Bonn-Aachen International Center for Information Technology (b-it)
University of Bonn

Master Thesis

Data Science Approaches for Identification of Genetic Impact on Cholesterol Levels in Parkinson's Disease

Submitted as fulfillment of the requirements
towards the achievement of the degree Master of Science in Life Science Informatics

By
Thomas Lordick

1. Supervisor: Prof. Dr. Holger Fröhlich
2. Supervisor: Prof. Dr. med. Dipl. Phys. Peter Krawitz

Tuesday 12th January, 2021

Acknowledgement

I'd like to thank Prof. Dr. Holger Fröhlich for making this project possible and guiding me through the process and challenges this project brought. Special thanks go out to Tamara Raschka for doing the best job a daily supervisor can do by always providing me her assistance and new ideas to improve this work from start to finish without exceptions. Big thanks go out to Meemansa Sood und Phillipp Wendland. I highly appreciate your expertise and help in establishing the VAMBN model. Thanks to Meike Knieps for doing all the paperwork for my start, showing me the facilities and getting rid of annoying Microsoft Teams problems. Thanks to André Gemünd for solving all my IT problems and being patient with me. At the end, I'd like to thank the entire AI-DAS team. Your good organization, even in times of the pandemic, made it as easy as possible for me to work on my thesis.

Best,
Thomas

Contents

List of Figures

List of Tables

Abbreviations

BIC Bayesian Information Criterion.

CADD Combined Annotation–Dependent Depletion.

eQTL Expression Quantitative Trait Loci.

GQ Genome Quality.

GWAS Genome-wide Association Study.

HDL-C High-density Lipoprotein Cholesterol.

HI-VAE Variational Autoencoder for Heterogeneous and Incomplete Data.

IPD Idiopathic PD.

KL Kullback-Leibler.

LD Linkage Disequilibrium.

LDL-C Low-density Lipoprotein Cholesterol.

LMM Linear Mixed Modeling.

MBN Modular Bayesian Network.

PCA Principal Components Analysis.

PCs Principal Components.

PD Parkinson's Disease.

PheWAS Phenome-wide Association Study.

PRS Polygenic Risk Score.

SKAT Sequence Kernel Association Test.

SKAT-O Optimized Sequence Kernel Association Test.

SNP Single-Nucleotide Polymorphism.

VAE Variational Autoencoder.

VAMBN Variational Autoencoder Modular Bayesian Networks.

VCF Variant Call Format.

1 Introduction

One of the most common and investigated neurodegenerative disorders after Alzheimer's disease is Parkinson's disease (PD). Up to date, neither an ultimate cause nor cure for PD has been found. Several types of PD exist, with idiopathic PD as the most common and still, least understood type. [1]. The disease outbreak and pathogenesis of idiopathic PD is thought to originate from a complex combination of both genetic and environmental risk factors. Environmental risk factors such as the exposure to certain metals, industrial solvents or traumatic head injuries are highly addressed and discussed in the scientific PD community, but remain untouched here [2, 3]. In this project, the attention is predominantly drawn to the genetic aspects of PD.

Moreover, recent studies indicate that blood cholesterol levels are associated with the pathogenesis and risk of developing PD. Multiple studies with moderate to high numbers of participants could demonstrate an association of lower low-density lipoprotein cholesterol (LDL-C) levels and LDL-lowering substances to higher PD prevalences within the respective study cohorts [4–6]. Study designs were predominantly case-control studies with logistic regression predicting the occurrence of PD by the cholesterol measurements including the adjustment for confounding and population effects. Also, a time-dependent association study using cox models to access PD hazard ratios was performed [7].

Minding these indications, this project comes with a different, genetic based, approach of investigating the conceivable cholesterol-PD association. Using a 106 samples case-control data set from the *AETIONOMY* project [1], PD-related single nucleotide polymorphisms (SNP) from GWAS [8] will be used to predict LDL-lowering phytosterol levels. On varying biological feature levels, ranging from individual SNPs over gene and biological mechanisms to a genome-wide level, different associating testing methods will be applied. Including the idea of polygenic risk scores, burden-, and

[1] https://www.aetionomy.eu

kernel-machine based tests plus an alternative scoring method, this project aims to investigate in the **genetic part** of PD to cholesterol in contrast to the above mentioned study approaches. A special focus will be on testing methods on gene- and mechanisms level, as they provide a deeper and more intrinsic biological interpretive scope than the standard approach on individual SNP level. Conclusion may be useful for prospective clinical decisions in terms of patient treatment and more.

Because sample size is an essential factor in association testing, another major focus of this project is the simulation of realistic patient data based on the original *AETIONOMY* data set. Several phenotype- and genotype simulation tools come with the requirement of customizing the inner data structures (for instance covariance across the modalities) by setting parameters in order to simulate data [9, 10]. Furthermore, genotype data in form of genetic variant genotypes is often the foundation of simulated phenotype features and needs to be simulated a priori. Standalone genotype simulation in particular, is often simplified by disregarding inner SNP-SNP interactions and covariances [11] or does not consider multi-modal interactions between phenotype and genotype features as it is the case in PD for example. In this thesis, the alternative machine learning approach VAMBN will be used to simulate virtual patient cohorts with varying sample sizes [12] on both genotype and phenotype level to overcome the rigidity of standard approaches. Those virtual patient cohorts will be used in the association testing methods to obtain more meaningful and unbiased results without the constraints that small sample sizes bring along. Virtual SNP genotypes in particular will be additionally compared to simulated SNP genotypes from the resampling-based framework of the tool *sim1000G* [13].

2 Theoretical Background

2.1 Parkinson's Disease

Idiopathic PD is a common and complex, neurodegenerative disorder that affects 1 - 2 % of the population with an age of 65 years and older. The term *idiopathic* refers to the fact that the cause is unknown. It typically comes with motor system symptoms as tremor, rigidity, slowness of movement and walking troubles. Non-motor symptoms such as cognitive impairment and sleep disturbance occur as well. In general, symptoms intensify as the disease progresses and women are more likely to be affected than men [14].

Through the curse of PD, dopaminergic neurons in the substantia nigra that are responsible for the Dopamine secretation become harmed and further on die. In addition, the presence of Lewy bodies (accumulations of alpha-synuclein) located in many of the remaining neurons are traditional characteristics of the disease. Up to date, no available treatments or medications exists that can cease or reverse the neurodegenerative processes [1]. An ultimate cause of idiopathic PD still remains unknown. It is no longer regarded, that the risk of developing PD is predominantly due to so called environmental factors alone, and rather tend to result from a combination of both environmental and genetic factors affecting numerous cellular processes and mechanisms [2, 3]. The latter, genetic risk factors, which will be the main focus, while environmental factor may be mostly disregarded for this study.

One has to distinguish between monogenic forms of PD that are caused by a single rare mutations in a small set of inherited genes (such as *SNCA* and *LRRK2*) [15], and forms which underlie the complex structure and interaction between several more common mutations at multiple risk loci. Just a small percentage of PD cases are of monogenic forms, variable in genetic penetrance, but better understood in the clinical field [14]. Mono-

genic forms account for approximately 30 % of the familial (cases that have a first-degree relative who has PD) and approximately 5% of the sporadic cases [14]. Still, more and more genetic variants across multiple loci are now recognized as risk factors in PD and therefore, support the assumption that PD may occur from polygenic origin. More evidence from GWAS and clinical studies supports, that idiopathic PD and genetic forms of PD are more or less the same entity as well. Those complex interaction of multiple risk variants may affect and determine penetrance, age at onset, severity and more phenotypic traits in the progression of PD [16, 17].

Until now, 19 risk loci have been identified and the underlying mutations have been identified in eleven of them [18]. With the rising advance in next-generation sequencing technology, the methods to identify risk loci with their underlying mutations have become a major tool in PD research. Genome wide association studies significantly contributed to the identification of risk variants in PD. Some identified SNPs are present in the same loci that are normally associated with monogenic PD forms. Additional ones are found in loci also associated with other disorders, such as the glucocerebrosidase gene *GBA* or microtubuleassociated protein tau gene *MAPT* [15, 19]. The latest GWAS, consisting of 37 688 PD cases, 18 618 PD proxies and over 1 400 000 controls has identified 92 different risk variants that potentially contribute to the risk of PD [8]. In this project, the main focus will be on those single-nucleotide polymorphisms (SNPs), found in GWAS and therefore indicate a potential role in the polygenic-based pathogenesis of PD.

The Role of Cholesterol in PD

Cholesterol is a lipid present in cell membranes of all animal tissues and is transported within the blood plasma. While the majority of the cholesterol is synthesized by the body cells, it can also be included in the food and therefore of dietary origin. In densely-packed cell membranes, such as, in liver, which synthesizes most of the cholesterol, but also in the spinal cord and specially brain cells, lipid levels are high and therefore cholesterol

levels as well. In many biochemical processes, such as the composition of cell membranes, with its essential function by organizing the biophysical properties of the phospholipid bilayers and by being a precursor for steroid molecules, cholesterol plays a major role in the metabolism of animals [20, 21].

To date, two major forms of cholesterol are known:

- Low-density lipoprotein (LDL-C)
- High-density lipoprotein (HDL-C)

HDL, sometimes referred to as the "good cholesterol", is often correlated with a good cardiovascular health, while LDL, the "bad cholesterol", is largely known to have negative impacts on the humans health when levels are too high. For example, too much LDL can pose an increased risk of cardiovascular and artery diseases. Superabundances of LDL form so called "plaques" in the arteries and inhibits normal blood flow to the heart. In the worst case, this causes heart attacks. However, the association between cholesterol and neurodegenerative disease risks is rather unknown in general, but debated in the scientific community [20].

In some studies, it has been observed that high serum cholesterol may increase the risks of Alzheimer's disease [22, 23] and ischemic stroke [24]. Little is known about the association between cholesterol and PD risk. Multiple case-control studies were carried out, testing if there is a significant difference in cholesterol levels between PD patients and controls. Some of them, with either small or rather large sample sizes, did not find any significant association [25]. While in other studies, low LDL levels and LDL-lowering substances could be linked to higher occurences of PD [6]. According to a case-control study from China, PD patients are linked with lower LDL and HDL levels [4] and according to another case-control study [5] with 234 participants, indications were found that low LDL may be associated with higher occurrence of PD as well. According to a time-dependent association study from 2018, higher levels of LDL could be associated to lower PD risk among men [7].

Even though, multiple indications exist, the entire cholesterol PD linkage still warrants further investigations which is one of the main emphases is this project. What distinguishes this project from the above mentioned case-control studies is the approach of testing the genetic impact in PD on cholesterol. As described in section 3.1.1, five different phytosterols which are recognized for having a proven LDL-lowering effect [26], will be examined.

2.2 Association Testing

With the rise of GWAS and the huge amount of freely available GWAS summary statistics in the last years, new life has been brought to the field of association testing. With the growing interest in investigating the influence of genetic risk factors on human disease phenotypes and the easy access to technical tools, this has attributed to an increase of possibilities for research in the field of medicine and genetics [27]. The results of a successful GWAS typically reveals SNPs that significantly contribute to the trait of a certain investigated phenotype, for example PD or other human diseases of complex genetic origin. The strength of association in each investigated SNP can be measures with effect sizes like odds ratios (OR) or regression coefficients (betas) and corresponding p-values of the underlying summary statistic. In upstream analyses, this information can be utilized for further association analyses. With the aim of explaining certain disease phenotypic traits such as the levels of cholesterol in the human blood serum or age of the disease outbreak, causal SNPs from GWAS can reveal great functionality to find a statistical linkage in those disease associated phenotypes [28]. The following sections will provide insights in the methods that utilize those SNPs and will be applied during this project.

Standard Approach

The most baseline approach would be the so called *standard approach* that models each SNP individually to predict the phenotypic trait. This is usually done within the framework of a linear regression model.
Confounding effects, sometimes called external exposures, that affect the phenotypic trait being studied, have to be controlled and would be included in the regression model. By doing that, the unbiased association of the SNP effect can be accessed and evaluated [29]. Another source of systematic bias would be population stratification and sample structure. This is, in essence, the presence of multiple subpopulations or clustering due to relatedness within the subjects being studied. Population stratification can lead to false positive associations and mask true associations, when allele frequencies differ between subpopulations in the study [30, 31]. There are several methods to correct for population stratification. Two of the most common methods were applied in this thesis: 1. Principal component analysis (PCA) and 2. Random effects modeling within a linear mixed model. The former, PCA, is probably the most widely used method to address population stratification [30]. This method uses genotype data to estimate principal components (PCs) that can be used as covariates in the linear models in the association analysis. The idea is, that PCs represent features of genomic ancestry of the subjects. Since the setting of each SNP differs along each axis of ancestry, the PCA approach corrects not only for false positives but also for false negatives. On the other side, linear mixed modeling (LMM), offers a different strategy to account for population stratification [32]. LMMs can be defined as an extension of fixed effects models (such as linear regression) by the incorporation of a random effects term. Considering Y as the dependent variable, X as a nxp matrix for p independent variables, a standard LMM is defined as:

$$\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \tag{2.1}$$

By doing that, we account for more than one source of random variability,

genetic variability, in the data in contrast to linear regression. That random effect term u is defined in the following way.

$$u \sim \mathcal{N}(\mu, \sigma^2 R) \tag{2.2}$$

R defines a genetic relatedness matrix which accounts for the pairwise genetic similarity between the subjects. Pairwise genetic similarities between subjects encapsulate correlations due to population structure [28, 33]. Details on the application of these methods will be further elaborated on in the Material & Methods chapter.

Still, such methods that model SNPs individually, can be statistically underpowered. Small to moderate effect sizes of the SNPs make it hard to explain the variance in the phenotypic trait of interest [34]. It usually requires large sample sizes to detect causal SNPs many complex traits such as PD [35]. Moreover, individual testing might lead to an inaccurate measurement of the influence of the SNPs on the phenotype, if certain SNPs show effects only by the interaction with other SNPs. Not only SNP interactions, also identifying sets of SNPs, which belong to the same region, gene or biological mechanism that are associated with the disease is not feasible in the standard approach [36]. More generally spoken, individual testing ignores the multivariate structure of the data being studied. Hence, by replacing individual SNPs by SNPs sets, statistical power may increase and opportunities to shed light on the biological background with a more medical-orientated interpretability become feasible. Therefore, the interest in developing methods that aggregate the individual effect of SNPs into sets leads to more sophicasted methods of testing. The following paragraphs elucidate common alternatives to individual SNP testing on different genetic levels that will be taken into account in this project.

Polygenic Risk Score

Polygenic risk scores (PRS) analyses do not test individual variants. Instead, such analyses aggregate the genetic risk across the whole genome

in an individual score that can be used for regressing on a phenotypic trait of interest. Usually, it requires a large base data set, often referred to as the discovery data set, which is in substance the GWAS summary statistic or multiple GWAS summary statistics that investigate the same phenotype. This data set determines the effect sizes of each SNP that is expected to contribute to the PRS of a trait. In combination with the base data, the target data set, which determines the genome wide genotype of the SNPs of the subject being studied, will be used to calculate the PRS. Based on genetic disposition and weights from the base data set, score calculations result in individual risk presentations of each subject [36, 37]. Even though, PRS may be not potent enough to predict disease risk on the individual patient level [38], significant associations both within and across traits could be successfully shown. For instance, GWAS summary statistics for schizophrenia could be utilized to find significant association in phenotypic traits such as bipolar disorder, level of creativity, and risk of immune disorders [39–41]. In Alzheimer's disease and PD, the PRS revealed significant association to the diseases' age at onset [42, 43]. Besides age at onset, disease status, cognitive decline and motor progression could be associated to the PRS in PD [44].

All these methods design risk scores on a genome-wide level. In an analogous manner, a PRS design on PD-related biological mechanism level is theoretically feasible and will be tested in this project beside the genome-wide level approach. Biological mechanisms in this context can be defined by SNPs sets, in which SNPs can be linked to mechanisms via the mapping on genes that are involved in these mechanisms. For instance, by encoding proteins that play a functional role in a mechanism of interest. This idea potentially provides a deeper genetic and medical interpretability that regular PRS studies do not provide. Details on the PD-related mechanisms will further elaborated on later in this chapter.

Burden and Kernel-Machine based approaches

In contrast to PRS, that implements the idea of scoring on genome-wide level, the cumulative effects of SNPs in a *SNP set*, such as genetic regions or genes, offers another way in association testing. Burden tests, also referred to as *Mutational Load* test in this project, collapses SNPs in a predefined genetic region into a burden variable. That burden variable can be the ratio of effective alleles (e.g. the number of mutated alleles in the SNP) by the overall count of SNPs in the respective SNP set. Mutational load testing basically assumes that all SNPs of a set are causal and therefore, influence the phenotype in the same direction. When these assumptions are violated, this method becomes less meaningful [45].

In the kernel-machine based approach on the other hand, also called SKAT method (sequence kernel association test) the individual test statistics scores of the SNPs will be aggregated into pre-defined SNP sets and SNP-set level p-values will be efficiently computed [46]. That method overcomes the limitation that burden score brings, when a set contains both causal and non-causal SNPs. SNPs are weighted based on their prevalence (up-weighting rare SNPs, downweighting common SNPs) in a linear weighting scheme. The SKAT approach comes with the additional feature of a kernel matrix which aggregates the associations between SNPs and the phenotypic trait (Details in Material and Methods section). This method allows for potential SNP-SNP interactions, e.g epistatic effects, that will not be taken into account in the standard, burden and PRS approach. Still, if the assumption of a burden test are present in the defined gene sets, the burden test usually performs better [47]. Hence, the SKAT-O test, which is a linear combination of both kernel-based and burden tests was developed and will be applied in this project to regress phytosterol levels on PD-related genes as well. SKAT-O tries to find an optimal solution by adapting to the conditions of the data. When a burden test is more powerful than the kernel-based test, it behaves like the burden test and vice versa. It also shows good performance on small-sample size cohorts, which is for instance the case in the project's data set [47]. It has success-

fully applied in gene level testing of multiple complex diseases such as inflammatory bowel disease or the anthracycline-induced cardiotoxicity in childhood cancer [48] [49].

GenePy

A modern way to convert SNPs level risk on a gene level can be realized by calculating *GenePy* scores [50]. While methods like burden and kernel-machine based tests predominantly focus on the allele frequencies, *GenePy* comes with a more sophicasted approach, by additionally incorporating data observed zygosity and a pre-defined variant deleteriousness metric. The deleteriousness metric underlies a machine learning approach to score variants on pathogenicity, allecic diversity and disease severity [51]. In this project, the combined annotation–dependent depletion (CADD) metric has been chosen. CADD is a trained support vector machine to differentiate common real human variants from simulated variants. "C scores" for the possible human variants enable scoring of SNPs and correlate with allelic diversity, functionality, pathogenicity etc. Those scores significantly distinguish known pathogenic variants within individual genomes from simulated and more harmless variants. [51]. *GenePy* utilizes this scoring scheme and hence, improves the incorporation of intrinsic biological information compared to burden and SKAT based approaches. Even though *GenePy* scores are not intentionally proposed to be utilized in association testing, this project will still try to integrate them in the statistical analyses.

2.3 NeuroMMSig

As described in the PRS section 2.2, association testing on biological mechanisms (also called pathways) level requires the definition of an accurate SNPs to biological mechanisms projection (Which SNPs map on which mechanism). Furthermore, the right framework has to be chosen that se-

lects disease specific mechanisms that relate to the pathophysiology of PD.

In that context, the mechanisms enrichment server NeuroMMSig [1] [52], a neurodegenerative disease specified alternative to the original MSigDB [53] implementation, has been chosen to select PD-related mechanisms and their underlying genes. Each NeuroMMsig subraph has been enriched with well curated, disease related data such as literature evidence, image or related drug information. This modeling across multiple modalities allows to infer deep causal and correlative relationships among e.g disease related genes and therefore gather disease related mechanisms.

In this project, 64 PD-related mechanisms and their gene sets were extracted from NeuroMMsig and only those were used in the mechanism and gene based association based approaches. SNPs were mapped onto mechanisms through their respective genes. Details on the mapping can be taken from the Material & Methods section.

2.4 Synthetic Patient Generation

A major part of this project is the simulation of synthetic patients. Because sample size is crucial in genetic association testing [54] and the underlying data set containing 106 subjects is not sufficient enough to conclude meaningful results, the simulation of both phenotype- and genotype data based on the real data will be carried out.

Common simulation tools for association testing studies, such as *PhenotypeSimulator* or *phenosim*, are dependent from pre-simulated or real genotype data [9, 10]. In essence, simulated phenotypic data will be based on a priori defined variant genotypes and a certain degree of association has to be predefined. Those approaches are well-suited for a proof of concept methodology. When it is the aim to test the statistical power of association methods under various conditions provided by flexible phenotypic data

[1] https://neurommsig.scai.fraunhofer.de

generation, those tools may be the right choice. In this project in contrast, no assumption of association between the SNPs and the phytosterols can be set, and therefore, an alternative approach has to be considered. Regarding variants genotypes, common tools like *plink* [11] or the built-in genotype simulator *simulateGenotypes* in *PhenotypeSimulator* draw simple bi-allelic SNPs from a binomial distribution with equal probabilities for user given allele frequencies. These methods disregard the multivariate structure within the SNPs such as correlation due to linkage disequilibrium (LD). SNP dependencies that are present in realistic data simply can not be simulated with these tools. Resampling-based approaches on the other side, such as *sim1000G* or *HapGen2*, are more suitable for the simulation of case-control data sets (as in this project), because of their ability to regard LD structure and multivariate dependencies along the SNPs [13] [55]. The term *resampling* refers to the ability of simulating SNPs based on user-given reference SNPS genotype samples. Simulated SNPs will therefore be similar in terms of LD structure and covariance accross the SNPs to the user-given reference data sets. *HapGen2* in particular, is designed to work with publicly available reference variant data as part of the HapMap or 1000 Genomes project and can therefore not be applied to personal data sets straight away, such as the *AETIONOMY* data in this project. *sim1000G* is better suited for this purpose, because it comes up with the possibility to simulate SNPs based on any user-given SNPs genotype data. By simply inputting a personal VCF file, *sim1000G* extracts SNPs and their covariance structure across them. This covariance structure is used to mimic the LD structure in the simulated SNPs of related or unrelated subjects. *sim1000G* works within the framework of genomic-regions up to chromosome length which means only SNPs mapping on those regions can be properly simulated.

In this project, a machine learning approach will be considered for the simulation of virtual PD data sets. The generative modeling framework of VAMBN [12] will be used to learn the structure and multi-modal dependencies between both genotype and phenotype data, in order to synthe-

size virtual patient cohorts along multiple sample sizes. The idea behind VAMBN is the combination of two types of generative modeling methods: A variational autoencoder (VAE) [56] and a modular bayesian network (MBN) [57]. In the VAE, the multivariate distributions of the original features, such as demographics, phytosterol levels and genotypes of SNPs, will be encoded into a low-dimensional latent distribution (=latent space, typical multi Gaussian distribution) module-wise by defining proper modules. Those modules contain features that can be combined according to the study conditions (details on module design are elaborated on in the Material & Methods section). The encoding is realized by multiple consecutive steps through the layers of a neural network within the VAE model. In the modular Bayesian network step, the conditional dependencies and relationships between these modules will be learned in a directed acyclic graph structure. Those statistical dependencies in the MBN allow for drawing samples (=virtual patients) from it by following the topological order of the graph structure. Because drawn samples are represented in their module-dependent latent distribution, virtual patients will be decoded again through the VAE in order to get the original feature distributions. Following that scheme, virtual patient cohorts can be generated with unlimited increasing sample sizes.

A key difference to traditional approaches, for instance the *PhenotypeSimulator*, is that the statistic dependencies of the real data will be authentically represented in the the virtual patient cohorts without the need for predefining associations and dependencies by rigid parameter settings a priori. The machine learning aspect of VAMBN results in a more sensible and deeper data structure access than traditional approaches. Even though, resampling-based genotype simulation tools like *sim1000G* access LD and SNP-SNP interaction structure as well, VAMBN comes with the additional feature of modeling genotype and phenotype dependencies within the MBN framework and does not rely on genomic regions when simulating SNP genotypes. SNPs genotypes can be simulated without the constraint of genetic regions as genes or chromosomes, which will be realized in this

project by simulating SNPs sets genotype that map on PD-related NeuroMMsig mechanisms (see Material & Methods for details).

Here we additionally compare the SNP genotype simulation results of VAMBN and a resampling-approach based tool *sim1000G* in order to outline key differences in both methodologies.

3 Material & Methods

3.1 Data

3.1.1 Parkinson's Disease Patients Data

The analyzed data is part of the *AETIOMONY* project [58]. It contains cohort of 106 subjects including 67 idiopathic Parkinson's disease (IPD) patients and 39 controls that have not been diagnosed with IPD. Along with basic demographic information including features like age, gender, alcohol consumption, smoking status and Levodopa treatment (drug used to increase dopamine concentrations in the treatment of Parkinson's disease), blood serum levels of 19 different cholesterol, cholesterol-derivatives and cholesterol-influencing substances are collected (hereinafter called for simplicity outcomes).

The above mentioned demographic data will be included as confounding factors in the models because they potentially influence the levels of the outcomes. Except for smoking status and alcohol consumption no data point was missing. Since both features are ordinal, the two missing data points were imputed with the most frequent value technique using the *scikit-learn* library in python [59]. Table 3.1 briefly describes the demographic information including the p-value of a two-sided Welch test which tests whether the features reveal identical average values between the groups. P-values below $\alpha = 0.05$ indicate that the feature has significantly different mean values between IPD and control patients.

Table 3.1: *Demographic information of the AETIONOMY data. Smoking denotes the smoking status of the subject ranging from 0 (never), 1 (ex) to 2 (current). Alcohol denotes the drinking behaviour of alcoholic liquids ranging from 0 (never) to 5 (strong drinking habit)*

Feature		**No IPD**	**IPD**	**P-value**
Age	mean	59.7	62.5	0.11
Smoking	mean	0.67	0.40	0.03
Alcohol	mean	2.5	3.0	0.04
Sex (F/M)	count	25 / 14	21 / 46	0.001
Levodopa (Yes/No)	count	1 / 38	19 / 48	5.8e-5

To focus on potentially interesting outcomes, only those were selected which reveal a significant difference in measurements between healthy and IPD individuals. For this purpose, it was tested whether the binary feature *Group* (1 for IPD, 0 for control) has a significant impact (α = 0.05) on each outcome among the confounding features in a simple univariate regression model using the *statsmodel* package [60]. P-values are based on a likelihood ratio test.

As a result, five different phytosterols (*Campestanol*, *Campesterol*, *Stigmasterol*, *Sitostanol*, *Sitosterol*) were selected. Figure 3.1 illustrates the quantile normalized distributions of each selected outcome in dependence of their group membership ($1 = IPD, 0 = NoIPD$).

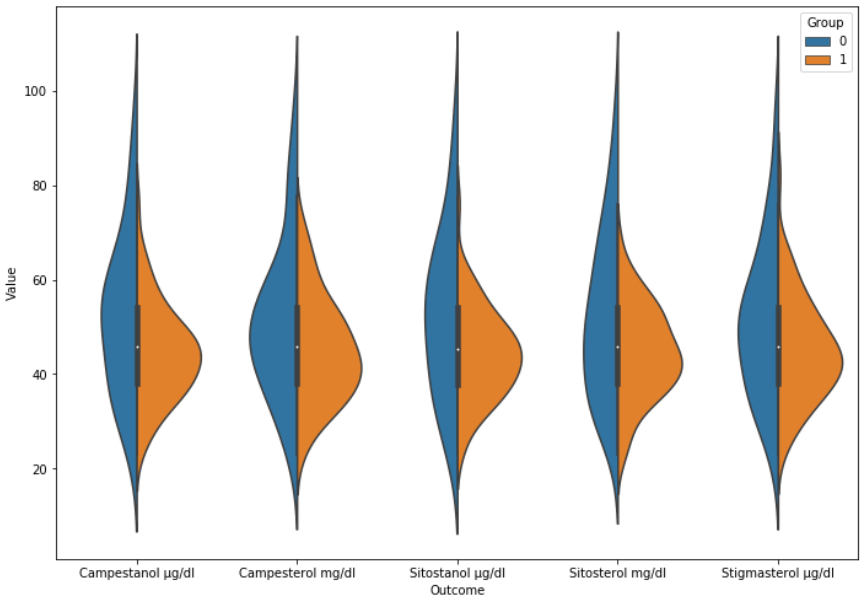


Figure 3.1: Distribution of the five selected phytosterols colored according to their group membership.

Another data source in the *AETIONOMY* data set are gene sequences for the genotypes of each patient. Those are stored in a Variant Call Format file (VCF), version 4.2. Missing variants were a priori phased and imputed for missing genotypes using *IMPUTE2* [61]. Furthermore, the VCF file was built using the human genome assembly GRCh37 as reference. Since all rs identifiers (variant IDs) used in the upstream analyses are based on the human genome build hg38 and do not tend to be stable between the different build versions, the whole VCF file was lifted over to hg38 using *LiftOver* [62] in a 2-folded fashion (GRCh37 to hg19 to hg38). Chain files were directly collected from the UCSC Genome Browser [63]. After that, missing variant identifiers were annotated using *bcftools* [64]. The reference annotation vcf file based on genome build hg38 was downloaded from the

UCSC Genome Browser[1]. Further quality control steps of the VCF file will be described in the upcoming sections, as those differ slightly between the corresponding tools which use the VCF file as input.

3.1.2 NeuroMMSig Mechanism Mapping

In this section, the architecture of the SNPs to mechanisms projection will be described. In the first step, SNPs are being mapped on genes and after that, those genes will be mapped on the NeuroMMSig mechanisms related to PD. The underlying foundation of the mapping is defined as a set of PD associated single-nucleotide polymorphisms (SNPs). While taking all types of SNPs into account, including those that fall in coding sequences of genes, non-coding regions of genes, or in the intergenic regions in the first place, only those that can be significantly associated with PD from GWAS and PheWAS (Phenome Wide Association Studies) were considered in the end.

Two major databases were used to gather PD associated SNPs: PheWAS Catalog [2] [65] and DisGeNet [3] [66]. On PheWAS Catalog, 121 PD associated SNPs were gathered (Summer 2020). On DisGeNet, 321 additional PD associated SNPs could be collected. This set of, in total 442 SNPs, which will be called "seed SNPs", is further extended by SNPs which are in Linkage Disequilibrium with these seed SNPs. To achieve that, *HaploReg* (Version 4.1) was used [67]. Setting the LD threshold to 0.8, the set of seed SNPs could get extended by 6887 SNPs ending up in a total of 7329 SNPs.

To map the SNPs to their corresponding genes, two major ways were defined:

[1] https://genome.ucsc.edu
[2] https://phewascatalog.org
[3] https://www.disgenet.org

- Chromosomal location of the gene
- Via expression quantitative trait loci

The latter, via expression quantitative trait loci (eQTL), will be the case if certain SNPs have an impact on the transcription of another gene, even if they are not directly located on those. To find the amount of SNPs that act as eQTLs plus their corresponding gene, each SNP was queried on the public resource GTEx Portal [4] [68]. Only brain tissue cis-eQTL were taken into account. Cis-eQTLs are located near the gene of origin. The SNP to gene mapping is highly surjective and counts 10329 SNPs mapped onto 588 different genes.

In the final step of the mapping, the NeuroMMSig knowledge base [52] was used to find those genes that can be linked to PD-related biological mechanisms, which are mostly involved into metabolism and signaling conduction. In the PD NeuroMMSig graph, 394 genes are associated to 64 different PD mechanisms. 18 genes from the above outlined "SNPs to Genes" mapping were found in the NeuroMMSig based genes. Putting the intersection of the two mappings together, 1817 final SNPs could be mapped onto 39 (out of 64) PD-related mechanisms via 18 genes. For the remaining mechanisms, no SNP could be mapped. Figure 3.3 illustrates the outcome of the final SNPs to Mechanisms mapping.

[4] https://gtexportal.org/home/

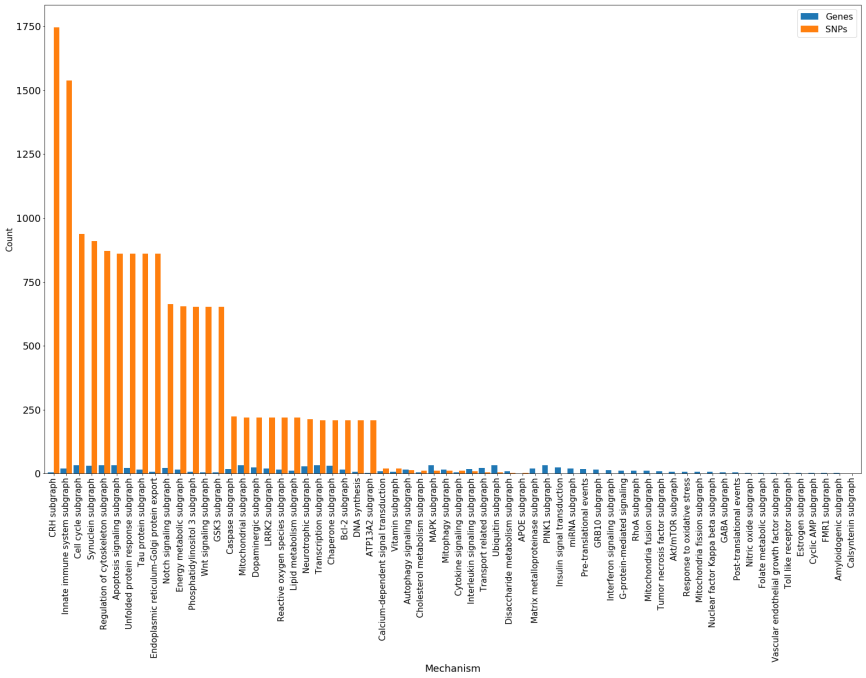


Figure 3.2: Overall count of mapped SNPs and corresponding genes in dependence of PD-related mechanisms from the NeuroMMSig knowledge base sorted by their SNPs count. All 64 mechanisms are shown here. Only 39 of those count more than zero SNPs. SNP counts range from nearly 1750 SNPs per mechanism (CRH subgraph) to two SNPs per mechanism (APOE subgraph).

Because gene lengths differ significantly within the NeuroMMSig gene set, SNP counts were normed by their corresponding gene lengths they map on (= SNPs/Bp ratio). To illustrate that, the following figure shows the SNPs/Bp ratios in dependence of NeuroMMSig mechanisms along with their accompanying genes. The SNP rates were scaled (x 1000) in order to match the scale of the gene counts.

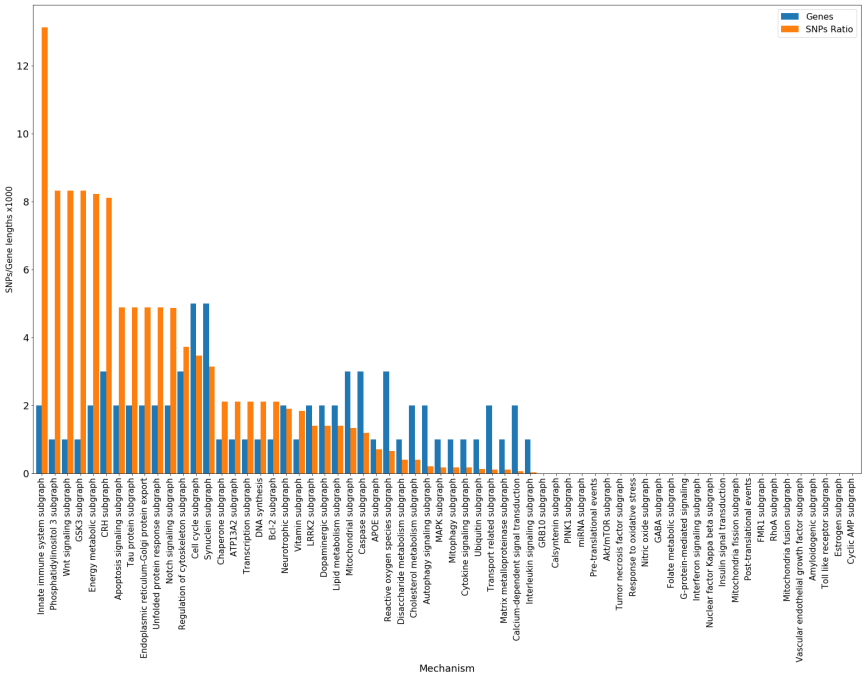


Figure 3.3: SNPs rates per bps (scaled) and corresponding genes in dependence of PD-related mechanisms from the NeuroMMSig knowledge base sorted by their SNPs count.

Table 3.2 reports the number of SNPs mapped on the 18 genes used in the mapping.

Table 3.2: *Genes and corresponding number of SNPs mapped on the respective genes*

Genes	**SNPs count**
GPX1	1
TOMM40	1
CPLX1	1
APOE	2
GBA	2
MAP1LC3A	2
CASP8	4
STX1B	4
GDF5	4
CACNB2	10
FDFT1	10
LRRK2	11
SLC41A1	21
GPNMB	31
GAK	34
SNCA	208
MAPT	653
CRHR1	885

Important to mention is, that 1583 from these 1817 SNPs were found in the VCF file of the *AETIONOMY* patients and finally used for upstream scoring and testing methods. The missingness is due to the fact that not 100 % of missing variants could be imputed with *IMPUTE2* and annotated with *bcftools*.

3.2 Scoring Methods

Based on the above describes SNPs to mechanisms projection, scores for the genes and mechanisms were computed using the underlying characteristics of the SNPs with respect to their occurrence in the data set individuals. Each method differs in their computational design including differences in quality control of the underlying SNPs sets. The final derived

scores were then used as features in linear models whose design and architecture will be explained in the further sections, as well as, the differences in each of the several methods.

3.2.1 Polygenic Risk Score

The polygenic risk score analyses was carried out using the tool *PRsice-2* [37]. According to the guidelines described in [69], the original VCF file, also called target data, has been converted into plink binaries (.fam, .bim, .bed) using *plink2* [11], as the first step. In the next step, the following quality control steps were applied to the target data:

- Removing all SNPs with minor allele frequency less than 0.01.
- Removing all SNPs with low P-value ($p \leq 1e-6$) from the Hardy-Weinberg Equilibrium Fisher's exact test
- Excluding all SNPs that are missing in a high fraction of subjects
- Excluding subjects that have a high rate of genotype missingness

In the following step, highly correlated SNPs were removed (a method called clumping). This is a procedure in which only the most significant SNPs (lowest p-value) in each LD block (window size 200 bp) is identified. Applying this, only SNPs with the strongest statistical evidence will be retained.

The GWAS summary statistics, called base target, which include information about the effect sizes of each PD-related SNP was obtained from the Nalls et al. meta-study in 2019 [70]. It includes approximately 12 million analyzed variants. To avoid systematic errors, almost 2 million ambiguous variants were excluded. Those with complementary alleles, either C/G or A/T, in particular. Due to unknown information about the genotyping

chips used in the creation it will be unknown whether the base and target data are referring to the same allele.

The polygenic risk score was computed on whole-individual level and mechanism level, respectively. The latter made use of the *PRsice* extension *PR-set* (paper under development [5]), exactly made for such pathway based analysis. Besides the base (GWAS summary statistics) and target data (plink binaries), mechanisms and corresponding genes were further used as input, in order to determine a polygenic risk score for each mechanism. The human hg38 GTF (gene transfer format) obtained from the UCSC table browser [63] was used to set the genomic boundaries of each gene. Either way, polygenic risk scores (PRS) were calculated in an additive way by summing up the effect sizes S (log(OR)) multiplied by the number of effective alleles observed G (0,1 or 2). Risk scores were normalized by the number of effective alleles M observed in the j'th subject.

$$PRS_j = \sum_i \frac{S_i \times G_{ij}}{M_j} \tag{3.1}$$

3.2.2 GenePy

Pathogenicity on gene level can be estimated by using the tool *GenePy* (version 1.2) [50]. To improve score consistency, the following quality control filters were a priori applied to the VCF file (according to Tom et al. [71]) by using *vcftools* [72]:

- Keep only bi-allelic SNPs,
- Remove SNVs with missing rate > 30 %, and
- Retain high confidence calls only by setting GQ < 20 as missing

[5] https://www.prsice.info/prset_detail/

Furthermore, the VCF file was filtered for SNPs from the NeuroMMSig mapping only, resulting in keeping 1583 final SNPs.

Following the guidelines on the GenePy GitHub repository[6], all SNPs were annotated against the CADD (version 1.3) whole genome metrics (deleteriousness metrics) [51] for all missense variants using *annovar* [73]. Scores were calculated for the NeuroMMSig genes of interest and normalized by their gene lengths. For 12 out of 18 genes, scores were determined. Six remaining genes couldn't be used according to missingness of mapped SNPs in the CADD deleteriousness metric.

3.2.3 Burden and Kernel-machine based Scores

Burden Scores

A less sophisticated approach of scoring gene level pathogenicity can be defined by calculating so called "Mutational Loads" for each gene of interest. The mutational load is defined by the ratio of the number of effective alleles (mutated/alt alleles) divided by the overall sum of mapped alleles for each corresponding gene. Assuming that N SNPs map on the k'th gene, the mutational load M for the j'th subject of the k'th gene will be calculated as follows:

$$M_j k = \sum_{i=1}^{N} \frac{G_i}{N_k} \tag{3.2}$$

The same quality control filters for calculating GenePy scores were applied a priori to the VCF file for purposes of comparison. Filtering on SNPs from the mapping yielded mutational load scores for all 18 genes then.

[6] https://github.com/UoS-HGIG/GenePy

Kernel-machine based Scores

To apply kernel-machine based testing on gene level, the tool *SKAT* [74] was used. By specifying the optimized SKAT (SKAT-O) method [47], we optimize generalized SKAT by combining both burden and the non-burden sequence kernel association test (generalized SKAT). Given the SNPs to gene mapping, individual score test statistics were aggregated for a given SNPs set and an overall p-value in an upcoming association test via linear regression was computed. In order to establish a contrast between rare and common variants, the individual squares of the test statistics were weighted based on their minor allele frequencies (MAF's). The type of kernel was set to linear weighted. Given G as the genotype matrix (of the given SNP set) and W as the diagonal weight matrix, the final kernel matrix is defined as follows:

$$K = GWWG \tag{3.3}$$

3.3 Association Testing Methods

In this section, the tools and underlying models for conducting association tests for the different approaches (standard, gene-, mechanism- and overall level) will be explained. Methods that use the same model design and tool are grouped into the same paragraph (for example GenePy and Burden Scores). While the details differ between each approach, all methods generally regress each outcome separately on the level-dependent feature(s) while adjusting for the confounding effects and the sample structure in a linear model (univariate testing). Outcomes were quantile normalized according to the guidelines in standard association testing as outlined in 3.1.1. The confounding variable "Age" was scaled beforehand by removing the mean and scaling to unit variance using *scikit-learn*. To account for sample structure and potential clustering due to population

stratification among the *AETIONOMY* subjects, either a kinship matrix was computed and modeled in the random effects in a linear mixed model [75] (applies for standard association testing with *GEMMA*, gene-level testing with *lmekin*) or principal components of the genotypes were determined and modeled as further fixed effects among the confounding variables [30]. That applies for kernel-machine based testing with *SKAT*, polygenic risk score testing with *PR-sice*. P-values of the final model fits are then based on likelihood-ratio tests that test the effect of the genetic feature against the null-Model containing confounders only. Moreover, methods to control for cumulative type I error (due to multiple testing) differ between the methods and will be elaborated within the individual sections.

3.3.1 Standard Association Testing

For applying the standard association that models each variant separately, the univariate linear mixed model (LMM) utility of the *GEMMA* [28] tool package was used. According the *GEMMA* guideline, each outcome was quantile normalized . For consistency between the methods, only the 1583 PD-related SNPs from the mapping were analyzed. A centered relatedness matrix (kinship matrix) from the filtered input VCF file was computed using the *GEMMA* utility and used to model random effects in a n-vector u, drawn from joint n-dimensional multivariate normal distribution with mean 0 and fixed relatedness matrix (n = 106). The confounding variables, age, sex, alcohol, smoking, levodopa treatment, and group, were defined as covariates (W) and modeled as fixed effects. According to the documentation, *GEMMA* fits the following univariate mixed model with y as the outcome vector:

$$y = W\alpha + x\beta + u + \varepsilon \tag{3.4}$$

Here, x is the vector of PD-related SNP genotypes with β as corresponding effect sizes. ε defines the error term.

For each SNP, the alternative hypothesis $\beta \neq 0$ is tested against the null hypothesis $\beta = 0$. To account for cumulative type I error due to multiple testing, the bonferroni correction was used to correct for it. Therefore, each null hypothesis meeting the following condition, was rejected:

$$p \leq \frac{\alpha}{m} \tag{3.5}$$

with $\alpha = 0.05$ as the significance level and $m = 1583$ for the number of tests.

3.3.2 GenePy and Burden Scores Testing

GenePy and Burden scores were tested independently using the *lmekin* function of the *coxme* package in R [76]. The reason why this tool was chosen, is based on the fact that it fits univariate linear mixed models with random genetic effects straight forward by inputting a kinship matrix (similar to *GEMMA*). Each of the five outcomes was tested against the GenePy scores and Mutational load scores on gene level in an iterative fashion (only one gene per model). Scores and confounding effects were modeled as fixed effects, while the computed kinship matrix (section 3.3.1) was modeled as random effect (same way as in standard association testing). The model formula can be adapted from the standard association testing with the only difference that x is a n-vector of corresponding scores from the GenePy or Mutational load method. Accordingly, a bonferroni correction to avoid cumulative Type 1 error was applied as well (See formula 3.5).

3.3.3 Kernel-Machine based Testing

The above describes SKAT-O scores were tested in a simple linear regression model which utility is included in the *SKAT* package. Confounding variables and the first three principal components were fitted into the model to account for sample structure in a similar fashion as in the polygenic risk score analysis as describes in the next section. The number of PCs had to be reduced from ten to three because of computational problems. This accounts for an explained variance of 85 %. To control for cumulative type I error, an empirical p-value via resampling was computed in the same way as for the polygenic risk score analysis (section 3.3.4, formula 3.6).

3.3.4 Polygenic Risk Score Testing

As outlined in section 3.2.1, polygenic risk scores were calculated using *PRSice-2*. Fitting those scores in linear regression models was carried out by *PRsice-2* after the risk score computation (included feature of the package). Because *PRsice-2* does not support a linear mixed model utility to fit a kinship matrix (in contrast to *GEMMA* and *lmekin*), sample structure had to be modeled by including the first ten principal components (PCs) [77] from the genotypes as fixed covariates. The principal component analysis was carried out using *Plink*. The first ten PCs cover ∼73 % of the explained variance and will be illustrated in figure 3.4.

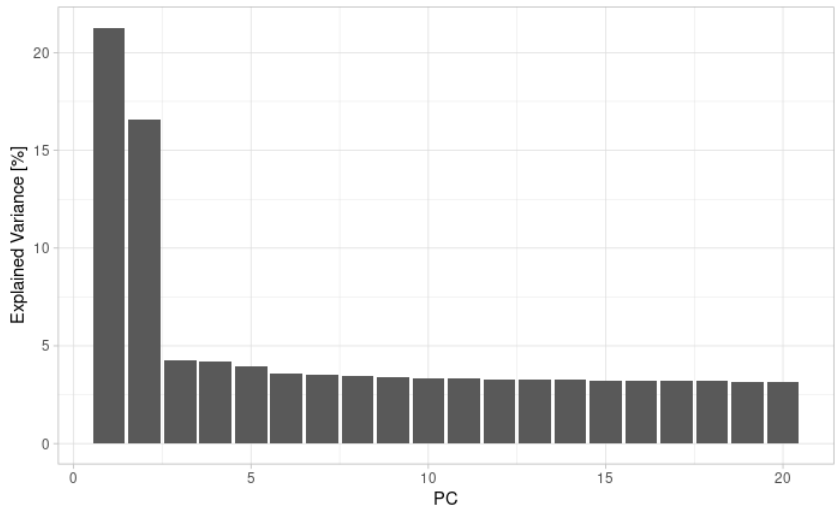


Figure 3.4: Proportion of explained variance of the first 20 principal components.

In consistency with the remaining methods, confounding variables were fitted into the model as well. The linear regression models for the risk scores on overall level (= determining a single score for each individual subject) were built on multiple p-value thresholds from the GWAS base data based on p-values for association (not to be confused with p-values of the model fit). The lower bound p-value was set to p=5e-08. With a step size of 5e-05, only SNPs that match those thresholds (p-value $\leq$ threshold) will be included in the computation of the score and therefore included in the current regression model. The upper bound was set to p=0.5. After this iterative procedure, the best model fit was selected by testing the alternative hypothesis $\beta \neq 0$ against the null hypothesis $\beta = 0$ for the effect sizes β of the score. Furthermore, an empirical p-value based on a resampling test was computed in order to control the cumulative type I error and account for overfitting. After obtaining the best p-value threshold and the p-value of the best model fit P_o, the phenotype was randomly shuffled and the analysis was repeated, resulting in a best p-value P_{null} of the model fit. That step has been repeated N times. Setting N to 10 000, the empirical

p-value was then computed as follows:

$$Empirical - P = \frac{\sum_{n=1}^{N} I(P_{null} < P_o) + 1}{N + 1} \tag{3.6}$$

I is defined as the indicator function, returning 0 if the inner condition ($P_{null} < P_o$) is true.

The polygenic risk score analysis on mechanism level with *PR-set* was conducted in the same fashion except for not including the p-value thresholding while fitting a model. Because it is unclear whether the set is associated with the phenotype when the best-threshold contained only a small amount of SNPs mapping on the gene set (=mechanism), p-value thresholding was omitted. All SNPs that are present in the base data and map on PD-related gene sets which map on a particular mechanism were included. Mapping on PD-related genes was defined by the direct physical location on a gene and indirect mapping through linkage disequilibrium (LD) with a SNP that is physically located on a gene. The LD threshold is set to $LD > 0.8$.

3.4 Synthetic Patient Generation

3.4.1 VAMBN

Due to the limited sample size of the *AETIONOMY* data set, the machine learning approach VAMBN (Variational Autoencoder Modular Bayesian Network) [12] was applied to generate larger sample size cohorts for demographic, genotype and derived score (GenePy and polygenic risk scores) data based on the original data.

Association methods on gene-, mechanism- and overall level were applied to the synthetic data sets in order to conclude results on larger sample sizes

and compare those to the original data set containing only 106 samples.

The VAMBN workflow follows four major steps:

1. Definition of modules that summarizes original input features
2. Encoding of modules into lower dimensional latent distributions (multivariate Gaussians) via a variational autoencoder for heterogeneous and incomplete data (HI-VAE)
3. Structure learning of a modular bayesian network (MBN) between the encoded modules
4. Simulating virtual patients by drawing samples from the MBN and and decoding into the original feature space through the HI-VAE

In the first step, modules were defined. They contain sets of original features that can be grouped together in accordance with the study design. The following table defines each module's composition.

Table 3.3: *Modules and their original feature composition*

Modules	**Features**
Outcomes	Campestanol, Campesterol, Stigmasterol, Sitostanol, Sitosterol
Demographics	Age, Gender, Alcohol, Smoking
GenePy Scores	APOE, CACNB2, CASP8, MAPT, LRRK2, FDFT1, SLC41A1
Polygenic Risk Scores	For each of the five outcomes
Mechanism Polygenic Risk Scores	Scores for eight mechanisms
Genotype Matrices	SNPs genotypes for 13 mechanisms

The features *Group* and *Levodopa* were used as standalone modules in the MBN. The, so called, genotype matrices contain mapped SNPs on the NeuroMMSig mechanisms (see section 3.1.2) with their corresponding allele values encoded into 0 (0/0), 1 (1/0, 0/1) and 2 (1/1). Initially all 39 Mechanisms from the mapping were considered to be encoded into 39 modules, but the training loss turned out to be too high and didn't converge properly (see Supplementary Material for all training stats). 13 mechanisms, covering overall 644 SNPs, could be successfully encoded into 13 genotype modules. For consistency in the structure learning process, only those *GenePy* scores of which underlying genes map on the 13 mechanisms were taken into account and combined into one module. Accordingly, polygenic risk scores for eight out of the 13 mechanisms were encoded into one module. For the remaining five mechanisms, no risk score could be calculated due to the lack of GWAS summary information in the used base data set for the mapped SNPs (see section 3.2.1).

For each module, a HI-VAE was trained using the python package *tensorflow* [78]. Hyperparameter optimization involved a grid search over the following set of parameters:

- y-dimension (latent space): 1,2,3
- Learning rate: 0.05, 0.01, 0.001
- Mini-batch size: 16, 32
- Activation function: none, ReLu, Tanh

Via a 3-fold cross-validation, each parameter combination was evaluated using the reconstruction loss as objective function.

In advance of the structure learning step, causal conditions on possible edges between modules had to be determined in order to reduce the search space, as MBN structure learning grows super-exponentially with the number of nodes [79]. According to the constraints imposed in [12], a black list which determines edges between modules that violate against these constraints was built. Hence, those edges will be removed from the possible

search space. In principle, all possible edges influencing the demographics module were blacklisted. Furthermore derived score modules including the *GenePy*, both polygenic risk score modules and the outcomes module were restricted to influence all genotype matrices. Also, the Levodopa to Group edge was blacklisted. The final blacklist counts overall 72 edges.

In the structure learning process, a modular bayesian network was built applying the R package *bnlearn* [80]. The learning algorithm was set to Hill-Climbing with a Bayesian Information Criterion (BIC) score function for mixed data ("bic-cg", mixed categorical and normal variables). The final MBN counts 17 edges (see Supplementary Material for the edges).

In a final step, independent virtual patient cohorts were drawn from the MBN and the encoded modules were decoded by the HI-VAE into their original feature space. Overall, three virtual patient cohorts with increasing sample sizes were created (original sample size $n = 106$).

- 2x n: 212 samples
- 5x n: 530 samples
- 10x n: 1060 samples

The final virtual patient cohorts were compared in terms of their marginal distributions and correlation structure (distribution of correlation coefficients) to the real data.

3.4.2 sim1000G

Virtual SNP genotypes from the simulated VAMBN model were compared to simulated genotypes using the tool *sim1000G* [13]. The exact set of SNPs that were simulated in the mechanisms framework in VAMBN were chromosome-wise simulated with *sim1000G*. This has been done by filtering the original VCF file on the 644 SNPs and splitting this file chromosome-

wise using *vcftools*. Chromosome reference files (genetic maps) were automatically downloaded from an online database on GitHub [7] when running *sim1000G*. Using the function *generateUnrelatedIndividuals()*, genotype cohorts of 212, 530 and 1060 individuals, consistent VAMBN simulated genotypes, were simulated and the SNP genotype matrices were retrieved applying *retrieveGenotypes()*.

SNP genotypes from VAMBN and *sim1000G* were compared in terms of their marginal distributions, inner correlation structure among all SNPs and Kullback-Leibler divergency to the real ones.

3.5 Plots

Plot and diagrams were generated using the python libraries *seaborn* [81] and *matplotlib* [82].

3.6 Statistical Hypothesis Testing

Statistical hypothesis tests including the Wilcoxon rank and t-test were carried out using the *statsmodel* package in python [60].

[7] https://github.com/adimitromanolakis/geneticMap-GRCh37

4 Results

4.1 Synthetic Patient Generation

In this section, results of the synthetic patient generation will be presented. Applying the VAMBN worklow, virtual patient cohorts for each module including demographics, outcomes, derived genetic scores (polygenic risk scores, *GenePy* scores) and genotypes of the SNPs were simulated. Also the two standalone variables *Group_IPD* and *Levodopa* are included. For purposes of comparison to alternative approaches, SNP genotypes were additionally compared to *sim1000G* simulated ones.

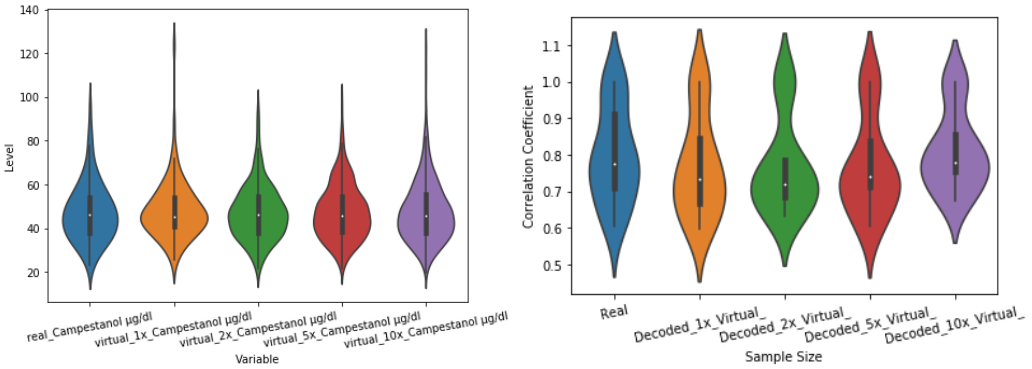
Virtual patients will be compared to the real patients in terms of marginal distributions of individual variables and correlation structures. Simulated SNP genotypes in particular will be compared to the real genotypes and quality is further accessed by calculating the Kullback-Leibler divergency between real and simulated genotypes for both approaches (VAMBN and *sim1000G*).

Because the number synthetic features is simply to vast to display all of them in this section, only selected ones will be presented. Omitted results in this section can be received from the Supplemental Material or compiled by request.

4.1.1 Outcomes, Demographics & Genetic Scores

The following violin plots display the marginal distribution of the real and simulated outcome Campestanol, decoded from the module *Outcomes* for different sample size cohorts. The Pearson correlation between all outcomes was accessed and distributions of correlation coefficients are shown. For validation purposes, the relative error of the Frobenius norm

of each correlation matrix is accessed by dividing the norm of the difference between real and simulated by the norm of the real matrix ($Rel.Error = \frac{Frob(Real-Simulated)}{Frob(Simulated)}$). Respective results are shown in the table underneath.



(a) (b)

Figure 4.1: **(a)** Marginal distributions of Campestanol for real and simulated data. Sample sizes for virtual patients range from 106 (1x) to 1060 (10x). **(b)** Distribution of Pearson correlation coefficients between all marginal variables of the module *Outcomes* for real and virtual data.

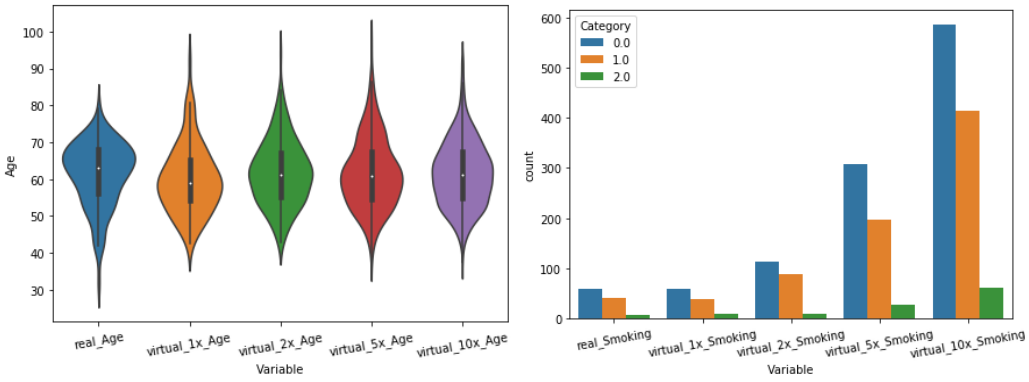
Table 4.1: *Frobenius Norms and relative errors of Pearson correlation matrices from the decoded Outcomes* variables.

Type	**Norm**	**rel. Error**
Real	4.04	-
Virtual 1x	3.90	0.099
Virtual 2x	3.88	0.097
Virtual 5x	3.90	0.106
Virtual 10x	4.09	0.075

Synthetic patient cohorts reveal similar distributions along each sample size cohort in terms of mean and quartile ranges (displayed within the violin plots). The same accounts for the correlation structures as seen in (b). The relative error of the Frobenius norm is comparatively small with a minimum of $\approx$ 7.5% for the 1060 patients cohort and $\approx$ 10.6% at most regarding the virtual patient cohort of 530 individual (5x).

In the following plots, two selected features of the *Demographic* module originally consisting of the variables *Sex*, *Alcohol*, *Smoking* and *Age* are

shown. Except Age, those variables are categorical and distributions will be therefore presented by the use of count plots. Correlation structure was accessed using the Spearman's rank correlation, because of its better performance on the associations of ordinal data in contrast to Pearson.



(a) (b)

Figure 4.2: Marginal distributions of the real and decoded Demographics module variables **(a)** Age and **(b)** Smoking. Sample sizes for virtual patients range from 106 (1x) to 1060 (10x).

The following table represents the frequencies of ordinal values in the smoking variable (b).

Table 4.2: *Frequencies of data points in the smoking variable with respect to their virtual cohort.*

Value	**Real**	**Virtual 1x**	**Virtual 2x**	**Virtual 5x**	**Virtual 10x**
0 (never)	0.552	0.556	0.537	0.581	0.551
1 (ex)	0.390	0.367	0.415	0.369	0.389
2 (current)	0.057	0.075	0.047	0.049	0.058

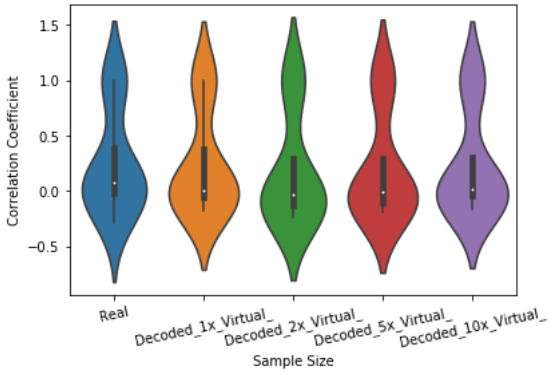


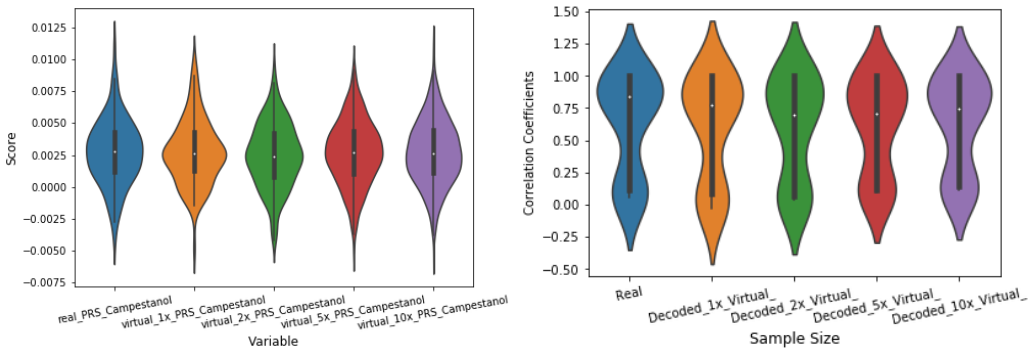
Figure 4.3: Distributions of Spearman rank correlation coefficients between all marginal variables of the module "Demographics" for real and virtual data.

Table 4.3: *Frobenius Norms and relative errors of Pearson correlation matrices of variables from the decoded* ***Demographics*** *module.*

Type	**Norm**	**rel. Error**
Real	2.07	-
Virtual 1x	2.02	0.209
Virtual 2x	2.05	0.254
Virtual 5x	2.03	0.230
Virtual 10x	2.01	0.216

Similar to the outcomes module, virtual demographic variables are as well realistic in terms of their marginal distributions and correlation structure along each cohort. Relative errors are on the average slightly higher than for the outcomes module (Table 4.3), but do remain stable with increasing sample size.

To depict the realistic structure of simulated genetic scores, the following figure illustrates the distributions of the polygenic risk score designed for the outcome *Campestanol* in addition with the distributions of Pearson correlation coefficients of the entire polygenic risk score module. See the upcoming table for the relative error of the correlation matrices.



(a) (b)

Figure 4.4: **(a)** Marginal distributions of the real and decoded polygenic risk scores for the outcome Campestanol. **(b)** Distribution of Pearson correlation coefficients between all marginal variables of the polygenic risk score module for real and virtual patients

Table 4.4: *Frobenius Norms and relative errors of Pearson correlation matrices of variables from the decoded* ***Polygenic Risk Scores*** *module.*

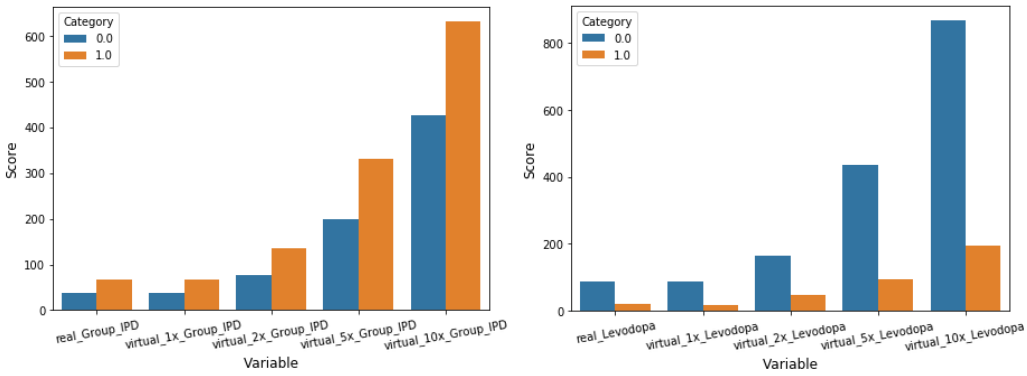
Type	**Norm**	**rel. Error**
Real	3.66	-
Virtual 1x	3.58	0.098
Virtual 2x	3.53	0.110
Virtual 5x	3.52	0.104
Virtual 10x	3.59	0.089

The distributions turn out to be realistic and reveal uniform correlation structures along every cohort. The relative errors are relatively low with a minimal value of 8.9 % for the 1060 virtual patients cohort as seen in Table 4.4.

In an analogous manner, realistic results were obtained from the remaining genetic scores based on *GenePy* and polygenic risk scores on mechanisms level and can be reviewed in the Supplementary Material chapter.

4.1.2 Standalone Features

Two categorical features, the *Group_IPD* and the *Levodopa* treatment variable, were not combined into a module by the HI-VAE, but modeled standalone in the Bayesian network. Both are binary features. So, the following count plots illustrate the distributions for the different virtual cohorts. Respective frequencies of data points are shown in the respective tables for a better comprehension. Both features are realistic in each cohort in comparison to the real data.



(a) (b)

Figure 4.5: Distributions of the real and decoded standalone variables **(a)** Group and **(b)** Levodopa. Sample sizes for virtual patients range from 106 (1x) to 1060 (10x).

Table 4.5: *Frequencies of data points in the Levodopa variable with respect to their virtual cohort.*

Value	**Real**	**Virtual 1x**	**Virtual 2x**	**Virtual 5x**	**Virtual 10x**
0 (No)	0.811	0.830	0.783	0.822	0.816
1 (Yes)	0.189	0.170	0.217	0.178	0.184

Table 4.6: *Frequencies of data points in the Group variable with respect to their virtual cohort.*

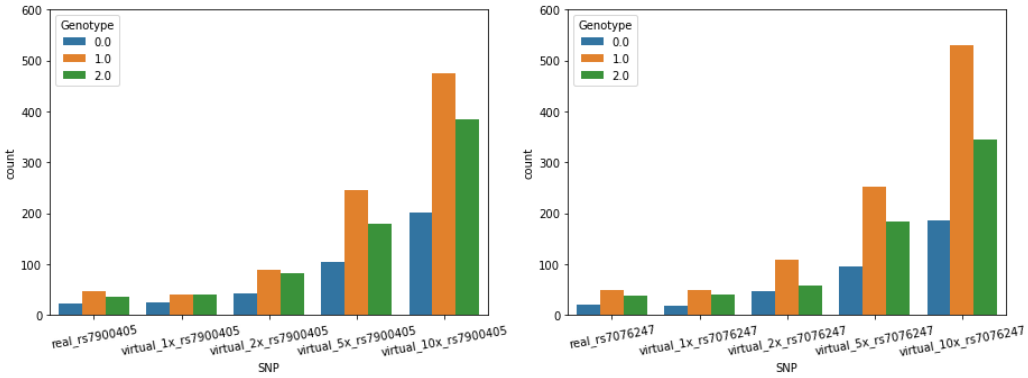
Value	**Real**	**Virtual 1x**	**Virtual 2x**	**Virtual 5x**	**Virtual 10x**
0 (control)	0.368	0.368	0.359	0.374	0.403
1 (IPD)	0.632	0.632	0.641	0.626	0.597

4.1.3 Genotypes

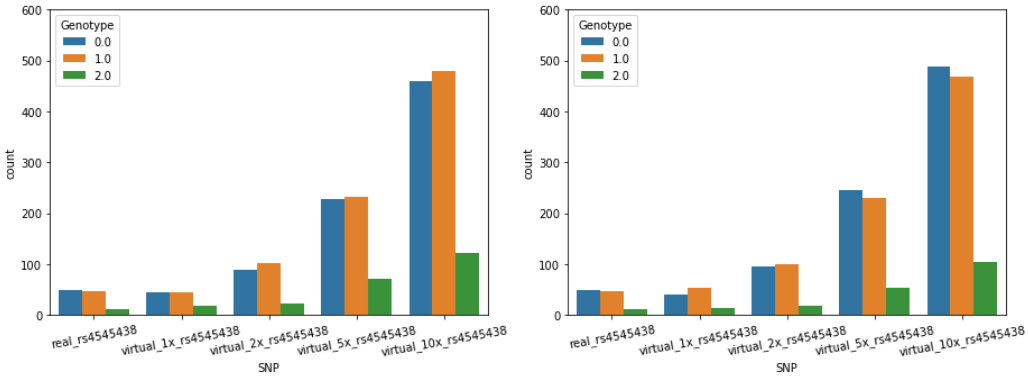
Within the VAMBN approach, 13 genotype matrices modules were generated. Modules are typical PD-related mechanisms from NeuroMMsig with underlying SNPs sets that map onto those mechanism via PD-related genes (Details of the mapping are shown in the Material & Methods section). These modules cover an overall amount of 644 PD-related SNPs. To demonstrate the performance of the simulation, simulated SNPs of the virtual patients cohorts (1x, 2x, 5x and 10x) for the **Interleukin signaling pathway** will be shown and compared to real patients.

While VAMBN simulates SNPs within the framework of underlying mechanisms (multivariate approach), *sim1000G* takes a different approach and rather models simulated genotypes in the framework of genomic regions (chromosome-wise). We contrast the results of both approaches by comparing marginal distributions of SNPs that fall onto the Interleukin mechanisms plus access the correlation structures of the simulated SNPs. Further, the Kullback-Leibler divergence will be applied to demonstrate levels of conformity between real and simulated marginal SNPs distributions for all 644 SNPs.

Ten SNPs fall onto the Interleukin signaling mechanism. In the following, marginal distributions of two selected SNPs genotypes (rs7900405 and rs4545438) are shown via count plots. Their corresponding genotype frequencies (relative amount of genotype 0,1 and 2) are listed in the table underneath.



(a) rs7900405 **left** VAMBN, **right** sim1000G



(b) rs4545438 **left** VAMBN, **right** sim1000G

Figure 4.6: Simulated genotypes of two SNPs mapping onto the Interleukin signaling subgraph for virtual patients cohorts with samples sizes of 106 (virtual_1x), 212 (virtual_2x), 530 (virtual_5x) and 1060 (virtual_10x) patients.

Table 4.7: *Frequencies of genoytypes for* ***rs7900405*** *in comparison between VAMBN and sim1000G*

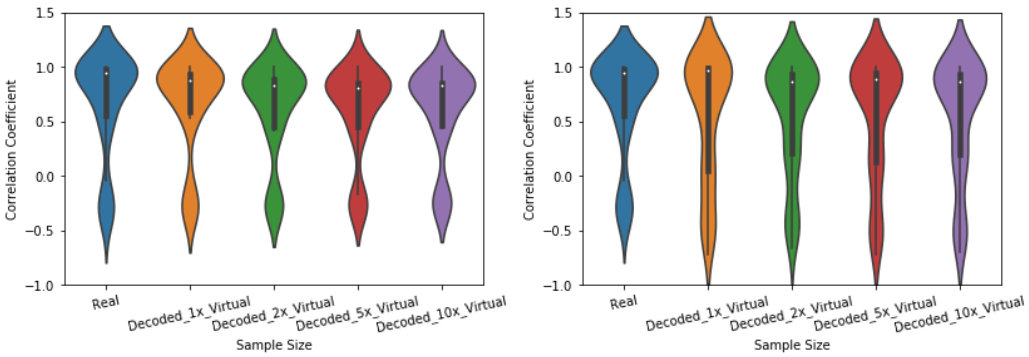
VAMBN/sim1000G	**Real**	**Virtual 1x**	**Virtual 2x**	**Virtual 5x**	**Virtual 10x**
0	0.216	0.226 / 0.169	0.198 / 0.216	0.198 / 0.200	0.189 / 0.180
1	0.443	0.386 / 0.471	0.419 / 0.537	0.462 / 0.483	0.447 / 0.504
2	0.339	0.386 / 0.358	0.382 / 0.245	0.339 / 0.316	0.363 / 0.315

Table 4.8: *Frequencies of genoytypes for **rs4545438** in comparison between VAMBN and sim1000G*

VAMBN/sim1000G	Real	Virtual 1x	Virtual 2x	Virtual 5x	Virtual 10x
0	0.462	0.415 / 0.367	0.415 / 0.448	0.430 / 0.464	0.433 / 0.461
1	0.433	0.424 / 0.500	0.476 / 0.466	0.437 / 0.433	0.451 / 0.441
2	0.103	0.160 / 0.132	0.108 / 0.084	0.132 / 0.101	0.115 / 0.097

Both simulation approaches return realistic results. The variant rs7900405 can be considered as a more common SNP, while rs4545438 is rather rare. VAMBN simulated genotypes for rs7900405 (Figure 4.6a) turn out to be closer to the real ones and tend to reveal more consistent frequencies along each sample size cohort, while *sim1000G* genotypes tend to fluctuate more. Simulated genotypes of the rare variant rs4545438 (b) turn out to be slightly closer to the real one by the *sim1000G* approach for higher sample sizes cohorts (5x and 10x).

The following figure compares the Spearman rank correlation structures of the ten SNPs mapping onto the Interleukin signaling subgraph between VAMBN and *sim1000G* simulated cohorts.



(a) (b)

Figure 4.7: Distribution of Spearman rank correlation coefficients between **(a)** VAMBM **(b)** sim1000G simulated genotypes

Both approaches reveal stable correlation structures along each virtual sample patient cohort. The VAMBN approach appears to be more consistent with the real data. That can be quantified by the relative error of the

Frobenius norms of the Spearman's rank correlation matrices as depicted in the following table.

Table 4.9: *Frobenius Norms and relative errors of Spearman's rank correlation matrices of the ten SNPs from the decoded the* ***Interleukin signaling subgraph*** *module.*

VAMBN / sim1000G	**Norm**	**rel. Error**
Real	8.19	-
Virtual 1x	7.89 / 7.81	0.075 / 0.310
Virtual 2x	7.44 / 7.53	0.111 / 0.223
Virtual 5x	7.24 / 7.69	0.135 / 0.253
Virtual 10x	7.28 / 7.60	0.127 / 0.235

Relative errors of *sim1000G* based correlation matrices turn out be larger along each virtual patients cohort. Most likely justified by the fact, that more common SNPs fall onto this mechanisms which multivariate structure can be captured more accurately by the VAMBN approach.

To access the quality of **all** simulated SNPs, the Kullback-Leibler divergences (KL) between the real and virtual patients cohort with sample size 106 was computed pairwise for each of the 644 variant distributions. The following figure compares the distributions of KL divergences for VAMBN vs. Real and *sim1000G* vs. Real for the 106 sample size cohorts.

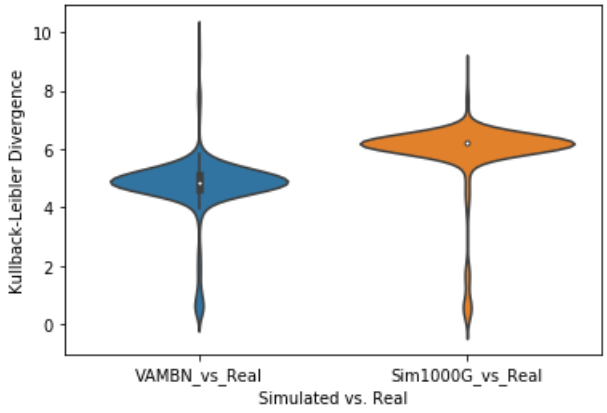


Figure 4.8: Distribution of KL divergences for VAMBN and *sim1000G* simulated variants.

On the average, the KL divergences for the VAMBN approach ($mean = 4.74$) are lower than for the sim1000G approach ($mean = 5.86$), speaking for a more realistic and closer simulation of SNPs applying VAMBN (Two-sampled t-test: p=9.83e-94, reject null hypothesis that means have identical expected values with significance threshold $\alpha = 0.05$). Still, the KL divergence in the VAMBN vs. real comparison has a higher maximum of 9.80 in contrast to 8.54 in the *sim1000G* vs. real. Applying a Wilcoxon rank test, it can be further quantified that the two distribution do not stem from the same distribution. Setting the significance threshold to $\alpha = 0.05$ and calculating a p-value of $2.8e - 87$, the null hypothesis can be rejected and hence, both distributions are significantly different.

4.2 Association Testing

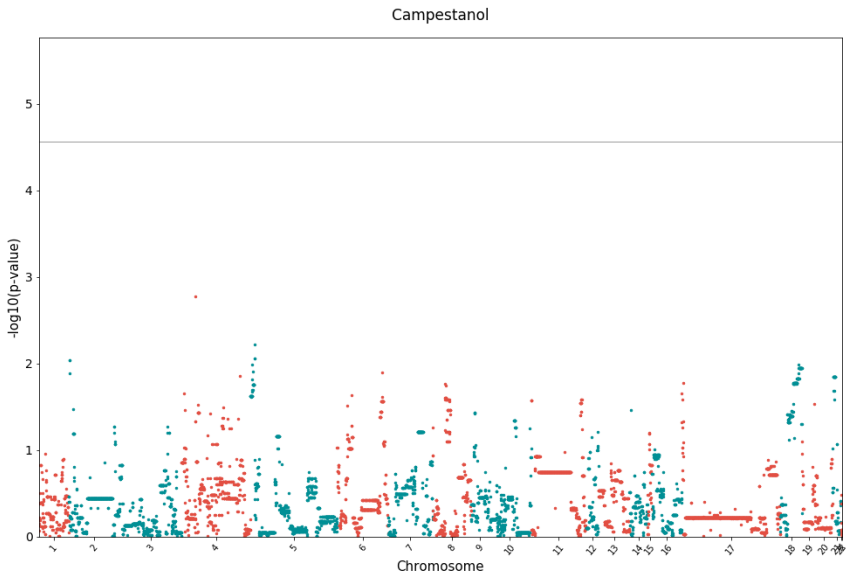
In this section, results of the different association testing approaches will be presented. Starting with the standard approach that tests individual SNPs, to more sophisticated techniques on gene, mechanism and genome-wide level. Since the number of models is vast, the focus will be on the single outcome *Campestanol*. Models predicting *Stigmasterol*, *Campesterol*, *Sitostanol* and *Sitosterol* can be obtained from the Supplemental Material if interested.

The approaches were applied to both real patients (sample size 106) and VAMBN based simulated patients cohorts with sample sizes of 212, 530 and 1060.

4.2.1 Standard Model

The results of the standard models with *GEMMA* which test the effect sizes of each of the individual markers (PD-related SNPs) separately predicting

Campestanol are shown in the following Manhattan plots. On the X-axis, markers in order of their chromosome membership are displayed (1-22). The negative logarithm of the association p-value (corrected for sample structure and confounders) for each marker is displayed on the Y-axis. The Bonferroni corrected significance level with $\alpha = 0.05$ and m = 1583 (number of tests) is set to $\alpha_{Bonferroni} \approx 2.8e-05$ which is illustrated as the grey line in the plots (-log($\alpha_{Bonferroni}$) $\approx$ 4.6). Markers passing the significance threshold with a p-value below $\alpha_{Bonferroni}$ (located above the grey line) can significantly be associated to the respective outcome.



(a)

Figure 4.9: *GEMMA* results for the outcome Campestanol for the real patients. The significance threshold ($\alpha_{Bonferroni} \approx 2.8e-05$) is labeled as grey line.

Clearly, none of the 1583 markers can be significantly associated to Campestanol. For instance, the lowest p-value model yields $1.7e-03$ with a negative logarithm of $\approx 2,77$, far away from the significance threshold $p = 2.8e-05$. Model results for the remaining outcomes (Campestero, Sitostanok,

Sitosterol and Stigmasterol) reveal similar results with no major improvement in terms of model fits and p-values (Plots are shown in the Supplemental Material).

Applying the standard approach to virtual cohorts could not be achieved with *GEMMA* due to the shortcoming of computing a kinship matrix from genome-wide synthetic genotypes. That would require to simulate genome-wide variants which was not feasible within the applied simulation framework.

4.2.2 Gene Level Models

Association testing methods on gene level comprise *GenePy*, Burden testing (=Mutational Load) and kernel-machine based testing with the SKAT-O method. Their results will be combined in this section. *GenePy* and mutational load scores were modeled with the *lmekin* linear mixed model utility, while kernel-machine based testing was carried out using the *SKAT* package which includes an ordinary linear regression approach (detailed description provided in the Material & Methods section). The following bar plot summarizes p-values from model fits of the corresponding gene scores predicting Campestanol corrected for sample structure and confounders. Overall, 18 PD-related genes were tested with differing numbers of SNPs mapped (Details see Material and Methods section). The Bonferroni corrected significance threshold with $\alpha = 0.05$ and $m = 54$ (18 genes tested 3 times each) is set to $\alpha_{Bonferroni} \approx 1e-03$.

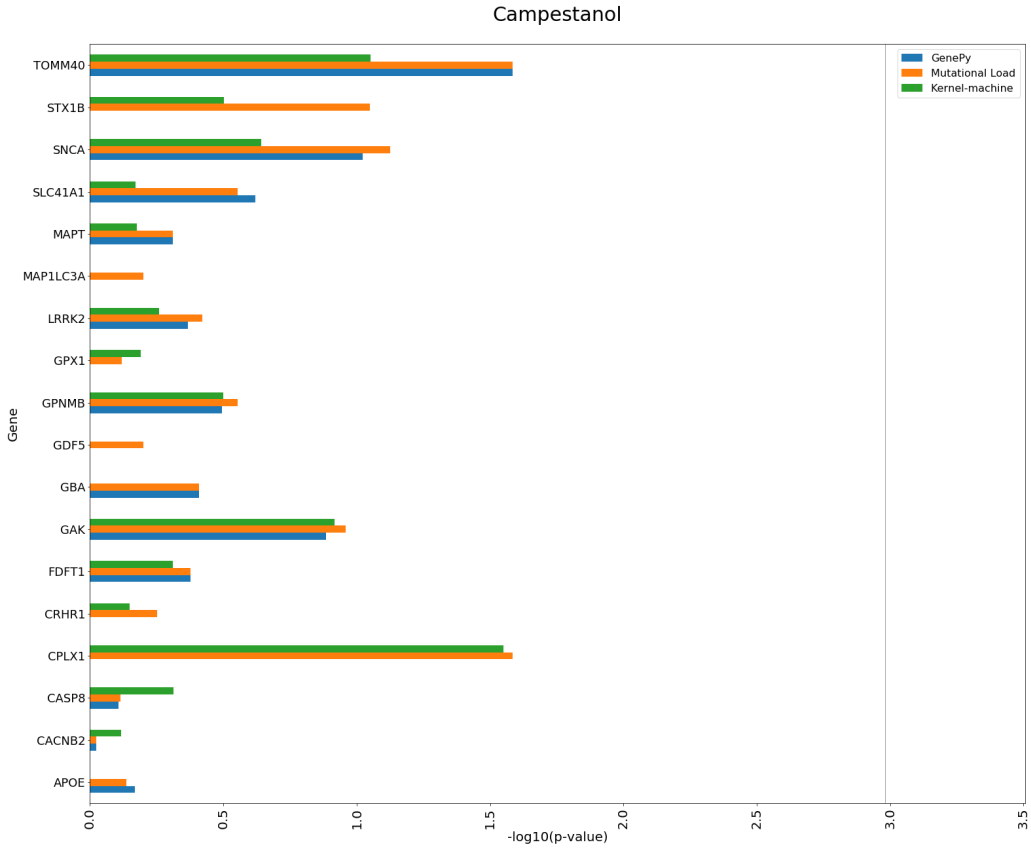


Figure 4.10: Gene-level association using *GenePy*, mutational load and kernel-machine based scores (applying *SKAT* predicting Campestanol in the real patients data. The significance threshold ($\alpha_{Bonferroni} \approx 1e-03$) is labeled as grey line.

Six (*CPLX1*,*CRHR1*,*GDF5*,*GPX1*,*MAP1LC3A*,*STX1B*) out of 18 genes could not get scored with *GenePy*, because of missingness in the deleteriousness metric (see Material & Methods). Moreover, p-values do not reach the significance threshold (grey line) for any of the genes. The lowest p-value of 0.026 could get obtained from the *TOMM40* gene with *GenePy* and Burden testing, while SKAT-O reveals a p-value of 0.087. *CPLX1* has no GenePy score, but reveals a p-value of 0.026 in the Burden test and 0.028 applying SKAT-O. Regarding the empirical p-value that has been calculated within the *SKAT* pipeline, none of the genes show significant associations.

Regarding models based on the virtual patient cohorts, eight out of 18 genes were simulated. Realistic synthetic *GenePy* scores, mutational load scores and kernel-machine based scores derived from simulated SNP genotypes were built. Linear regression models were built including the simulated confounders. None of the genes reveal a stronger trend in association for any larger sample size cohort.

Similar results were observed predicting the remaining four phytosterols (for details see Material and Methods).

4.2.3 Polygenic Risk Score Models

Genome-wide Models

The PRS analysis on genome-wide level by *PRsice-2* is shown in the following bar plot. The bars present model fits at broad p-value thresholds (p-values from GWAS summary statistics, no model fits) and the highlighted threshold for the best model fit. High resolution plots can be obtained from the Supplement. P-values of the model fits are displayed on top of each bar. They refer to the effect size of the PRS corrected for confounding effects and sample structure. The explained variance R^2 of the PRS in the regression model is displayed on the y-axis.

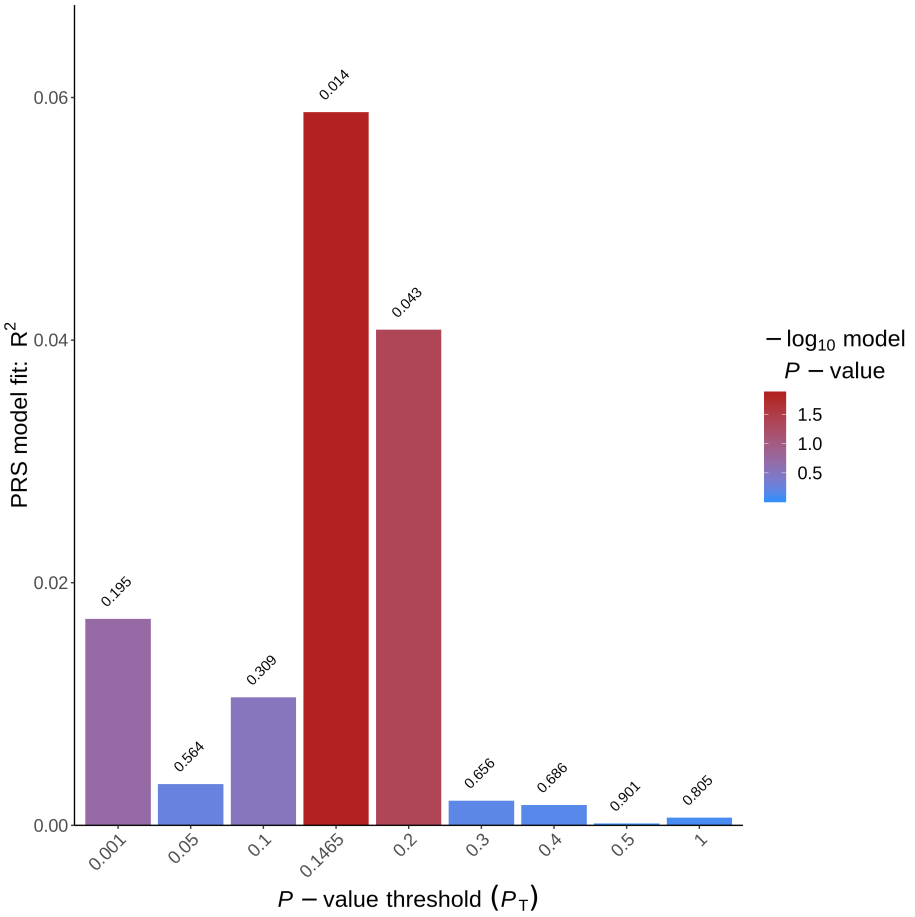


Figure 4.11: Bar plots showing regression model fits at broad p-value thresholds for the PD polygenic risk score predicting Campestanol

The p-value threshold for the best model fit predicting Campestanol is 0.1465 and leads to a best model fit p-value of 0.014. 1524 SNPs were included in the calculation of the respective PRS. The explained variance R^2 by the PRS measures 5.9%. The full model including confounders and principal components is 23.2%.

The empirical p-value that ensures a family-wise error rate (FWER) control is 0.29. Hence this association can not be stated as significance if we control for cumulative type I errors due to multiple testing with a significance threshold of $\alpha = 0.05$.

Synthetic PRS are based on the score that result in the best model fit as

described above. The linear regression results on synthetic PRS did not turn out significant as well. For the 1060 virtual patients cohort, the model fit of the PRS measures a R^2 of 6.3%. The corresponding p-value is 0.051 (corrected for synthetic confounders).

Model results for the remaining outcomes reveal much higher p-values with similarly low R^2 performances. Hence, no significant associations can be concluded for both real and simulated data with more samples as well.

Mechanism Level Models

The PRS analysis on mechanism level will be presented in a similar fashion as in the genome wide PRS analysis. The following bar plot presents the best ten mechanisms ordered by their strength of association predicting Campestanol in terms of the p-value of the model fit (lowest p-value displayed on the bottom). Since the mechanisms based analysis does not implement the p-value thresholding only one model including all SNPs fallen on the respective mechanism was built for each mechanism.

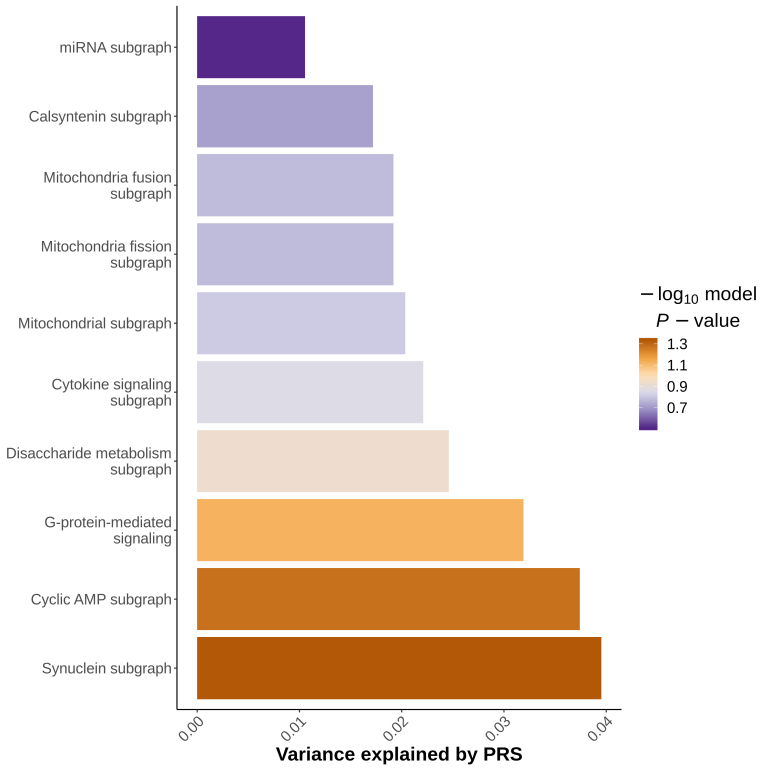


Figure 4.12: *PR-set* regression results predicting Campestanol of the 10 best mechanism model fits. A p-value below 0.05 could be obtained The Synuclein subgraph ($p = 0.046$)

Polygenic risk scores for 33 out of 39 mechanisms from the mapping could be calculated. For the remaining mechanisms, mapped-on SNPs were not available in the base data. No more than at most eleven SNPs were considered in calculation of the mechanism based scores which is as well due to the unavailability in the base data. The majority of the mechanisms contains ≤ 5 SNPs in their score composition. The PRS of the best performing mechanism Synuclein subgraph takes 3 SNPs into account. The respective model fit (best performing PRS, see figure 4.12) reveals a p-value of 0.046 with a R^2 of 3.9%. The full model measures a R^2 of 21.2%. In consistency with the PRS analysis on overall-level, empirical p-values were additionally computed. Regarding the best performing mechanism, Synuclein subgraph, an empirical p-value measures 0.18 and therefore no significant association exists when we control for cumulative type I errors due to mul-

tiple testing.

In the synthetic patient generation, eight realistic mechanism based polygenic risk scores including the Synuclein subgraph were simulated. While adjusting for confounding effects, none of those reveal trends to stronger associations to Campestanol for any of the virtual cohorts.

For the remaining outcomes, no significant association could be observed as well.

5 Conclusion

In a way, this project can be separated into two major subjects. Firstly, association testing for a deeper investigation and understanding of the debated cholesterol-PD association from a genetic perspective and secondly, the simulation of realistic patients cohorts.
Despite the fact that the latter stem from the necessity of having larger sample size cohorts to obtain meaningful results in association testing in the first place, the simulation of virtual patients process can still be seen as a subproject on its own. Therefore achievements and limitations for both association testing and patients simulation will be discussed separately and as a whole with respect to the entire project.

The synthetic patient generation with VAMBN worked out fine. Virtual patients are realistic in terms of their feature distributions and inner correlation structure among the modules. Worth mentioning, VAMBN has not been applied to generate more virtual samples than inputted real samples until this project. Furthermore, VAMBN generated SNP genotypes outperform the simulated genotypes based on the resampling-approach realized in *sim1000G* with respect to the data used in this project. The VAMBN approach allows for simulating SNPs genotypes within the framework of biological mechanisms and is not necessarily bound to the constraints of pre-defined genomic regions as it is the case in *sim1000G*. The ability to capture the multivariate structure of SNPs and their covariances unlimited by genetic regions makes this approach unique. Phenotype and SNP genotype data is set into a probabilistic model within the Bayesian network and therefore, the conditional dependencies between SNP genotypes and phenotypes modules are set further into relation. That ensures a simulation of features with respect to other features that might influence each other and therefore, captures the biological relation of phenotype and genotype data. In contrast to established phenotype simulation tools like *PhenotypeSimulator*, VAMBN can authentically capture the underlying data structure as it is, which allows for upstream analysis with higher sample sizes. One has

to mention that tools like the *PhenotypeSimulator* are designed for a more flexible data simulation including the customization of genetic variants, infinitesimal genetic effects (=population structure) and covariates such as confounding effects that are combined into the final phenotype. Therefore, a direct comparison of VAMBN to those tools is not reasonable. The features of VAMBN rather extend the the current possibilities that established tools offer.

On the average, marginal distributions of all 644 analyzed SNPs tend to be closer to the original genotypes than in *sim1000G* as presented by comparing the Kullback-Leibler divergences. Still, regarding computational power, this approach lacks efficiency when it comes to high numbers of SNPs. While *sim1000G* shows good performances simulating 1000 of SNPs at a time, VAMBN is up to this point limited by too many SNP combined within a module. On the average, modules containing more than 20 SNPs did not converge properly with a high loss in the encoding step of the HI-VAE. This aspect led to the drawback that no realistic population structure effect based on the effect of genome-wide variants could be generated in this context. A possible work-around could be the generation of synthetic principal components based on a kinship matrix from the real patients that has not been realized in this project. Furthermore, the simulation of rare variants turns out to be marginally more realistic by *sim1000G*. The fact that most SNPs are rather common than rare in the underlying *AETIONOMY* patients makes VAMBN still more suited than *sim1000G* in these circumstances. From a user point of view, *sim1000G* offers a more user-friendly application and additional pedigrees can be generated by modeling recombination events. VAMBN is limited by generating synthetic data based on the exact structure of the real data and no derived pedigrees of related patients can be generated. Important to mention, these further applications are not part of the intended scope for which VAMBN was implemented for initially, but could be adapted for future versions focused on SNP genotype generation eventually.

The implementation of the different association testing methods could be

successfully realized. From individual SNP testing to genome wide polygenic risk score modeling, all methods came with individual challenges and limitations regarding the computational establishment and could be solved or a suited work-around could be found. Still, none of the applied approaches could strengthen an indicated risk factor of LDL and LDL-lowering substances in the pathogenesis of PD. A major limitation in the first place is the fact that no direct LDL and HDL measurments were available. This study is therefore limited by the work-around of linking the genetic features to LDL only indirectly through LDL-lowering phytosterol measurements. Either way, the standard approach of modeling risk SNPs individually with *GEMMA* was expected to be statistically too underpowered to work out properly. Still, the failure to apply this method to a larger sample size cohort due to limitation of simulating genome-wide variants for a realistic virtual random effects computation, no serious conclusion can be allowed to be made on the standard approach with *GEMMA*. Only the application of case-control cohorts with a sufficiently large number of individuals allows for meaningful result interpretations.

The testing on gene level could not strengthen an association between genetic risk factors and LDL-lowering phytosterols, respectively. This could be due to multiple reasons. In the first place, not all known risk loci for PD were tested due to purposes of comparison between the methods. Only those genes were taken into account that are present in the NeuroMMSig knowledge base of PD related mechanisms and in addition, only those genes were tested, for which PD related SNPs were present in the set of SNPs extracted from PheWAS catalog and DisGeNet. As mentioned in the Material & Methods chapter (Section 3.1.2) only 18 out of 394 PD related genes from NeuroMMsig were tested and therefore, no conclusion can be made on the remaining ones. From a computational perspective, genes are defined as simple SNPs sets which individual SNPs project either by physical location or indirectly by eQTL or SNP-SNP interactions on certain genes. This resulted in SNP sets with counts ranging from 1 up to 748 respective SNPs that were actually used (Theoretically up to 885 but not

all SNPs were found in the VCF file). The more SNPs fall into certain sets, the higher redundant correlated effects exist in the respective sets just by statistical chance. In the applied methods on gene level, including burden, SKAT-O and *GenePy*-based tests, all SNPs within a set were taken into account and no further control method on shrinking inner SNP-SNP correlation was carried out. Although, SNP sets were designed to maximize the number of SNPs, those inner correlation effects might have led to a poorer predictive ability. A technique to reduce correlation among SNPs is called "clumping", which is an essential step in polygenic risk score analyses. If more time would be available, I would have applied that to the gene level methods as well to retain SNPs with the strongest statistical evidence only. *GenePy*-based scores in particular can be seen as the most sophisticated method on gene-level and although, scores for PD related genes discriminate PD patients from control, they are probably not suitable for association tests in this context as proposed in the paper [50] anyway.

The polygenic risk analyses turned out the be most promising on both genome-wide and mechanism level with *PRset*. It also incorporated the most advanced QC steps including clumping and p-value thresholding that has been disregarded on the above mentioned methods. The best model fit on genome-wide level regressing Campestanol measures a p-value of 0.014. Although, this could not be stated as significant after controlling for type I error for the real *AETIONOMY* data set of 106. Contrary to the expectation, on higher sample size cohorts synthesized with VAMBN, no stronger association could be achieved even for the 1060 virtual patients cohort. This could be due to multiple reason. Polygenic risk scores were directly simulated based on the real scores from the best model fit and not from underlying synthesized SNP genotypes. Although, QC steps as clumping are independent from the number of individuals, p-value thresholding can certainly be dependent. On several p-value thresholds, PRS models are being built based on the included SNPs meeting that threshold. With different sample size cohorts, p-value thresholds for best

model fits can theoretically vary and hence, the final polygenic risk score distribution. In this project, the PRS for virtual patients cohorts is directly based on the PRS that fits best the 106 samples from the *AETIONOMY* data set. A further step to investigate this, could be a PRS design based on simulated SNP genotypes by VAMBN. Of course, another cause explaining the predictive weakness of the PRS for virtual sample sizes can simply be the non-existence of statistical linkage between PD risk scores and phytosterol levels. Still, regarding the realistic feature distributions of simulated patients plus the PRS results for *AETIONOMY* data set, further investigation has to be done before holding to this trivial explanation.

The PRS analysis on mechanism level with *PR-set* turned out to be promising for some of the NeuroMMsig mechanisms as well. Still, on larger sample size cohorts, the good trends couldn't be confirmed and yet no strong association was found even for the 1060 virtual patients cohort. Interestingly, the PRS design for the mechanisms did not include p-value thresholding and SNPs in LD outside of the genetic regions were included in contrast to the PRS analysis on genome-wide level. Needless to say, fewer SNPs than in the genome-wide approach were included as well. That led to less continuous distributions over the scores and respective distributions with few, cumulated numeric values over a restricted range of values (pseudo-categorical). Although, simulated scores are realistic in terms of distribution and correlation structure (see Supplemental Material), those scores did not turn out to work for virtual patients regressing the phytosterols. Reasons for this requires further investigation. A PRS design on mechanism level based on the actual SNP genotypes rather than the scores for the 106 samples *AETIONOMY* data set for larger sample size cohorts could potentially lead to more realistic distributions and might be worthy for future steps. Although, simulated mechanisms based scores do not rely on the underlying SNP genotypes as strongly as for genome-wide scores due to no p-value thresholding. Similar to the genome-wide PRS analysis, calling the non-existence of statistical association between phytosterols and mechanism-based PD polygenic risk scores demand for

more investigation beforehand and can not be stated after this project.

A future improvement, independent from all the association testing techniques, can be a multivariate testing approach. All tests in this project are based on a univariate testing approach for each phytosterol separately. Since all phytosterols have a LDL-lowering effect, the multidimensional effect of this could be captured more accurately and enhance the predictive power to detect linear associations from SNP to genome-wide level in a multivariate fashion [83] [84]. Multivariate testing could not be realized on the real data set (106 samples) due to computational restrictions. Too many parameters with too little number of samples led to convergence issues in the linear (mixed) model designs and univariate testing was therefore realized as a more parsimonious alternative with less parameters. With higher-sample sizes cohorts, the multivariate testing approach was disregarded due to consistency reasons and purposes of comparison.

With the existence of larger sample size cohorts now, it would be interesting to apply the approaches in a multivariate framework and compare the results to the univariate models.

6 Supplementary Material

Because the amount of supplemental figures and stats would be too large in terms of size restrictions of this thesis, supplemental material can be obtained from my public GitHub repository:

https://github.com/TeaByrd/Supplemental_Material

Instruction regarding the structure are displayed in the respective Readme file.

Bibliography

[1] C. A. Davie. A review of Parkinson's disease. *British Medical Bulletin*, 86(1):109–127, 04 2008.

[2] C. Klein and A. Westenberger. Genetics of Parkinson's disease. *Cold Spring Harb Perspect Med*, 2:109–127, 2012.

[3] J. Tran, H. Anastacio, and C. Bardy. Genetic predispositions of Parkinson's disease revealed in patient-derived brain cells. *npj Parkinsons Dis*, 6, 2020.

[4] G. Xiaoyan, S. Wei, C. Ke, C. XuePing, Z. Zhenzhen, C. Bei, H. Rui, Z. Bi, W. Ying, and S. Hui-Fang. The serum lipid profile of parkinson's disease patients: a study from china. *International Journal of Neuroscience*, 125(11):838–844, 2015. PMID: 25340257.

[5] X Huang, H. Chen, and W.C. et al. Miller. Lower low-density lipoprotein cholesterol levels are associated with parkinson's disease. *Mov Disord.*, 22(3):377–382, 2007.

[6] G. Scigliano, M. Musicco, P. I. Soliveri, P. G. Ronchetti, and F. Girotti. "Reduced risk factors for vascular disorders in Parkinson disease patients: a case-control study. *Stroke*, 37(5):1184–1188, 2006.

[7] V. Rozani, T. Gurevich, N. Giladi, B. El-Ad, J. Tsamir, B. Hemo, and C. Peretz. Higher serum cholesterol and decreased parkinson's disease risk: A statin-free cohort study. *Movement Disorders*, 33(8):1298–1305, 2018.

[8] M.A. Nalls, C. Blauwendraat, C.L. Vallerga, K. Heilbron, S. Bandres-Ciga, D. Chang, M. Tan, D.A. Kia, A.J. Noyce, A. Xue, J. Bras, E. Young, R. von Coelln, J. Simón-Sánchez, C. Schulte, M. Sharma, L. Krohn, L. Pihlstrøm, A. Siitonen, H. Iwaki, H. Leonard, F. Faghri, J.R. Gibbs, D.G. Hernandez, S.W. Scholz, J.A. Botia, M. Martinez, J.C.

Corvol, S. Lesage, J. Jankovic, L.M. Shulman, M. Sutherland, P. Tienari, K. Majamaa, M. Toft, O.A. Andreassen, T. Bangale, A. Brice, J. Yang, Z. Gan-Or, T. Gasser, P. Heutink, J.M. Shulman, N.W. Wood, D.A. Hinds, J.A. Hardy, H.R. Morris, J. Gratten, P.M. Visscher, R.R. Graham, A.B. Singleton, 23andMe Research Team, System Genomics of Parkinson's Disease Consortium, and International Parkinson's Disease Genomics Consortium. Parkinson's disease genetics: identifying novel risk loci, providing causal insights and improving estimates of heritable risk.

[9] H. V. Meyer and E. Birney. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics*, 34(17):2951–2956, 03 2018.

[10] T. Günther, I. Gawenda, and K. Schmid. Phenosim - a software to simulate phenotypes for testing in genome-wide association studies. *BMC bioinformatics*, 12:265, 06 2011.

[11] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.AR. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81(2759–2772), 2007.

[12] L. Gootjes-Dreesbach, M. Sood, A. Sahay, M. Hofmann-Apitius, and H. Fröhlich. Variational autoencoder modular bayesian networks for simulation of heterogeneous clinical study data. *Frontiers in Big Data*, 3:16, 2020.

[13] Dimitromanolakis, Apostolos, Xu, Jingxiong, Krol, Agnieszka, Briollais, and Laurent. sim1000g: a user-friendly genetic variant simulator in r for unrelated individuals and family-based designs. *BMC Bioinformatics*, 20(1):26, Jan 2019.

[14] J.C. Greenland and R.A. Barker. The differential diagnosis of parkin-

son's disease. *Parkinson's Disease: Pathogenesis and Clinical Aspects [Internet]*, Chapter 6, 2018.

[15] L. Correia Guedes, T. Mestre, T. F. Outeiro, and J. J. Ferreira. Are genetic and idiopathic forms of parkinson's disease the same disease? *Journal of Neurochemistry*, 152(5):515–522, 2020.

[16] S Lesage and A. Brice. Parkinson's disease: from monogenic forms to genetic susceptibility factors. *Human Molecular Genetics*, 18(R1):R48–R59, 04 2009.

[17] A.C. Belin and M. Westerlund. Parkinson's disease: a genetic perspective. *FEBS J.*, 275(7):R48–R59, 04 2008.

[18] D.G. Hernandez, X. Reed, and A.B. Singleton. Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. *J Neurochem.*, Suppl 1(7):59–74, 2016.

[19] E. Hong and J. Park. Sample size and statistical power calculation in genetic association studies. *Genomics informatics*, 10:117–22, 06 2012.

[20] LDL and HDL: Bad and Good Cholesterol. *Centers for Disease Control and Prevention*.

[21] National Center for Biotechnology Information. Pubchem compound summary for cid 5997, cholesterol, 2020.

[22] I.L. Notkola, Sulkava R., J. Pekkanen, T. Erkinjuntti, C. Ehnholm, P. Kivinen, J. Tuomilehto, and A. Nissinen. Serum total cholesterol, apolipoprotein e FC12e4 allele, and alzheimer's disease. *Neuroepidemiology*, 17(1):14–20, 1998.

[23] M. Kivipelto, E. L. Helkala, and M. P. et al. Laakso. Apolipoprotein e epsilon4 allele, elevated midlife total cholesterol level, and high midlife systolic blood pressure are independent risk factors for late-life alzheimer disease. *Annals of Internal Medicine*, 137(3):149–155, 2002.

[24] S. Lewington, G. Whitlock, R. Clarke, P. Sherliker, J. Emberson, J. Halsey, N. Qizilbash, R. Peto, and R. Collins. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55000 vascular deaths. *LANCET*, 370:292–292, 2007.

[25] M. Sohmiya, M. Tanaka, and N.W. et al. Tak. Redox status of plasma coenzyme Q10 indicates elevated systemic oxidative stress in Parkinson's disease. *Journal of the Neurological Sciences*, 223(2):161–166, 2004.

[26] EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA) . Scientific opinion on the substantiation of health claims related to plant sterols and plant stanols and maintenance of normal blood cholesterol concentrations (id 549, 550, 567, 713, 1234, 1235, 1466, 1634, 1984, 2909, 3140), and maintenance of normal prostate size and normal urination (id 714, 1467, 1635) pursuant to article 13(1) of regulation (ec) no 1924/2006. *EFSA Journal*, 8(10):1813, 2010.

[27] G. R. Svishcheva, N. M. Belonogova, I. V. Zorkoltseva, A. V. Kirichenko, and T. I. Axenovich. Gene-based association tests using GWAS summary statistics. *Bioinformatics*, 35(19):3701–3708, 03 2019.

[28] X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet.*, 44(7)(821-824), 2012.

[29] M.A. Pourhoseingholi, A.R. Baghestani, and M. Vahedi. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench*, 15(2):79–83, 01 2021.

[30] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 12 2006.

[31] H. Zhao, N. Mitra, P.A. Kanetsky, K.L. Nathanson, and T.R. Rebbeck. A practical approach to adjusting for population stratification in genome-wide association studies: principal components and propensity scores. *Stat Appl Genet Mol Biol.*, 17(2):79–83, 12 2018.

[32] F. Zamudio, R. Wolfinger, and B. et al. Stanton. The use of linear mixed model theory for the genetic analysis of repeated measures from clonal tests of forest trees. I. A focus on spatially repeated data. *Tree Genetics Genomes*, 4:299, 12 2008.

[33] D.E. Runcie and L. Crawford. Fast and flexible linear mixed models for genome-wide genetics. *PLoS Genet*, 15(2), 02 2019.

[34] T. Andries, H. de K. Marees, S. Stringer, F. Vorspan, E. Curis, C. Marie[U+2010]Claire, and E.M. Derks. A tutorial on conducting genome[U+2010]wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res.*, 27(2), 06 2018.

[35] J. Gratten, N. Wray, and M. et al. Keller. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat Neurosci 17*, 17:782–790, 2014.

[36] H. Schwender, I. Ruczinski, and K. Ickstadt. Testing SNPs and sets of SNPs for importance in association studies. *Biostatistics*, 12(1):18–32, 01 2011.

[37] S.W. Choi and P. F. O'Reilly. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, 8(7), 07 2019. giz082.

[38] N. Wray, J. Yang, and B. et al. Hayes. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet*, 14:507–515, 2013.

[39] R. Power, S. Steinberg, and G. et al. Bjornsdottir. Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat Neurosci*, 18:953–955, 2015.

[40] The International Schizophrenia Consortium, Manuscript preparation, and S. et al. Purcell. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460:748–752, 2009.

[41] S. Stringer, R. S. Kahn, L. D. de Witte, R. A. Ophoff, and E. M. Derks. Genetic liability for schizophrenia predicts risk of immune disorders. *Schizophrenia Research*, 159(2):347 – 352, 2014.

[42] W.W. Li, Z. Wang, D.Y. Fan, Y.Y. Shen, D.W. Chen, H.Y. Li, L. Li, H. Yang, Y.H. Liu, X.L. Bu, W.S. Jin, Z.Q. Zeng, F. Xu, J.T. Yu, L.Y. Chen, and Y.J. Wang. Association of Polygenic Risk Score with Age at Onset and Cerebrospinal Fluid Biomarkers of Alzheimer's Disease in a Chinese Cohort. *Neurosci Bull*, 37(7):696–704, 2020.

[43] L. Ibáñez, U. Dube, B. Saef, J. Budde, K. Black, A. Medvedeva, J. Del-Aguila, A. Davis, J. Perlmutter, O. Harari, B. Benitez, and C. Cruchaga. Parkinson disease polygenic risk score is associated with parkinson disease status and age at onset but not with alpha-synuclein cerebrospinal fluid levels. *BMC Neurology*, 17:198, 11 2017.

[44] K.C. Paul, J. Schulz, J.M. Bronstein, C.M. Lill, and B.R. Ritz. Association of polygenic risk score with cognitive decline and motor progression in parkinson disease. *JAMA Neurol.*, 75(3):360–366, 03 2018.

[45] S. Morgenthaler and W. G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (cast). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1):28 – 56, 2007.

[46] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.*, 89(1):82–93, 07 2011.

[47] S Lee, M.J. Emond, M.J. Bamshad, Barnes K.C., M.J. Rieder, D.A. Nickerson, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, D.C. Christiani, M.M. Wurfel, and X. Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*, 91(2)(16):224–237, 08 2012.

[48] S. Takahashi, G. Andreoletti, and R. et al. Chen. De novo and rare mutations in the HSPA1L heat shock gene associated with inflammatory bowel disease. *Genome Med*, 9(8), 2017.

[49] S. Ruiz-Pinto, G. Pita, A. Patiño-García, J. Alonso, and A.J. et al. Pérez-Martínez, A. Cartón. DExome array analysis identifies GPR35 as a novel susceptibility gene for anthracycline-induced cardiotoxicity in childhood cancer. *Pharmacogenet Genomics.*, 27:445–453, 2017.

[50] E. Mossotto, J.J. Ashton, and L. et al. O'Gorman. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinformatics*, 20(254), 2019.

[51] P. Rentzsch, D. Witten, G.M. Cooper, J. Shendure, and Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*, 47(D1):D886–D894, 2019.

[52] D. Domingo-Fernández, A.T Kodamullil, A. Iyappan, M. Naz, M. A. Emon, T. Raschka, R. Karki, S. Springstubbe, C. Ebeling, and M. Hofmann-Apitius. Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics*, 33(22):3679–3681, 06 2017.

[53] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, S.L. Paulovich, A.and Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[54] E. Hong and J. Park. Sample size and statistical power calculation in genetic association studies. *Genomics informatics*, 10:117–22, 06 2012.

[55] Z Su, J. Marchini, and P. Donnelly. Hapgen2. *Bioinformatics*, 27(16):2304–2305, August 2011.

[56] D. P. Kingma and M. Welling. An introduction to variational autoencoders. *CoRR*, abs/1906.02691, 2019.

[57] D. Heckerman. Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1:79–119, 1997.

[58] Aetionomy. https://www.aetionomy.eu.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[60] Perktold Josef Seabold, Skipper. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[61] Howie B., Fuchsberger C., Stephens M., Marchini J., and Abecasis G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics 44(8): 955-959, 2012.

[62] Bioconductor. liftover: Changing genomic coordinate systems with rtracklayer::liftover. r package version 1.14.0, 2020.

[63] WJ. Kent, CW. Sugnet, TS. Furey, KM. Roskin, TH. Pringle, AM. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome Res.*, Jun;12(6).:996–1006, 2002.

[64] H. Li, B. Handsaker, and A. Wysoker. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[65] J. Lenny, L. Bastarache, and M. et al. Ritchie. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, 31:1102–1111, 2013.

[66] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 11 2019.

[67] Lucas D. Ward and Manolis Kellis. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, 40(D1):D930–D934, 11 2011.

[68] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nat Genet.*, 45(6):580–585, 11 2013.

[69] S.W. Choi, T.SH. Mak, and P.F. O'Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*, 15(2759–2772), 07 2020.

[70] M.A. Nalls, C. Blauwendraat, C.L. Vallerga, K. Heilbron, S. Bandres-Ciga, D. Chang, M. Tan, D.A. Kia, A.J. Noyce, A. Xue, J. Bras, E. Young, R. von Coelln, J. Simón-Sánchez, C. Schulte, M. Sharma, L. Krohn, L. Pihlstrøm, A. Siitonen, H. Iwaki, H. Leonard, F. Faghri, J.R. Gibbs, D.G. Hernandez, S.W. Scholz, J.A. Botia, M. Martinez, J.C. Corvol, S. Lesage, J. Jankovic, L.M. Shulman, M. Sutherland, P. Tienari, K. Majamaa, M. Toft, O.A. Andreassen, T. Bangale, A. Brice, J. Yang, Z. Gan-Or, T. Gasser, P. Heutink, J.M. Shulman, N.W. Wood, D.A. Hinds, J.A. Hardy, H.R. Morris, J. Gratten, P.M. Visscher, R.R. Graham, A.B. Singleton, 23andMe Research Team, System Genomics of Parkinson's Disease Consortium, and International Parkinson's Disease Genomics Consortium. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.*, 18(1091-1102), 2019.

[71] J.A. Tom, J. Reeder, and W.F. et al. Forrest. Identifying and mitigating batch effects in whole genome sequencing data. *BMC Bioinformatics*, 18(351), 2017.

[72] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 06 2011.

[73] K Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, 07 2010.

[74] I. Ionita-Laza, S. Lee, V. Makarov, J.D. Buxbaum, and X. Lin. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet*, 92(6)(16):841–853, 06 2013.

[75] G.E. Hoffman. Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PLoS One*, 8(12).

[76] Terry, T. and Mayo, C. the lmekin function.

[77] Y. Yao and A. Ochoa. Testing the effectiveness of principal components in adjusting for relatedness in genetic association studies. *bioRxiv*, 2019.

[78] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. et. al Isard. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[79] D. Maxwell Chickering, D. Heckerman, and C. Meek. Large-sample learning of bayesian networks is np-hard. *ArXiv*, abs/1212.2468, 2004.

[80] M. Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

[81] M. Waskom and the seaborn development team. mwaskom/seaborn, September 2020.

[82] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[83] J. Chung, G.R. Jun, and J. et al. Dupuis. Comparison of methods for multivariate gene-based association tests for complex diseases using common variants. *Eur J Hum Genet*, 27:811–823, 2019.

[84] Q. Yang, H. Wu, C.-Y. Guo, and C. S. Fox. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic Epidemiology*, 34(5):444–454, 2010.

7 Statement of Authorship

I declare that I completed this thesis on my own and that information which has been directly or indirectly taken from other sources has been noted as such. Neither this nor a similar work has been presented to an examination committee up to this point.

Bonn, Tuesday 12th January, 2021