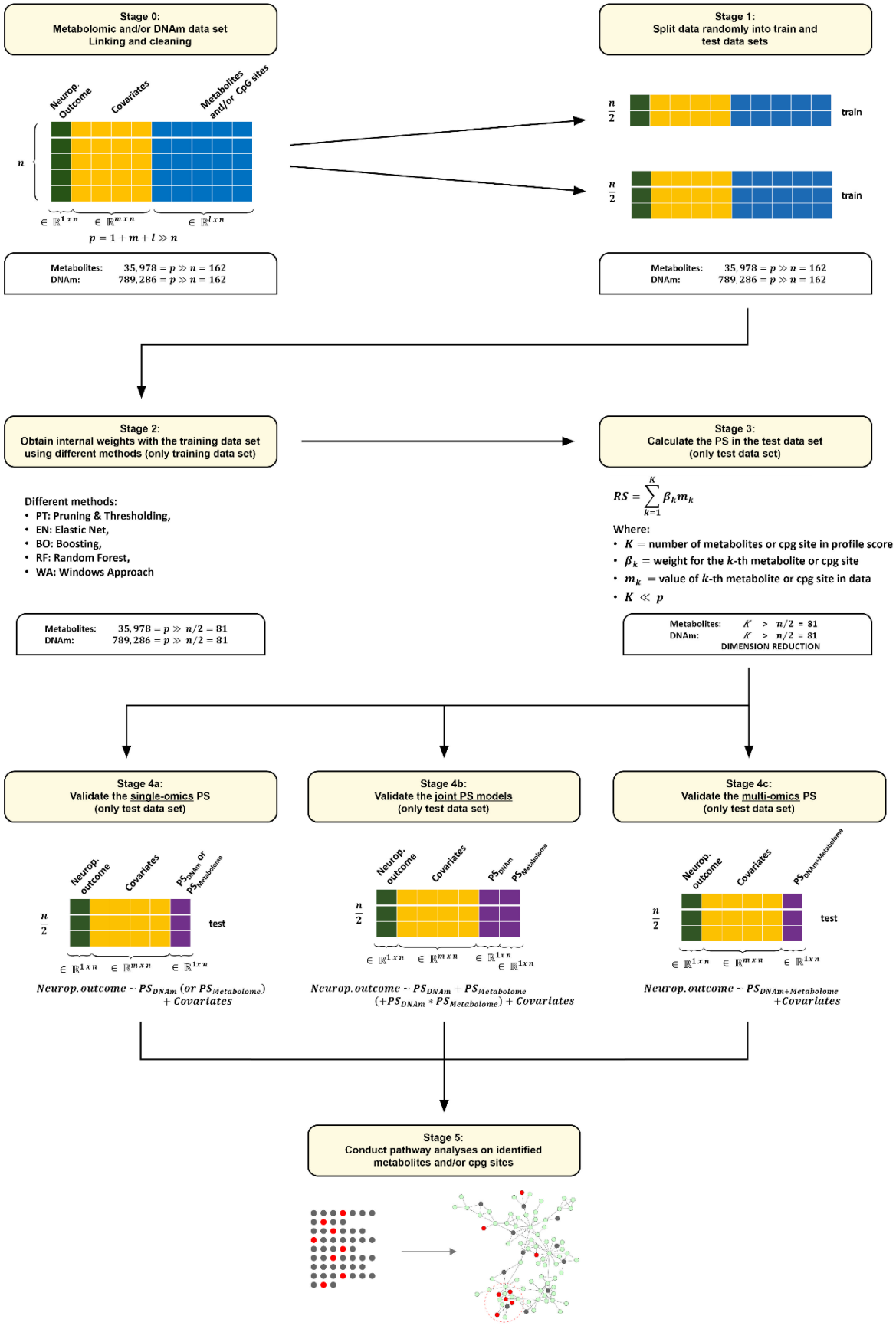
Stage 0:
Metabolomic and/or DNAm data set
Linking and cleaning
Neurop. Outcome
Covariates
Metabolites and/or CpG sites
n
$\in \mathbb{R}^{1 x n}$
$\in \mathbb{R}^{m x n}$
$\in \mathbb{R}^{l x n}$
$p = 1 + m + l \gg n$
Metabolites: $35,978 = p \gg n = 162$
DNAm: $789,286 = p \gg n = 162$
Stage 1:
Split data randomly into train and test data sets
$\frac{n}{2}$
train
$\frac{n}{2}$
train
Metabolites: $35,978 = p \gg n = 162$
DNAm: $789,286 = p \gg n = 162$
Stage 2:
Obtain internal weights with the training data set using different methods (only training data set)
Different methods:
• PT: Pruning & Thresholding,
• EN: Elastic Net,
• BO: Boosting,
• RF: Random Forest,
• WA: Windows Approach
Metabolites: $35,978 = p \gg n/2 = 81$
DNAm: $789,286 = p \gg n/2 = 81$
Stage 3:
Calculate the PS in the test data set
(only test data set)
$RS = \sum_{k=1}^{K} \beta_k m_k$
Where:
• K = number of metabolites or cpg site in profile score
• β_k = weight for the k-th metabolite or cpg site
• m_k = value of k-th metabolite or cpg site in data
• $K \ll p$
Metabolites: $K > n/2 = 81$
DNAm: $K > n/2 = 81$
DIMENSION REDUCTION
Stage 4a:
Validate the single-omics PS
(only test data set)
Neurop. outcome
Covariates
PS_{DNAm} or $PS_{Metabolome}$
$\frac{n}{2}$
test
$\in \mathbb{R}^{1 x n}$
$\in \mathbb{R}^{m x n}$
$\in \mathbb{R}^{1 x n}$
$Neurop.outcome \sim PS_{DNAm}\ (or\ PS_{Metabolome}) + Covariates$
Stage 4b:
Validate the joint PS models
(only test data set)
Neurop. outcome
Covariates
PS_{DNAm}
$PS_{Metabolome}$
$\frac{n}{2}$
$\in \mathbb{R}^{1 x n}$
$\in \mathbb{R}^{m x n}$
$\in \mathbb{R}^{1 x n}$
$\in \mathbb{R}^{1 x n}$
$Neurop.outcome \sim PS_{DNAm} + PS_{Metabolome} (+PS_{DNAm} * PS_{Metabolome}) + Covariates$
Stage 4c:
Validate the multi-omics PS
(only test data set)
Neurop. outcome
Covariates
$PS_{DNAm+Metabolome}$
$\frac{n}{2}$
test
$\in \mathbb{R}^{1 x n}$
$\in \mathbb{R}^{m x n}$
$\in \mathbb{R}^{1 x n}$
$Neurop.outcome \sim PS_{DNAm+Metabolome} + Covariates$
Stage 5:
Conduct pathway analyses on identified metabolites and/or cpg sites

Figure 1. Overview of the statistical procedure for determining the weights and for calculating and validating the individual PS. First, the respective data are prepared and linked to relevant demographic covariates (age at death, race, gender, Post Mortal Index, Education and Area Deprivation Index) and outcomes (neuropathologic scores such as ABC score, Braak Stage and CERAD score). Then the data is split into training and test data. Various methods are used to determine the required internal weights based on the training data set. The following methods are used here: PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach and combinations of these methods. The PS is now determined on the test data by means of the specific weighting of the respective PS. We validated the PS by regression against the outcomes for the single-omics PS (for each data set) and the multi-omics PS (for both data sets in one model optional with an interaction term). Finally, an optional metabolic pathway analysis using the cpg-sites (DNAm) or features (metabolomics) identified by the respective PS can be performed to evaluate relevant biological pathways.

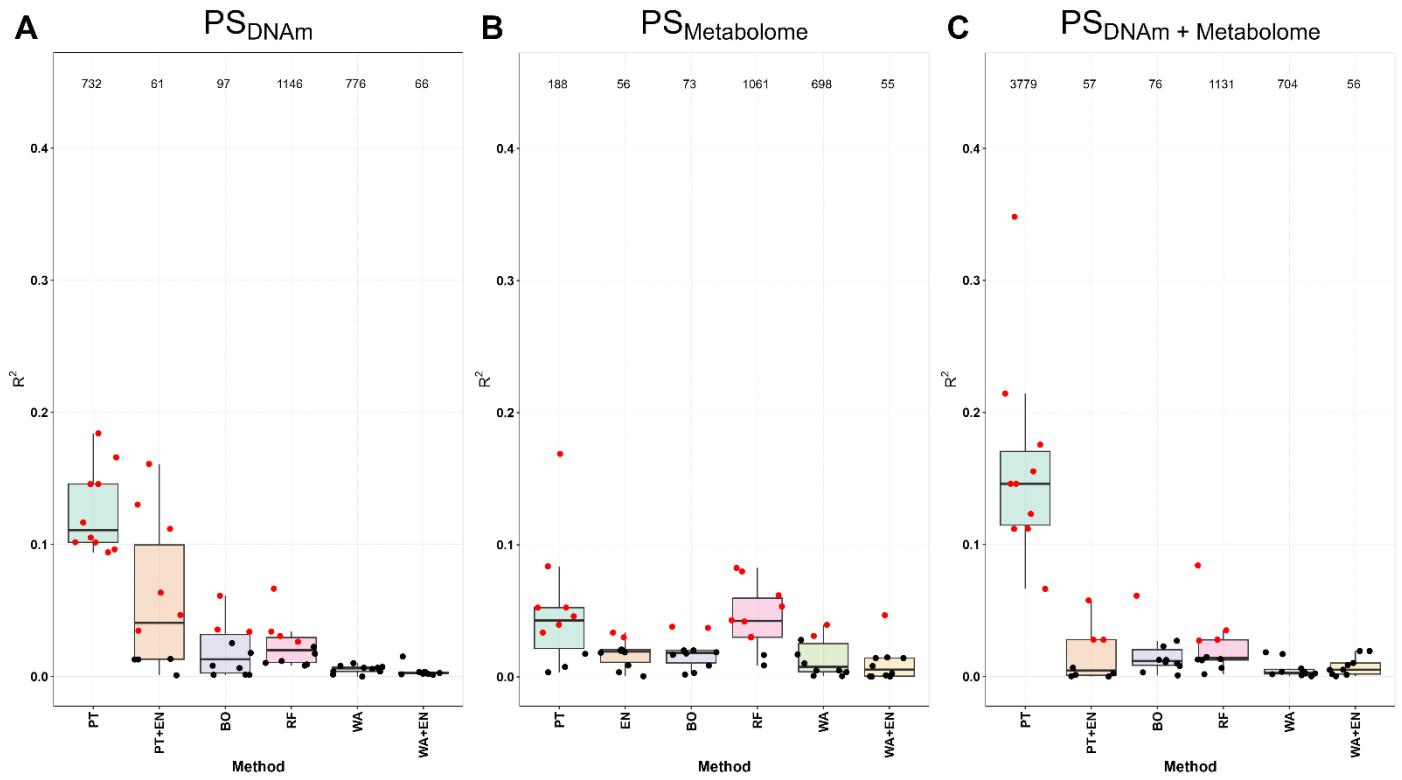

Figure 2. Overview of the results of PS calculation using the various methods and their accuracy of fitting from the individual and combined data sets (McFadden R^2 and p-value) on ABC score outcome. All three sub-graphs are structured in the same way and therefore the explanation can be made on one graph and apply to all three graphs: A) DNAm data set (single-omics PS), B) metabolome data set (single-omics PS) and C) combination of both individual data sets (multi-omics PS). The following 6 methods are shown on the x-axis: PT: Pruning & Thresholding, EN: Elastic Net or EN+PT, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN (due to the high dimension, EN can only be applied to the smaller data set and then PT is used before for dimension reduction). McFadden R^2 is shown on the y-axis, so that for each method a boxplot is shown for the 10 iterations, with the individual results shown as black (not significant) and red (significant) points. At the top are the recorded number of features or cpg-sites included in the PS (mean value across the 10 iterations).

Figure 3. Results from single- and multi-omics PS each individual graph shows the R^2 for the PS of the individual and combined datasets, with the method from the DNAm dataset shown on the x-axis and the metabolome dataset on the y-axis (ABC score). The methods are PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN. McFadden R2 is shown

on the y-axis, so that for each method the median with standard error (SE) is drawn as bar plot with SE bars. The values in a subgraph are from the following data sets or combinations: Joint PS models with interaction term (green), Joint PS models without interaction term (orange) and multi-omics PS from both combined data sets (golden). Since the single PS are at most as good (median) as the joint PS, we have omitted these for better consideration and drawn only the PS with the highest value as a line in the subgraph. If the single PS from DNAm data set is higher than the PS from the metabolome data set, we took the value of the DNAm PS and drew a blue line (vice versa for the single PS from metabolome data set purple). Due to the calculation, the golden boxplots are only present in the subgraphs that use the same methods in both data sets.

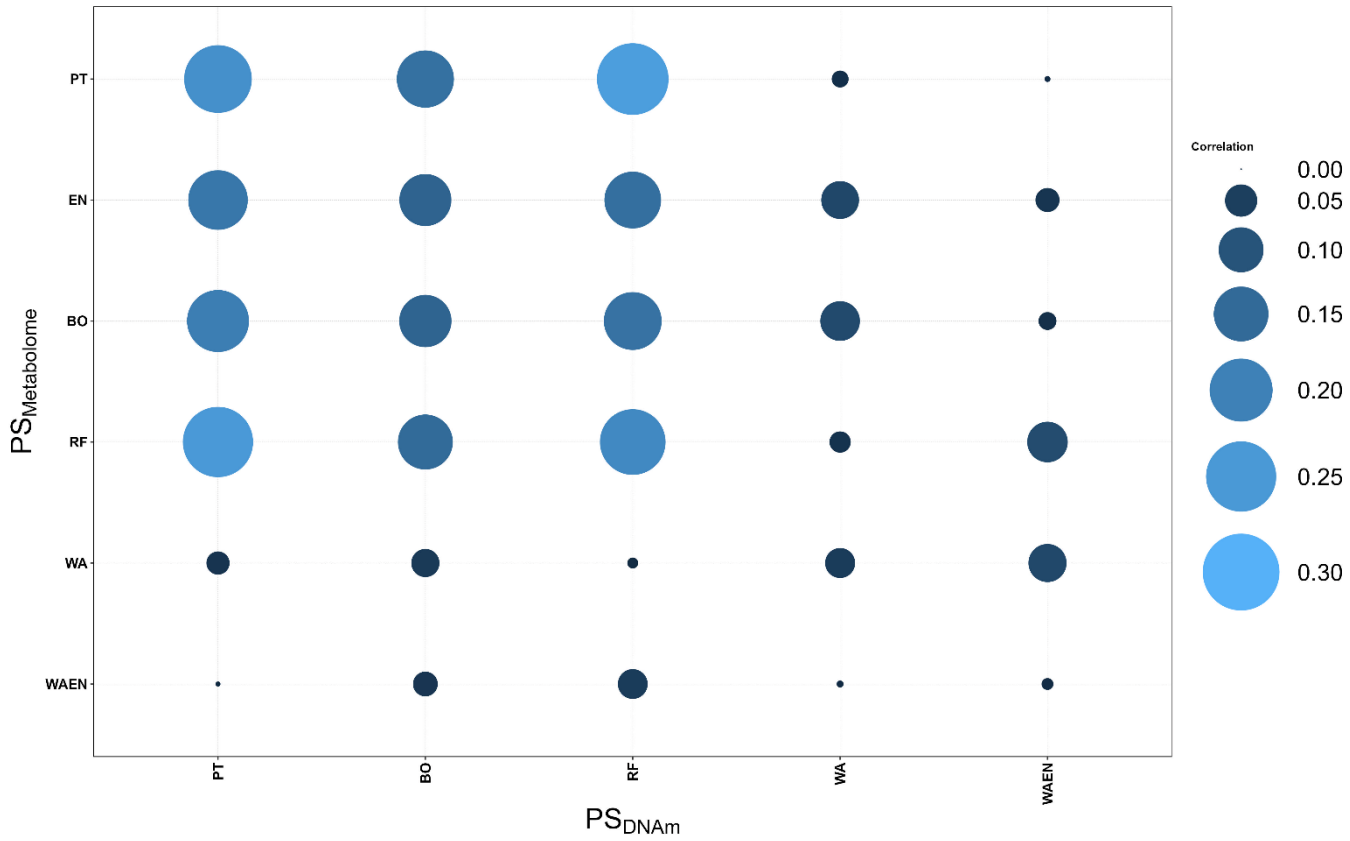

Figure 4. Pearson Correlation structure between the different methods and PS of each data set. On the x-axis are the PS of the different methods based on DNAm data set and on the y-axis the PS of the different methods based on the metabolome data set. The methods are PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN. The color intensity and the size of the dots indicate the respective correlation coefficient according to Bravais-Pearson.

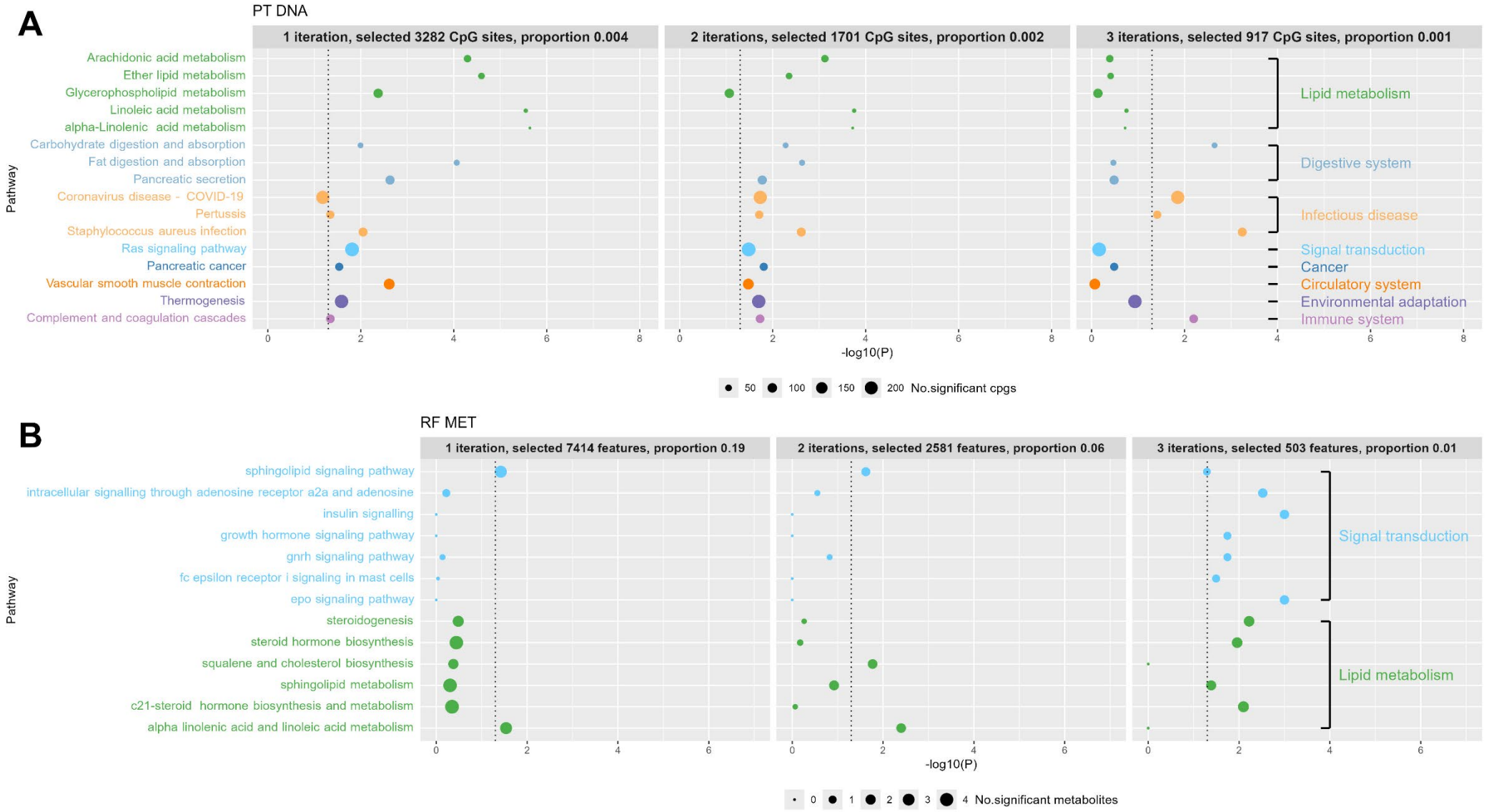
A
PT DNA
1 iteration, selected 3282 CpG sites, proportion 0.004
2 iterations, selected 1701 CpG sites, proportion 0.002
3 iterations, selected 917 CpG sites, proportion 0.001
Pathway
Arachidonic acid metabolism
Ether lipid metabolism
Glycerophospholipid metabolism
Linoleic acid metabolism
alpha-Linolenic acid metabolism
Carbohydrate digestion and absorption
Fat digestion and absorption
Pancreatic secretion
Coronavirus disease - COVID-19
Pertussis
Staphylococcus aureus infection
Ras signaling pathway
Pancreatic cancer
Vascular smooth muscle contraction
Thermogenesis
Complement and coagulation cascades
Lipid metabolism
Digestive system
Infectious disease
Signal transduction
Cancer
Circulatory system
Environmental adaptation
Immune system
-log10(P)
50 100 150 200 No.significant cpgs
B
RF MET
1 iteration, selected 7414 features, proportion 0.19
2 iterations, selected 2581 features, proportion 0.06
3 iterations, selected 503 features, proportion 0.01
Pathway
sphingolipid signaling pathway
intracellular signalling through adenosine receptor a2a and adenosine
insulin signalling
growth hormone signaling pathway
gnrh signaling pathway
fc epsilon receptor i signaling in mast cells
epo signaling pathway
steroidogenesis
steroid hormone biosynthesis
squalene and cholesterol biosynthesis
sphingolipid metabolism
c21-steroid hormone biosynthesis and metabolism
alpha linolenic acid and linoleic acid metabolism
Signal transduction
Lipid metabolism
-log10(P)
0 1 2 3 4 No.significant metabolites

Figure 5. For the DNAm data set (A) with the PT method scatter plots of KEGG pathway enrichment analysis, where at least in (1), (2) or (3) the weighted CpG-sites were found. For the Metabolome data set (B) with the RF method scatter plots of KEGG pathway enrichment analysis, where at least in (1), (2) or (3) iterations the weighted features were found. Only pathways classes, which had at least one pathway with significant p-values in more than one iteration. The number of significant CpG-Sites/ features in the pathway is indicated by the circle area, and the circle color represents the predefined pathway classes. On the x-axis the p-value (as $-\log_{10}(p)$) is represented with the significance level of 0.05 (dotted line). The proportion refers to the number of selected variables divided by the total number of variables (789,286 for DNAm; 35,978 for metabolomics). We display on the y-axis the different pathway terms enriched by KEGG database; the single pathway stands on the y-axis and the classes of them in the last plot. The following pathways classes are not presented in the figure because the corresponding pathway class was only significant in one iteration: A) For DNAm: Carbohydrate metabolism (pentose phosphate pathway), Metabolism of terpenoids and polyketides (terpenoid backbone biosynthesis), Information processing in viruses (virion – flavivirus and alphavirus) and Infection disease: parasitic (Chagas disease); B) For metabolome: Cancer (pathways in cancer), Endocrine system (ovarian steroidogenesis), Infection disease: parasitic (African trypanosomiasis), Biosynthesis of other secondary metabolites (pterine biosynthesis), Digestive system (mineral absorption), Amino acid metabolism (arginine and proline metabolism) and Xenobiotics biodegradation and metabolism (metabolism of xenobiotics by cytochrome p450; drug metabolism – other enzymes)

Supplement Files:

A) Additional figures

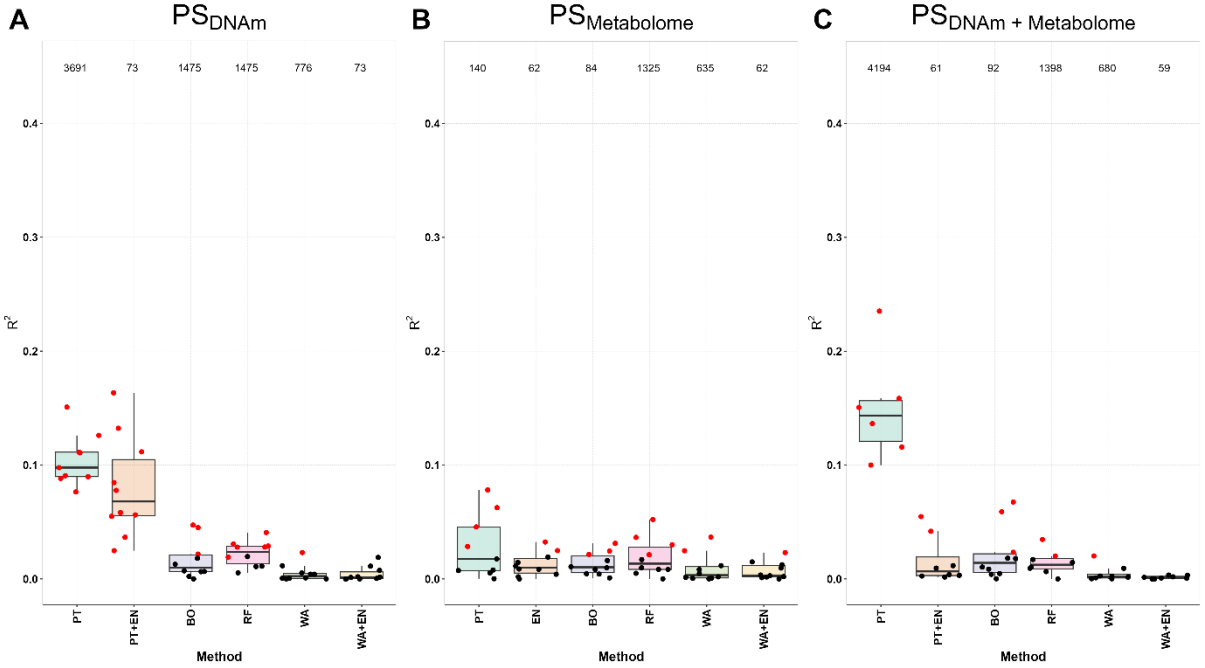

Figure S1. Overview of the results of PS calculation using the various methods and their accuracy of fitting from the individual and combined data sets (McFadden R^2 and p-value) on Braak stage outcome. All three sub-graphs are structured in the same way and therefore the explanation can be made on one graph and apply to all three graphs: A) DNAm data set (single-omics PS), B)) metabolome data set (single-omics PS) and C) combination of both individual data sets (multi-omics PS). The following 6 methods are shown on the x-axis: PT: Pruning & Thresholding, EN: Elastic Net or EN+PT, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN (due to the high dimension, EN can only be applied to the smaller data set and then PT is used before for dimension reduction). Partial McFadden R^2 is shown on the y-axis, so that for each method a boxplot is shown for the 10 iterations, with the individual results shown as black (not

significant) and red (significant) points. At the top are the recorded number of features or cpg-sites included in the PS (mean value across the 10 iterations).

Figure S3. Results from single- and multi-omics PS each individual graph shows the R^2 for the PS of the individual and combined datasets, with the method from the DNAm dataset shown on the x-axis and the metabolome dataset on the y-axis (Braak stage). The methods are PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN. McFadden R2 is shown on the y-axis, so that for each method the median with standard error (SE) is drawn as bar plot with SE bars. The values in a subgraph are from the following data sets or combinations: Joint PS models with interaction term (green), Joint PS models without interaction term (orange) and multi-omics PS from both combined data sets (golden). Since the single PS are at most as good (median) as the joint PS, we have omitted these for better consideration and drawn only the PS with the highest value as a line in the subgraph. If the single PS from

DNAm data set is higher than the PS from the metabolome data set, we took the value of the DNAm PS and drew a blue line (vice versa for the single PS from metabolome data set purple). Due to the calculation, the golden boxplots are only present in the subgraphs that use the same methods in both data sets.

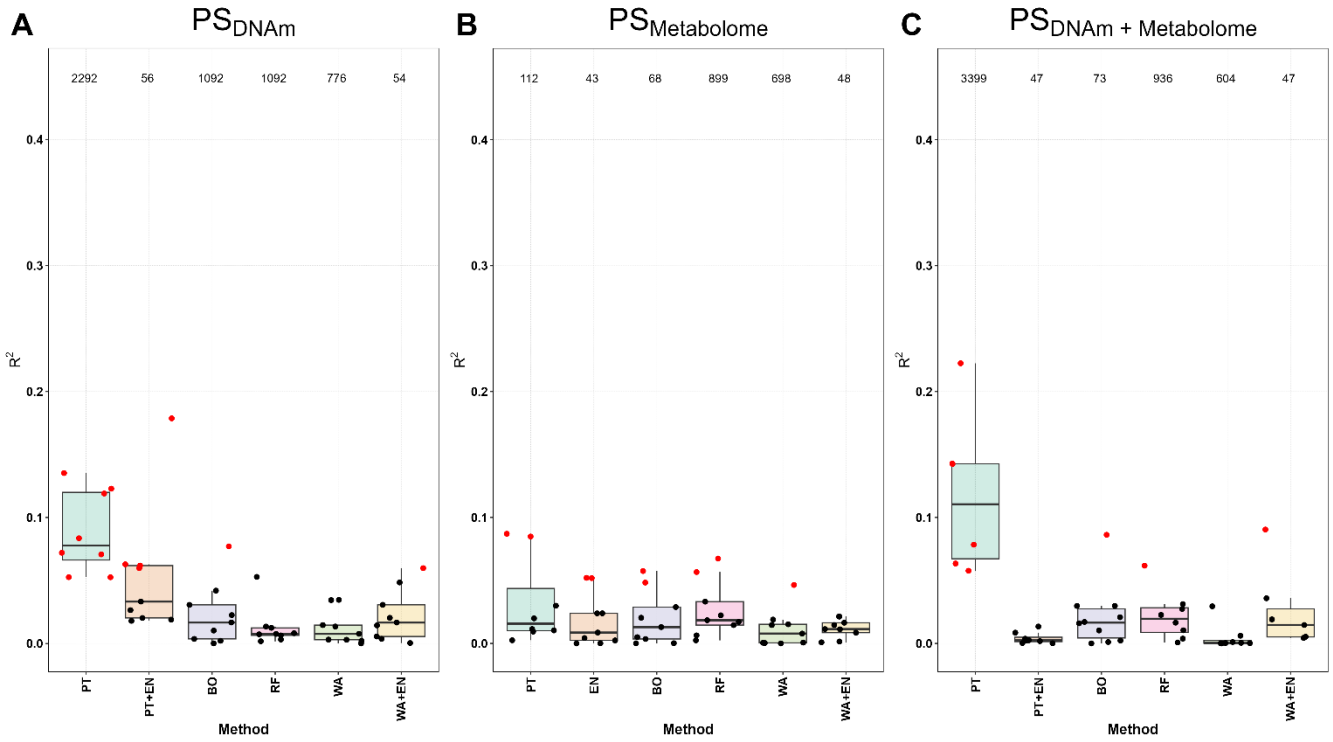

Figure S2. Overview of the results of PS calculation using the various methods and their accuracy of fitting from the individual and combined data sets (McFadden R^2 and p-value) on CERAD outcome. All three sub-graphs are structured in the same way and therefore the explanation can be made on one graph and apply to all three graphs: A) DNAm data set (single-omics PS), B) metabolome data set (single-omics PS) and C) combination of both individual data sets (multi-omics PS). The following 6 methods are shown on the x-axis: PT: Pruning &

Thresholding, EN: Elastic Net or EN+PT, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN (due to the high dimension, EN can only be applied to the smaller data set and then PT is used before for dimension reduction). McFadden R^2 is shown on the y-axis, so that for each method a boxplot is shown for the 10 iterations, with the individual results shown as black (not significant) and red (significant) points. At the top are the recorded number of features or cpg-sites included in the PS (mean value across the 10 iterations).

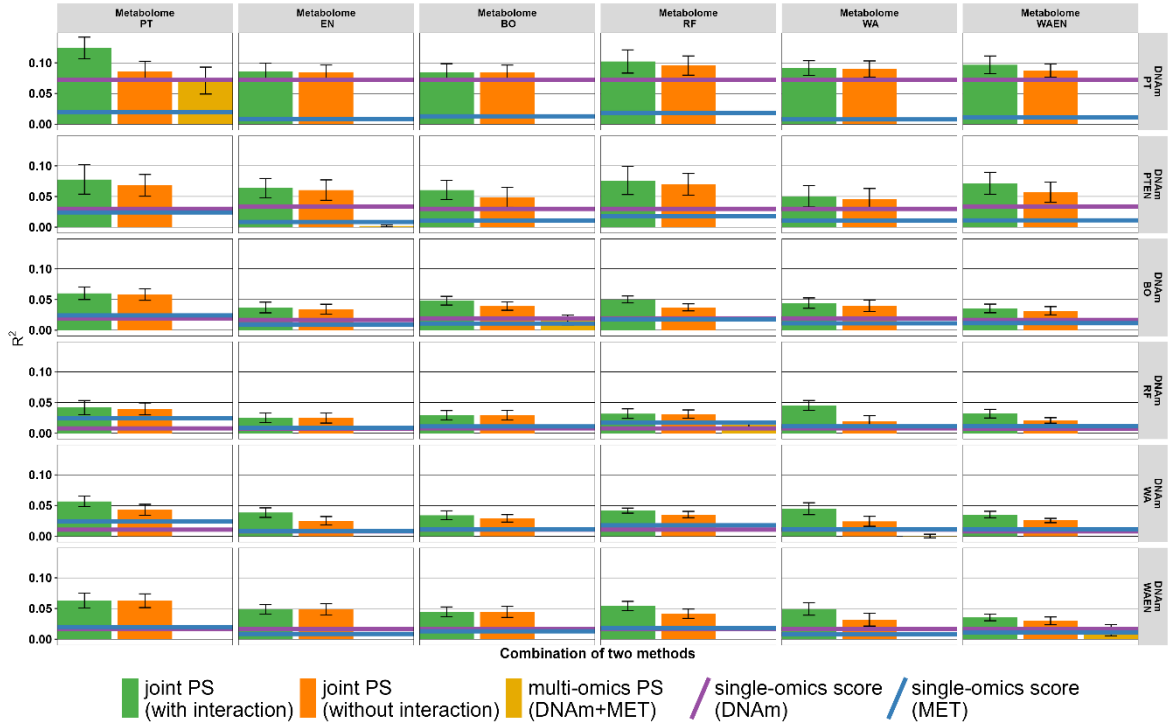

Figure S4. Results from single- and multi-omics PS each individual graph shows the R^2 for the PS of the individual and combined datasets, with the method from the DNAm dataset shown on the x-axis and the metabolome dataset on the y-axis (CERAD score). The methods are PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN. McFadden R2 is shown on the y-axis, so that for each method the median with standard error (SE) is drawn as bar plot with SE bars. The values in a subgraph are from the following data sets or combinations: Joint PS models with interaction term (green), Joint PS models without interaction term

(orange) and multi-omics PS from both combined data sets (golden). Since the single PS are at most as good (median) as the joint PS, we have omitted these for better consideration and drawn only the PS with the highest value as a line in the subgraph. If the single PS from DNAm data set is higher than the PS from the metabolome data set, we took the value of the DNAm PS and drew a blue line (vice versa for the single PS from metabolome data set purple). Due to the calculation, the golden boxplots are only present in the subgraphs that use the same methods in both data sets.

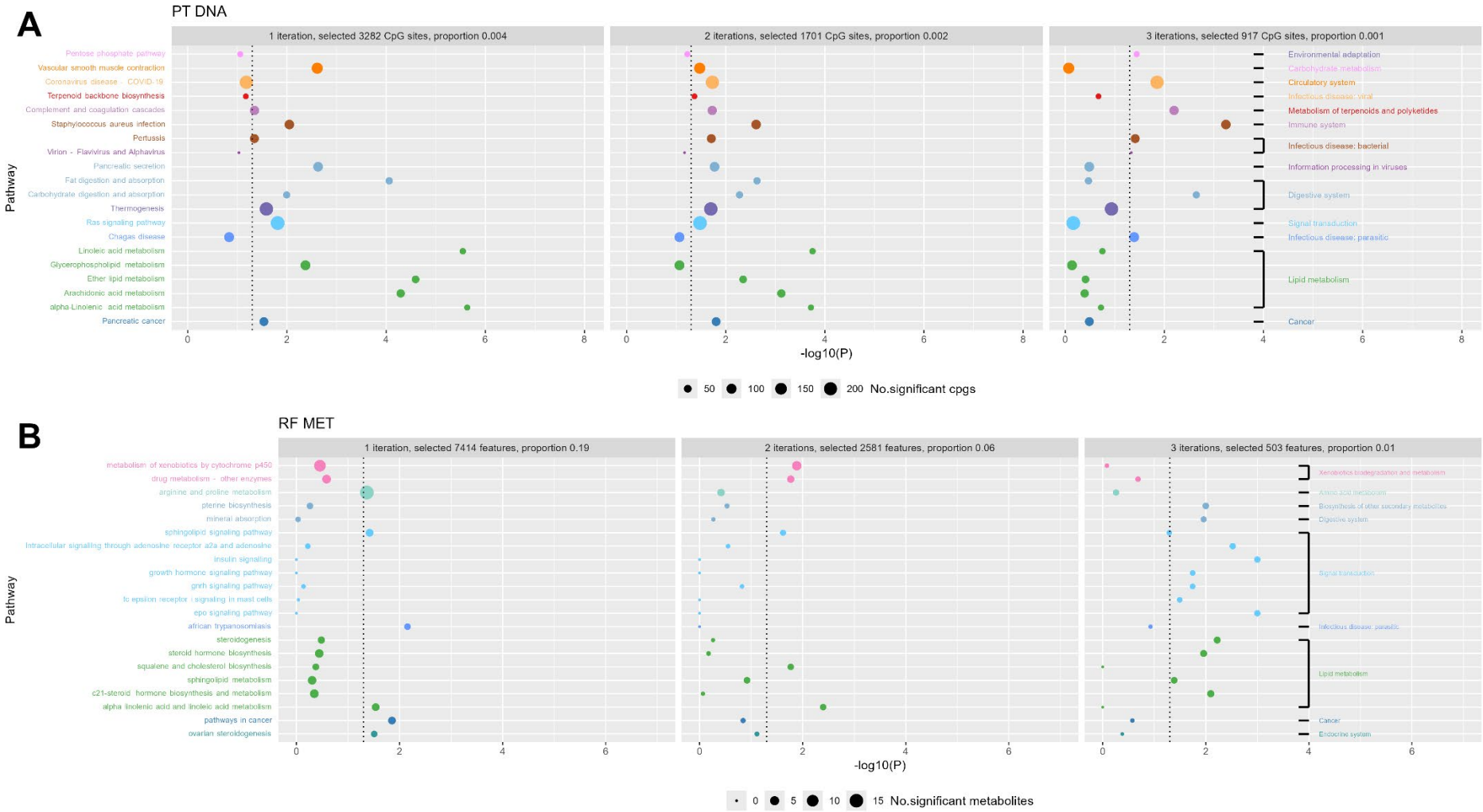

Figure S5. For the DNAm data set (A) with the PT method scatter plots of KEGG pathway enrichment analysis, where at least in (1), (2) or (3) the weighted CpG-sites were found. For the Metabolome data set (B) with the RF method scatter plots of KEGG pathway enrichment analysis, where at least in (1), (2) or (3) iterations the weighted features were found. Only pathways classes, which had at least one pathway

with significant p-values in more than one iteration. The number of significant CpG-Sites/ features in the pathway is indicated by the circle area, and the circle color represents the predefined pathway classes. On the x-axis the p-value (as -log10(p)) is represented with the significance level of 0.05 (dotted line). The proportion refers to the number of selected variables divided by the total number of variables (789,286 for DNAm; 35,348 for metabolomics). We display on the y-axis the different pathway terms enriched by KEGG database; the single pathway stands on the y-axis and the classes of them in the last plot.

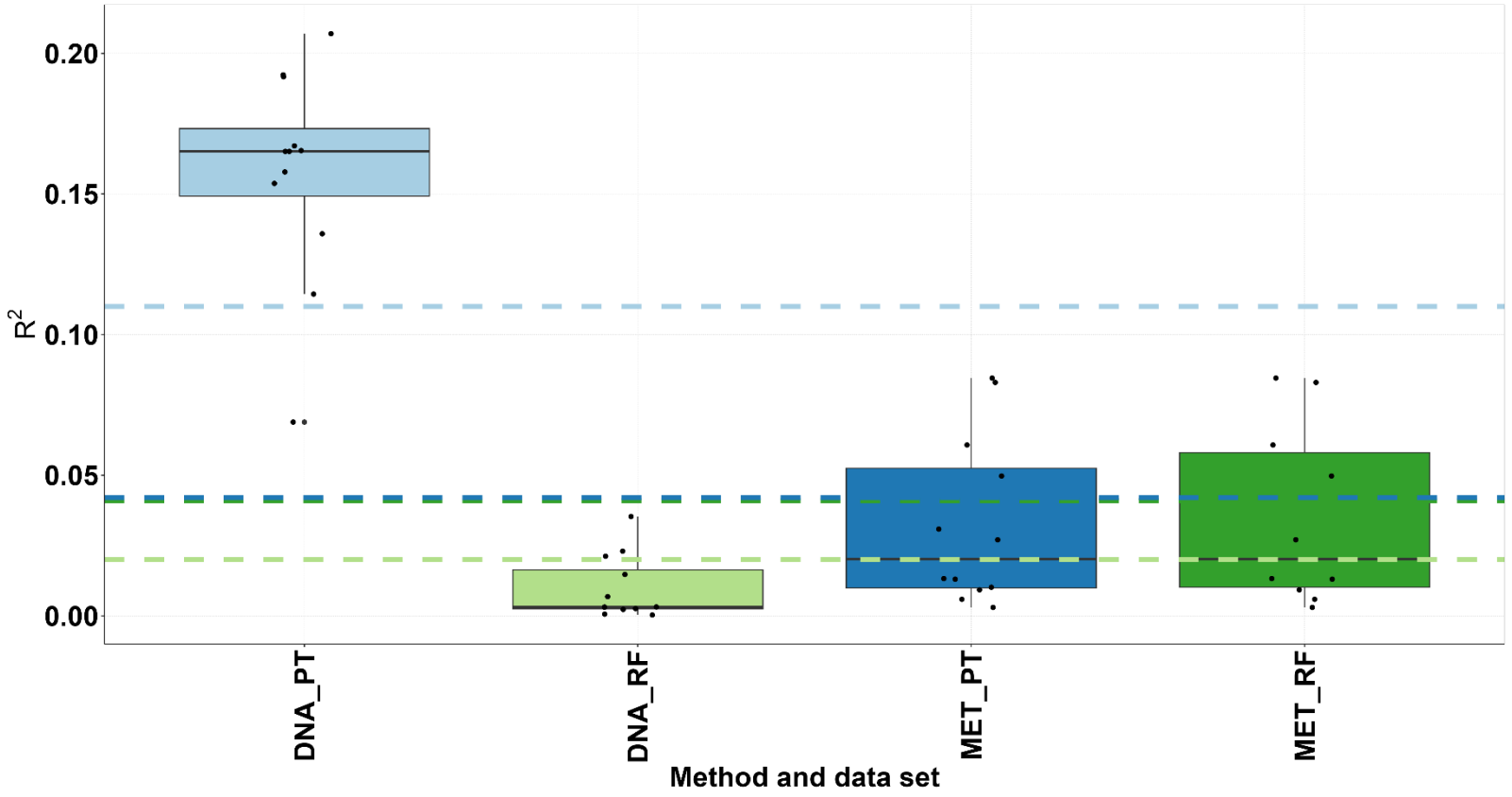

Figure S6. Comparison of the two best methods in the DNAm and metabolome datasets of the R^2 of different sample sizes. In the previous analysis of the single-omics PS, the maximum number of observations of the respective data sets were taken, i.e. DNAm (n=154) and metabolome (n=141) and R^2 determined. Now the joint observations were used as in the joint or multi-omics PS determination of n=138. We ran RF and PT as best methods for both data sets with 10 iterations. The methods and corresponding data sets are plotted on the x-axis and the R^2 on the y-axis. For comparison with the previous analyses, the medians were entered as dashed lines.

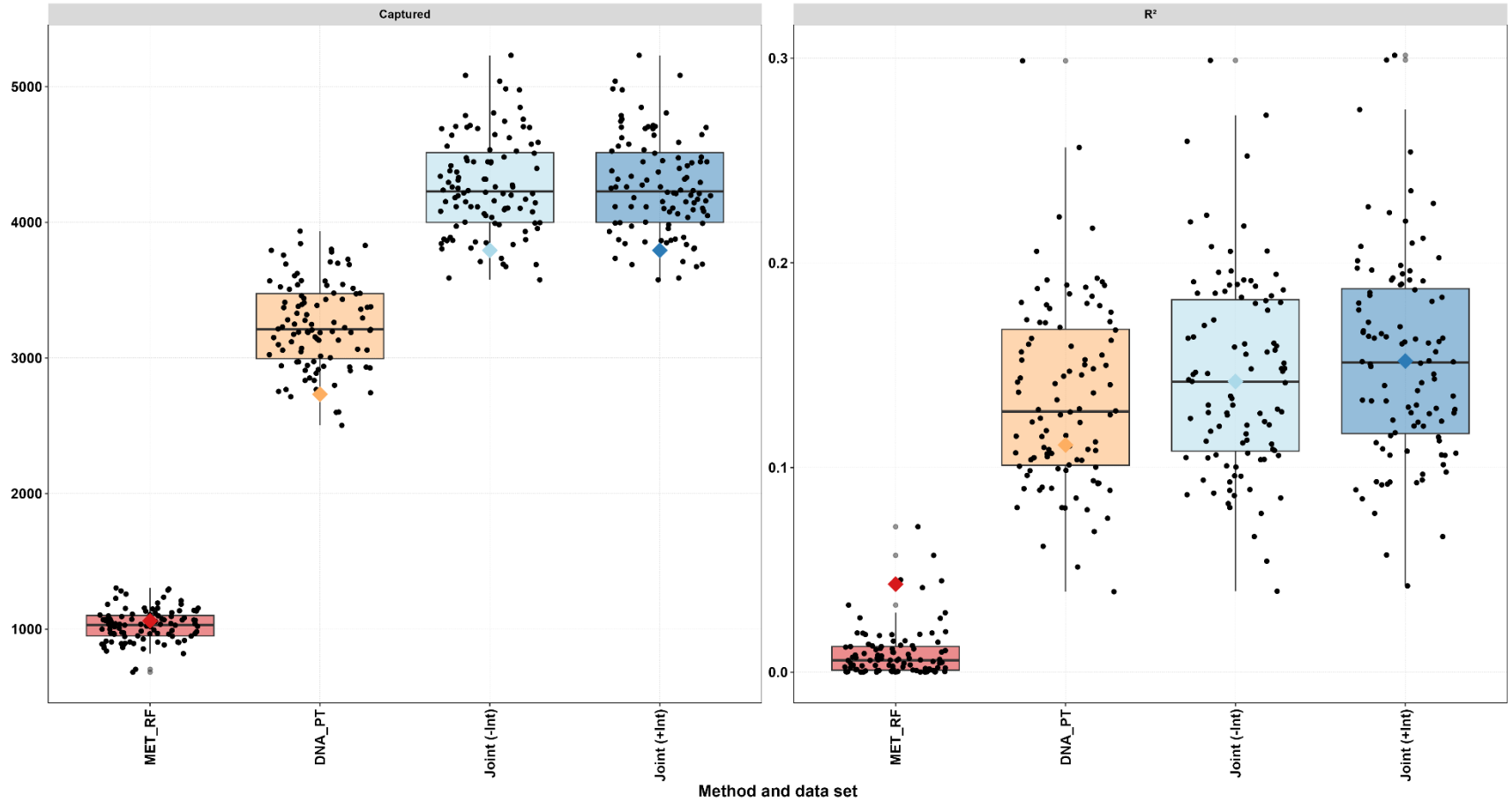

Figure S7. Comparison of the results of the PS for the best methods in the DNAm (PT) and metabolome (RF) datasets and the joint PS with and without the interaction term in 100 iterations. In the left figure we can see the numbers of captured features / CpG sites (y-axis) for the four different PS (x-axis) as a boxplot with 100 jittered points. In the right figure the R^2 for the four methods and datasets is showing for the 100 iterations. The colored star is the mean value of the 10 iterations from the main part of this work for both values (captured features/CpG sites and R^2).

B) Additional tables

Table ST1. Comparison of R^2 values using Median (Q1: 1st, Q3: 3rd quartile and IQR: interquartile range) for the combination of both data sets (DNAm and metabolome data set) in the joint model with and without interaction term (equation (4)).

	additive with interaction (N:10)	additive without interaction (N:10)
DNA_BO vs. MET_BO		
Median (Q1, Q3)	0.008 (0.004, 0.033)	0.004 (0.001, 0.012)
IQR	0.029	0.011
DNA_PT vs. MET_BO		
Median (Q1, Q3)	0.017 (0.007, 0.024)	0.006 (0.001, 0.021)
IQR	0.017	0.019
DNA_PTEN vs. MET_BO		
Median (Q1, Q3)	0.022 (0.015, 0.023)	0.015 (0.014, 0.016)
IQR	0.008	0.002
DNA_RF vs. MET_BO		
Median (Q1, Q3)	0.020 (0.016, 0.028)	0.011 (0.003, 0.017)
IQR	0.012	0.015
DNA_WA vs. MET_BO		
Median (Q1, Q3)	0.019 (0.008, 0.025)	0.006 (0.002, 0.008)
IQR	0.017	0.006
DNA_WAEN vs. MET_BO		
Median (Q1, Q3)	0.009 (0.003, 0.018)	0.003 (0.002, 0.004)
IQR	0.014	0.003
DNA_BO vs. MET_EN		
Median (Q1, Q3)	0.009 (0.004, 0.035)	0.004 (0.001, 0.012)
IQR	0.031	0.011
DNA_PT vs. MET_EN		
Median (Q1, Q3)	0.016 (0.008, 0.025)	0.007 (0.001, 0.018)
IQR	0.017	0.016
DNA_PTEN vs. MET_EN		
Median (Q1, Q3)	0.020 (0.014, 0.025)	0.016 (0.011, 0.017)
IQR	0.011	0.006
DNA_RF vs. MET_EN		
Median (Q1, Q3)	0.021 (0.017, 0.028)	0.011 (0.004, 0.016)
IQR	0.011	0.012
DNA_WA vs. MET_EN		
Median (Q1, Q3)	0.019 (0.008, 0.024)	0.006 (0.002, 0.009)
IQR	0.016	0.006
DNA_WAEN vs. MET_EN		
Median (Q1, Q3)	0.008 (0.003, 0.016)	0.003 (0.001, 0.004)
IQR	0.013	0.003
DNA_BO vs. MET_PT		
Median (Q1, Q3)	0.019 (0.006, 0.020)	0.005 (0.001, 0.013)

	additive with interaction (N:10)	additive without interaction (N:10)
IQR	0.014	0.012
DNA_PT vs. MET_PT		
Median (Q1, Q3)	0.027 (0.005, 0.058)	0.005 (0.001, 0.024)
IQR	0.053	0.023
DNA_PTEN vs. MET_PT		
Median (Q1, Q3)	0.043 (0.012, 0.055)	0.032 (0.010, 0.043)
IQR	0.044	0.033
DNA_RF vs. MET_PT		
Median (Q1, Q3)	0.019 (0.015, 0.026)	0.008 (0.006, 0.012)
IQR	0.012	0.006
DNA_WA vs. MET_PT		
Median (Q1, Q3)	0.014 (0.004, 0.026)	0.003 (0.002, 0.004)
IQR	0.023	0.002
DNA_WAEN vs. MET_PT		
Median (Q1, Q3)	0.004 (0.003, 0.006)	0.003 (0.002, 0.004)
IQR	0.003	0.003
DNA_BO vs. MET_RF		
Median (Q1, Q3)	0.017 (0.008, 0.027)	0.006 (0.002, 0.013)
IQR	0.020	0.010
DNA_PT vs. MET_RF		
Median (Q1, Q3)	0.029 (0.024, 0.032)	0.016 (0.009, 0.030)
IQR	0.009	0.021
DNA_PTEN vs. MET_RF		
Median (Q1, Q3)	0.025 (0.019, 0.030)	0.019 (0.012, 0.025)
IQR	0.011	0.013
DNA_RF vs. MET_RF		
Median (Q1, Q3)	0.022 (0.015, 0.023)	0.010 (0.006, 0.012)
IQR	0.008	0.006
DNA_WA vs. MET_RF		
Median (Q1, Q3)	0.013 (0.008, 0.015)	0.005 (0.001, 0.010)
IQR	0.008	0.010
DNA_WAEN vs. MET_RF		
Median (Q1, Q3)	0.005 (0.002, 0.005)	0.003 (0.001, 0.004)
IQR	0.003	0.004
DNA_BO vs. MET_WA		
Median (Q1, Q3)	0.008 (0.005, 0.014)	0.004 (0.002, 0.011)
IQR	0.009	0.010
DNA_PT vs. MET_WA		
Median (Q1, Q3)	0.015 (0.009, 0.025)	0.009 (0.007, 0.022)
IQR	0.016	0.015
DNA_PTEN vs. MET_WA		
Median (Q1, Q3)	0.010 (0.009, 0.015)	0.006 (0.002, 0.010)
IQR	0.006	0.007
DNA_RF vs. MET_WA		

	additive with interaction (N:10)	additive without interaction (N:10)
Median (Q1, Q3)	0.008 (0.007, 0.021)	0.007 (0.005, 0.016)
IQR	0.014	0.011
DNA_WA vs. MET_WA		
Median (Q1, Q3)	0.010 (0.007, 0.018)	0.003 (0.001, 0.005)
IQR	0.010	0.004
DNA_WAEN vs. MET_WA		
Median (Q1, Q3)	0.015 (0.010, 0.020)	0.003 (0.001, 0.005)
IQR	0.010	0.004

C) Further statistical description

i) Stage 4

Stage 4a (Single-Omics PS): Here, each omics type (e.g., DNAm or metabolomic PS) is associated independently with the outcome variable. Models are adjusted for relevant covariates, such as sex, race, age at death, educational attainment, PMI and ADI, to accurately assess the association. For an ordinal outcome Y with L ordered categories ($Y \in \{1,2, \dots, L\}$), the cumulative probability is modeled as:

$$P(Y_i \leq k) = \Phi(\theta_k - \eta_i), k = 1, \dots, L-1,$$

where $P(Y_i \leq k)$ is the cumulative probability that Y falls in category k or below, Φ cumulative distribution function (e.g., logistic or normal), θ_k are thresholds separating the categories and η_i are linear combination of predictors $\eta_i = \beta_0 + \beta_1 \cdot PS_i + \sum_{j=1}^{p} \gamma_j * X_{ij}$ [1]. The most widely used model for ordinal outcomes is the ordinal logistic regression model, which assumes proportional odds:

$$\log\left(\frac{P(y_i \leq k \mid X_i)}{P(y_i > k \mid X_i)}\right) = \theta_k - \eta_i, \qquad (1)$$

where $\frac{P(Yi>k)}{P(Yi\leq k)}$ are the cumulative odds of being in category k or below, θ_k category-specific threshold $\eta_i = \beta_0 + \beta_1 \cdot PS_i + \sum_{j=1}^{p} \gamma_j * X_{ij}$ [2].

McFadden's R^2, also known as the pseudo-R^2, is a measure of goodness of fit for logistic regression models, including ordinal and multinomial logistic regression. For a given model, McFadden's R^2 is calculated as:

$$R^2 = 1 - \frac{ln\ L(full\ model)}{ln\ L(null\ model)},$$

where $ln\ L(full\ model)$ is the log-likelihood of the fitted model, including all predictors and $ln\ L(null\ model)$ is the log-likelihood of the null model, which includes only the intercept. McFadden's R^2 is typically lower than traditional R^2 because it is based on the improvement in log-likelihood rather than the explained variance, reflecting the model's predictive power for ordinal outcomes rather than continuous variance. A value between 0.2 and 0.4 indicates a meaningful improvement over the null model and is considered a good fit in this context. Unlike traditional R^2, which explains the proportion of variance, McFadden's R^2 should be interpreted as a relative measure of model fit rather than a direct percentage [3]. While the standard McFadden R^2 provides an overall measure of model fit, it does not indicate the unique contribution of specific predictors. This is where the partial McFadden R^2 becomes valuable.

The partial McFadden R^2 quantifies the contribution of a set of predictors by comparing two models: Restricted model that includes all covariates except for the predictors whose effect is being isolated and the full model that includes all covariates, including

the predictors of interest. By evaluating the improvement in log-likelihood when the predictors of interest are added, the partial McFadden R^2 provides insight into their explanatory power. The formula for the partial McFadden R^2 is given by: $R^2_{partial} = \frac{lnL_{restricted} - lnL_{full}}{lnL_{restricted}} = 1 - \frac{\ln L_{full}}{\ln L_{restricted}}$, where $\ln L_{restricted}$ is the log-likelihood of the model that excludes the predictors of interest and $\ln L_{full}$ is the log-likelihood of the model that includes the predictors of interest.

The partial McFadden R^2 can be viewed as analogous to a partial R^2 in linear regression, quantifying the incremental improvement provided by the additional predictors. Since this value is derived from a ratio of log-likelihoods, it reflects the relative explanatory power of the added predictors compared to the restricted model. In practice, researchers and practitioners often use this measure to decide whether the inclusion of extra predictors justifies the added model complexity.

Stage 4b (Joint-Omics PS): This step involves integrating multiple omics Risk Scores into a joint model to test if they collectively improve predictive power Assuming that we have two PS from different omics data the regression equation (see (1)) is extended to $\eta_i = \beta_0 + \beta_1 \cdot PS_{1,i} + \beta_2 \cdot PS_{2,i} + \sum_{j=1}^{p} \gamma_j * X_{ij}$. This is the combined integrated model for both omics data sets. In ordinal outcome models, interaction terms are used to assess whether the effect of one predictor depends on the level of another predictor, resulting in the following equation: $\eta_i = \beta_0 + \beta_1 \cdot PS_{1,i} + \beta_2 \cdot PS_{2,i} + \beta_3 \cdot (PS_{1,i} * PS_{2,i}) + \sum_{j=1}^{p} \gamma_j * X_{ij}$.

Incorporating an interaction term between the two single-omics PS allows for a more nuanced understanding of relationships between predictors and ordinal outcomes. For the goodness of fit for logistic regression models we use like in the single-omics PS the partial McFadden's R^2 [3].

Stage 4c (Multi-Omics PS):

As mentioned above here we combine both data sets (DNAm and metabolome) with some preprocessing steps like z-score transform. With this combined new data set the calculation stages are the same as in Stage 4a for the single-omics PS was described.

ii) Methods

Statistical and machine learning methods for variable selection & estimation of weights

1. Pruning & Thresholding (PT)

Pruning involves systematically reducing the number of features, or CpG-sites in a model by removing those that are highly correlated or redundant. By eliminating these

correlated markers, pruning simplifies the model, reduces multicollinearity, and decreases computational demands without significantly compromising prediction accuracy [4]. Pruning generally involves iteratively removing variants/variables based on their correlations without relying on association significance, focusing primarily on reducing multicollinearity across the entire set of variants. In the present analysis, pruning was integrated into hierarchical clustering. Hierarchical clustering with pruning is a technique that applies pruning principles to hierarchical clustering to reduce redundancy in high-dimensional data. By combining clustering with pruning, it is possible to create a streamlined clustering result that retains the most informative clusters and minimizes noise or redundancies, particularly useful in DNAm, metabolomic, and other "omics" datasets. Hierarchical clustering organizes data points into a tree-like structure (dendrogram) based on their pairwise distances. It begins with each data point as its own cluster and iteratively merges the two closest clusters until a single cluster encompassing all data points is formed. We use one of two main types of hierarchical clustering: the Agglomerative (bottom-up) clustering: Start with each data point as a separate cluster and merge them iteratively. Complete-linkage clustering is a corresponding method for agglomerative clustering where the distance between two clusters is defined as the maximum distance between any pair of points in the two clusters. For the distance metric in hierarchical clustering, we use the Pearson correlation to calculate the distance between the clusters. When applying pruning to hierarchical clustering, the goal is to simplify the dendrogram by removing clusters or data points that add little new information, thus focusing on clusters with distinct, meaningful information. In high-dimensional data, each data point (e.g., CpG sites, metabolites features) has an associated importance score (e.g., association strength, p-value). During hierarchical clustering, prune variables within each cluster that fall below a set importance threshold, reducing redundancy by keeping only the top variables within each cluster.

Thresholding is a technique commonly used in high-dimensional data analysis, particularly in the context of genetic, epigenetic, and metabolomic studies, to filter out weak associations and focus on the most relevant variables. By applying a statistical or p-value-based cutoff, thresholding helps to reduce noise, improve interpretability, and enhance the predictive power of models by including only those variables that meet a predefined level of importance or association [4].

Thresholding can be formally described as follows:

1. Define a Threshold t: The first step is to define a threshold t, which could be a p-value, a correlation coefficient, or any other measure of association or importance.

2. Apply the Threshold: For each feature x_i (e.g., CpG site, metabolite) with an associated score $s(x_i)$, retain x_i if and only if $s(x_i) \geq t$. This score $s(x_i)$ could represent statistical significance (e.g., p-value), effect size, correlation

coefficient, or another measure relevant to the context. Mathematically: x_i is retained if $s(x_i) \geq t$.

3. Binary Masking: Thresholding can be represented by a binary indicator function, where each feature is assigned a value of 1 (included) or 0 (excluded) based on whether it meets the threshold:

4. $I(x_i) = \begin{cases} 1 & if\ s(x_i) \geq t \\ 0 & if\ s(x_i) < t \end{cases}$

This binary masking function $I(x_i)$ can then be used to filter variables in downstream analyses, retaining only those that meet the threshold.

Together, pruning and thresholding provide complementary techniques to manage the vast number of variables, leading to more streamlined and interpretable prediction models [5].

2. Elastic Net (EN)

Elastic Net (regression) is a regularization technique that linearly combines the penalties of Lasso (L1) and Ridge (L2) regression to address some of the limitations of each approach individually. It is particularly useful when dealing with datasets with many predictors or when predictors are highly correlated. Elastic Net aims to retain the feature selection ability of Lasso and the stability provided by Ridge in a single model [6].

The Elastic Net objective function is defined as:

$$\min_{\beta} \left(\frac{1}{2N} \sum_{i=1}^{N} (y_i - X_i\beta)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 \right) \right),$$

where N is the number of observations, p is the number of variables, y_i is the target outcome, X_i is the vector of variables for the i-th observation, β is the vector of coefficients to be estimated, λ is the regularization parameter controlling the overall strength of the penalty and α is the mixing parameter that balances the L1 (Lasso) and L2 (Ridge) penalties.

The Elastic Net penalty term consists of two components: 1. L1 Penalty (Lasso): $\alpha \sum_{j=1}^{p} |\beta_j|$, which encourages sparsity by driving some coefficients to zero. 2. L2 Penalty (Ridge): $\frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2$, which helps handle multicollinearity by shrinking coefficients without necessarily setting them to zero. Both key parameters can be explained in the following manner: The regularization parameter λ controls the extent of regularization. Higher values of λ increase the penalty on the coefficients, reducing model complexity. Mixing Parameter α determines the balance between Lasso and Ridge penalties: For $\alpha = 1$ the model behaves like Lasso, focusing on feature selection. For $\alpha = 0$ the model

behaves like Ridge, focusing on coefficient shrinkage without setting them to zero. And between $0 < \alpha < 1$ the model applies a combination of both L1 and L2 penalties, achieving a trade-off between sparsity and stability.

Elastic Net is advantageous for datasets where predictors are highly correlated because it can select groups of correlated variables together [6]. In contrast, Lasso tends to select only one feature from a group of correlated variables, while Elastic Net encourages joint selection.

When applying Elastic Net regression to an ordinal outcome (i.e., a dependent variable with ordered categories, such as rating scales from "low" to "high"), additional considerations are necessary to account for the ordered nature of the outcome. Standard Elastic Net regression is designed for continuous outcomes, but with ordinal outcomes, ordinal regression models (e.g., ordinal logistic regression) can be adapted with Elastic Net regularization. This approach ensures that the model respects the ordinal structure while applying regularization to manage multicollinearity and improve prediction accuracy.

The Elastic Net regularized ordinal regression objective function becomes:

$$\min_{\beta}\left(-\sum_{i=1}^{N}\log P(y_i\,|X_i\beta) + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2\right)\right),$$

where $P(y_i\,|X_i\beta)$ is the probability of observing the ordinal outcome y_i for the i -th observation, given the predictors X_i and coefficients β. The probability $P(y_i\,|X_i\beta)$ is often modeled using an ordinal logistic (proportional odds) or probit function, respecting the ordinal structure by assuming that outcomes fall within specific ranges of a latent continuous variable.

Example: Proportional Odds Model with Elastic Net Regularization

For ordinal logistic regression, the proportional odds assumption assumes a common slope for each category of the ordinal outcome but different intercepts, defined as:

$$\log\left(\frac{P(y_i \le k\,|X_i)}{P(y_i > k\,|X_i)}\right) = \theta_k - \eta_i,$$

where k indexes the thresholds between ordinal categories, θ_k are the threshold parameters that separate the ordinal categories and $X_i\beta$ is the linear predictor with the Elastic Net penalty applied to β as in the Elastic Net objective function.

This formulation, incorporating Elastic Net regularization, encourages a sparse or grouped solution depending on the values of α and λ and is advantageous when predictors are numerous or correlated [2].

3. Boosting (BO)

Gradient boosting [7]is a state-of-the-art ensemble method for constructing predictive models in an iterative fashion. The idea behind gradient boosting is performing functional gradient descent in the linear span of simple models (such as simple linear regression models or shallow decision trees). More precisely, in each iteration,

1. the gradient of the current loss (between the outcome and the current model predictions) is computed,
2. a simple model is fit to the gradient, and
3. the ensemble model is updated by adding the new model.

Thus, the resulting model $f(\boldsymbol{X}) = \rho_0 + \sum_{b=1}^{B} f_b \rho_b(\boldsymbol{X})$ is a sum of *B* simple models f_b weighted by boosting coefficients ρ_b.

In high-dimensional applications in which only a fraction of predictors influence the outcome, gradient boosting can be particularly efficient for constructing multiple linear regression models, as the computational complexity is given by $O(Bpn)$ (see, e.g., [8]), where *B* is the (pre-specified) number of boosting iterations/maximum number of terms, compared to ordinary linear regression or regularized procedures such as elastic net that exhibit a complexity of $O(p^2 n)$ [9].

Gradient boosting can be also employed for other outcome types that belong to the exponential family [10]. In this paper, an ordinal outcome $Y \epsilon \{1, \dots, K\}$ is studied. Hence, an ordinal boosting algorithm is used that models the outcome with the commonly used proportional odds assumption

$$P(Y \leq k) = \frac{1}{1 + \exp(f(X) - \theta_k)}, \qquad k = 1, \dots, K,$$

for category thresholds $\theta_1, \dots, \theta_K$. Schmid et al. [11] derived the necessary negative log-likelihood loss function with corresponding gradient for applying gradient boosting to ordinal outcomes and proposed optimizing the thresholds in every boosting iteration with respect to the current model. More details can be found in [11]. We implemented the algorithm in R and did not use standard boosting libraries, as their application crashed when attempting to fit models to our high-dimensional data set (*p* = 789*,286*).

4. Random Forest (RF)

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and robustness, making it effective for both classification and regression tasks [12]. The model builds a "forest" of trees, where each tree is grown on a random subset of the data, helping to reduce overfitting and improve generalizability.

The Random Forest algorithm follows these main steps:

1. Bootstrap Sampling: From the original dataset of size N, multiple samples are drawn with replacement to create bootstrap samples. Each sample will serve as the training set for a single decision tree.
2. Feature Subset Selection: At each node within each tree, a random subset of variables (of size m, where $m < p$, with p as the total number of variables) is selected. The best split among these variables is chosen, a process that helps reduce the correlation between trees and thus improves ensemble diversity.
3. Tree Growth: Each decision tree is grown without pruning, allowing it to reach its maximum depth until a stopping criterion (e.g., minimum node size) is met. This ensures high variance among individual trees.
4. Aggregation (Voting or Averaging): For a classification problem, the Random Forest prediction is given by: $\hat{y} = mode(\{T_b(x)\}_{b=1}^{B})$, where: $\hat{y}$ is the predicted class label, $T_b(x)$ is the prediction from the b -th tree in the ensemble, given input x and B is the total number of trees.

For regression, the Random Forest prediction is the mean of the predictions from each tree:$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$,where: $\hat{y}$ is the predicted class label, $T_b(x)$ is the prediction from the b -th tree in the ensemble, given input x and B is the total number of trees.

Since each tree in RF is trained on a bootstrap sample, roughly one-third of the original samples are left out (not included in the training sample). These are known as Out-of-Bag (OOB) samples and can be used to estimate the model's error without requiring a separate test set [12]. Random Forest provides a measure of feature importance based on how much each feature improves the split criterion (e.g., Gini impurity or mean squared error) when used in nodes across all trees. By combining multiple trees and using random feature selection at each split, Random Forest reduces the risk of overfitting compared to individual decision trees [13].

To identify statistically significant variables in a Random Forest model, a variable importance test was developed. The work from Janitza et al. [14],introduces a computationally efficient approach designed for high-dimensional data, where traditional permutation-based tests are often too resource-intensive. The proposed method leverages a modified version of the permutation Variable Importance Measures (VIMs), inspired by cross-validation, to estimate an empirical null distribution based on non-positive importance scores. Compared to standard permutation-based tests, it

maintains Type I error control while achieving similar or even higher power at a significantly lower computational cost. The function var.sel.vita() implements the Vita approach for variable selection, ensuring that feature importance is statistically validated rather than merely relying on raw importance scores. By constructing an empirical null distribution from non-positive VIMs, as described by Janitza et al. [14], and utilizing the importance_pvalues function from the ranger package [15], it assigns p-values to each variable. This process enables the identification of statistically significant predictors, retaining only truly relevant features and enhancing model interpretability while mitigating overfitting risks.

5. (Sliding) Windows approach (WA)

Another method, particularly suitable for large amounts of data, is the so-called cross leverage score (CLS) [16]. These were originally developed for variable selection in genome-wide studies but can also be applied to other types of data, such as omics data. We will use these scores for a pre-selection of variables with which we will later calculate the risk scores. The CLS indicate for each variable its leverage on the outcome variable. Theoretically, each CLS is equal to its corresponding parameter in a least squares solution up to a small bounded additive error. At the same time, the score also contains information on whether the corresponding variable is part of a significant interaction effect, even if there is no significant main effect of the variable. The great advantage of the CLS is that these scores can also be calculated streamwise. The data stream algorithm allows data to be read in sequentially, resulting in a very fast and effective calculation that uses only a fraction of the computer's memory.

The approach is motivated by dimension reduction methods, whereby a $n \ll p$ problem has to be addressed here. Therefore, we consider the transposed matrix

$$\tilde{X} = [X, y]^T \in \mathbb{R}^{\tilde{p}\, x\, n}$$

with $\tilde{p} = p + 1$ and p the number of variables and n the number of observations. $X \in \mathbb{R}^{n\, x\, p}$ is the data matrix and $y \in \mathbb{R}^n$ denotes the response. The CLS are given by the off-diagonal entries of the hat matrix $H = QQ^T$ of $\tilde{X}$. So, we need to calculate the QR-decomposition $\tilde{X} = QR$. Since we are only interested in the CLS between the variables $j \in \{1, \dots, p\}$ and the response y, we have to calculate the dot products of rows $Q_{j.}$ with row $Q_{\tilde{p}.}$:

$$c_{j\tilde{p}} = \langle Q_{j.}, Q_{\tilde{p}.} \rangle, \qquad j \in \{1, \dots, p\}\,.$$

This score provides information on the mutual influence of x_j and y [16]. To avoid computing the QR decomposition with running time $O(n^2 p)$ [17], which is prohibitively slow for p very large (e.g. millions of omics variables), we consider a streamwise computation, here the sliding window approach [16]. We calculate the QR-decomposition for many small matrices instead of one QR-decomposition for a very

large matrix and merge the results a suitable way. We the select the $\mathrm{q} = \lceil n \log n \rceil$ variables that have the most extreme CLS for subsequent analyses. Theoretically, this is motivated by the coupon collector's problem [18] , which requires oversampling by a $\log n$ factor so that the selection contains at least as many variables to ensure that the submatrix preserves the full rank n of the original matrix [19].

References (Supplement only)

[1] McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*.

[2] Simon, N. *et al.* (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*.

[3] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior.

[4] Wu, J. *et al.* (2013). Strategies for Developing Prediction Models From Genome-Wide Association Studies. *Genetic Epidemiology*. https://doi.org/10.1002/gepi.21762.

[5] Privé, F. *et al.* (2019). Making the Most of Clumping and Thresholding for Polygenic Scores. *The American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2019.11.001.

[6] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[7] Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*. https://doi.org/10.1214/aos/1013203451.

[8] Lau, M. *et al.* (2025). Boosting interaction tree stumps for modeling statistical interactions. .

[9] Efron, B. *et al.* (2004). Least angle regression. *The Annals of Statistics*. https://doi.org/10.1214/009053604000000067.

[10] Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*. https://doi.org/10.1214/07-STS242.

[11] Schmid, M. *et al.* (2011). Geoadditive regression modeling of stream biological condition. *Environmental and Ecological Statistics*. https://doi.org/10.1007/s10651-010-0158-4.

[12] Breiman, L. (2001). Random Forests. *Machine Learning*. https://doi.org/10.1023/A:1010933404324.

[13] Tin Kam HoRandom decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*. https://doi.org/10.1109/ICDAR.1995.598994.

[14] Janitza, S. *et al.* (2015). A computationally fast variable importance test for random forests for high-dimensional data.

[15] Wright, M.N. and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v077.i01.

[16] Teschke, S. *et al.* (2024). Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores. *Biometrical Journal*. https://doi.org/10.1002/bimj.70014.

[17] Golub, G.H. and Van Loan, C.F. (1996). *Matrix Computations* , Johns Hopkins University Press.

[18] TROPP, J.A. (2011). IMPROVED ANALYSIS OF THE SUBSAMPLED RANDOMIZED HADAMARD TRANSFORM. *Advances in Adaptive Data Analysis*. https://doi.org/10.1142/S1793536911000787.

[19] Erdős, P. and Rényi, A.. (1961). *On a Classical Problem of Probability Theory.A Magyar Tudományos Akadémia Matematikai. Kutató Intézetének Közleményei* 6.

Eidesstattliche Versicherung (Affidavit)

Teschke, Sven
Name, Vorname
(Surname, first name)

165423
Matrikel-Nr.
(Enrolment number)

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden, § 63 Abs. 5 Hochschulgesetz NRW.

Die Abgabe einer falschen Versicherung an Eides statt ist strafbar.

Wer vorsätzlich eine falsche Versicherung an Eides statt abgibt, kann mit einer Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft werden, § 156 StGB. Die fahrlässige Abgabe einer falschen Versicherung an Eides statt kann mit einer Freiheitsstrafe bis zu einem Jahr oder Geldstrafe bestraft werden, § 161 StGB.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:

Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offence can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offences of this type is the chancellor of the TU Dortmund University. In the case of multiple or other serious attempts at deception, the candidate can also be unenrolled, Section 63, paragraph 5 of the Universities Act of North Rhine-Westphalia.

The submission of a false affidavit is punishable.

Any person who intentionally submits a false affidavit can be punished with a prison sentence of up to three years or a fine, Section 156 of the Criminal Code. The negligent submission of a false affidavit can be punished with a prison sentence of up to one year or a fine, Section 161 of the Criminal Code.

I have taken note of the above official notification.

Dortmund, 15.04.2025
Ort, Datum
(Place, date)

Unterschrift
(Signature)

Titel der Dissertation:
(Title of the thesis):

Variable Selection Methods for Detecting Interactions in Large Scale Data

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht.
Die Arbeit hat in gegenwärtiger oder in einer anderen Fassung weder der TU Dortmund noch einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegen.

I hereby swear that I have completed the present dissertation independently and without inadmissible external support. I have not used any sources or tools other than those indicated and have identified literal and analogous quotations.

The thesis in its current version or another version has not been presented to the TU Dortmund University or another university in connection with a state or academic examination.*

*Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the PhD thesis is the official and legally binding version.

Dortmund, 15.04.2025
Ort, Datum
(Place, date)

Unterschrift
(Signature)