THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Development and application of omics data analysis tools to examine molecular associations linking complex exposures to cardiovascular disease

YINGXIAO YAN

Food and Nutrition Science

Department of Life Sciences

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2025

Development and application of omics data analysis tools to examine molecular associations linking complex exposures to cardiovascular disease

Yingxiao Yan
ISBN 978-91-8103-236-9

Acknowledgements, dedications, and similar personal statements in this thesis, reflect the author's own views.

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5694
ISSN 0346-718X

Department of Life Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)768611733

Cover:
[Illustration is made by Yingxiao Yan, and with BioRender.com]

Printed by Chalmers digitaltryck
Gothenburg, Sweden 2025

Development and application of omics data analysis tools to examine molecular associations linking complex exposures to cardiovascular disease

YINGXIAO YAN

Department of Life Sciences
Chalmers University of Technology, Gothenburg, Sweden 2025

ABSTRACT

Preventing cardiovascular disease (CVD) requires better understanding of underlying metabolic perturbations to identify targets for intervention. Most studies have investigated exposures separately, and tools to analyze health effects of multiple exposures jointly are largely lacking. The work in this thesis aims to develop tools for complex omics analysis and apply them to population-based cohorts to investigate omics patterns linking combined exposures including diet, persistent organic pollutants (POPs), and gut microbiota to CVD risk.

This thesis includes development of two tools presented as R packages that make advanced omics data analysis accessible to a broader research community: 1) **MUVR2** provides supervised machine learning with nested cross-validation to mitigate overfitting and false discovery combined with variable selection in high-dimensional data. MUVR2 includes elastic net which allows covariate adjustment thus enhancing epidemiological modeling. 2) **TriplotGUI** is a user-friendly tool integrating omics data reduction with meet-in-the-middle and mediation analyses to explore exposure-omics-outcome associations through intuitive visualizations.

Using earlier versions of these tools on the Swedish Mammography Cohort revealed two distinct omics sub-patterns linking POPs to CVD. The first involved perturbed lipid metabolism and inflammatory pathways associated with higher levels of organochlorine compounds, lower levels of per- and polyfluoroalkyl substance and higher myocardial infarction (MI) risk. The second involved carnitines and possible mitochondrial dysfunction and associated with OCs and stroke.

MUVR2 and TriplotGUI were applied to discover and replicate metabolites associated with diet, POPs, gut microbiota, and CVD incidence using four Nordic cohorts. Notable findings supporting metabolites mediating exposure-outcome associations included: An association between nuts and dried fruit and reduced MI risk, possibly mediated by pipecolate. Moreover, associations between fish intake and reduced MI risk, possibly mediated by phosphatidyl-ethanolamine(P-16:0/22:6) and an unknown metabolite. Importantly, only few exposure-metabolite-outcome associations were reproduced across cohorts, stressing the importance of replication for generalizable conclusions.

This thesis contributes advanced, accessible methods for linking environmental exposures to health outcomes through omics-based mediators, with MUVR2 and TriplotGUI improving the identification and interpretation of molecular signatures. Application of these tools to CVD enabled characterization of molecular signatures linking diet to health and underscored the necessity of rigorous external validation to minimize spurious associations.

Key words: omics, diet, gut microbiota, persistent organic pollutants, cardiovascular disease, machine learning, molecular epidemiology, meet-in-the-middle analysis, mediation analysis, cross-cohort design

List of publications

This thesis is based on the work covered in the following manuscripts and published papers, referred to by Roman numerals in the text:

I. **Yan Y**, Schillemans T, Skantze V & Brunius C. Adjusting for covariates and assessing modeling fitness in machine learning using MUVR2. *Bioinformatics Advances*. 2024;4(1):vbae051.

II. Yan Y, Schillemans T, Ribbenstedt A, Brunius C. Software Application Profile: TriplotGUI, A Molecular Epidemiology Toolbox for Investigating Associations between Exposures, Omics and Outcomes (submitted)

III. Schillemans, T, **Yan, Y**, Ribbenstedt, A, Donat-Vargas, C, Lindh, C.H, Kiviranta, H, ... & Brunius, C. OMICs Signatures Linking Persistent Organic Pollutants to Cardiovascular Disease in the Swedish Mammography Cohort. *Environmental Science & Technology.* 2024;58(2), 1036-1047.

IV. **Yan Y**, Schillemans T, Toubon G, Ribbenstedt A, Åkesson A, Johansson I, Bergdahl I, Brunius C. Metabolites linking diet, persistent organic pollutants and microbiota composition to cardiovascular disease (manuscript)

Paper I: Yingxiao Yan (YY) participated in the conceptualization of the MUVR2 tool, wrote the algorithm and tutorial. YY analyzed data and performed testing of the tool and interpreted results. YY wrote the first draft of the manuscript and revised the manuscript together with co-authors. (published)

Paper II: YY participated in the conceptualization of the TriplotGUI tool, wrote the algorithm and tutorial. YY analyzed data and performed testing of the tool and interpreted results. YY wrote the first draft of the manuscript and revised the manuscript together with co-authors. (submitted)

Paper III: YY is the second author of this paper for which he performed data cleaning and analysis. YY participated in writing the methods and results of the first draft of the manuscript and revised the manuscript together with co-authors. (published)

Paper IV: YY participated in the conceptualization of the project, coordinated data acquisition, performed data cleaning, analysis and interpretation of results. YY wrote the first draft of the manuscript and revised the manuscript together with co-authors. (manuscript)

Contents

Abbreviations

ACE	Average causal effect
ANN	Artificial neural network
ASV	Amplicon sequence variant
BioDiVa	the BioDiVa subcohort of the Västerbotten Intervention Programme
COSM	Cohort of Swedish Men
COSM -C	Cohort of Swedish Men - Clinical
CV	Cross-validation
CVD	Cardiovascular disease
DASH	Dietary Approaches to Stop Hypertension
DE	Direct effect
DHA	docosahexaenoic acid
EN	Elastic net
EPA	U.S. Environmental Protection Agency
GLM	Generalized linear models
HDL-C	High-density-lipoprotein cholesterol
IE	Indirect effect
LDL-C	Low-density-lipoprotein cholesterol
MAX	the MAX subcohort of the Diet, Cancer and Health-Next Generations (DCH-NG) cohort
MI	Myocardial infarction
ML	Machine learning
MS	mass spectrometry
NCD	Non-communicable disease
NDE	Natural direct effect
NIE	Natural indirect effect
NR	Natural receptor
NMR	nuclear magnetic resonance
OC	Organochlorine compound
OTU	operational taxonomic units
PC	Phosphatidylcholine
PCA	Principal component analysis
PCB	Polychlorinated biphenyls
PE	Phosphatidylethanolamine
PFAS	Per and poly-fluoroalkyl substances
PLS	Partial least squares
POP	Persistent organic pollutants
rdCV	Repeated double cross-validation
RF	Random Forest
SCFA	Short-chain fatty acid
SHAP	Shapley additive explanations
SIMPLER	Swedish Infrastructure for Medical Population-Based Life-Course and Environmental Research
SIMP-UC	SIMPLER-Uppsala Clinical
SIMP-VC	SIMPLER-Västerås Clinical
SMC	Swedish Mammography Cohort
SMC-C	Swedish Mammography Cohort - Clinical
SNP	Single-nucleotide polymorphism
SVM	Support vector machine
TE	Total effect
TG	triglyceride
TMAO	trimethylamine N-oxide
T2D	Type 2 diabetes
VIP	Västerbotten Intervention programme
WGCNA	Weighted correlation network analysis

1. Introduction

Non-communicable diseases (NCDs), including cardiovascular disease, cancer, chronic respiratory disease and diabetes, are the leading causes of death worldwide[1]. NCDs are a key focus of the United Nations' Sustainable Development Goal 3.4[1], which aims to reduce by one-third premature mortality from NCDs through prevention and treatment and promote mental health and well-being. Among NCDs, cardiovascular disease (CVD), including conditions such as myocardial infarction and stroke, accounts for a major proportion of NCD deaths and the number of CVD deaths has increased in most countries, straining healthcare systems and economies[2].

While some factors, such as genetic disposition, age, sex, ethnicity and family history of diseases, are non-modifiable[3,4], major modifiable behavioral risk factors include diet, physical activity, tobacco use and alcohol consumption[5,6] . Environmental factors like air pollution have also been identified as contributors to CVD risk[5,7]. Additionally, gut microbiota is gaining increasing interest as a modifiable risk factor for CVD pathogenesis[8]. Understanding the modifiable determinants of CVD risk and their underlying pathophysiological mechanisms is crucial for effective prevention strategies[9,10].

Numerous studies have investigated individual risk factors for CVDs and their mechanisms[11]. However, several CVD risk factors, including diet[12], persistent organic pollutants (POPs)[7,13], and microbiota[8], are highly complex and consist of a high number of variables. Additionally, these risk factors may interact with each other, adding further complexity: For example, POPs may accumulate through food intake and alter the metabolic activity of gut microbiota. There is a notable scarcity of research systematically investigating the combined effect of multiple joint exposures on CVD risks. Moreover, causal mechanisms linking combined exposures to CVD outcomes remain largely unexplored since methodology for evaluating the effect of multiple exposures on metabolic perturbations is lacking.

Omics technologies have emerged as powerful tools for capturing the cumulative effect of environmental and endogenous factors over time[14,15]. These technologies have enabled molecular epidemiology to mechanistically link exposures to outcomes via molecular data that can potentially act as mediators or effect modifiers[16,17]. By identifying biological features, such as metabolites, proteins, and genes, that are associated with both exposures and outcomes, omics approaches can help elucidate underlying mechanisms related to CVD risk. Approaches like meet-in-the-middle analysis [18,19] or mediation analysis [6,20], addressed later in the thesis, are advantageous for this purpose.

Applying omics technologies, previous studies have investigated molecular associations of diet, POPs and microbiota, separately, with CVD[13,21–23]. However, a more comprehensive approach is needed to holistically understand how multiple exposures interact to influence CVD risks via molecular mediators. Moreover, robust methodologies for identifying such molecular mediators are lacking. In addition, studies on exposure-CVD association via omics have relied on data from single cohorts[24] and only few have provided replication[21,22], which has likely inflated false discovery and reported cohort-specific associations that are difficult to generalize.

Omics data often contain a large number of features that may reflect either biologically relevant information or "noise"[25] irrelevant to the research question of interest (e.g., not associated with exposure or outcome). Selecting informative features of interest, using univariate modelling or machine learning approaches[26], may benefit downstream data analysis. However, machine learning algorithms are prone to overfitting[27] and typically lack the possibility to adjust for

covariates[28]. Thus, methods without overfitting, that are capable of both unbiased variable selection and covariate adjustment are needed. Nevertheless, even after selecting relevant molecular features, challenges remain in interpreting the complex interactions between multiple exposures, omics variables and outcomes. Holistic approaches that capture these relationships while enhancing interpretability are critical for advancing our understanding of exposure-omics-outcome associations.

In summary, novel tools for omics feature selection and approaches to enhance interpretability for exposure-omics-outcome associations are highly warranted. While studies have explored mechanisms linking complex exposures (e.g., diet, pollutants, gut microbiota) to CVD via omics[13,21–23], the field lacks advanced, yet accessible methodologies for identifying mechanistically relevant omics features. Moreover, a holistic understanding of the combined effect of multiple complex exposures on CVD risk remains elusive. Additionally, the validation of findings has not been consistently prioritized across studies. Therefore, comprehensive investigations that simultaneously examine multiple exposures, employ rigorous feature selection methodologies and emphasize result validation are highly desired.

This thesis addresses these critical gaps in molecular epidemiology and omics research by developing novel tools for variable selection and interpretable visualization and applying these tools to investigate the molecular associations between diet, POPs, and gut microbiota and CVD. The research consists of four studies, progressing from methodological development to practical application across multiple cohorts, enabling more robust, generalizable insights into the metabolic perturbations underlying CVD risk.

2. Objectives

The overall aim of the thesis was to develop tools for omics data analysis and to apply these tools to investigate molecular associations linking combined environmental exposures (diet, persistent organic pollutants, microbiota) to cardiovascular outcomes (myocardial infarction and stroke) across four population-based cohorts (SIMPLER-Uppsala Clinical (SIMP-UC); SIMPLER-Västerås Clinical (SIMP-VC); the BioDiVa subcohort of the Västerbotten Intervention Programme (VIP) and the MAX subcohort of the Diet, Cancer and Health-Next Generations (DCH-NG) cohort).

The thesis consists of four papers that follow a progression from methodological development (**Papers I** and **II**) to cross-cohort applications (**Papers III** and **IV**). The methodological development aims to enhance the analytical and interpretative capacity for research involving omics and molecular epidemiology by establishing analytical frameworks for identifying molecular associations linking complex exposures to outcomes. The cohort applications aim to unravel such molecular links between combined environmental exposures (diet, persistent organic pollutants, microbiota) and cardiovascular disease.

The work in the thesis follows these the following objectives coinciding with specific papers:

1) To develop an enhanced predictive multivariate modeling algorithm (MUVR2) that can associate molecular omics data to exposures or outcomes using machine learning, and also perform variable selection without introducing bias, enable covariate adjustment, and systematically assess prediction performance and overfitting (**Paper I**).
2) To create a molecular epidemiology toolbox (TriplotGUI) that integrates and visualizes relationships between exposures, omics features, and outcomes for enhanced interpretation, based on omics data reduction, meet-in-the-middle analysis and mediation analysis (**Paper II**).
3) To identify genetic, proteomic, and metabolic biomarkers linking persistent organic pollutants and cardiovascular disease in the SIMP-UC cohort and to explore underlying mechanisms using meet-in-the-middle analysis (**Paper III**).
4) To identify metabolites linking diet, persistent organic pollutants, and gut microbiota to cardiovascular disease using all four cohorts in a cross-cohort discovery and replication design, and to explore underlying mechanisms through meet-in-the-middle and mediation analyses (**Paper IV**).

3. Background

This chapter provides essential background for understanding the research presented in this thesis. It begins with an overview of cardiovascular disease and key complex exposures – persistent organic pollutants, diet and gut microbiota – followed by an introduction of omics technologies and their applications in investigating exposure-CVD associations. These topics lay the groundwork for the analyses described in **Papers III** and **IV**. The chapter then introduces the main analytical methods and frameworks, including unsupervised and supervised machine learning, meet-in-the-middle and mediation analyses, which underpin the methodological developments described in **Papers I** and **II**.

3.1 Cardiovascular disease

Cardiovascular disease (CVD) refers to a group of disorders of the heart and blood vessels[5], such as coronary artery disease, peripheral arterial disease and cerebrovascular diseases. CVDs are the leading cause of death globally and are a major contributor to reduced quality of life[29]. Ischemic heart diseases, such as myocardial infarction (MI) and cerebrovascular diseases, such as stroke, account for the majority of CVD mortality[30] and are associated with high rates of recurrent cardiovascular events[31].

Myocardial infarction, colloquially known as heart attack, occurs when blood flow to the heart is severely restricted or blocked, causing oxygen deprivation and subsequent cardiac tissue death[5,32,33]. Most cases of MI result from atherosclerotic plaque rupture, leading to thrombosis[34]. Symptoms of MI include chest pain, shortness of breath, sweating, nausea and vomiting[35]. Stroke can be ischemic, caused by a blockage in a blood vessel supplying the brain, typically due to a blood clot, or hemorrhagic, caused by bleeding into or around the brain, usually from a ruptured blood vessel[36,37]. Both types may lead to lasting brain damage and long-term disability[38]. Ischemic stroke accounts for about 80% of strokes and can be caused by thrombosis, embolism or stenosis[39]. Symptoms of stroke include sudden numbness in the face or limbs on one side of the body, difficulty in speaking, visual disturbance, loss of coordination and severe headache[40].

CVD risk is shaped by a combination of non-modifiable factors, including genetics, age, sex and ethnicity[3,4], as well as a broad range of modifiable risk factors, such as diet, physical activity, tobacco use and alcohol consumption[6,41,42]. Socioeconomic and psychosocial factors like lower education and symptoms of depression may also increase CVD risks[43]. Environmental toxicants, such as those originating from air pollution, have also been recognized as contributors to CVD risk[44]. Furthermore, recent research highlights gut microbiota as a modifiable risk factor for CVD pathogenesis[8]. These risk factors interact with each other and may influence CVD through various metabolic processes, including dysregulated lipid metabolism, mitochondrial dysfunction, inflammation, and oxidative stress[45–47].

Non-pharmacological prevention interventions for CVD have focused on lifestyle changes such as adopting a healthy diet, increasing physical activity levels, ceasing smoking and reducing alcohol consumption[48–50]. It is estimated that over 70% of CVD may be preventable through addressing modifiable risk factors[51]. Both single- and multifactorial lifestyle interventions have demonstrated effectiveness in reducing CVD mortality and morbidity[48]. Multifactorial interventions have shown benefits in lowering intermediate risk factors, such as blood pressure, total cholesterol, BMI and waist circumference[48,52]. While multifactorial interventions may show effectiveness, how different risk factors in combination influence CVD risk remains underexplored. The Framingham risk score[53] was developed to predict individual risk by

considering multiple factors. However, there is a gap in understanding the mechanistic links between complex risk factors in combination and CVD outcomes.

3.2 Exposome and cardiovascular disease

Observational epidemiological research has transitioned from studying associations between single exposures and outcomes to exposure-wide and outcome-wide studies[54,55]. This shift aligns with the exposome framework, first proposed by Wild in 2005[56]. The exposome concept represents the totality of environmental exposures from conception onwards, complementing the genome in understanding disease etiology. Investigating the simultaneous effects of exposures on CVD risks may provide a more holistic understanding of the complex underlying mechanisms.

Within the exposome, diet[12], POPs[7,13], and gut microbiota[8] represent critical exposures that influence CVD risk. These exposures are composed of hundreds to thousands of variables, reflecting the multifaceted nature of dietary components (food items, nutrients, dietary patterns), the diverse chemical variability of POPs and the vast taxonomic and functional diversity of the microbiome. These exposures accumulate over the life course and interact, further contributing to complexity[57–60]. While studying each domain separately has yielded valuable insights, the combined effects and interactions among these exposures can likely better represent real-world exposure scenarios and their impact on cardiovascular health. However, despite their importance, studies investigating multiple simultaneous exposures remain scarce, and robust methodology to evaluate the exposome in relation to CVD risk is lacking, particularly regarding approaches addressing high dimensionality, temporal variations, and complex correlation structures among exposures.

3.2.1 Persistent organic pollutants

Persistent organic pollutants (POPS) are resistant to environmental degradation and are extremely widespread, imposing a long-lasting impact on human health[61–64]. POPs include several large groups of organic compounds, for instance, organochlorine compounds (OCs), such as pesticides (e.g., dichlorodiphenyltrichloroethane, p,p'-DDT), hexachlorobenzene (HCB), dioxins, and polychlorinated biphenyls (PCBs), as well as per- and poly-fluoroalkyl substances (PFAS)[61,65,66]. This thesis focuses on OCs and PFAS due to their environmental prevalence, established health impacts and data availability while excluding other classes like polybrominated diphenyl ethers (PBDEs) and polycyclic aromatic hydrocarbons (PAHs).

Organochlorine compounds

Organochlorine compounds are hydrocarbons with one or more chlorine atoms in their structure[67]. OCs encompass a large group of chemicals, such as dichlorodiphenyltrichloro-ethane (DDT), chlordane, lindane, endosulfan, and dieldrin[68]. Several OCs were produced as pesticides to increase crop yields and they gained popularity, particularly in the mid-20th century[69].

OCs are characterized by their high lipophilicity and environmental persistence [70], being resistant to microbial and chemical degradation[68]. They commonly enter the environment through industrial discharge, pesticide use, and urban waste, especially from chlorination and combustion processes[71]. Despite restrictions on OC pesticides in many countries since the 1970s, they continue to be detected in living organisms and remain a subject of environmental concern[72]. Their environmental effect on human health has been associated with food chain

accumulation[71]. Insects, amphibians and some reptiles that are carriers of OCs through water are prey for organisms including birds and mammals [71]. Ultimately, humans absorb OCs through fat-rich animal products such as meat and milk, where OCs are accumulated due to their lipophilicity [68,70]. OCs act as neuroendocrine-disrupting chemicals that stimulate the central nervous system, causing symptoms such as convulsions, tremors, nausea, and mental confusion at higher doses[68–70]. Additionally, OCs have been linked to reproductive suppression, immune dysfunction, dysregulated lipid biosynthesis and cancer[69,70]. Epidemiological studies have associated OCs with cardiometabolic conditions such as dyslipidemia, atherosclerosis [73,74], hypertension [75–77], and increased cardiovascular disease (CVD) risks [78–80].

Mechanistically, OCs may contribute to CVD through multiple pathways. Cell line studies report that OCs may disrupt endocrine function and act as activators for the aryl hydrocarbon receptor pathway, leading to oxidative stress[81], altered lipid metabolism and inflammation[82,83]. Additionally, animal studies using zebrafish models have reported that OCs impair glucose metabolism and mitochondrial function[84,85]. An observational study in humans has linked lipid-adjusted p,p'-DDT and HCB levels with up-regulated fatty acids, glycerophospholipids, monoglycerides and sphingolipids[85], which may in turn disturb global lipid metabolism by making available more reactants necessary for lipid biosynthesis and thus propagating an elevation of blood lipids [86]. However, exact mechanisms relating OCs to CVD risk remain unclear and a deeper understanding is crucial for establishing causality.

Per- and poly-fluoroalkyl substances

PFAS are defined by the Organization for Economic Co-operation and Development (OECD) as fluorinated substances containing at least one fully fluorinated methyl (-CF_3) or methylene (-CF_2) carbon atom, excluding those with hydrogen, chlorine, bromine, or iodine attached to these moieties[87]. PFAS have been produced and used in a broad range of consumer products and industry applications since the 1940s[88]. Currently, 4,730 PFAS compounds have been identified on the OECD focus list 2018[89]. The Toxic Substances Control Act (TSCA) 2021 definition from the U.S. Environmental Protection Agency (EPA) resulted in 6,648 PFAS chemicals[90].

Due to their high persistence and widespread use, PFAS are found ubiquitously in the environment, animals and humans[91,92]. PFAS from production waste streams travel through air or water into the ground or drinking water and then accumulate in fish, aquatic animals and humans[61]. PFAS are absorbed in the gastrointestinal tract but are not metabolized. Instead, they accumulate in organs such as the liver [93]. Adverse health effects associated with PFAS exposure include elevated total cholesterol and low-density lipoprotein cholesterol (LDL-C) level and dyslipidemia; reduced antibody production following vaccination in children; liver toxicity; fetal growth restriction and kidney and testicular cancer[61,94,95].

Evidence suggests that PFAS exposure is linked to increased CVD risk [96–98] due to its association with elevated blood lipid levels [99,100]. Human and rodent studies suggest that PFAS exposure may disrupt lipid metabolism and liver functions[101]: Nuclear receptors (NRs) are receivers for intracellular signals, hormones or dietary lipids, regulating lipid metabolism[102]. Given that PFAS's chemical structure resembles fatty acids, they may bind to the NRs for fatty acids and induce metabolic changes[103–105]. For example, PFAS may suppress hepatocyte nuclear factor 4α (HNF4α) expression, which reduces cholesterol 7α-hydroxylase (CYP7A1) expression and cholesterol's conversion to bile, increasing cholesterol level[104]. Additionally, PFAS exposure has been shown to induce oxidative stress and mitochondrial dysfunction *in vitro* and *in vivo* [106].

Contrarily, some studies have shown null or inverse associations between PFAS exposure and CVD[107–109]. Particularly, an inverse association between PFAS and myocardial infarction was reported in observational studies[13,110]. The possible mechanisms here can also be related to NRs. Rodent models showed that PFAS may activate peroxisome proliferator-activator receptor α (PPARα), which leads to increased lipoprotein lipase (LPL) expression and corresponding triglyceride clearance, decreasing triglyceride level[104,111]. Moreover, animal evidence suggests that PFAS may exert antioxidant and anti-inflammatory effects via activation of PPARα[112] or upregulation of phosphatidylcholine synthesis[101]. These findings highlight the need for stringent studies to further investigate the association between PFAS exposure and CVD risk while elucidating underlying mechanisms.

OC and PFAS concentrations are typically measured from human biological matrices, such as blood serum or plasma and adipose tissue[113,114]. Blood samples are usually collected after overnight fasting and are centrifuged for separation before storage[113,114]. The analysis of OCs and PFAS can be conducted using gas/liquid chromatography-mass spectrometry (GC/LC-MS). In this thesis, OC and PFAS concentrations were measured using plasma samples collected after overnight fasting and analyzed by GC-MS [113,114].

3.2.2 Diet

Dietary composition plays a critical role in modulating CVD risk[115,116]. Extensive research has elucidated cardioprotective or detrimental effects of specific foods, nutrients, and dietary patterns[117]. This section covers several dietary components that are relevant to the results of the cohort applications presented in this thesis.

Individual food items

Higher intake of fruits and vegetables has consistently been associated with reduced CVD incidence[118], in part attributed to their content of dietary fiber, vitamins, alpha-linolenic acid, and antioxidants[119]. Wholegrains are rich in fibers, vitamins and polyunsaturated fatty acids[120] and are associated with decreased risk of CVD[118]. In addition, nuts and seeds are rich in unsaturated fatty acids and bioactive compounds such as vitamin E. Frequent consumption of nuts and seeds is associated with a lower risk of CVD [118], likely due to their cholesterol-lowering effect and the beneficial effect on oxidative stress, inflammation and vascular reactivity[121,122].

Fatty fish consumption has been linked to reduced CVD risk in observational studies[123,124], primarily due to omega-3 fatty acids[125] like docosahexaenoic acid (DHA) and eicosapentaenoic acid (EPA)[126]. These fatty acids have shown lipid-lowering effects by reducing triglyceride (TG) and non-high-density lipoprotein cholesterol (non-HDL-C)[126] and can reduce platelet activation and aggregation as well as regulate blood pressure and heart rate[126]. Conversely, excessive alcohol intake adversely affects cardiovascular health through a combination of hypertension, oxidative stress, inflammation, endothelial dysfunction and structural damage to the heart [127].

Nutrients

Among nutrients, dietary fat including trans-fatty acids, and saturated, monosaturated or polyunsaturated fatty acids, can influence CVD risk[128]. Studies have shown that trans fatty acids are consistently associated with increased risk of CVD[129] and the replacement of saturated fatty acid with mono- or polyunsaturated fatty acids is associated with reduced CVD risk[128]. The type of carbohydrates consumed also matters for cardiovascular health. Diets high in refined carbohydrates and sugars exacerbate CVD risk, while fiber-rich foods are protective[130]. Plant-

based proteins have shown benefits compared to animal proteins for CVD risk reduction[130]. Vitamins (e.g., C and E) may play crucial roles in cardiovascular health by mitigating oxidative stress and managing blood pressure and inflammation[130].

Dietary patterns

Investigating food items and nutrients in relation to CVD risk has provided insights into our food choices. However, food items or nutrients are not consumed in isolation and their consumption may interact or be intercorrelated with each other. Dietary patterns may be more representative of the eating habits of free-living individuals. A broad range of indices representing dietary patterns have been developed[131]. Several of these indices representing healthy diets, such as the Healthy Nordic Diet, Mediterranean Diet and Dietary Approaches to Stop Hypertension (DASH), have been investigated in relation to diet-related diseases and have been reported to reduce e.g., CVD risk[132–134].

Diet is most often assessed by self-reported methods, such as diet history, food frequency questionnaire (FFQ) and 24-hour dietary recall (24-HDR)[135]. In this thesis, the FFQ method was used to assess dietary intake. These methods have inherent limitations including recall bias, social desirability bias, and challenges in portion size estimation[136].

3.2.3 Gut microbiota

The gut microbiota refers to the diverse community of microorganisms residing in the gastrointestinal tract, including bacteria, viruses, fungi and archaea, with bacteria being the most abundant[137–139]. These microorganisms perform essential physiological functions, including fermenting indigestible carbohydrates and dietary fibers[140,141]; synthesizing vitamins and hormones[142,143]; regulating immune homeostasis[144], maintaining gut barrier integrity[140]; metabolizing drugs and xenobiotics[140] and preventing the overgrowth of pathogens[140].

The gut microbiota has emerged as a significant factor influencing CVD risk through multiple interrelated mechanisms[145]. Dysbiosis, i.e., an imbalance in microbial communities, induces loss of diversity[146] and has been reported as a CVD-predisposing condition and associated with several CVD risk factors, including atherosclerosis, hypertension and heart failure[8]. Furthermore, accumulating evidence has suggested that microbial metabolites have been mechanistically linked to CVD pathogenesis[8,147]. For example, elevated trimethylamine N-oxide (TMAO) has been associated with an increased risk of adverse cardiovascular events[148–150], while short-chain fatty acids (SCFAs) derived from fiber fermentation have shown protective effects against hypertension and inflammation[148,150,151].

After the collection of fecal samples, gut microbiota composition is typically analyzed by either of two DNA-based approaches: 16S ribosomal ribonucleic acid (rRNA) gene sequencing or whole-genome shotgun sequencing: The 16S rDNA is a bacterial-specific genetic marker encoding the small subunit of the ribosome, containing 9 hypervariable regions V1–V9 unique to microbial groups, flanked by conserved sequences shared across species[152]. These variable regions enable taxonomic classification by comparing the obtained DNA sequences to reference databases, with sequencing workflows often grouping sequences into operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) based on similarity thresholds.

Shotgun metagenomics instead sequences the entire genomic content of a specific environment[153,154], enabling high-resolution taxonomic profiling down to species and strain level and can provide insights into the functional potential of the microbiome, such as metabolic

pathways [153,154]. Although shotgun metagenomics sequencing has gained popularity in the last years[153,154], the 16S approach remains widely adopted for large-scale epidemiological studies due to its cost-effectiveness and simpler bioinformatics requirements. However, 16S rRNA sequencing provides limited resolution compared to shotgun metagenomics, particularly for species-level identification, and short-read assays cannot assess functional potential, as they only target specific regions of the 16S rRNA gene rather than the entire genomic content[155].

In this thesis, both shotgun metagenomics and 16S rRNA methods were employed to characterize microbiota composition across study cohorts, with the 16S method targeting the V3 and V4 regions with taxonomic resolution at genus level.

3.2.4 Interaction between exposures

The interplay between diet, persistent organic pollutants (POPs), and gut microbiota represents a critical nexus in understanding how combined exposures influence cardiovascular health.

Diet exerts profound effects on gut microbiota composition[57], as evidenced by enterotype classifications, where distinct microbial community clusters have been linked to specific dietary habits[156]. These enterotypes are frequently reported as driven either by *Bacteroides*, *Prevotella* and *Ruminococcus/Ruminococcaceae* [156]. The *Bacteroides* enterotype has been associated with animal protein, certain amino acids and saturated fats[157,158], while the *Prevotella* enterotype has been associated with carbohydrate-enriched diets and higher intake of fruits and eggs[159]. The *Ruminococcus/Ruminococcaceae* enterotype has been characterized by higher intake of vegetables, seaweed, nuts and seeds as well as dietary fibers[159].

A major exposure pathway for POP exposures occurs through diet due to the accumulation in food, especially in the general population that is not occupationally exposed to POPs[160]. Numerous studies have monitored the presence of OCs and PFAS in many different foods, including fish, dairy products, meat, fruits and vegetables[160–168]. For OCs, associations between fatty fish consumption and PCBs and HCB appear to be the most consistent, followed by dairy products and PCBs[160,166,167,169]. For PFAS, fish and other seafood have been reported as major sources of exposure from diet [162]. Food packaging materials may also contain PFAS as well as the drinking water in certain areas, contributing to PFAS exposure from food and water intake [162]. In addition, although fish consumption provides cardioprotective omega-3 fatty acids, it is also a major source of OC and PFAS contamination and overall effects on cardiovascular health risk may be uncertain[170].

Exposure to POPs, which occurs primarily through diet, has also been shown to alter the composition and metabolic activity of gut microbiota[161,171,172]. For example, exposure to the OC 2,3,7,8-tetrachlorodibenzofuran (TCDF) was shown to induce shifts in the ratio of *Firmicutes* to *Bacteroidetes* and inhibit nuclear receptor signaling in mice, triggering inflammation and host metabolic disorders[161]. PFAS exposure may alter gut microbiome composition at different taxonomic levels, including *Turicibacter, Allobaculum, Bacteroides acidifaciens, and Dehalobacteriaceae*, associating with perturbations of host glucose and lipid metabolism[171,172].

The interplay between exposures to human health is intricate, where diet contributes to POP exposure and influences gut microbiota compositions. Concurrently, POPs can alter microbiota compositions, in turn potentially affecting the impact of diet on health. Additionally, other risk factors, such as lifestyle factors (e.g., physical activity, smoking) and geographic disparities in food contamination levels may strengthen or counteract these interactions, introducing further variability in cardiovascular outcomes (**Figure 1**). Consequently, addressing cardiovascular risk

often requires simultaneous consideration of these interconnected exposures to better understand their combined effects on health outcomes.

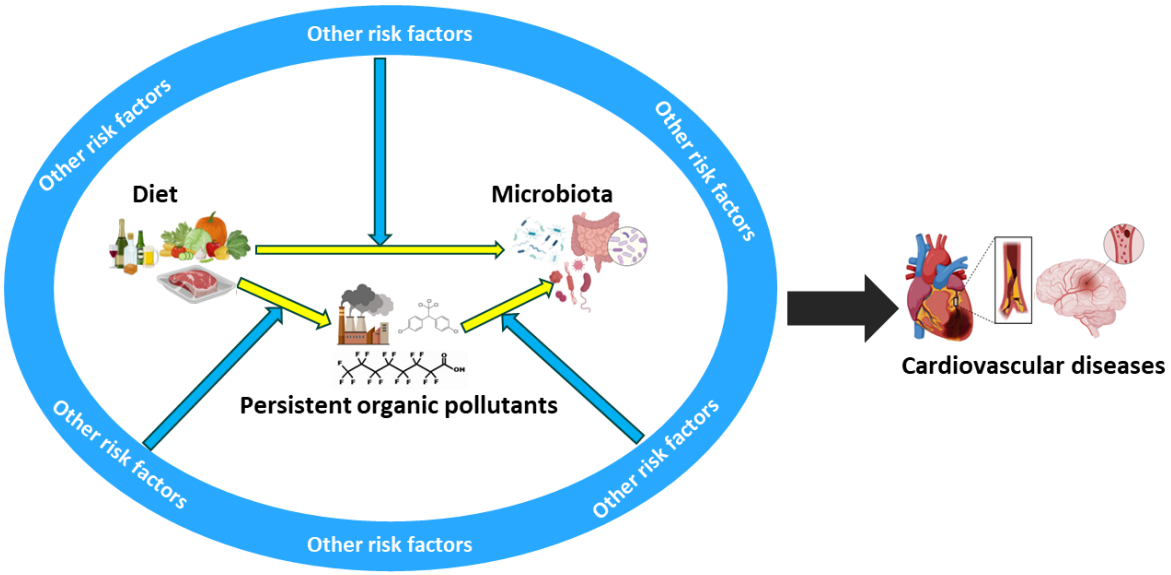


Figure 1. Conceptual framework illustrating directional interactions between diet, persistent organic pollutants, and gut microbiota. Other risk factors may strengthen or counteract these interactions, introducing further variability in cardiovascular outcomes. Figure created with BioRender icons in PowerPoint. Yan, Y. (2025)

3.3 Omics

Further improving inference with respect to causality in exposure-outcome relationships requires elucidating molecular associations connecting environmental factors to health effects. The rapid development of omics technologies has enabled systematic investigation of mechanisms by identifying molecular markers associated with both exposures and outcomes[16,173].

The omics approaches provide insights into biological systems at different molecular levels, including genomics, transcriptomics, proteomics and metabolomics[174,175]. Living systems are intricately organized across these molecular layers, each governed by distinct molecular components that work together to maintain health and respond to environmental challenges[176]. At the foundation are the genes, DNA segments that contain instructions for building and regulating all cellular processes[177]. The complete set of genes in an organism is called genome, and the study of genomes and their functional roles is known as genomics[178,179]. This field can examine genetic variations and their influence on susceptibility to environmental exposures and disease outcomes.

Genes exert their effects through transcription into RNA[178,179], with transcriptomics being the large-scale study of RNA molecules, providing insights into which genes are active under specific conditions and how gene expression changes in response to stimuli or during disease progression[180].

At the following level, proteins are synthesized according to the instructions carried by mRNA[181]. Proteins function as enzymes, structural components and signaling molecules driving mechanistic responses to exposures[181]. Proteomics investigates the structure, function, and interactions of proteins, shedding light on how they carry out cellular functions and how their dysregulation can lead to disease[182].

Finally, metabolites are commonly defined as small molecules (<1500Da) involved in metabolism and present in biological samples such as cells, biofluids, tissues or organisms[15,183]. Metabolomics aims to measure either the entire metabolome or specific portions of it[221,222], providing a real-time snapshot of physiological and metabolic states and reflecting biochemical activity[184,185]. Metabolomics is crucial for understanding how genetic and environmental factors influence metabolism, and for identifying biomarkers of disease or responses to interventions such as diet or drug treatment[186,187].

Together, these omics layers reflect the central dogma of molecular biology: Genetic variations influence transcriptional activity, which translates into protein expression and function, ultimately shaping metabolic regulations. Environmental exposures (e.g., diet, pollutants, lifestyle, stress) modulate molecular processes across the omics layers, altering gene expression, transcription, translation and metabolism (**Figure 2)**

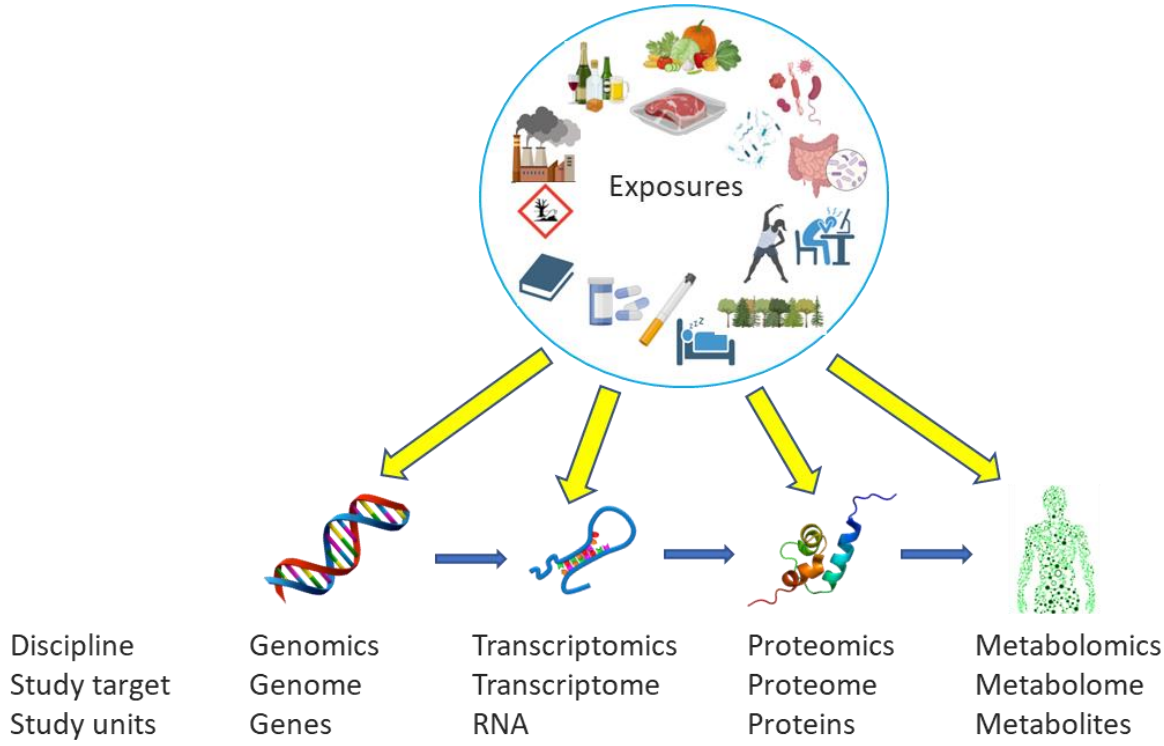


Figure 2. Omics studies include genomics, transcriptomics, proteomics and metabolomics. Their corresponding study targets are the genome, transcriptome, proteome and metabolome, which respectively focus on genes, RNA, proteins and metabolites as study units. Genetic variations influence transcriptional activity, which translates into protein expression and function, ultimately shaping metabolic regulations. Environmental exposures (e.g., diet, pollutants, education, medication, smoking, sleep, green space exposure, physical activities, stress, gut microbiota) influence molecular processes across all omics layers. Figure created with BioRender icons in PowerPoint. Yan, Y. (2025)

3.3.1 Metabolomics

Metabolites are the intermediate and end products derived from the gene cascade interacting with the cumulative effect of life-course exposures (e.g., diet, POPs, gut microbiota)[183,188–193], providing a functional readout of both endogenous metabolic processes and exogenous influences [178,179]. Therefore, the metabolite composition provides the closest representation of molecular phenotypes in health and disease[15,193]. Metabolite profiling, the systematic identification and quantification of metabolites, provides a “snapshot” of the human

metabolome and is frequently used to gain mechanistic insights into the effects of exposures such as dietary intake or medication use, or pathways involved in health and disease trajectories[194–197]

Measurement

Metabolomics can be categorized into targeted and untargeted approaches[197]. In targeted metabolomics, a known subset of metabolites is quantitated or semi-quantitated, where the metabolite coverage (or assay) is typically based on predetermined research questions or hypotheses or, alternatively, based on method availability[185]. Untargeted metabolomics instead aims to measure all detectable metabolites in a sample, without an a priori metabolite coverage based on previous knowledge or assumptions[198]. This approach is often used for biomarker discovery, identifying novel associations and mechanisms linking metabolites to exposures or health outcomes[199]. This lack of pre-selection of metabolite coverage makes untargeted approaches particularly well-suited for hypothesis-generating and data-driven research. In this thesis, untargeted metabolomics was used to enable comprehensive exploratory analysis of metabolic perturbations linking environmental exposures to cardiovascular outcomes without being limited by prior hypotheses about specific metabolites.

From an analytical chemistry perspective, metabolomics analyses are typically performed using either nuclear magnetic resonance spectroscopy (NMR) or mass spectrometry (MS). NMR measures the magnetic resonances at chemical shifts[200], while MS measures the mass-to-charge (m/z) ratio of analytes, typically coupled with upstream separation techniques, such as liquid chromatography (LC) or gas chromatography (GC) to provide additional resolution for distinguishing analytes[201]. In LC, the technique employed to generate the data used in this thesis, analytes are separated by polarity using either reversed phase (RP) chromatography to investigate more hydrophobic analytes or hydrophilic interaction liquid chromatography (HILIC) to investigate less hydrophobic analytes[202].

After chromatographic separation, analytes are ionized for MS analysis, typically using electrospray ionization or other soft ionization techniques to minimize fragmentation. Both positive and negative ionization can be applied to accommodate analytes prone to specific charges, resulting in the measurement of m/z ratio using mass analyzers such as quadrupoles, orbitraps and time-of-flight[203].

The combination of chromatography and MS ionization polarity results in different analytical modes, where analytes are differentiated by their retention time (rt) on the column and the specific m/z ratios and referred to as metabolite features. In this thesis, untargeted metabolomics was performed using LC-MS with reversed-phase separation, both positive and negative ionization polarities and high-resolution m/z detection using quadrupole-time-of-flight mass analyzers.

Metabolite annotation

The identification of what compound (e.g., a metabolite or chemical or biological exposure) an analytical feature corresponds to, is based on matching MS fragmentation obtained from tandem MS (referred to as MS/MS or MS2) to reference databases. The first MS (usually a quadrupole) fragments the metabolite feature, while the second MS (usually a high-resolution mass analyzer, such as orbitrap or time-of-flight) measures the obtained fragments to generate a molecular fingerprint matched against available libraries (in-house or online databases). However, only a

limited number of metabolites are registered in these databases, constituting one of the major roadblocks in untargeted LCMS-based metabolomics[204].

In this thesis, feature identification follows the Schymanski scale: Level 1 corresponds to matching rt, m/z and MS2 data with authentic standards. Level 2 indicates matches between MS2 data and library spectra. Level 3 provides evidence for a potential structure but lacks sufficient evidence for determination. Level 4 represents known molecular formulas but without structural information. Level 5 reports exact rt and m/z[205,206]. This standardized approach for annotation enables transparent reporting of identification confidence and facilitates meaningful comparisons across studies. In this thesis, mechanistic interpretations focused primarily on metabolite features with higher confidence levels (Levels 1–3). Features with lower confidence levels were treated as unannotated and were not included in mechanistic analyses.

3.3.2 Metabolomics for investigating exposure-outcome associations

Metabolomics has been widely utilized over the last two decades to investigate various diseases, including cancer[207–209], diabetes[210,211] and cardiovascular disease[212,213]. Recent studies have demonstrated the utility of metabolomics in identifying markers for CVD risks. Associations have been reported for specific metabolites, including acylcarnitines, branched-chain amino acids and certain lipids, such as phosphatidylcholine (PC), lysophosphatidylcholines (LPC) and sphingomyelins with various cardiovascular conditions [214–216].

Metabolomics has also frequently been linked to exposures, such as diet, POPs and gut microbiota. Studies have identified metabolites that reflect specific food exposures[183,217,218] and dietary patterns[219–221], assisting in elucidating disease-related metabolic processes influenced by diet[222–224]. POPs have been shown to induce substantial changes in metabolites, due to their toxicity. For example, PFAS exposure has been shown to alter the levels of amino acids, fatty acids, glycerophospholipids, glycolipids, phosphosphingolipids, bile acids, ceramides, purines, and acyl carnitines[225,226]. OC exposure has been linked to increased levels of saturated fatty acids and decreased levels of polyunsaturated acids in the liver[227,228]. Microbial metabolites are produced or converted by microbiota and they can enter the bloodstream and impact human health[229]. Gut microbiota have been reported to explain up to 46% of the variance in individual plasma metabolites[230]. These microbial metabolites are involved in molecular signaling related to gut-brain communication and metabolic reactions[192,231]. A wide range of gut microbiota-derived metabolites has been identified as central regulators in metabolic disorders, including TMAO, SCFAs, bile acids, branched-chain amino acids, tryptophan and indole derivatives[231,232].

Although numerous studies have investigated metabolites associated with single exposures or outcomes, few have comprehensively applied omics approaches to explore exposure-outcome associations[13,233–236]. Additionally, existing omics studies have largely investigated these associations in isolation, overlooking interactive effects between exposures. This thesis bridges this gap by holistically investigating the combined effect of complex exposures (diet, POPs, and gut microbiota) on cardiovascular outcomes (myocardial infarction and stroke). Moreover, the methodological advances described in subsequent sections of this thesis enable a more robust identification of metabolite features that may mechanistically link exposures to outcomes. This approach aligns with the "meet-in-the-middle" concept in molecular epidemiology, where molecular markers associated with both upstream exposures and downstream disease outcomes serve as evidence for potential causal pathways[20].

3.4 Machine learning

Omics data present unique analytical challenges due to their high dimensionality and inherent noise[237]. Machine learning (ML) has emerged as a critical tool for addressing these complexities, enabling data reduction, variable selection and predictive modeling [238]. In the context of this thesis, ML approaches are essential for extracting meaningful patterns from complex exposure-omics-outcome relationships and identifying relevant metabolic features associated with cardiovascular outcomes.

ML methodologies can be broadly categorized into unsupervised, supervised and reinforcement ML. Unsupervised machine learning is intended for the analysis of unlabeled data to discover hidden patterns, which is particularly useful in exploratory omics research. Supervised ML methods, on the other hand, are designed to learn from labeled data with the possibility of making predictions on new data. Reinforcement learning solves problems by learning new experiences through trial and error, but is not covered in the scope of this thesis[238].

Cross-validation techniques are commonly used to assess model performance and generalizability in supervised ML[239]. This thesis focuses primarily on supervised ML with a particular emphasis on the MUVR algorithm, which performs variable selection within repeated double cross-validation, which is a gold standard for minimizing overfitting[240].

Before applying unsupervised and supervised ML, data preprocessing steps, such as centering, scaling and transformation, can be important for robust and interpretable results, depending on the specific method used [241]. Centering adjusts data by subtracting the mean of each variable, aligning the data distribution around zero. Scaling ensures that all variables are on comparable ranges, preventing variables with inherently larger values from dominating analysis[242]. Two common scaling methods are: Standardization (Z-score normalization), which transforms data to have a mean of 0 and a standard deviation of 1. This method preserves the shape of variables' distributions and is ideal for algorithms assuming linearity and normality; Min-max normalization, which rescales data to a fixed range. This method is useful for algorithms sensitive to the ranges of variables [242]. Transformations like log-transformation address skewed distributions common in biological data, stabilizing variance and approximating normality[243]. Additionally, for omics applications, preprocessing also includes handling missing values through removal or imputation, managing outliers by exclusion or winsorization, and encoding categorical variables[244]. Ultimately, the choice of preprocessing steps should be tailored to the characteristics of the data, the requirements of the ML methods, and the specific analytical goals[245].

3.4.1 Unsupervised methods

Unsupervised ML methods identify latent patterns and hidden structures in high dimensional data without predefined outcomes or targets[246,247]. While unsupervised machine learning can cover a range of topics, such as clustering, association, anomaly detection and latent representation using autoencoders[248], among the most common techniques used in omics research are clustering (such as K-means clustering and hierarchical clustering) and data reduction (such as principal component analysis and factor analysis)[246]. In this thesis, two unsupervised data reduction methods are covered, namely principal component analysis (PCA) and weighted correlation network analysis (WGCNA).

Principal component analysis

Principal component analysis (PCA) is a conventional technique for dimensionality reduction. It transforms correlated variables into uncorrelated principal components that capture the overall variance in the data through orthogonal linear transformations[249,250]. The procedure involves decomposing the covariance matrix to obtain eigenvectors that define new orthogonal axes (principal components) that approximate the overall variance using fewer variables. The first principal component explains the largest proportion of variance followed by successive orthogonal components[249,250]. This enables efficient data compression[249,250]. In the context of molecular biology, PCA allows for the reduction of thousands of metabolites into a small set of interpretable components that can be associated with exposures and outcomes.

Weighted correlation network analysis

Weighted correlation network analysis (WGCNA) was originally developed for identifying clusters of highly correlated expressed genes, hence the original name "Weighted Gene Co-Expression Network Analysis" [251], and is a frequently used technique for data reduction in omics. WGCNA quantifies not only the correlations between individual pairs of variables but also the extent to which these variables share the same neighbors[252,253].

WGCNA first calculates pairwise correlation coefficients between omics variables and then constructs a weighted network through a soft-thresholding transformation, amplifying strong correlations while diminishing weak ones[251,252]. The transformed correlations create an adjacency matrix that reflects the strength of connections between variables, resulting in clusters of related variables called modules[251,252]. A PCA is performed on each module with the first principal component of each cluster selected as representative[252] (**Figure 3**). WGCNA poses advantages for omics data analysis compared to other methods, since it focuses on identifying co-regulated modules that may represent shared biological pathways or functional relationships[252]. In this thesis, WGCNA is implemented as an alternative data reduction approach in the TriplotGUI tool (**Paper II**) to capture biologically meaningful patterns in omics data. However, WGCNA is both more computationally intensive and more sensitive to parametrization compared to PCA[252,253].

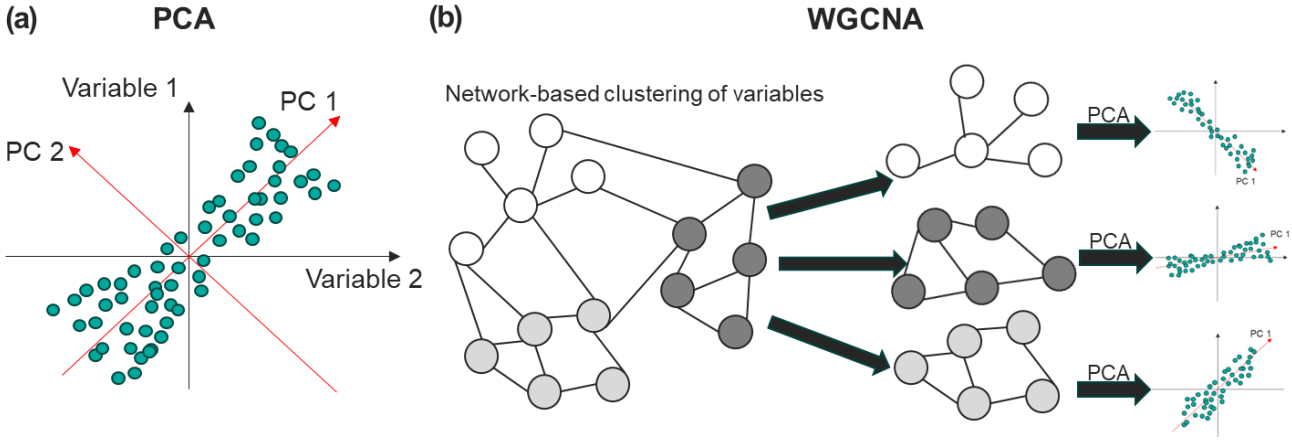


Figure 3. Conceptual comparison of PCA and WGCNA approaches for dimensionality reduction of omics data. **(a)** PCA creates components that maximize the captured variance across all variables. **(b)** WGCNA identifies modules of correlated variables before applying PCAs to represent each module.

3.4.2 Supervised methods

Supervised ML learns from existing data to forecast/classify a specific outcome of interest[254]. These methods develop mapping functions between input data (predictors; typically numerical data) and labeled outcomes of interest (target variables; either numerical or categorical data) and use the mapping functions to predict the target variables for new data[255]. The categorical or numerical nature of the target variable determines whether the supervised ML constitutes either classification or regression analysis[256].

In omics applications, supervised ML has been widely applied to predict disease outcomes[257]. Given the high dimensionality and noisiness of omics data, selecting relevant features can improve the performance of predictive models[240]. In addition, feature selection is often of primary interest when using supervised ML for biomarker discovery[258]. For some supervised ML methods, such as elastic net, feature selection is performed implicitly, while others require variable importance ranking to assess feature contributions, with the possibility to perform variable selection based on these rankings[259]. In this thesis, supervised ML methods were employed to identify metabolite features associated with exposures (diet, POPs, gut microbiota) and outcomes (cardiovascular disease), supporting the investigation of metabolic perturbations linking these exposures to health outcomes.

For clarity and consistency in this thesis, predictor variables will be denoted as **X**, while target variables, also often referred to as labels or responses, will be denoted as **Y**. The term "samples" refers to individuals or observations in the dataset. "Training data" refers to the **X** and **Y** samples used to construct the model for prediction, while "testing data" refers to the **X** and **Y** samples used to evaluate prediction performance (i.e., whether predicted **Y**, often referred to as $\widehat{\boldsymbol{Y}}$, corresponds to actual **Y**)**.** Several other terminologies and concepts are briefly described here to facilitate the reading of this thesis:

Parameters and hyperparameters: Parameters are internal variables that a model learns directly from the training data, which can be optimized during the training process to minimize prediction errors on training data, such as weights of neurons in artificial neural networks or beta coefficients in linear regression[260]. Hyperparameters, on the other hand, are variables that cannot be directly estimated from data training and must be manually specified or tuned[260], i.e., optimized by minimizing prediction errors on validation data [260]. Examples include the number of trees in the random forest and the regularization strength in the elastic net.

Overfitting and underfitting: Overfitting is a common pitfall in machine learning, occurring when a model trained on the available data fails to generalize well to new testing data[261]. One major cause of overfitting is an overly complex model relative to the amount of noise in the data, which gives the model excessive flexibility and leads it to learning the training data too well, capturing noise and random fluctuations as if they were meaningful patterns[262]. As a result, the model performs better on training data than on testing data. Besides model complexity, another source of overfitting is data leakage, which occurs when information that should remain unavailable during training, such as testing data, is used to train or optimize the model. This leads to artificially optimistic prediction performance and poor generalization[263,264]. Although overfitting is widely recognized as a problem in supervised ML, established methods for systematically assessing or comparing overfitting between models are sorely lacking. Underfitting, on the contrary, occurs when a model is too simple to capture underlying patterns in the data[262], leading to poor performance both on training and testing data.

Bias-variance trade-off: While bias in epidemiology refers to systematic error, machine learning employs a conceptually similar definition, i.e., errors caused by overly simplistic assumptions in modeling, leading to underfitting. In contrast, variance in machine learning describes errors resulting from a model's sensitivity to noise in the training data, caused by overfitting[265,266].

Increasing model complexity, for example by including more complexity-increasing hyperparameters such as the number of layers, nodes, or polynomial degree, can reduce bias but typically increases variance, making the model more sensitive to fluctuations in the training data and less able to generalize[267]. Data leakage, i.e., where the same data is reused for training and testing, can artificially reduce bias without increasing variance, but this comes at the cost of misleadingly optimistic performance estimates[263,264,268]. Finding an optimal balance between bias and variance in supervised ML is critical for creating models that generalize well and make accurate predictions on new, unseen data.

Cross-validation: Cross-validation (CV) techniques are widely used in supervised ML to mitigate inflated prediction performance caused by data leakage[269]. The most widely used CV techniques include k-fold CV and leave-one-out CV[270,271]. In k-fold CV, the entire dataset is divided into k subsets: k-1 subsets are merged and used for training, while one subset is held out and used for testing in each iteration (**Figure 4**). This ensures all samples are predicted without being used in training. For categorical target variables, randomization of samples into subsets is typically stratified to maintain consistent class proportions across training and testing datasets[272]. Leave-one-out CV instead sets one sample aside as testing data while using all remaining samples for training during each iteration[271].

While CV reduces overfitting and can provide error estimates for predictions, it fails to fully address some pitfalls. For example, while hyperparameter tuning in CV may address overfitting from model complexity, several software packages introduce data leakage by using the same data for evaluation as for training[273].

Double (or nested) cross-validation is a more advanced CV scheme, which employs two nested CV loops, where the outer loop splits data into calibration and testing sets for assessing model performance, similar to a k-fold CV, and the inner loop further divides calibration sets into training and validation sets for hyperparameter tuning[274] (**Figure 4**). This structure ensures that hyperparameter optimization is performed independently for each outer loop fold, mitigating the risk of data leakage and overfitting[275]. However, nested CV comes at the cost of increased computational cost, as the number of models trained and tested grows rapidly with the number of folds chosen for each loop[275].

Despite these improvements, CV remains sensitive to data partitioning strategies, since the segmentation of samples into subsets is a stochastic process[276]. Repeated cross-validation further addresses this stochasticity by performing CV multiple times with different random splits, thereby reducing variance in performance estimates and providing more stable results[277]. This approach is particularly valuable for small datasets, where single segmentations may inadequately represent data distributions. Similarly, repeated double cross-validation (rdCV) extends this concept by executing multiple iterations of nested CV with randomized splits, further enhancing robustness[239] (**Figure 4**).

While rdCV effectively reduces overfitting and enhances evaluation reliability, integrating it with advanced functionalities like variable selection introduces methodological challenges,

stemming from balancing algorithmic integration, computational efficiency and bias reduction, which remain active frontiers in ML research[240,278].

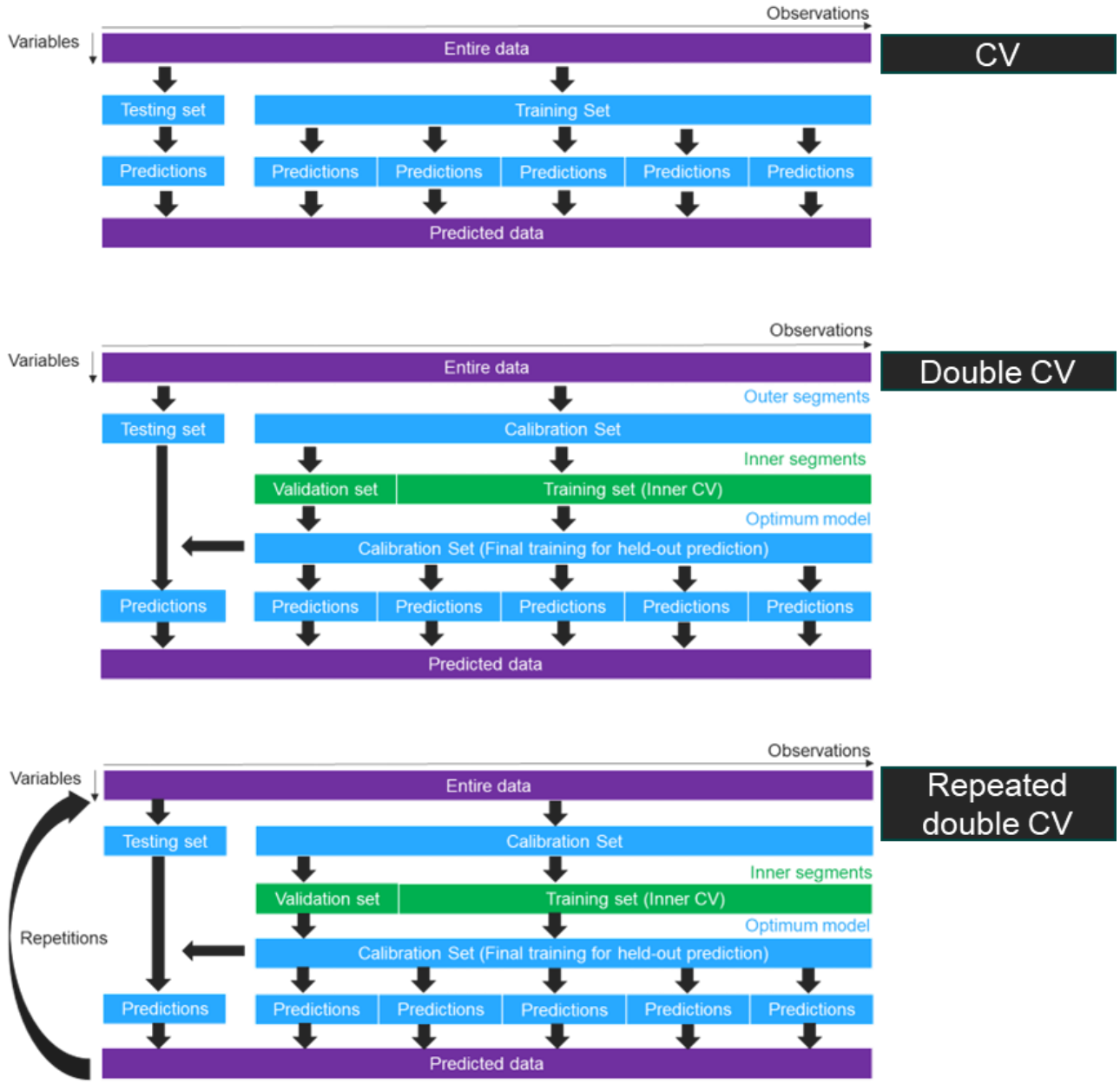


Figure 4. Schematic illustration of different cross-validation strategies, including CV, double CV, and repeated double CV (rdCV). The rdCV approach forms the basis for the MUVR approach further developed in this thesis into MUVR2. CV, cross-validation.

In this thesis, five supervised ML methods were investigated: partial least squares (PLS), random forest (RF), elastic net (EN), support vector machine (SVM) and artificial neural network (ANN).

Partial least squares

Partial least squares (PLS) uses a projection-to-latent space approach to model the linear covariance structure between two matrices (**X** and **Y**). The projection is performed to maximize the covariance of **X** and **Y**[258,279,280]. The newly constructed predictor variables in PLS are referred to as components or latent variables, similar to PCA. PLS discriminant analysis (PLSDA) extends

the PLS framework specifically for classification problems, transforming categorical **Y** to dummy-coded variables[281]. While PLSDA is based on standard PLS regression, its objective shifts toward maximizing class separation rather than covariance prediction[274,282]. PLS and PLSDA are particularly useful when collinearity exists among predictor variables, but are not equally suitable for data with non-linear relationships or significant outliers[283,284]. For preprocessing, PLS is typically performed with standardization to ensure that all variables can contribute equally to the covariance maximization[285].

Variable selection in PLS and PLSDA can be performed e.g. by examining the variable importance in the projection score[285], which quantifies each variable's contribution to the model, with higher scores indicating greater importance. The number of latent variables is the sole hyperparameter to tune in PLS and PLSDA models[286]. In this thesis, PLS serves as one of the core methods in the MUVR2 algorithm, providing a linear approach to identify molecular features associated with either exposures or outcomes (**Paper I**).

Random Forest

Random Forest (RF) operates on the principles of decision tree aggregation and bootstrap sampling[279,280]. A random subset of training samples is selected with replacement (bootstrapping), followed by the creation of a decision tree based on this subset. At each split in the decision tree, a random subset of variables is used. This process generates multiple, partially independent decision trees. Final predictions are determined by aggregating individual tree predictions through majority voting for classification or averaging for regression tasks. The "out-of-bag error" is calculated by aggregating predictions only from trees where the sample was excluded during bootstrapping (i.e., out-of-bag). RF does not assume Gaussian distribution of predictor variables and is particularly effective for handling complex non-linear relationships and mixed variable types[287,288]. For preprocessing, RF is scale-invariant but may benefit from outlier removal[289].

Variable selection in RF is performed using metrics such as mean decrease in impurity or mean decrease in accuracy, which quantify each predictor's contribution to model performance [258,290]. The hyperparameters to tune in RF include the number of trees, maximum depth of trees, minimum samples required to split a node and the number of features considered at each split. In this thesis, RF is implemented as an alternative core method in the MUVR2 algorithm, offering advantages in capturing non-linear relationships between molecular features and either exposures or outcomes (**Paper I**).

Elastic net

Elastic net (EN) is a regularized extension of generalized linear models (GLMs) that combines L1 (lasso) and L2 (ridge) penalties to address limitations of classical regression in high-dimensional settings[291]. While GLMs encompass linear, logistic, and Poisson regressions to model various target variable distributions, they falter with omics data when the number of predictors far exceeds the number of samples, and also struggle with collinearity among predictors, which frequently occurs in omics data[292]. EN addresses these limitations by integrating the sparsity-inducing L1 penalty, which drives trivial coefficients to exactly zero for variable selection, with the L2 penalty, which shrinks and groups correlated predictors to mitigate collinearity [291]. This dual regularization enables EN to perform simultaneous variable selection and coefficient shrinkage, making it particularly suitable for complex biological datasets with many correlated variables [291].

For preprocessing, standardization of predictors is usually performed to ensure that the regularization penalties are applied equally across all variables[291].

Variable selection in EN is achieved through the L1 penalty by identifying variables with non-zero beta coefficients. The variable importance can be assessed based on the magnitude of retained standardized beta coefficients, with larger absolute coefficients indicating more influential predictors. The key hyperparameters to tune include α, which controls the balance between L1 (α=1) and L2 (α=0) penalties, and λ, which controls the overall regularization strength. In this thesis, EN is incorporated into the MUVR2 algorithm, essential for addressing potential confounding in exposure-outcome relationships.

Support vector machine

Support vector machine (SVM) works by mapping data into higher-dimensional space to identify an optimal hyperplane that maximizes the separation between classes (classification) or minimizes prediction errors (regression)[280]. In classification, the hyperplane (i.e., support vector classifier) applies a soft margin that allows controlled misclassification of samples, balancing maximizing margin width and minimizing misclassification (bias-variance trade-off) [293]. In regression, SVM applies a similar principle by fitting a hyperplane within a margin of tolerance, allowing for prediction errors while penalizing larger deviations[293]. In both cases, the observations that lie on the edge or within the margin are called support vectors. SVM employs kernel functions such as linear, polynomial, or radial basis functions to systematically find support vectors and is suitable for binary or multiclass classification problems and complex nonlinear separable data[294,295]. For preprocessing, SVM is usually performed with normalized variables for optimal kernel performance[296].

SVM does not provide built-in methods for ranking variables or performing variable selection. However, variable importance can be inferred from absolute weights calculated by SVM orevaluated using filtering approaches[297] or general methods such as Shapley additive explanations (SHAP)[298] or sensitivity analysis techniques. Key hyperparameters include regularization parameter C, which controls margin violations; kernel types and kernel-specific parameters. Although SVM was investigated as a potential method for integration into the MUVR2 framework, computational demands limited its practical implementation in the final algorithm, as detailed in Section 5.1.1.

Artificial neural network

Artificial neural network (ANN) is based on interconnected nodes organized into layer architectures, comprising an input layer (**X**), one or more hidden layers and an output layer (**Y**) [280]. Each node receives inputs, multiplies them by connection weights, adds a bias term, and then applies an activation function such as linear functions, unit step activation functions, sigmoid functions, hyperbolic tangent, and rectified linear unit functions (ReLU)[299]. The learning process involves backward propagation using gradient descent to iteratively adjust weights to minimize differences between predicted and actual outputs[300]. ANN excels at modeling complex nonlinear relationships and has been widely applied in tasks such as image recognition and natural language processing[301]. For preprocessing, ANN is usually performed with input normalization for stable gradient descent[302].

Similar to SVM, ANN does not have direct methods for variable selection due to its complex structure. However, variable importance can be inferred by examining absolute values of connection weights between input and hidden layers using methods such as Garson's or Olden's

algorithms[303] or general methods such as SHAP[298] .Hyperparameters in ANN include the number of layers and nodes, learning rate, activation function type, batch size and epochs during training processes. As with SVM, ANN was explored for potential integration into the MUVR2 framework but was ultimately not implemented due to computational constraints and challenges in extracting consistent variable importance metrics within the repeated double cross-validation structure.

The supervised ML methods described above offer complementary approaches to analyze relationships between exposures, metabolite features, and outcomes. While PLS and RF form the core methods in the original MUVR algorithm, the addition of EN in MUVR2 represents a methodological advancement that enables adjustment for covariates, a critical feature for molecular epidemiological studies where confounding factors must be considered. The methodological developments section details a comparative analysis of these approaches, evaluating their performance in handling high-dimensional metabolomics data and their suitability for real-world study designs.

3.4.3 The MUVR algorithm

The MUVR algorithm is a supervised ML method designed to identify associations between predictor variables (a matrix **X** of continuous variables) and a target variable (a vector **Y**, continuous for regression or categorical for classification)[240]. It is particularly suited for high-dimensional datasets with few observations, such as omics data[240]. The MUVR algorithm automates the selection of informative variables that contribute to the prediction of **Y** and has been shown to demonstrate improved prediction performance and reduced overfitting, compared to rdCV. The MUVR algorithm is implemented in the *MUVR* R package [240].

The MUVR algorithm incorporates minimally biased variable selection through recursive elimination within a repeated double cross-validation framework [240]. Four computational loops are embedded in the MUVR algorithm for reducing stochastic effects (repetition loop), assessing prediction performance (outer loop), variable ranking and selection (variable selection loop) and hyperparameter tuning (inner loop) (**Figure 5**). MUVR supports PLS and RF as core ML methods, accommodating regression (continuous **Y**), classification (categorical **Y**), and multilevel (utilizing a dummy **Y** and an effect matrix **X**, calculated from data with sample dependency, e.g., before/after or cross-over interventions) modeling [240].

In MUVR, variable ranking in PLS uses variable importance in projection scores, while RF employs the mean decrease in impurity. MUVR enables three different variable selections based on recursive feature elimination, corresponding to different research question perspective[240]: minimal-optimal ('min') selection, which identifies the smallest subset of highly informative variables-of-interest and is useful for the discovery of predictive biomarkers; all-relevant ('max') selection, which retains all variables-of-interest contributing to associations between predictors and target variables and is useful for biological interpretation and mechanistic investigation. MUVR additionally offers a 'mid' selection that corresponds to the geometric mean of the number of variables from the minimal-optimal and all-relevant selection, retaining strong predictors but with some redundancy. The 'Mid' selection may thus be useful in situations where alternatives to the strongest individual predictors are desirable.

Beyond its core algorithm, the *MUVR* package provides tools for permutation analysis and a variety of visualization tools for model interpretation (e.g., stability assessments of variable selection) [240]. The MUVR algorithm has been widely applied in metabolomics research to uncover

metabolic signatures linked to diseases or environmental exposures[13,304,305]. However, the original MUVR algorithm is limited to PLS and RF as core ML methods and lacks functionality for covariate adjustment, which represents a common limitation in many machine learning approaches when applied to study data where confounding factors must be considered.

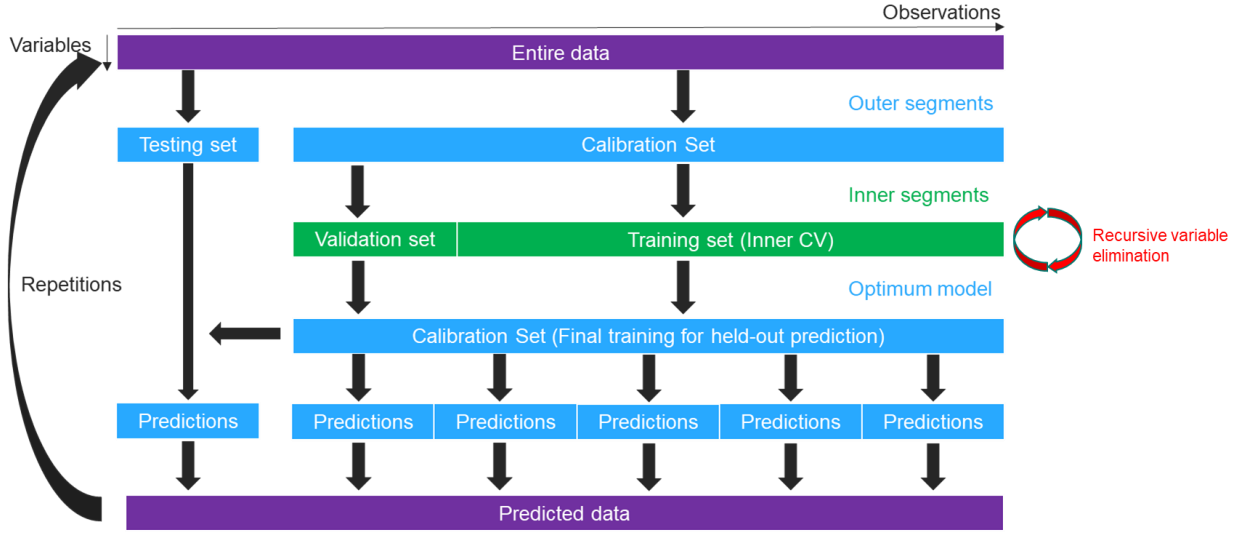


Figure 5. Graphical representation of the MUVR algorithm. The repetition, outer, variable selection and inner loop are presented in purple, brown, blue and green colors. The figure is adapted from the tutorial material from Shi et al 2019[240].

3.5 Exposome-omics-outcome analysis

While supervised ML leverages omics data to predict exposures or outcomes and identifies relevant molecular features, molecular epidemiology aims to investigate molecular associations between exposures, omics, and outcomes through causal frameworks. Establishing causality in observational studies requires careful consideration of study design, temporal relationships, and potential confounding factors. Causal diagrams, typically constructed as directed acyclic graphs (DAGs), constitute powerful visual tools to represent and analyze these complex relationships. These diagrams consist of nodes representing variables (e.g., exposures, omics data, and outcomes) and directed arrows indicating hypothesized causal relationships between them (the variable at the arrow's origin influences the variable at its endpoint).

When omics features serve as intermediates that connect exposures to outcomes, this establishes a temporal sequence: Complex exposures such as diet, pollutants and microbiota may influence metabolic regulations, which may subsequently affect cardiovascular trajectories, influencing e.g., myocardial infarction and stroke (**Figure 6**). Among the omics technologies, metabolomics is particularly suited to represent such metabolic perturbations due to its proximity to phenotypic expression[15,193]. In epidemiological terms, metabolite features can function as potential molecular mediators between exposures and outcomes. However, this requires careful consideration of temporal relationships, as metabolites must be influenced by exposures before affecting outcomes. Confounders may also be present, affecting exposure-mediator and mediator-outcome associations (**Figure 7**). Further contributing to the complexity, important sources of unmeasured confounding may include genetic factors, lifestyle variables not captured in the study, and unmeasured environmental exposures. Two analytical approaches have been commonly employed to investigate potential mechanistic relationships linking

exposures, molecular mediators and outcomes, namely meet-in-the-middle analysis and mediation analysis.

To support mechanistic investigations, the molecular epidemiology tool triplot was developed to enhance visualization and interpretation of exposure-omics-outcome relationships by integrating dimensionality reduction with meet-in-the-middle analysis [306].

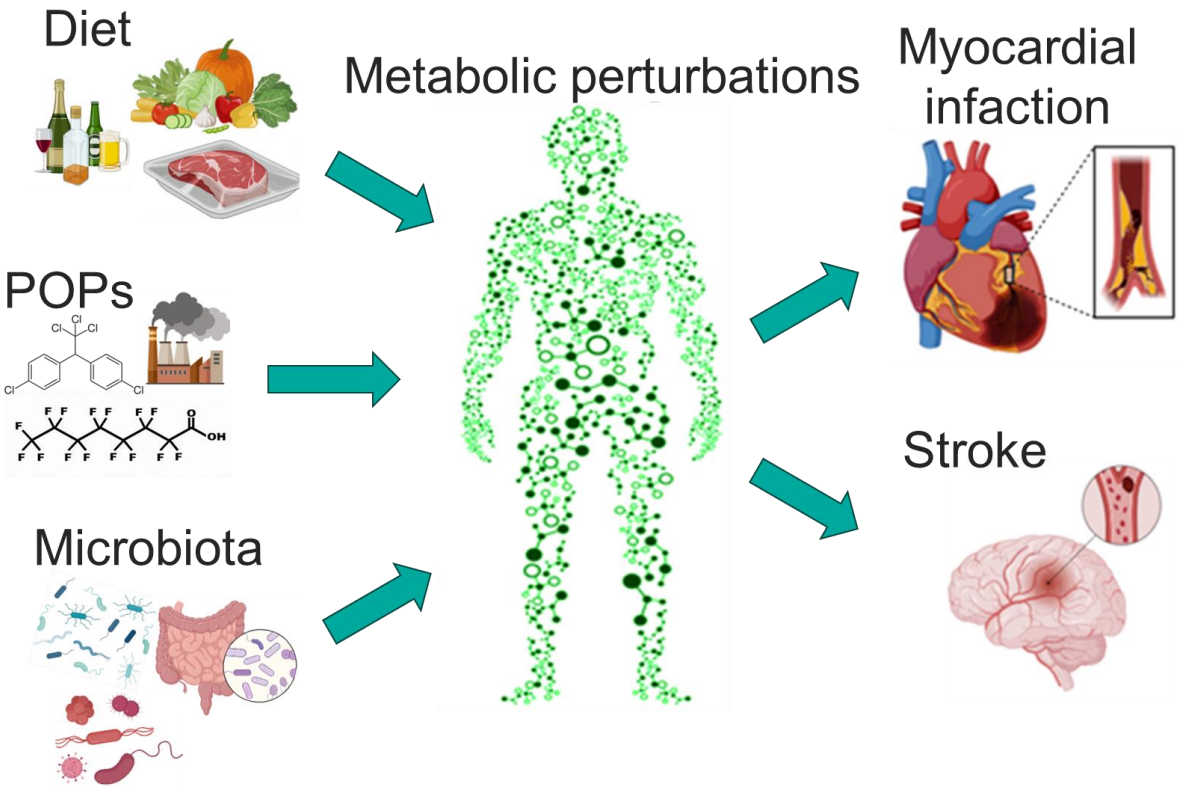


Figure 6. The combined effect of exposures (including diet, persistent organic pollutants and gut microbiota) may produce metabolic perturbations, which can affect cardiovascular health outcomes. Figure created with BioRender icons in PowerPoint. Yan, Y. (2025)

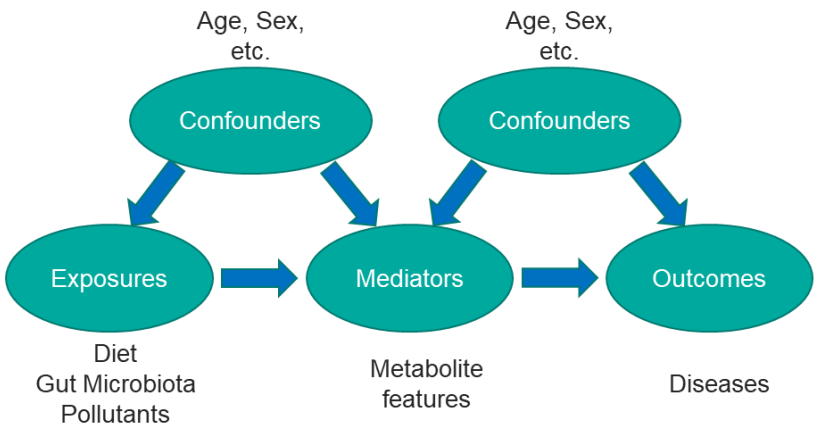


Figure 7. Causal diagram in the form of a directed acyclic graph (DAG), representing hypothesized directional relationships between exposures, molecular mediators and outcomes in the presence of confounders.

3.5.1 The Meet-in-the middle approach

The concept of "meeting-in-the-middle" originates from observations in prospective studies that certain preclinical biomarkers associated both with specific exposures as well as with certain diseases, providing evidence for potential causal pathways between exposures and outcomes[20,307]. More generally, this approach identifies molecular signatures that overlap between exposure-related markers and outcome-predictive markers, thus effectively meeting in the middle, suggesting these overlapping intermediates as potential mechanistic links[20]. The strength of this approach relies on the assumption that molecular features genuinely mediate exposure-outcome relationships rather than simply correlate with both due to unmeasured factors.

By analyzing omics data, the meet-in-the-middle approach[20] investigates metabolic perturbations through which exposures (e.g., diet, POPs, gut microbiota) may affect disease outcomes via molecular mediators. These mediators are selected based on their dual associations with both exposures and outcomes. However, this approach has several limitations: It requires an adequate temporal sequence between exposure, mediator, and outcome which however may be difficult to achieve in cross-sectional studies. It may identify spurious mediators due to unmeasured confounding; and it requires careful consideration of multiple testing when screening large numbers of molecular features. On the other hand, separating the exposure from the outcome analyses simplifies the causal structure leading to less strict assumptions and overall modeling complexity.

3.5.2 The Triplot tool

Triplot is a molecular epidemiology tool designed to visualize and interpret complex relationships between multiple exposures, omics intermediates, and health outcomes[306]. Triplot uses a meet-in-the-middle framework and integrates three layers of information: First, omics data are processed by unsupervised ML (e.g., PCA) to create interpretable components; Second, these component scores are correlated with single or multiple exposures; Third, the associations between component scores and single or multiple outcomes are calculated[306]. Triplot provides a co-visualization of these three layers to facilitate the interpretation of exposure-omics-outcome relationships, enabling holistic analysis in molecular epidemiology [306]. The triplot tool is implemented in the *triplot* R package.

Although triplot was applied in several studies exploring exposure-omics-disease relationships[13,308], it faces limitations. Methodologically, it lacks support for multiclass categorical variables, primarily suggests PCA for data reduction, and does not incorporate more statistically rigorous mediation analysis. Practically, the package requires manual execution in R and provides limited automation, demanding user familiarity with statistical and general coding.

3.5.3 Mediation analysis

Similar to meet-in-the-middle analysis, mediation analysis in epidemiological research aims to disentangle potential mechanistic pathways through which exposures affect outcomes[309]. However, mediation analysis specifically quantifies the extent to which the relationship between an exposure and an outcome may operate through hypothesized mediators[309]. The total effect (TE) represents the observed association between exposure and outcome. The indirect effect (IE) quantifies the portion of this relationship potentially mediated by the hypothesized pathway(s), while the direct effect (DE) captures residual association not explained through such mediator(s)[309,310]. This approach requires that the exposure precedes the mediator, which in turn

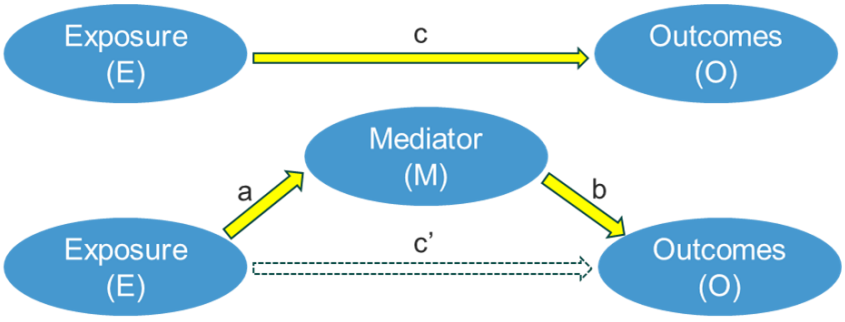
precedes the outcome, ensuring a plausible causal pathway for inference. For simplicity, we denote exposure as **E**, mediator/potential mediator as **M**, outcome as **O**, and confounders as **C**.

Four stringent assumptions must hold[311,312]: No unmeasured **E**-**O**, **E**-**M** or **M**-**O** confounding and no **M**-**O** confounders affected by **E**[311,313]. Despite widespread use, mediation analysis faces challenges in real-world settings and its issues and recommendations have been widely discussed [309,314–319] . For example, the violation of assumption 4 (exposure-induced confounding) is particularly problematic in longitudinal studies, where exposures can be measured at multiple time points and may influence later confounders of the mediator-outcome associations. To detect potential violation of assumptions, sensitivity analyses can be used to quantify robustness to unmeasured confounding. Mediation analysis can be conducted using both conventional regression-based as well as more modern counterfactual approaches[320].

Conventional mediation analysis

Baron and Kenny[318] established the foundational approach for conventional mediation analysis requiring four sequential criteria: 1) **E** associates with **O**; 2) **E** associates with **M**; 3) **M** associates with **O** conditional on **E**; 4) Attenuation of the **E**-**O** association when adjusting for **M**[321].

Two common estimation methods are the difference-in-coefficients and the products-of-coefficients[311,321,322]. In the difference-in-coefficients method, the mediation effect is estimated by comparing the associations between **E** and **O**, with and without **M** adjustment. In the product-of-coefficients method, two regressions are employed, where one regresses **O** on **E** and **M** (outcome model) and the other regresses **M** on **E** (mediator model) [311]. The indirect effect is the product of the exposure coefficient in the mediator model and the mediator coefficient in the outcome model[18,311] (**Figure 8**).



$O = i_1 + c\,E + \varepsilon_1$

$O = i_2 + c'\,E + b\,M + \varepsilon_2$

$M = i_3 + a\,E + \varepsilon_3$

Total effect (TE) = c = a * b +c'

Direct effect (DE) = c'

Indirect effect (IE) = a * b (product-of-cofficients method)

Indirect effect (IE) = c – c' (difference-in-coefficients method)

Figure 8. Causal diagrams and formulae for conventional mediation analysis, using the difference-in-coefficients and product-of-coefficients method. Confounders are not considered in this representation. Letters a, b, c and c' represent regression coefficients from linear models; i represents intercept and Ɛ represents residual. E, exposure; M, mediator; O, outcome.

Counterfactual mediation analysis

Counterfactual mediation analysis, rooted in the potential outcomes framework, provides a robust foundation for causal inference by comparing hypothetical scenarios under different exposure and mediator conditions. Unlike conventional mediation analysis, which estimates an average causal effect (ACE) across populations, counterfactual methods evaluate effects at the individual level, allowing for more rigorous causal interpretation.

In counterfactual mediation analysis, a potential outcome refers to the hypothetical value of an individual's outcome under specific exposure and mediator conditions. Consider a scenario with a single exposure, **E**, which can take the value **e** or **e*** (for example, unexposed versus exposed), and a mediator, **M**, which can have values **m** or **m***, corresponding to the value under exposure **e** or **e***. Although each individual can only be observed under one combination, either {**e, m**} or {**e*, m***}, two additional potential outcomes are imaginable: {**e, m***} and {**e*, m**}, resulting in four possible potential outcomes. Three causal effects are typically examined: The natural direct effect (NDE) quantifies the direct impact of **E** on **O** when **M** is set to the value it would naturally take under the reference exposure (such as the unexposed state, i.e., comparing $\mathbf{O}_{\{e,\, m\}}$ and $\mathbf{O}_{\{e^*,\, m\}}$); In contrast, the natural indirect effect (NIE) captures how changes in **M** induced by exposure (the natural values of **M** under exposed and unexposed scenarios) contribute to the effect of **E** on **O** (i.e., comparing $\mathbf{O}_{\{e^*,\, m\}}$ and $\mathbf{O}_{\{e^*,\, m^*\}}$); Finally, the controlled direct effect (CDE) quantifies the direct impact of exposure **E** on outcome **O** when the mediator **M** is fixed to a specific value $\mathbf{m}^{\dagger}$ independent of **E** (such as the population mean, i.e., comparing $\mathbf{O}_{\{e,\, m\dagger\}}$ and $\mathbf{O}_{\{e^*,\, m\dagger\}}$)[309,323] . Among these effects, NIE is the most important to establish mediation, as it quantifies the extent to which the association between exposure and outcome is explained by the mediator

In counterfactual mediation analysis, the potential outcomes are generated from models trained on the observed data. However, unlike conventional mediation analysis, which relies on linear models, the counterfactual framework operates without parametric constraints (e.g., predefined functional forms or estimation methods), making it inherently non-parametric[324] . This flexibility enables applications to diverse variable types (continuous, categorical, or mixed) and complex mediation scenarios involving non-linear relationships or interaction effects[323]. Crucially, the framework explicitly separates foundational causal assumptions—such as the requirement for no unmeasured confounding—from the choice of estimation techniques. While practical implementations frequently employ parametric models to estimate potential outcomes, the counterfactual framework inherently accommodates both linear and non-linear estimators, adapting to different data types and study designs.

While the analyses in this thesis focused on single-mediator scenarios, emerging methods for high-dimensional mediators and exposure-mediator interactions represent critical extensions. However, more complex causal structures simultaneously involving multiple mediators fall outside the current scope.

4. Material and methods

This chapter outlines the datasets and methods underpinning the research presented in this thesis. It describes the test datasets used for the development and validation of the MUVR2 algorithm (**Paper I**) and the TriplotGUI toolbox (**Paper II**), as well as the study populations, data collection procedures, and statistical methods applied in the cohort studies (**Papers III** and **IV**).

4.1 Test data for MUVR2 (Paper I)

Three datasets with complementary characteristics were employed to validate and showcase MUVR2 across different analytical scenarios: regression modeling (Freelive2), binary classification (Mosquito), and epidemiological case-control analysis (BioDiVa).

Freelive2

This dataset investigates metabolic profiles in relation to dietary exposures, adapted from the original MUVR Freelive dataset[240]. It includes 1147 urine metabolite features from 58 participants as predictors (**X**), with self-reported whole-grain rye consumption serving as the continuous target variable (**Y**).

Mosquito

The dataset is derived from a study of microbiota composition in mosquitoes across villages in Burkina Faso[240]. It includes 1678 16S rRNA operational taxonomic units (OTUs) as predictors (**X**), with 738 OTUs retained after filtering near-zero variance. The categorical target variable (**Y**) corresponds to the village of mosquito capture.

BioDiVa

This dataset is derived from a nested case-control study that profiled metabolic signatures associated with future type 2 diabetes onset in the Västerbotten Intervention Program[325]. Notably, one of the cohorts used in **Paper IV** (see section 4.3.2) is also derived from this nested case-control study. The cohort shares the name BioDiVa but differs in the selection of cases and controls. Predictors (**X**) comprise 24758 metabolite features from 842 participants (cases and individually matched controls), with prospective binary diabetes status as the target variable (**Y**). Additional covariate data, such as age and sex, are also available for analysis.

4.2 Test data for TriplotGUI (Paper II)

HealthyNordicDiet2

The HealthyNordicDiet2 data were employed to validate and showcase TriplotGUI. The dataset was simulated from authentic data used in a matched case-control study nested within the Swedish prospective Västerbotten Intervention Programme cohort. It contains information on 1000 synthetic observations (500 cases and 500 controls matched by gender and age), 17 exposure variables representing dietary indices and food intake; 31 omics variables and three outcomes: Future type 2 diabetes (T2D) as a binary variable, numeric body mass index (BMI) and categorical BMI (BMI_cat) with 4 levels (obese, overweight, normal, underweight). Additionally, five covariates (gender, age, education, physical activity level, smoking) were included as potential confounders, along with auxiliary data containing pairing information matching controls to cases. The simulated dataset preserves the statistical properties and correlation structure of the original data while ensuring privacy protection.

4.3 Study populations (Paper III, IV)

4.3.1 SIMP-UC (Paper III, IV) and SIMP-VC (Paper IV)

SIMPLER-Uppsala Clinical (SIMP-UC) and SIMPLER-Västerås Clinical (SIMP-VC) are part of the Swedish Infrastructure for Medical Population-Based Life-Course and Environmental Research (SIMPLER; https://www.simpler4health.se/)[326]

SIMP-UC: The Swedish Mammography cohort (SMC) was established in 1987, where women born during 1914–1948 residing in Västmanland County between 1987 and 1989 or living in Uppsala County between 1988 and 1990 (74% response rate, *n* = 61,433) were invited. From 2003 through to 2009, a subset of 5,022 women in SMC under 85 years old from Uppsala underwent health examinations, providing blood samples, anthropometric measurements (weight, height, waist/hip circumference), and questionnaire data. This population constitutes the SIMP-UC cohort.

SIMP-VC: The Cohort of Swedish Men (COSM) was established in 1997, where men born between 1918 and 1952 in Västmanland and Örebro counties (49% response rate, *n* = 45,906) were invited. From 2010 through to 2019, participants in COSM over 75 years old and living in Västerås were invited for health examination, using the same data collection protocols as SIMP-UC. When all men over the age of 75 had been invited, additional men in COSM living in Västerås County were invited, starting with the earliest birth year. In addition, between 2010 and 2019, women in SMC living in Västmanland county were invited for health examination. A total of 2,630 women and 4,754 men constitute the SIMP-VC cohort.

4.3.2 BioDiVa (Paper IV)

The Västerbotten intervention programme (VIP), is a subcohort of the Northern Sweden Health and Disease Study cohort and was launched in Västerbotten County in 1985 with the aim of reducing morbidity and mortality from CVD and diabetes [327]. In VIP, middle-aged individuals were invited when they reached 30, 40, 50 or 60 years of age for health examinations, including blood sampling and lifestyle surveys. Approximately 6,500-7,000 examinations take place each year and yearly participation rates range between 48-67%. The VIP cohort included 141,000 sampling occasions from 98,300 unique individuals by March 2015. The BioDiVa cohort includes 503 case-control pairs from a study of type 2 diabetes nested within the VIP cohort [325].

4.3.3 MAX (Paper IV)

The Diet, Cancer, and Health - Next Generations MAX study (DCH-NG MAX) is a validation subsample of the larger Diet, Cancer and Health-Next Generations (DCH-NG) cohort. The DCH-NG MAX study was conducted between August 2017 and January 2019, aiming to investigate dietary habits and their associations with health outcomes among participants aged 18 years and older[328]. It was designed as a one-year observational study with evaluations at baseline, six months, and 12 months[329], with a total of 720 individuals enrolled in the study. The MAX cohort includes 676 participants who completed at least one 24-hour dietary recall and health examinations including blood sampling [330].

Written informed consent was obtained from all study participants in all cohorts. Detailed information is available in **Papers III** and **IV**.

4.4 Omics assessment (Paper III, IV)

4.4.1 Metabolomics (Paper III, IV)

Untargeted metabolomics data across SIMP-UC, SIMP-VC, BioDiVa and MAX were acquired using LC-MS measurements on blood samples. Missing values were imputed prior to statistical analysis, using a random forest-based method (in-house R package "StatTools", https://gitlab.com/YingxiaoYan/StatTools). Detailed descriptions of the analytical procedures are provided in **Papers III and IV**.

SIMP-UC and SIMP-VC: Samples were analyzed in positive and negative ionization modes using a reversed-phase (RP) column. Study-specific quality controls (sQCs) and long-term quality controls (ltQCs) were used for drift correction and batch normalization[331,332]. In **Paper III**, features with a coefficient of variation >30% among quality control samples were removed. Features with putative POP annotations were excluded.

BioDiVa: Samples were analyzed in positive and negative ionization modes using both RP and HILIC columns. A constrained randomization scheme ensured paired case-control samples remained within the same batch, with sQCs and ltQCs similarly used for drift correction and batch normalization[325].

MAX: Samples were analyzed using the same methods as for SIMP-UC and SIMP-VC[331,332].

4.4.2 Proteomics and Genomics (Paper III)

In **Paper III**, proteomics and genomics data were generated in SIMP-UC from the same samples as for metabolomics analysis.

Proteomics

Proteins were quantified using high-throughput multiplex immunoassays: Olink® Proseek Multiplex CVDII, CVDIII and Metabolism (Olink Bioscience, Uppsala, Sweden) and yielded normalized protein expression values on a log2 scale standardized per analysis plate. Proteins with more than 25% values below limit of detection were removed prior to analysis and remaining missing values were imputed (StatTools R package).

Genomics

Genotyping was performed and single-nucleotide polymorphism (SNPs) were imputed using Haplotype Reference Consortium (HRC) v1.1 and 1000 Genomes project phase 3 reference panels. The results were then analyzed using GenomeStudio 2.0.3 from Illumina. The sample success rate was ≥98%. To prefilter the genetics data, SNPs associated with POP exposures were selected using linear model with additive effects at a cut-off of $p<0.000005$, balancing the stringency required to limit false positives with the need to retain sufficient variants for downstream analysis, using the Plink 2.0 software.

4.5 Outcomes, exposures and covariates assessment (Paper III, IV)

4.5.1 Outcome assessment (Paper III, IV)

First incident cases of MI and cases of ischemic stroke were ascertained via linkage of the SIMP-UC and SIMP-VC cohorts to the Swedish National Inpatient Register. The diagnoses were based

on the International Classification of Diseases, 10th Revision (ICD-10, WHO 2016), with codes I21 for MI and I63 for ischemic stroke, covering the period from baseline blood sampling through 2021. Follow-up time ranged from 0.0 to 16.0 years (median: 7.4 years) for MI cases; 0.0 to 16.5 years (median: 8.4 years) for stroke cases and 12.3 to 18.2 years (median: 14.6 years) for controls in SIMP-UC and from 0.0 to 10.5 years (median: 3.0 years) for MI cases; 0.0 to 10.9 years (median: 3.6 years) for stroke cases and 2.5 to 12.5 years (median: 6.8 years) for control in SIMP-VC. Individuals who had experienced MI or stroke before plasma sample collection or who lacked omics data were excluded from the analysis.

4.5.2 Dietary assessment (Paper IV)

Diet was assessed in SIMP-UC and SIMP-VC for **Paper IV** using validated food frequency questionnaires (FFQs). A range of dietary variables, including food items, nutrients and dietary scores were investigated as exposures. In SIMP-UC, participants received four different diet and lifestyle questionnaires at various time points: FFQ 1997, FFQ 2004V1, FFQ 2004V2, and the 2009 Lifestyle questionnaire. In SIMP-VC, the vast majority of participants completed the 2009 Lifestyle Month and Uppsala Home – Lifestyle Month questionnaires. Additionally, two dietary scores, i.e., a modified Mediterranean Diet score (mMed)[333] and the Dietary Approaches to Stop Hypertension score (DASH)[334], were calculated and included as dietary variables in the analysis.

Food items and nutrient variables with near zero variance (i.e., where unique values were present in less than 10% of all observations and where the ratio of the most common to the second most common value exceeded 95/5) or variables with >2/3 missing values were excluded. The remaining variables, including the 2 dietary scores, were included for individuals with available dietary and metabolomics data in the data analysis.

4.5.3 Persistent organic pollutants assessment (Paper III, IV)

POPs were measured in SIMP-UC in **Papers III and IV** and in BioDiVa in **Paper IV.**

SIMP-UC: OCs were measured in plasma samples by gas chromatography-triple quadrupole mass spectrometry (GC-MS/MS)[335]. Twenty-five compounds were analyzed: 13 Polychlorinated biphenyls (PCBs) (congeners 28, 52, 74, 99, 101, 118, 138, 153, 156, 170, 180, 183, and 187); 9 organochlorine pesticides or their metabolites: dichlorodiphenyltrichloroethane (p,p'-DDT), dichlorodiphenyldichloroethylene (p,p'-DDE), α-hexachlorocyclohexane (α-HCH), β-HCH, γ-HCH, pentachlorobenzene (PeCB), hexachlorobenzene (HCB), transnonachlor, and oxychlordane; and 3 polybrominated diphenyl ethers (BDE 47, 99, 153). PFAS were measured in plasma samples by targeted liquid chromatography-triple quadrupole mass spectrometry (LC-MS/MS)[336], covering eight PFAS compounds: perfluorohexanesulfonate (PFHxS), perfluoroheptanoic acid (PFHpA), perfluorooctanesulfonic acid (PFOS), perfluorooctanoic acid (PFOA), perfluorononanoic acid (PFNA), perfluorodecanoic acid (PFDA), perfluoroundecanoic acid (PFUnDA), and perfluorododecanoic acid (PFDoA).

BioDiVa: POPs were measured in plasma samples following similar methods as in SIMP-UC, albeit excluding 3 OCs, PCB (28, 52, 101) and additionally including 5 PFAS: Perfluorohexanoic acid (PFHxA), perfluorotridecanoic acid (PFTrA), perfluorotetradecanoic acid (PFTA), perfluoroheptane sulfonic acid (PFHpS) and perfluorodecane sulfonic acid (PFDS).

In **Paper III**, nine POP compounds were excluded prior to analysis due to >50% of values being below the limit of detection (LOD) (i.e., PFDoA, PeCB, α-HCH, γ-HCH, BDE 47, BDE 99, and BDE 153) or due to sample contamination (i.e., PFOA and PFHpA). Concentrations below LOD were

replaced by LOD/√2. In **Paper IV**, in both SIMP-UC and BioDiVa, the OCs above LOD were normalized using the standard formula for calculating total lipids (Total lipid = 2.27 × total cholesterol + triglycerides + 0.623 [g/L])[337]. For OC concentrations below the LOD, the corresponding lipid-normalized LOD/√2 value was assigned. For PFAS, the LOD/√2 value was used for concentrations below the LOD.

4.5.4 Gut microbiota assessment (Paper IV)

Gut microbiota composition was estimated for SIMP-VC and MAX and the details of the procedures are described in **Paper IV**.

SIMP-VC: After fecal sampling, DNA was extracted and Shotgun metagenomic sequencing of the gut microbiome was performed. For taxonomic composition, raw sequence data was processed with the StaG-metagenomic workflow collaboration[338,339], which includes host-DNA removal and taxonomic profiling using MetaPhlAn version 4.0.3 [340]. Microbiota compositions were represented by relative abundance, which was calculated as the proportion of reads assigned to each taxon relative to the total reads per individual at each taxonomic (i.e., kingdom, phylum, class, order, family, genus, species and strain) level. Alpha diversity indices (i.e., Shannon and inverse Simpson index[341]) were also calculated.

MAX: DNA extraction of fecal samples was performed and 16S rRNA gene sequencing was conducted. Following sequencing, data processing was carried out using the QIIME2 software, which involved removing primers, ambiguous sequences, and chimeric sequences through the dada2 method. Amplicon sequence variants (ASVs) were then assigned taxonomy using the Silva 16S rRNA database at a ≥97% identity threshold. ASVs that shared taxonomic identity were merged, while single-read phylotypes were excluded from further analysis [342]. ASVs in MAX were presented at kingdom, phylum, class, order, family and genus levels.

Participants with recent antibiotic use within six months or missing microbiota or metabolomics data were excluded. Microbiota features with near-zero variance or with prevalence <2/3 were excluded. Centered log-ratio transformation was then applied to the microbiota features at each taxonomic level to mitigate compositional data bias.

4.5.5 Covariate assessment (Paper III, IV)

Covariates were selected based on established confounding potential and directed acyclic graph (DAG) considerations. Variables that may impact both metabolic regulation and exposures (i.e., diet, POPs, gut microbiota) or cardiovascular outcomes were considered potential covariates, including age, sex, education (primary school, high school, university), smoking status (current, ever and never smoker), BMI and total energy intake (kCal/day). These variables were collected consistently across the cohorts via standardized questionnaires.

Additionally, the number of food items investigated varied between questionnaires within SIMP-UC, with potential impact on the calculated total energy intake. To address this issue, the questionnaire type was included as an additional covariate. In SIMP-VC, dietary assessment was homogeneous across participants, wherefore questionnaire type was not adjusted for.

In **Paper III**, missing values of covariates in SIMP-UC were imputed using a random forest method (R package “StatTools”) with omics data as predictors. In **Paper IV**, missing values of covariates in SIMP-UC and SIMP-VC were imputed separately, using MUVR2’s random forest-based algorithm with metabolomics data as predictors. The proportion of missing data for each covariate ranged from 0% to 5.67%.

4.6 Study design and statistical analysis (Paper III, IV)

To investigate the associations between exposures and outcomes via omics, meet-in-the-middle and mediation analyses were used to identify omics features associated with both exposures and outcomes. Combining machine learning and classical univariate analysis, **Paper III** adopted an exposure-centric approach, where omics features associated with exposures were first selected and their association with outcomes subsequently investigated; **Paper IV** instead adopted an outcome-centric approach, where omics features associated with outcomes were first selected and their associations with exposures subsequently investigated.

A comparison between the design and statistical analysis steps in **Papers III and IV** is summarized in **Table 1**.

4.6.1 Persistent organic pollutants, omics and cardiovascular disease risk in Swedish women (Paper III)

In **Paper III,** a nested case-control study design was applied within the SIMP-UC cohort, focusing on first incident cases of primary MI and ischemic stroke. Controls were randomly matched to cases based on age (±1 year) and sample date (±90 days), with 1 case to 2 controls for MI and 1 case to 1 control for stroke. Only subjects with available omics data were included (n=657).

To capture the variability among POP exposures and reduce the number of analyses, a principal component analysis with varimax rotation was performed on POPs, resulting in two components (57% of total variance) representing OCs and PFAS, respectively. Prior to statistical analysis, metabolite features were log-transformed to achieve approximate normality and standardized to unit variance.

Statistical analysis steps were performed as outlined in **Table 1**. At Step 1, omics features were associated with POPs cross-sectionally. Omics data were used as predictors and the OC and PFAS components were used as target variables, respectively, analyzed using random forest regression implemented in the MUVR algorithm. The MUVR algorithm employed repeated double cross-validation to minimize overfitting and provide robust variable selection. At Step 2, associations between POPs and POP-related omics features were adjusted for confounders using partial Spearman correlation. POP-related features with p-value<0.05 were retained for the next step. At Step 3, conditional logistic regression was performed for the prospective associations between POP-related omics features and CVD risk, including MI, stroke or composite CVD. At Step 4, relationships between POPs, selected molecular features and CVD outcomes were visualized using a triplot.

4.6.2 Exposome, omics and cardiovascular disease risk in Nordic adults (Paper IV)

In **Paper IV**, the study design expanded to include four cohorts, SIMP-UC (n=4876), SIMP-VC (n=7518), BioDiVa (n=402), and MAX (n=513), for cross-cohort discovery and replication (**Figure 9**). Prior to statistical analysis, metabolite features were standardized.

Table 1. Overview of designs and methods used in **Papers III** and **IV**.

	Paper III	**Paper IV**
Cohort	SIMP-UC (all female)	SIMP-UC (all female) SIMP-VC BioDiVa MAX
Design	Nested case-control study in a prospective cohort	Discovery and replication across cohorts
Outcome	Composite CVD/MI/stroke	MI/stroke
Omics	Genetics, Proteomics and metabolomics	Metabolomics
Exposure	POPs	Diet, POPs and gut microbiota
Covariates used in:		
Exposure-omics feature association	Age, sample year, education, healthy diet score and additionally BMI for OCs	Age, sex, education, smoking status, total energy intake, questionnaire type and additionally BMI for OCs
Omics feature-outcome association	Age, sample year, education, family history of CVD, smoking status, physical activity, healthy diet score, BMI, HDL, LDL, triglycerides, hypertension	Age, sex, education, smoking status.
Statistical analysis	Exposure-centric meet-in-the-middle analysis	Outcome-centric meet-in-the-middle and mediation analyses
Step 1	Select omics features that are cross-sectionally associated with POPs, using random forest (MUVR)	Select metabolite features that are prospectively associated with MI/stroke outcomes in discovery cohorts, using elastic net (MUVR2)
Step 2	Select POP-related omics features that were not a result of confounding, using partial Spearman correlation	Investigate the selected features' associations with outcomes in replication cohorts, using logistic regression
Step 3	Assess prospective associations between POP-related omics features and CVD/MI/stroke outcomes, using conditional logistic regression	Assess cross-sectional associations between exposure-outcome-related omics features in discovery cohorts, using partial Spearman correlation
Step 4	Visualize the selected POP- and outcome-related features with POPs and outcomes, using Triplot	Investigate exposure-outcome-related features' associations with exposures in replication cohorts, using partial Spearman correlation
Step5		Investigate the causal mediation effect of exposure- outcome-related features using conventional and counterfactual mediation analysis
Step 6		Visualize the selected exposure-outcome-related features with exposures and outcomes, using TriplotGUI

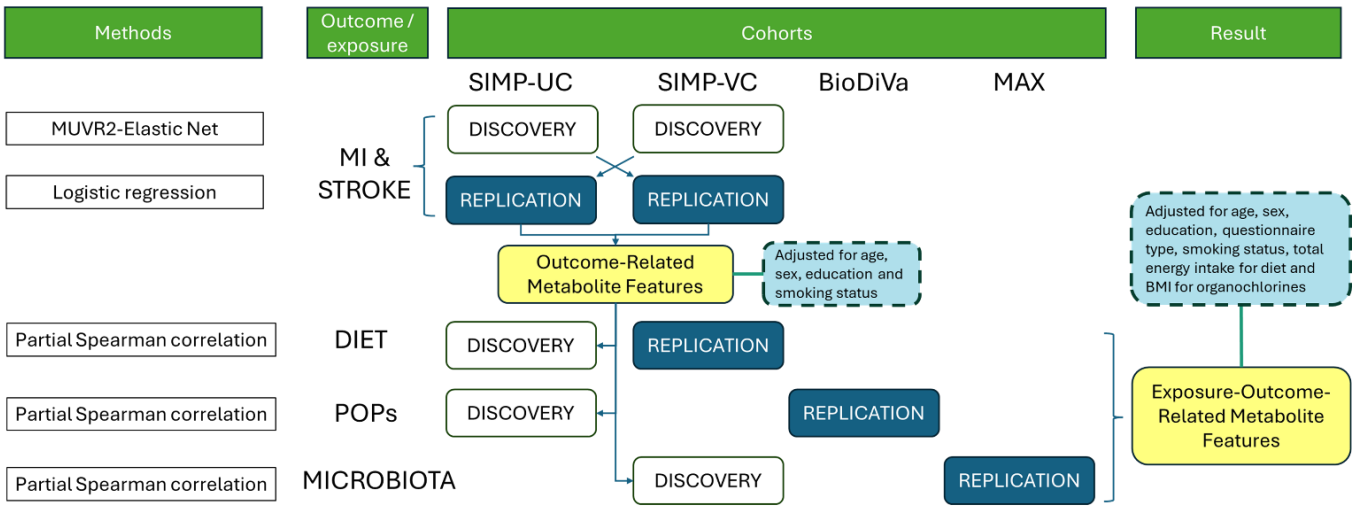


Figure 9. Schematic of the study design and the selection of metabolite features for **Paper IV**. Metabolite features associated with MI and stroke outcomes were first identified in the cross-cohort discovery and replication. These outcome-related features were then examined for their associations with exposures in the discovery and replication cohort to identify exposure-outcome-related features. MI, myocardial infarction; POP, persistent organic pollutants**.**

Statistical analysis steps were performed as outlined in **Table 1**. At Step 1, using the SIMP-UC cohort for discovery, metabolite features associated with MI or stroke were identified. For each outcome, cases were matched with an equal number of controls via random sampling. Elastic net modeling (MUVR2-EN) was applied to metabolite features as predictors and case-control status as the target variable. This process was repeated five times, and features selected in at least three iterations were retained for replication. At Step 2, using SIMP-VC for replication, logistic regression analysis was performed to replicate the associations between selected features and outcomes. Metabolite features with same effect direction and p-value<0.05 were considered replicated. The entire discovery and replication procedure was then repeated using SIMP-VC for discovery and SIMP-UC for replication. Only features that were associated with either MI or stroke in the same direction in both cohorts were considered outcome-related features for further analysis.

At Step 3, outcome-related features were linked to exposures using partial Spearman correlation in SIMP-UC as the discovery cohort for dietary and POP exposures and SIMP-VC as the discovery cohort for microbiota exposures. Associations with p-value<0.05 and coefficient $|\rho| > 0.2$ were retained for replication, balancing statistical significance with biological relevance. At Step 4, using partial Spearman correlation, SIMP-VC, BIODIVA and MAX were used as replication cohorts for diet, POPs and microbiota respectively. Replicated features (p-value<0.05) were considered exposure-outcome-related features.

At Step 5, conventional and counterfactual mediation analysis was performed to further explore potential causal associations for the selected exposure-metabolite-outcome associations, with mean + 1 standard deviation and mean - 1 standard deviation as contrasts for exposures in counterfactual mediation analysis[19]. A significant mediation effect was defined as an indirect effect with a p-value<0.05. However, we acknowledge that observational mediation analysis cannot establish definitive causal relationships and results should be interpreted as suggestive of potential mechanistic pathways requiring validation through experimental studies. At Step 6, relationships between exposures, selected exposure- and outcome-related features and CVD outcomes were visualized using TriplotGUI.

5. Summary of findings - Methodological development (Paper I, II)

This chapter presents the methodological contributions of this thesis, comprising two novel computational tools for molecular epidemiology research**. Paper I** describes the development of MUVR2, an enhanced machine learning framework for high-dimensional feature selection with covariate adjustment capabilities. **Paper II** presents TriplotGUI, a user-friendly toolbox for integrated analysis and visualization of exposure-omics-outcome relationships. Together, these methodological advances address critical gaps in molecular epidemiology by providing robust tools for variable selection and mechanistic interpretation through omics in complex multi-exposure/outcome studies.

5.1 MUVR2 (Paper I)

The MUVR2 algorithm represents a significant enhancement to the original MUVR framework, addressing key limitations in ML approaches for molecular epidemiology. Key improvements include: Incorporating elastic net analysis, which allows covariate adjustment; Providing a general framework for assessing prediction performance and overfitting and; Expanding the compatibility with diverse variable types such as categorical and mixed data. *MUVR2* is publicly available as an R package on the Comprehensive R Archive Network (CRAN) (https://cran.r-project.org/web/packages/MUVR2/index.html).

5.1.1 The MUVR2 Algorithm

Algorithm Development and Machine Learning Integration

The MUVR2 development process systematically evaluated the integration of three additional machine learning methods beyond the original PLS and RF implementations: elastic net (EN), support vector machines (SVM), and artificial neural networks (ANN). As MUVR2 performs recursive variable elimination, a variable importance rank is calculated in its repeated double cross-validation structure to guide variable selection. However, unlike PLS and RF, which internally compute variable importance scores useful for ranking variables, SVM and ANN do not inherently provide variable importance measures or perform variable selection. While EN enables variable selection by identifying variables with non-zero beta coefficients, it does not internally calculate a variable importance rank. To address these limitations, additional methods and R packages were investigated in MUVR2 to perform variable ranking and selection for SVM, ANN, and EN (**Table 2**). SHAP values were not used for variable importance ranking in MUVR2, due to their high computational demands.

Table 2. R packages involved in the MUVR2 for ML analysis, variable ranking and a brief comment on variable selection mechanism. PLS, partial least squares; RF, randomForest; EN, elastic net; SVM, support vector machine; ANN, artificial neural network; CV. cross-validation

	R package for ML analysis	R package for variable ranking	Variable selection mechanism
PLS	*pls*	*pls*	Built-in variable importance in projection scores
RF	*randomForest/ ranger*	*randomForest/ ranger*	Built-in variable importance
EN	*glmnet*	-	Variable importance based on the number of times having non-zero beta coefficients across repetitions and outer CV loops
SVM	*kernlab*	*rminer*	Variable rankings based on sensitivity analysis
ANN	*neuralnet/nnet*	*neuralNetTools*	Variable rankings based on connection weights using Olden and Garson's algorithms

Implementation Challenges and Solutions

Although SVM and ANN are theoretically compatible with the MUVR2 framework, their computational demands made them impractical for implementation within the repeated double cross-validation structure, particularly when combined with hyperparameter optimization. Computational complexity scales exponentially with the nested cross-validation loops, making SVM and ANN prohibitively resource-intensive for typical omics datasets. Consequently, the current MUVR2 implementation focused on EN integration, with two variable selection options based on variables with non-zero beta coefficients across repetitions and outer CV loops:

(1) Prediction performance is estimated as a function of the number of selected variables using locally weighted least square regression. This approach generates a curve analogous to variable selection for PLS and RF in the original MUVR algorithm, enabling variable selection based on optimal prediction thresholds (**Figure 10a**). (2) Only the quantiles of the number of variables selected across repetition and outer loops are used to define minimal-optimal and all-relevant variables (**Figure 10b**). The latter option is more suitable when the curve in the former exhibits instability.

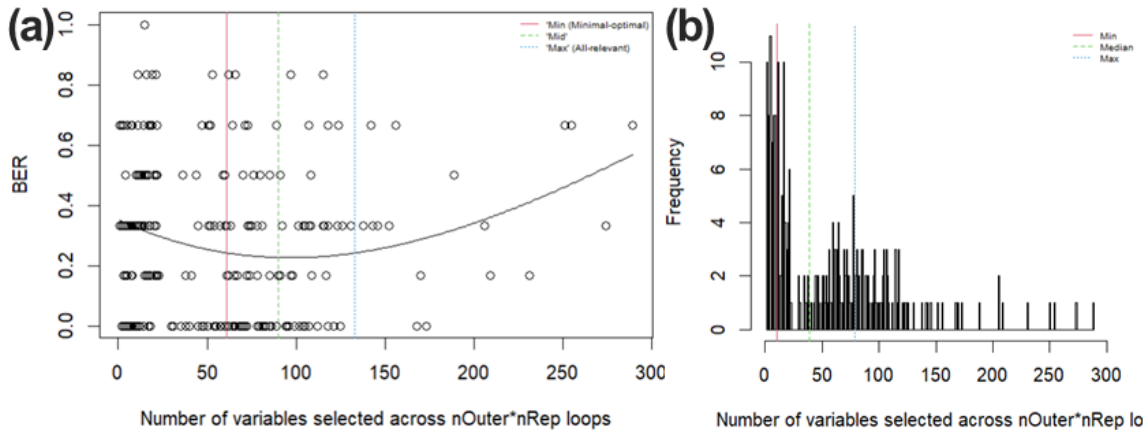


Figure 10. Example of validation plot and variable selection of MUVR2-EN in classification modeling using Mosquito data. (MUVR2-EN parameters: *nRep = 35, nOuter=8*). **(a)** Prediction performance is estimated as a function of the number of selected variables across repetition and outer loops. Variable selection is based on the fitted curve. **(b)** Variable selection is only based on the quantiles of the number of variables selected across repetition and outer loops. The figure is taken from Yan et al 2024[237] under a CC BY 4.0 license.

In addition to incorporating elastic net, MUVR2 introduces several enhancements beyond its predecessor. Covariate adjustment is supported in MUVR2 through elastic net. To systematically assess prediction performance and overfitting, MUVR2 provides both permutation and resampling tests as well as a reference distribution, enabling direct comparison of the model's actual prediction performance to what would be expected by chance. In addition, categorical predictors are allowed in MUVR2 through a one-hot-encoding procedure, transforming categorical variables into multiple binary numerical variables. Furthermore, prediction performance in MUVR2 can be evaluated by balanced error rate (BER) for classification tasks in addition to metrics previously employed in the original MUVR algorithm. A detailed presentation of the improvements offered by MUVR2 is presented in **Figure 11.** The MUVR2 functionality for covariate adjustment and assessing prediction performance and overfitting is described in sections 5.1.2 and 5.1.3.

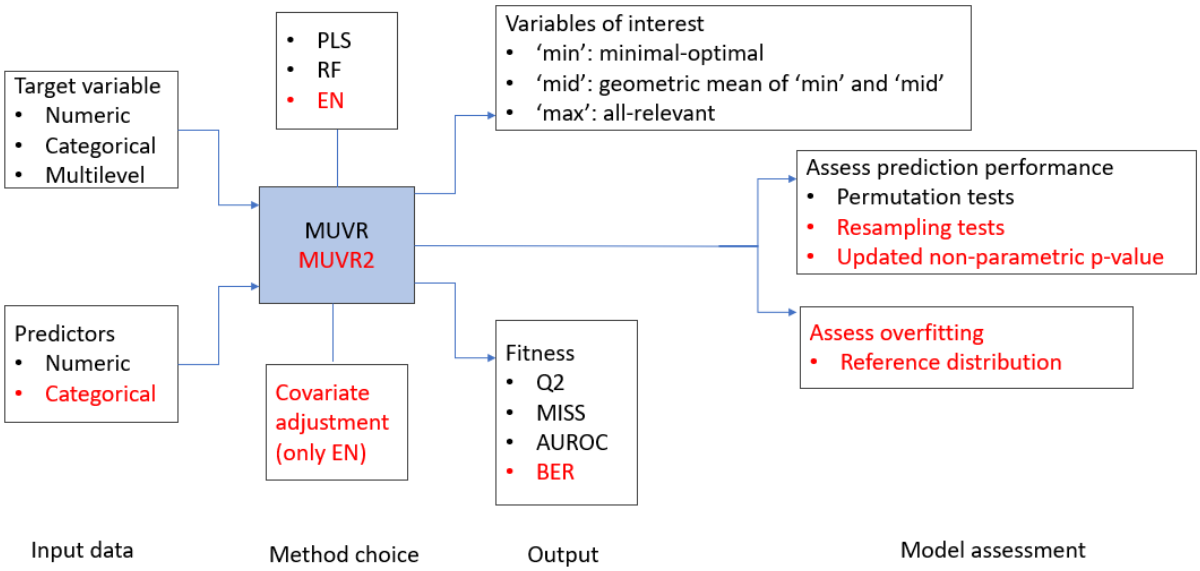


Figure 11. A comparison between the MUVR and MUVR2 algorithms for input data, method choice, output and model assessment. The new functionalities provided by MUVR2 are in red. PLS, partial least squares; RF, randomForest; EN, elastic net; MISS, misclassifications; AUROC, area under receiver operating characteristic curve; BER, balanced error rate. The figure is adapted from the tutorial material from Yan et al 2024[237].

5.1.2 Covariate adjustment

To address confounding in predictive modeling, several approaches to integrate covariate adjustment functionality in MUVR2 were investigated. The investigated hypotheses posited that covariates could be adjusted by either (1) systematically excluding them from recursive elimination in the MUVR2 PLS and RF framework or (2) suppressing their regularization in EN models. The expectation was that this would diminish the variable importance of the covariate-associated predictors, leading to their exclusion during variable selection.

Initial investigation using synthetic data confirmed successful covariate adjustment for EN but not for PLS or RF. This difference may be because, in MUVR2's recursive elimination, covariates are not deconvoluted but forced to compete with other predictors on equal terms. In contrast, EN ensures that the covariates are always fully incorporated in the regularized linear models, with their effects properly isolated from other predictors.

Further validation was performed using MUVR2-EN with real-world BioDiVa data. Sex was considered a potential confounder in metabolite-type 2 diabetes (T2D) associations. Three EN-based classification models were constructed:

- C_{none}: **X**=metabolite features, **Y**=T2D status
- C_{add}: **X**=metabolite features+sex, **Y**=T2D status, where sex is included but not adjusted for
- C_{keep}: **X**=metabolite features+sex, **Y**=T2D status, where sex is included and adjusted for

Comparisons between C_{none} and C_{add} (not shown) revealed nearly identical variable importance ranks, indicating that merely including sex without adjustment does not influence feature selection.

To assess covariate adjustment, the selection ratio (i.e., the frequency of features being selected across the nRep x nOuter models) of 286 sex-correlated features selected in C_{add} was compared to adjustment (C_{keep}). Under the null hypothesis that adjusting for sex would not reduce the selection of sex-correlated features, the change in the selection ratio would be normally distributed around zero. Conversely, a skewed distribution would be expected if adjusting for sex diminished the predictive contribution of sex-correlated features, thereby lowering their selection ratios in C_{keep}.

The observed skewed distribution (**Figure 12**) clearly demonstrated that covariate adjustment reduced the inclusion of sex-correlated features. Notably, features with stronger sex correlations showed greater reductions in selection ratios, while weakly correlated features remained largely unaffected. This finding suggests that covariate adjustment selectively filters variables whose associations with outcomes are confounded by covariates, thereby preserving features with more direct biological relevance.

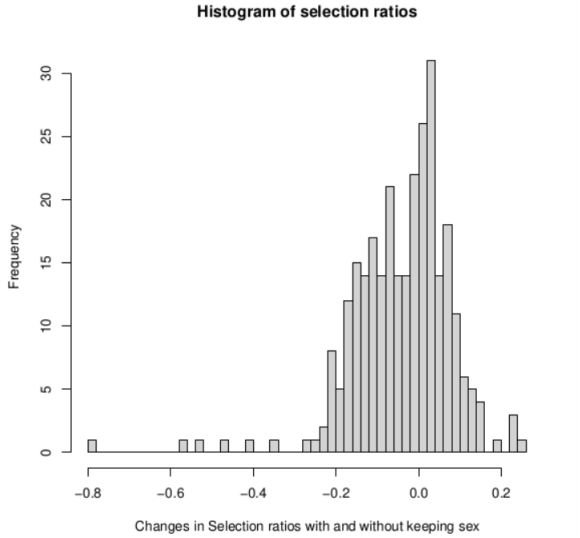


Figure 12. Change in selection ratios of 286 sex-correlated features, representing the difference in feature selection between adjusting for sex (C_{add}) and not (C_{keep}) in MUVR2-EN with 180 underlying models (*nRep = 30, nOuter = 6*). The left-skewed distribution highlights that sex-correlated features get selected less often when adjusted for in Elastic Net modeling. The figure is reproduced from the supplemental material from Yan et al 2024 [237] under a CC BY 4.0 license.

5.1.3 Assessing prediction performance and overfitting

MUVR2 introduces resampling tests and reference distributions as a general framework to systematically evaluate prediction performance and overfitting.

Permutation tests

In traditional permutation tests, the permuted target variable is obtained by randomly reassigning values from the actual target variable without replacement, thereby breaking associations between predictors and the target variable. When using predictors and the permuted target variable for modeling, the prediction performance should represent a scenario under the null hypothesis since the target variable is effectively unrelated to the predictors.

Permutation of the actual target variable can be performed many times, generating a distribution of prediction performance ($\mathbf{H0}_{\mathbf{modeled}}$) by modeling each permuted target variable as a function of the unpermuted predictors. If the actual prediction performance falls within this $\mathbf{H0}_{\mathbf{modeled}}$ distribution, it indicates that the model predicts random associations as effectively as true ones, suggesting that the actual model is unfit for inference. Thus, permutation tests evaluate whether a model's prediction performance exceeds what would be expected by chance, acting as a critical safeguard against inflated performance claims in machine learning.

Resampling tests

Resampling tests extend from permutation tests by simulating null-hypothesis target variables through random draws from the empirical distribution of the actual target variable, thereby introducing larger variability in the resampled target variable distribution compared to permutation tests. $\mathbf{H0}_{\mathbf{modeled}}$ is obtained as for permuted target variables, i.e., by modeling each resampled target variable as a function of the original, non-resampled predictors. Three types of statistical tests are provided to compare actual prediction performance to $\mathbf{H0}_{\mathbf{modeled}}$: (1) An empirical test, where p-values are obtained as the cumulative probability of the actual model within the smoothed $\mathbf{H0}_{\mathbf{modeled}}$; (2) A t-test, which assumes that $\mathbf{H0}_{\mathbf{modeled}}$ follows a t-distribution from which cumulative probability is calculated; (3) A non-parametric test, which uses the rank order of the actual prediction performance and $\mathbf{H0}_{\mathbf{modeled}}$, as implemented by Szymańska et al[343]. In general, the empirical test is recommended since $\mathbf{H0}_{\mathbf{modeled}}$ frequently does not follow a Gaussian distribution and the rank-order-based non-parametric test is limited in resolution and therefore unable to assess $p < (1 / \text{number of resamples})$.

Reference distribution

While permutation and resampling tests address the predictive performance on actual data in relation to a null hypothesis distribution obtained from breaking underlying associations between predictor and target variables, an important problem persists. Namely, depending on the suitability between the machine learning method used and the resampled (or permuted) data, the modeling itself may systematically affect the null-hypothesis distribution. To address this issue, a reference H_0 distribution ($\mathbf{H0}_{\mathbf{reference}}$) was introduced by calculating prediction performance directly from resampled target variables representing the predicted results. By excluding the modeling itself altogether, this approach effectively also excludes all modeling overfitting. By comparing actual prediction performance with $\mathbf{H0}_{\mathbf{modeled}}$ and $\mathbf{H0}_{\mathbf{modeled}}$ with $\mathbf{H0}_{\mathbf{reference}}$, overfitting can thus be assessed more systematically:

- A non-significant p-value indicates that the actual prediction performance is indistinguishable from the $\mathbf{H0}_{\mathbf{modeled}}$ distribution at the corresponding confidence level and is unfit for inference.
- The magnitude of the difference between $\mathbf{H0}_{\mathbf{modeled}}$ and $\mathbf{H0}_{\mathbf{reference}}$ represents the extent of systematic overfitting induced by the modeling strategy in combination with the data.

Using resampling tests and reference distributions, prediction performance and overfitting were evaluated across four CV modeling strategies for PLS, RF and EN: (1) Fit-predict: The entire dataset was used for both training and testing and hyperparameters were tuned through single cross-validation; (2) Single cross-validation (1CV); (3) Double cross-validation (2CV); (4) MUVR2. Results from regression analyses using the Freelive2 data (**Figure 13**) showed that the largest reduction in overfitting resulted from employing any CV and that more complex CV structures improved prediction performance. MUVR2 demonstrated bioinformatical advantages by achieving automated variable selection without compromising prediction performance or imposing bias, highlighting its utility in high-dimensional omics research. Interestingly, only PLS achieved $\mathbf{H0}_{\text{modeled}}$ corresponding to the $\mathbf{H0}_{\text{reference}}$ for the dataset investigated, indicating that PLS is particularly well-suited for prediction with this specific dataset, while other methods may not be as compatible.

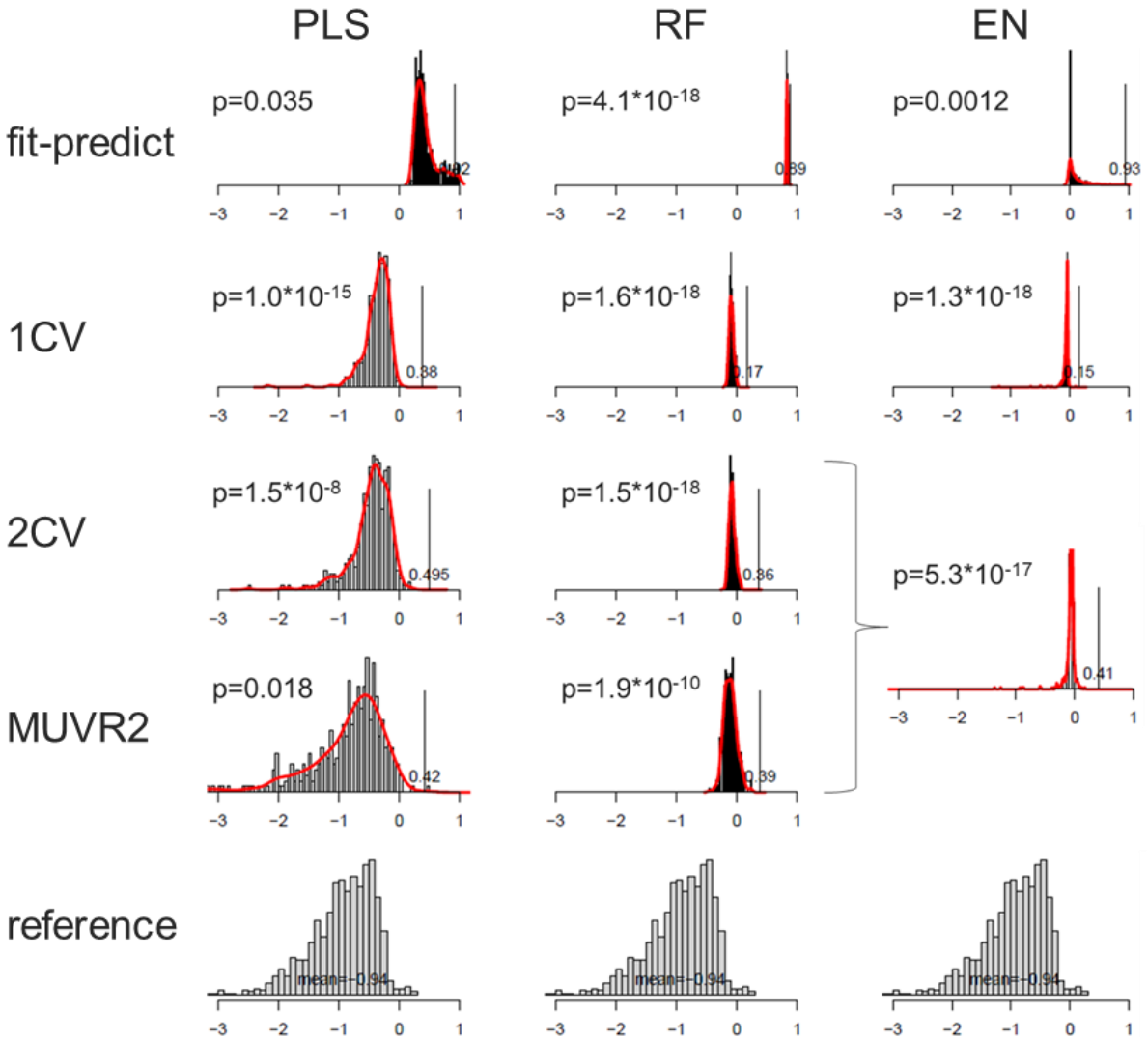


Figure 13. Predictive performance in regression (Q^2) for actual modeling (vertical lines) and resampling tests ($\mathbf{H0}_{\text{modeled}}$; histograms and smoothed curves) and reference distribution ($\mathbf{H0}_{\text{reference}}$; histograms at the bottom). Modeling was performed using PLS (left), RF (middle) and EN (right) with different validation strategies, including *fit-predict*, *1CV*, *2CV* and *MUVR2* (except for EN, since *2CV* is identical to *MUVR2*) using the Freelive2 data (58 observations, 1147 variables). P-values were generated from the smoothed curve of the $\mathbf{H0}_{\text{modeled}}$ distribution. PLS, partial least squares; RF, Random forest; EN, Elastic net; CV, cross-validation. The figure is reproduced from Yan et al 2024 [237] under a CC BY 4.0 license.

5.1.4 Discussion

The MUVR2 algorithm is a supervised ML algorithm that integrates variable selection and validation within a repeated double cross-validation framework, supporting regression, classification and multilevel analysis via PLS, RF and EN. Through validation using simulated and real-world data, MUVR2-EN has demonstrated the ability to perform covariate adjustment directly within ML modeling, thereby offering new opportunities for applying ML in epidemiological studies with minimal data leakage.

Moreover, MUVR2 introduces a comprehensive framework for systematic assessment of modeling prediction performance and overfitting, incorporating resampling tests and reference distributions. This framework enables researchers to quantitatively evaluate model reliability and overfitting, guiding the selection of robust modeling strategies for high-dimensional datasets. Compared to other CV strategies, MUVR2 demonstrates clear bioinformatics advantages by achieving competitive prediction performance while simultaneously performing variable selection without introducing additional overfitting.

Despite these advances, integrating SVM and ANN into MUVR2 remains challenging due to high computational demands, which currently limits the framework's comprehensiveness. Future development should focus on computational optimization – such as adopting faster programming languages[344], parallel processing, or approximation methods – to support advanced machine learning methods within the repeated double cross-validation structure. To demonstrate broader applicability of MUVR2's methodological innovations, further validation across diverse omics platforms, study designs, disease contexts, and other types of datasets is necessary.

5.2 TriplotGUI (Paper II)

The triplot tool was designed to enhance the visualization and interpretation of complex relationships between multiple exposures and outcomes using high-dimensional omics data as molecular intermediates. To expand its functionalities and applications, the TriplotGUI toolbox was developed. Compared to the original triplot package, TriplotGUI incorporates both PCA and WGCNA for data reduction of omics data and supports both mediation and meet-in-the-middle analyses. Notably, TriplotGUI features a user-friendly pipeline with a programming-free interface that automates data analysis and visualization. The toolbox is available as an R package at https://github.com/MetaboComp/TriplotGUI and its web application can be accessed at https://metabocomp.shinyapps.io/triplotgui

5.2.1 Design and Pipeline

TriplotGUI adopts a reactive stepwise modular design, consisting of five analysis steps and two additional functionalities (**Figure 14**). Each step and functionality corresponds to an independent tab in TriplotGUI's interface, providing a clear outline for data analysis and visualization. Color-coded boxes guide users through the following workflow: i) Uploading omics, exposure and/or outcome data; ii) Uploading covariate and auxiliary data; iii) Modifying settings for data analysis; iv) Modifying settings for visualization; v) Viewing figure and data output.

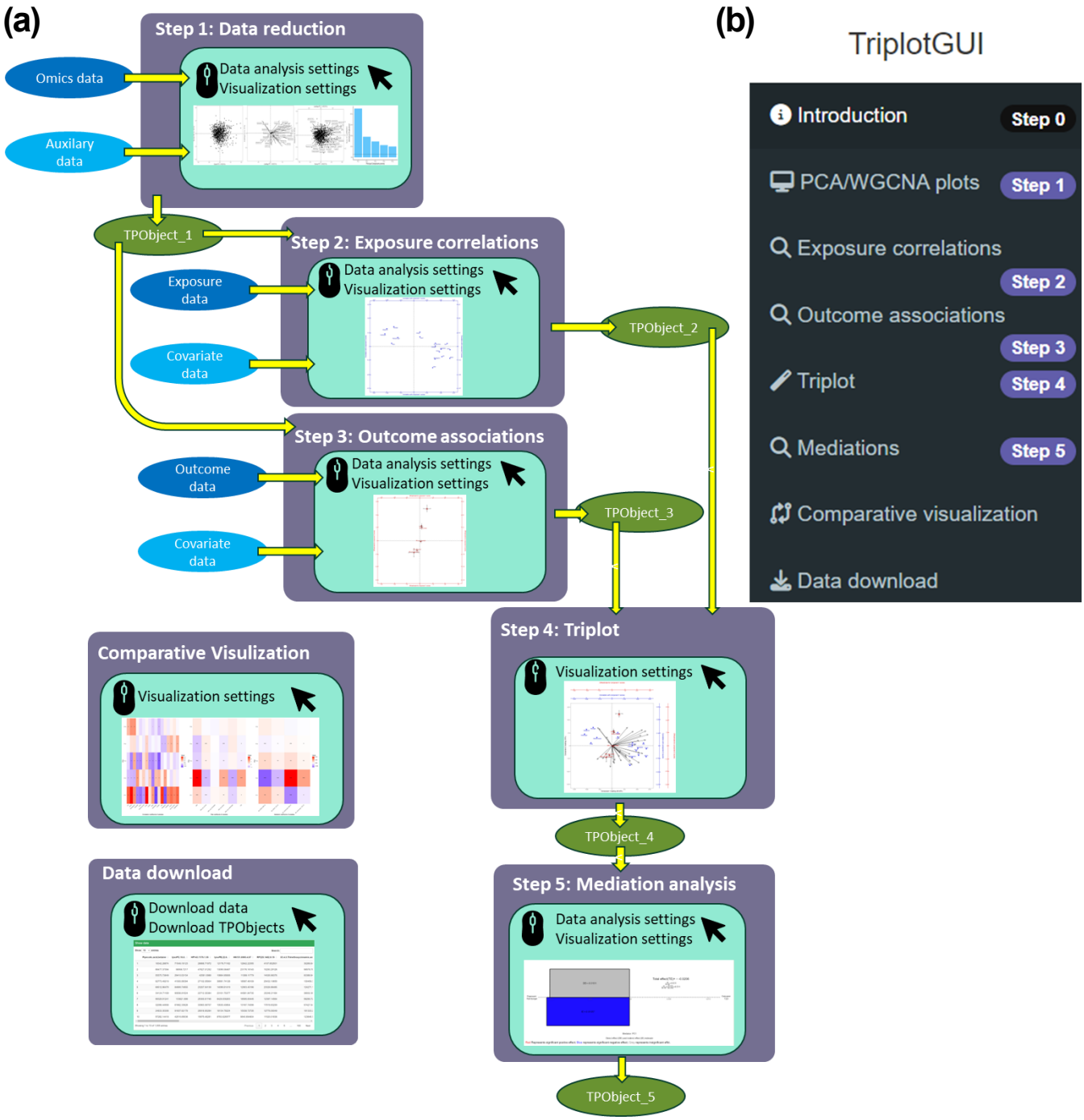


Figure 14. (a) Flow chart of the TriplotGUI structure. Each box corresponds to a tab in the TriplotGUI sidebar, representing one of the five analysis steps or two additional functionalities. Arrows connecting the boxes can be interpreted as data flows. Dark blue ovals represent omics, exposure and outcome input data, and light blue ovals represent covariate and auxiliary input data. Green ovals represent TriplotGUI objects ("*TPObjects*") generated in the different steps, containing the accumulated information gathered up to and including that step. **(b)** Sidebar from the TriplotGUI application that corresponds to the steps and functionalities visualized in the flowchart.

Step 1. Data reduction

High-dimensional omics data are uploaded and transformed into interpretable components using PCA or WGCNA. Users can visualize component loadings, scores and scree plots to assess

variance distribution across components. Auxiliary data may be uploaded to provide custom visualizations attributing individuals' characteristics to score color, size and shape. Pairing information for matched cases and controls can be provided as auxiliary data at this step and subsequently used for outcome associations in Step 3.

Step 2. Associations with exposures

Exposure data are uploaded and associated with component scores, optionally adjusting for confounders. For continuous exposures, correlation coefficients (Pearson, Spearman, or Kendall) are calculated. For categorical exposures, users may either apply one-hot encoding (converting categories to binary variables) or use linear models to regress components on categorical variables. Association strengths are quantified via analysis of variance (ANOVA) in the latter option and reported alongside p-values. Associations between exposures and components are visualized in 2D plots.

Step 3. Associations with outcomes

Outcomes are uploaded and associated with component scores, optionally adjusting for confounders, if provided in Step 3. Outcome associations are calculated differently depending on the outcome type (continuous, binary, or categorical (>2 classes)) and whether cases and controls are matched (**Table 3**). Associations between outcomes and components are visualized in 2D plots.

Table 3. Statistical methods and R packages (in parenthesis) used for outcome-component associations in Step 3, stratified by outcome type and study design.

Study design	**Outcome type**		
	Continuous	**Binary**	**Categorical (>2 class)**
Cases and controls not matched	Linear regression (*glm*)	Logistic regression (*glm*)	Using one-hot-encoding: Logistic regression (*glm*)
			Not using one-hot-encoding: Multinomial regression (*nnet*)
Cases and controls matched	Linear mixed model (*lmer*)	Conditional logistic regression (*survival*)	Using one-hot-encoding: Conditional logistic regression (*survival*)
			Not using one-hot-encoding: (*Not implemented*)

Step 4. Meet-in-the-middle Triplot Visualization

After uploading omics data in Step 1 and exposure and outcome data in Steps 2 and 3, omics components (loadings and optionally scores), exposure association estimates and outcome risk estimates (beta coefficients or odds ratios) can be co-visualized as three layers in a two-dimensional plot. This integrated visualization, termed a "triplot," offers a global perspective on relationships between exposures, outcomes, and omics variables. Examples of triplots and their interpretations are elaborated below.

Applying PCA and WGCNA to the omics data from the HealthyNordicDiet2 dataset, components 1 and 2 associated with exposures (i.e., individual food items and dietary indices) and outcomes (i.e., BMI, BMI_cat, T2D). Omics variable loadings (black arrows), correlation coefficients (blue circles) and risk estimates (red squares) are co-visualized in triplots (**Figure 15**). Using PCA for data reduction (**Figure 15a**), component 1 (x-axis) showed stronger overall correlation with

individual food items and dietary indexes, but weaker associations with BMI and T2D. This suggests that the features dominating component 1 may reflect metabolite patterns influenced by diet, but with limited associations to metabolic regulation in relation to health outcomes. In contrast, component 2 (y-axis) demonstrated stronger associations with BMI and T2D as well as with specific unhealthy exposures (e.g., sausage consumption), suggesting that certain food choices may affect health, potentially mediated through metabolite patterns represented by loadings that contribute to component 2.

Using WGCNA, a different perspective of data reduction is presented (**Figure 15b**). However, the overall associations remain consistent with PCA-derived trends (**Figure 15a**) although rotated counter-clockwise. This consistency underscores that both PCA and WGCNA can identify biologically relevant patterns for interpretations.

Step 5. Mediation analysis and visualization

Mediation analysis can be performed on user-specified exposures and outcomes using component scores as mediators. Covariates, if supplied in Steps 2-3 are automatically adjusted for in exposure-mediator and mediator-outcome associations, respectively. Both conventional and counterfactual mediation analyses are supported. In conventional mediation analysis, the product of coefficients approach is used to calculate the indirect effect. Counterfactual mediation analysis is performed using the *mediation*[19] package. Treatment and control contrast values are required to be specified by users. In both methods, categorical outcomes with more than 2 classes are one-hot-encoded to binary outcomes.

TriplotGUI reports several mediation estimates, including direct, indirect, and total effects, along with the proportion mediated and visualizes them as barplots, which highlights the direction and magnitude of effects. An example is demonstrated in **Figure 16**, using hamburger consumption as the exposure, BMI as the outcome and component 2 as the mediator. A significant positive indirect effect suggests that metabolite features represented by component 2 may mediate the association between hamburger consumption and increased BMI. This finding is consistent with the observed positive correlation between hamburger consumption and component 2, as well as the positive association of component 2 with higher BMI (**Figure 15**). The non-significant direct effect suggests that alternative molecular mechanisms either contribute minimally or cancel each other out.

In addition to these five steps, TriplotGUI also provides users with two additional functionalities:

Comparative visualization

TriplotGUI provides heatmaps to summarize exposure/outcome associations with components and mediation estimates across all analyzed variables. These heatmaps help users identify components of interest for further investigation via triplots at Step 4 or mediation analysis at Step 5 and prioritize these components and underlying omics variables for mechanistic explorations.

Data download

Data generated from the analysis pipeline, including association and mediation estimates, can be viewed and downloaded, allowing users to conduct further analyses. Additionally, a list object containing all relevant information passed through the steps is compiled into downloadable RDA formats, which is referred to as "*TPObject*" in TriplotGUI. This feature enables advanced R users to replicate results or perform additional analyses using this object.

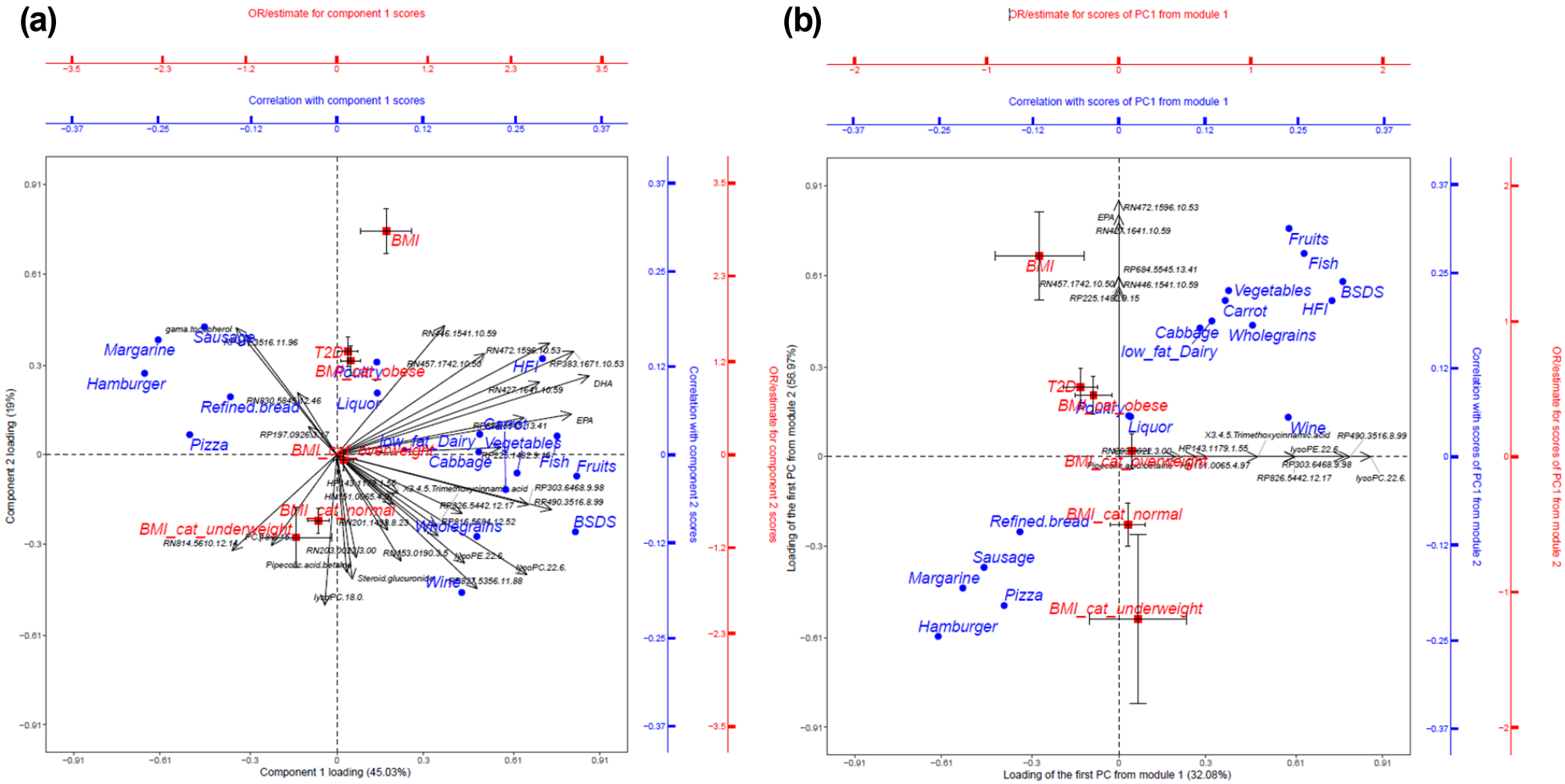


Figure 15: Visualization of triplots obtained in Step 4 of the TriplotGUI tool. 31 omics variables (black arrows), 17 exposure variables (blue dots) and 3 outcome variables (red squares) were used for the analysis. Reduction of the omics data was achieved using either **(a)** PCA or **(b)** WGCNA. Using PCA, both components show correlations with exposure variables while component 2 associates more strongly with the outcomes. Similar trends can be observed also using WGCNA for data reduction, albeit rotated counter-clockwise compared to the PCA. No covariates and auxiliary pairing information were used to adjust correlation or risk models. All other parameters used default settings. T2D, type 2 diabetes; BMI, body mass index; HFI, Healthy Food Index; BSDS, Baltic Sea Dietary Score; PCA, principal component analysis; WGCNA, weighted correlation network analysis.

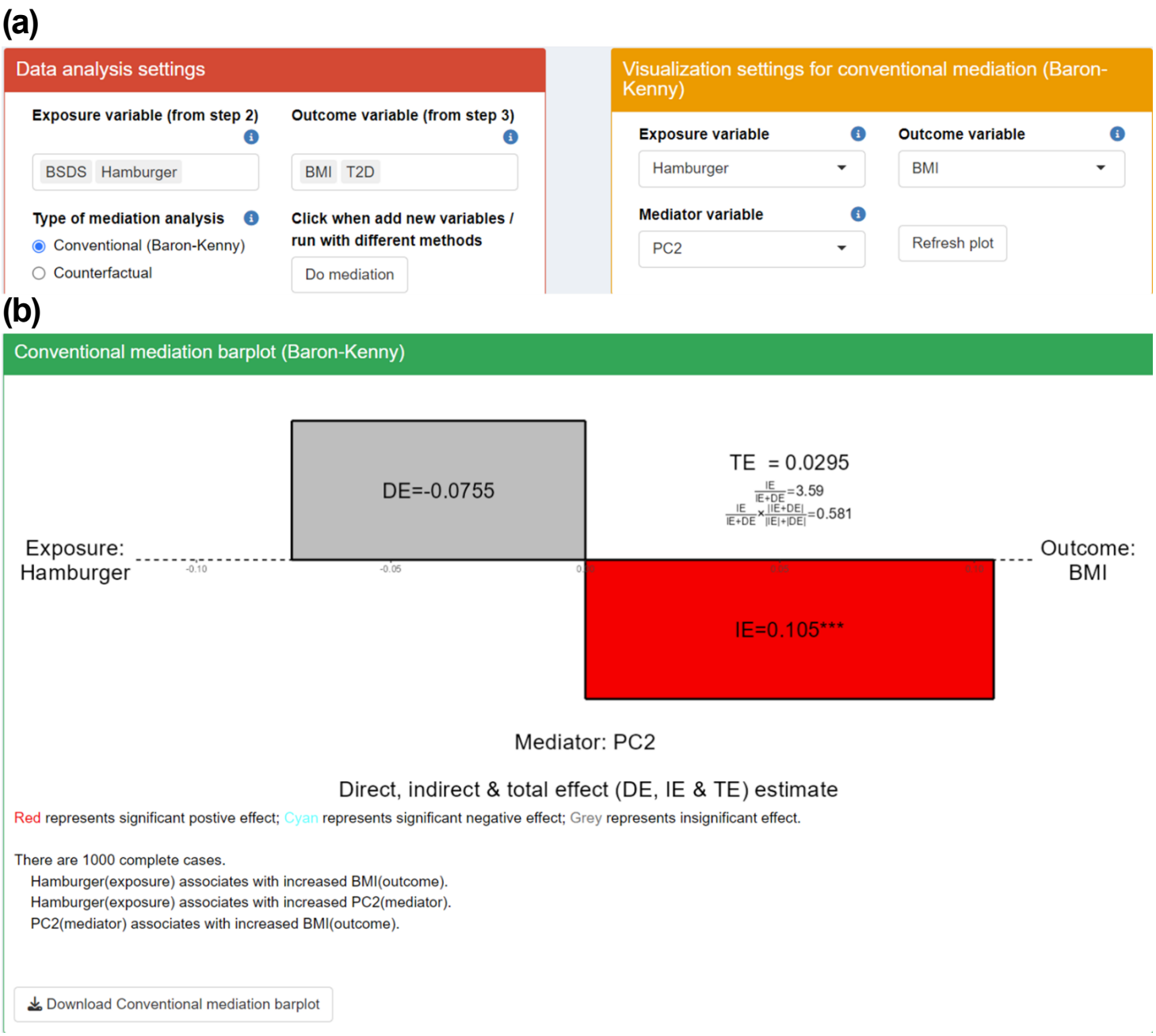


Figure 16. Mediation analysis in TriplotGUI at Step 5. **(a)** Data analysis and visualization settings. Mediation analysis with the conventional (Baron-Kenny) method is performed for the selected exposures and outcomes. A specific exposure-component-outcome combination, i.e., Hamburger consumption - component 2 - BMI was selected for visualization. **(b)** Mediation barplot using hamburger consumption as the exposure, component 2 as the mediator and BMI as the outcome. The positive indirect effect is consistent with directions for the correlation between hamburger and component 2 as well as the association between component 2 and BMI. In contrast, the direct and total effects were not significant. Red represents a positive effect (p-value < 0.05), cyan represents a negative effect (p-value < 0.05) and gray represents an effect with p>0.05. IE, indirect effect; DE, direct effect; TE, total effect; *, p value <0.05; **, p value <0.01; ***, p value <0.001.

5.2.2 Usage

TriplotGUI was developed in the R programming language and integrated with the Shiny package, providing a user-friendly interface for complex analysis pipelines. A detailed manual and use

cases are provided in an online tutorial available at [https://metabocomp.github.io/TriplotGUI_tutorial] (**Figure 17**), which includes step-by-step guidance for navigating the toolbox.

TriplotGUI offers flexibility through two deployment options: a web application accessible at [https://metabocomp.shinyapps.io/triplotgui] and a standalone R package available for download from [https://github.com/MetaboComp/TriplotGUI. The standalone version ensures data confidentiality by enabling local execution without external server transfers. Detailed setup instructions are provided in the "Setup" section of the tutorial.

TriplotGUI supports multiple formats for data input, but due to the diverse preprocessing requirements across different omics platforms and study designs, TriplotGUI does not include built-in preprocessing functionalities, such as missing value imputation, transformation, normalization, or outlier detection. Users should perform these steps independently or use existing tools[345,346]. Notably, missing values are not permitted in the omics data, which requires users to perform suitable imputations beforehand. For exposures, outcomes, and covariates, missing values are handled via complete case analysis by default.

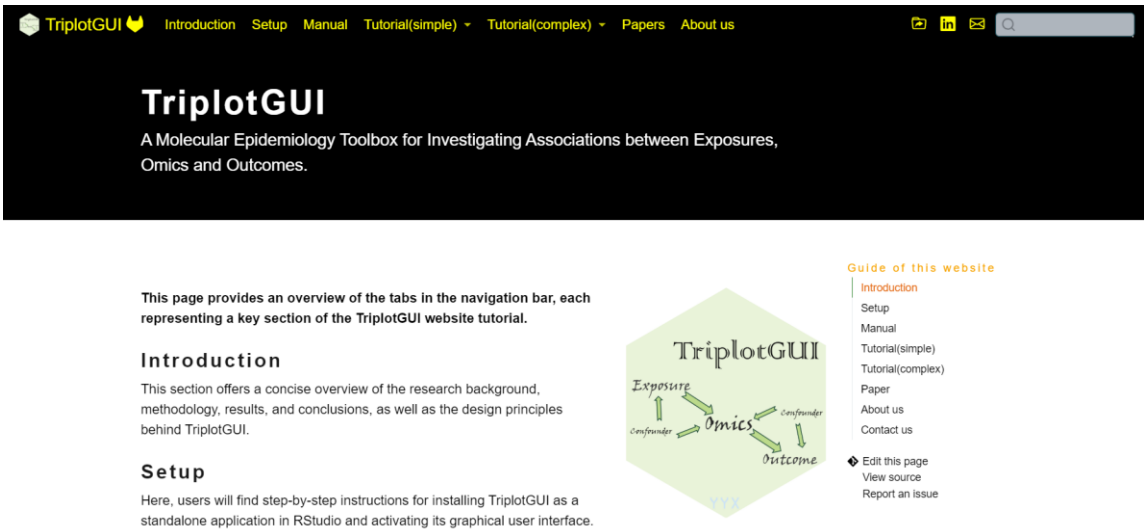


Figure 17. Online tutorial of TriplotGUI. A hexagon logo presents TriplotGUI as an R package. The screenshot is taken from the website created by Yan, Y (2024) under a CC BY-SA 4.0 license.

5.2.3 Discussion

The original triplot package provided a holistic framework for analyzing and visualizing relationships between exposures, omics data, and outcomes using a meet-in-the-middle approach. TriplotGUI extends this framework into a comprehensive molecular epidemiology toolbox that in addition to PCA incorporates WGCNA for data reduction and also integrates mediation analysis. The toolbox supports diverse data types, including continuous, binary, and multiclass categorical variables through tailored analytical strategies, such as linear mixed models for paired cases and controls and multinomial regression for categorical outcomes. Furthermore, its user-friendly interface automates complex workflows, thereby accelerating data analysis and reducing technical barriers for epidemiologists and data scientists while retaining flexibility for advanced users via direct customization of R functions.

TriplotGUI adopts a modular design for analysis and visualization, which facilitates future upgrades to meet evolving research needs as the field advances. For example, TriplotGUI can be expanded to include survival analysis for time-to-events outcomes, linear mixed-effects models for repeated measurements [347,348] and regularization techniques for multivariate modeling [349]. As methods and package advance, TriplotGUI will incorporate a broader range of options for data reduction[350], risk modeling and mediation analysis [351,352].

It should be noted that TriplotGUI operates under the assumption that omics variables act as potential molecular mediators between exposures and outcomes, which may require biological validation. While omics features underlying components-of-interest may link exposures to outcomes through biological mechanisms, domain-specific knowledge in epidemiology is required for formulating research questions, defining exposures and outcomes and conducting meaningful biological interpretation. While TriplotGUI adjusts for confounders in exposure-mediator and mediator-outcome associations, it does not address more complex confounding issues[309] or effect modification[353], which limits causal inference in scenarios with complex interactions. Additionally, the cross-sectional nature of many exposure-omics datasets may not reflect appropriate temporal relationships required for causal inference. A clear understanding of data collection procedures and potential directionality and temporality of exposure-omics associations is essential for accurately interpreting potential causal links.

6. Summary of findings - Cohort studies (Paper III, IV):

This chapter presents the empirical applications of the developed methodologies, focusing on two cohort studies that investigate complex exposure-omics-disease relationships using the advanced analytical frameworks developed in this thesis. In **Paper III**, omics signatures linking POPs to cardiovascular disease are investigated through a meet-in-the-middle analysis within a nested case-control study design, utilizing the MUVR and Triplot analytical frameworks. **Paper IV** extends this approach by examining metabolite features that link diet, POPs, and gut microbiota to cardiovascular disease through both meet-in-the-middle and mediation analyses, employing a cross-cohort discovery and replication design and utilizing the developed MUVR2 and TriplotGUI tools. Together, these studies demonstrate the robustness and broad applicability of the developed methodologies for identifying molecular signatures that characterize complex exposure-disease relationships through metabolic perturbations.

6.1 Persistent organic pollutants, omics and cardiovascular disease risk in Swedish women (Paper III)

When investigating metabolic perturbations underlying POP-CVD associations, 7 proteins and 22 metabolites were found to associate both with the OC component and the CVD outcomes (MI, stroke, composite CVD). Additionally, 12 metabolite features associated both with the PFAS component and the CVD outcomes, with none of them overlapping with the OC-related metabolites. No genetic polymorphisms associated with both exposures and outcomes.

The OC- and CVD-related proteins and metabolite features included tissue plasminogen activator (tPA); urokinase-type plasminogen activator receptor (uPAR), growth differentiation factor 15 (GDF-15), osteoprotegerin (OPG), low-density lipoprotein receptor (LDL-receptor), fibroblast growth factor 21 (FGF-21), interleukin 6 (IL-6), carnitines, glycerophospholipids, mono-, di- and triglycerides and hydroxy docosahexaenoic acid and unknown metabolites features. The PFAS- and CVD-related metabolite features included glycerophospholipids, cortisone-3-glucuronide, exogenous chemicals and unknown metabolites. The Triplot visualization of the 41 features revealed two subpatterns (**Figure 18**). Subpattern 1 associated with an increased risk of MI and correlated positively with the OC component and BMI but negatively with the PFAS component. Subpattern 2 associated with increased risk of stroke and correlated positively with the OC component and age.

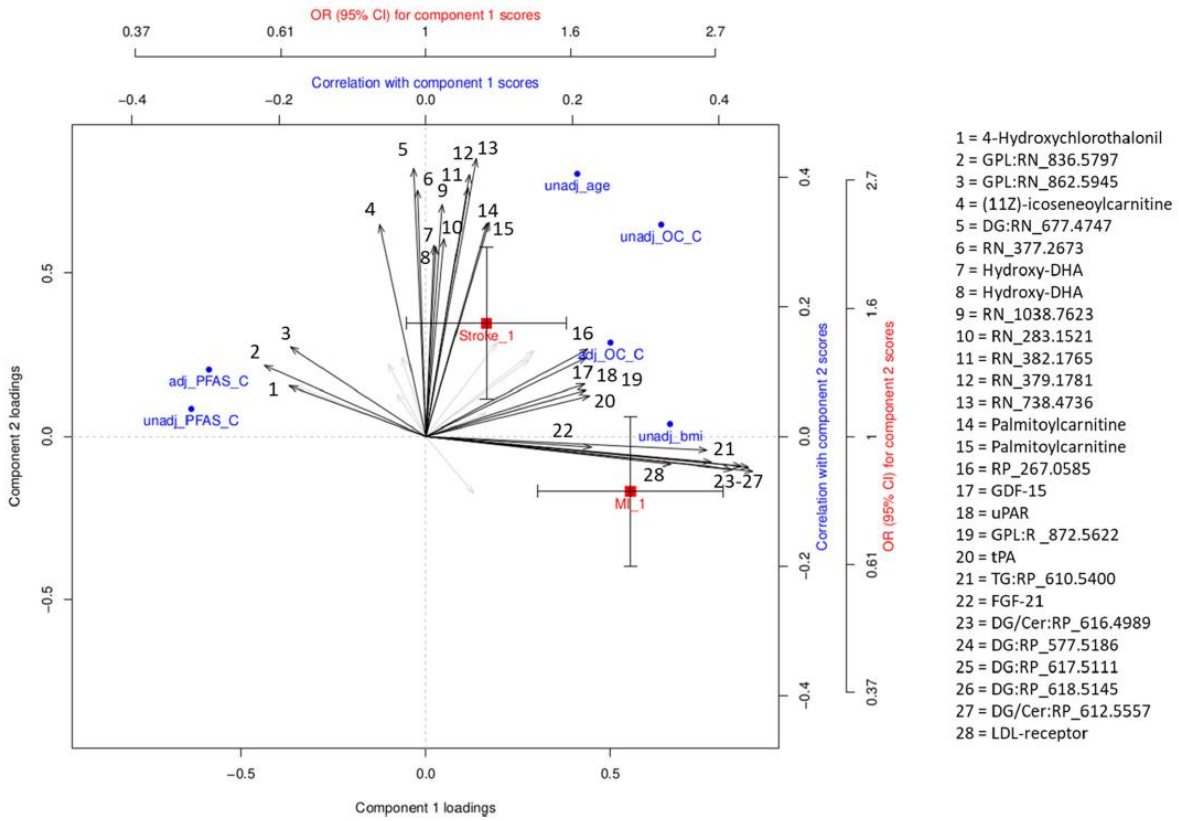


Figure 18. Triplot visualization of POP- and CVD-related omics features with exposure components (OC_C and PFAS_C), age, BMI and CVD outcomes. Each omics feature's contribution to the components is represented by loadings (arrows). The components' correlations with POP exposures, age and BMI (blue points) and the components' associations with MI and stroke risks (red squares) are co-visualized. Features, exposures and outcomes pulling in the same direction are associated, whilst opposite directions indicate inverse associations. Correlations are unadjusted or adjusted for age, sample year, education, healthy diet score, and BMI (for OC_C). Associations are presented as odds ratios with 95% confidence intervals adjusted for age and sample year, education, family history of CVD, smoking status, physical activity, and healthy diet score. Cer, ceramide; DG, diacylglycerol; DHA, docosahexaenoic acid; FGF-21, fibroblast growth factor 21; GDF-15, growth differentiation factor 15; GPL, glycerophospholipid; MI, myocardial infarction; OC_C, organochlorine compound component; PFAS_C, per- and poly-fluoroalkyl substance component; TG, triglyceride; tPA, tissue plasminogen activator; uPAR, urokinase-type plasminogen activator receptor. The figure is taken from Schillemans et al 2024[13] under a CC BY 4.0 license.

Our results demonstrate an inverse association between PFAS and MI, while OCs were associated with increased risk of both MI and stroke. These findings are consistent with several previous studies that have reported either null or inverse associations for PFAS[354–356] and positive associations between OCs and CVD risk factors and MI and stroke[82,357,358].

The subpatterns identified through triplot visualization, and the specific proteins and metabolite features underlying these patterns, provide additional mechanistic insights (**Figure 19**): Subpattern 1 is dominated by diglycerides, triglycerides, LDL-receptor, FGF-21, GDF-15, uPAR and tPA, suggesting the involvement of lipid metabolism pathways and inflammatory processes. PFAS exposure may activate PPARα, leading to lower triglyceride levels[359] and reduced inflammation[360]; Alternatively, PFAS may upregulate phosphatidylcholine synthesis[101], which

may be associated with lower MI risk[361,362]. In contrast, the positive association between OCs and MI through Subpattern 1 may involve endocrine disruption and activation of the aryl hydrocarbon receptor pathway, resulting in altered lipid metabolism and increased inflammation[82]. Additionally, the strong correlation between subpattern 1 and BMI may be attributed to the lipophilic nature of OCs, as their levels in the body can fluctuate based on adiposity, highlighting BMI as a potential important confounder in the OC-CVD association[363].

Subpattern 2 was characterized by carnitines and associated with stroke risk, showing positive correlations with both age and OC exposure. This subpattern may reflect mitochondrial dysfunction, incomplete fatty acid oxidation, and disruptions in carbohydrate and lipid metabolism[364]. The strong correlation between this subpattern and age, and the observation that age attenuates the correlation between this subpattern and OC exposure, suggest the possibility of confounding by age. This is further supported by the fact that OC concentrations tend to increase with age, likely due to historical periods of higher emissions, longer cumulative exposure, and age-related changes in metabolism[365–367].

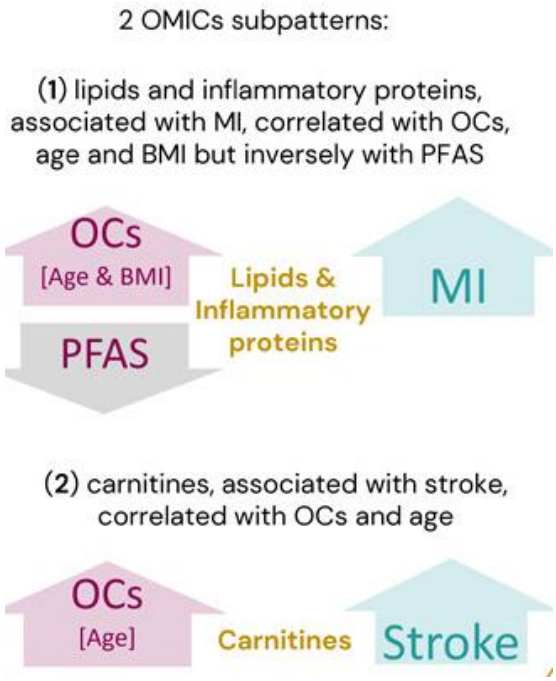


Figure 19. Summary of associations between persistent organic pollutants, two omic patterns and cardiovascular outcomes. OCs, organochlorine compounds; PFAS, per- and polyfluoroalkyl substances. The figure is taken from Schillemans et al 2024[13] under a CC BY 4.0 license.

6.2 Exposome, omics and cardiovascular disease risk in Nordic adults (Paper IV)

Through cross-cohort discovery and validation, a total of 33 features related to cardiovascular outcomes were selected. Among these, 8 features associated with one or more exposure variables in the discovery cohort. Of particular interest, 4 features (pipecolate, ω-hydroxylauric acid, phosphatidyl-ethanolamine(P-16:0/22:6) and an unannotated metabolite referred to as Unk1) exhibited consistent associations with dietary and POP exposures in both discovery and replication cohorts (**Figure 20**). Notably, all 4 features were inversely associated with MI but showed no association with stroke.

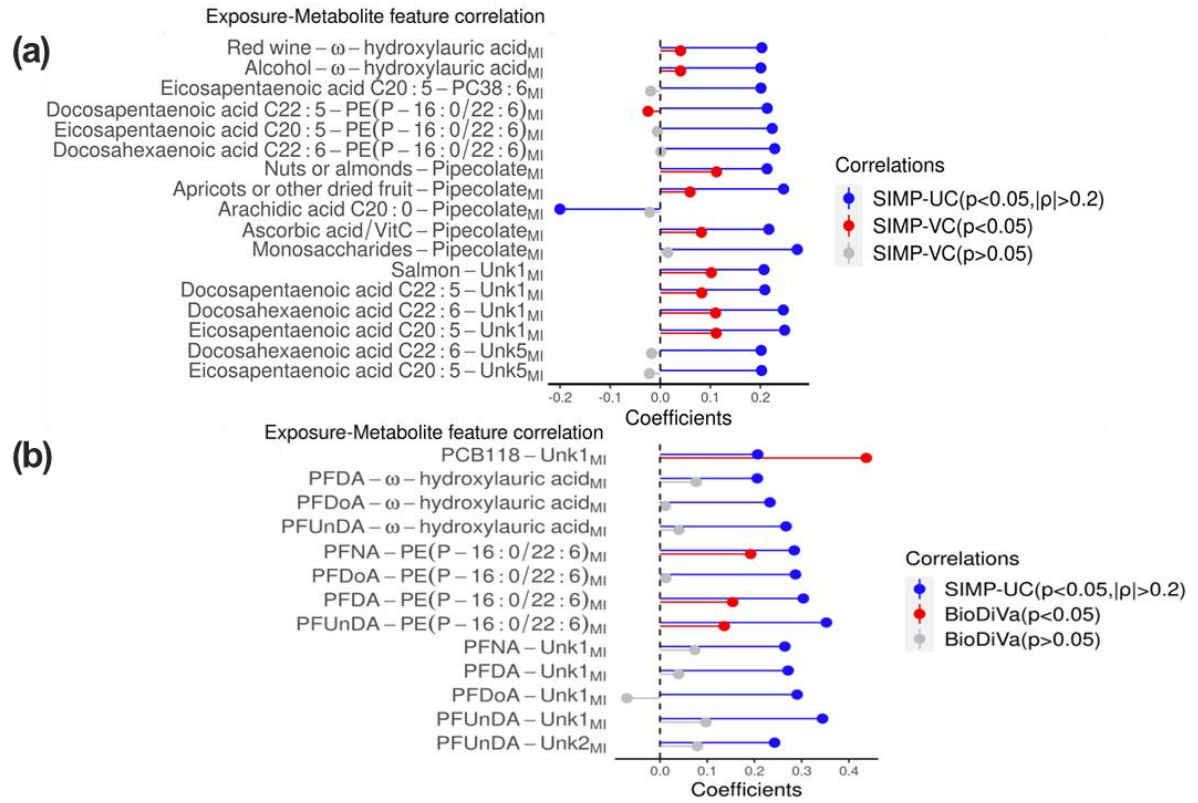


Figure 20. (**a**) Partial Spearman correlations between dietary exposures and outcome-related metabolite features with p-value < 0.05 and coefficient |ρ| >0.2 in the SIMP-UC discovery cohort that could also be matched in the SIMP-VC replication cohort. **(b)** Partial Spearman correlations between persistent organic pollutants exposures and outcome-related metabolite features with p-value < 0.05 and coefficient |ρ| >0.2 in the SIMP-UC discovery cohort that could also be matched in the BioDiVa replication cohort. Subscript MI denotes that the metabolite feature associated with myocardial infarction. The grey color denotes that the correlation did not reach p < 0.05 in the replication cohort. PE(P-16:0/22:6), Phosphatidyl-ethanolamine(P-16:0/22:6).

Among the annotated features, pipecolate showed consistent and positive correlations with the intake of *Nuts/almonds, Apricots/other dry fruits* and *Ascorbic acid/vitamin C*. ω-Hydroxylauric acid correlated positively with alcohol and red wine. The unannotated feature (Unk1) showed a consistent positive correlation with salmon consumption and ω-3-fatty acids, specifically *docosapentaenoic acid C22:5*, *docosahexaenoic acid C22:6*, and *eicosapentaenoic acid C20:5* (**Figure 20a**). In addition, PE(P-16:0/22:6) correlated positively with PFNA, PFDA and PFUnDA, and Unk1 associated positively with PCB118. Notably, in the SIMP-UC cohort, positive correlations were observed between several PFAS and Unk1, but these correlations were weak (|ρ| <0.1) in the BioDiVa cohort (**Figure 20b**). Only a single association involving microbiota features was identified in the SIMP-VC cohort: Hungatella hathewayi exhibited a negative correlation with another unannotated feature, Unk12, which in turn associated with an increased risk of stroke. However, since Unk12 could not be matched in the MAX cohort, replication of this finding was not feasible.

To provide insights into potential causal associations, both counterfactual and conventional mediation analyses were performed in the SIMP-UC cohort on the 4 replicated exposure-outcome-related features **(Table 4)**. Mediation analysis in SIMP-UC supported a possible mediation for pipecolate between *nuts/almonds*, *apricots/other dry fruit*, *ascorbic acid/vitamin C* and MI. Mediation was supported for ω-hydroxylauric acid between *red wine* and MI in

counterfactual but not in conventional mediation analysis. Mediation was supported for Unk1 between both *salmon* and docosahexaenoic acid C22:6 with MI. Furthermore, mediation was not supported for PE(P-16:0/22:6) between PFNA, PFDA, PFUnDA and MI. The results obtained from the two mediation approaches were largely consistent, indicating robustness of the identified mediation effects. Inconsistencies in the mediation results for ω-hydroxylauric acid could stem e.g., from modelling artifacts, casting doubt on the biological relevance of the association.

Table 4. Counterfactual and conventional mediation analysis results for exposure-outcome-related metabolite features with their associated exposures and outcomes in SIMP-UC. ↑ and ↓ indicate whether the mediators and outcomes increase or decrease with increased exposures. If the indirect effect of a mediation analysis has a p-value<0.05, the mediation effect is labeled as positive or negative based on the effect direction. - represents no discernible mediation effect.

Exposure	Potential mediator	Outcome	Conventional mediation effect	Counterfactual mediation effect
Nuts/almonds	pipecolate ↑	MI ↓	negative	negative
Apricots/other dry fruit	pipecolate ↑	MI ↓	negative	negative
Ascorbic acid/vitamin C	pipecolate ↑	MI ↓	negative	negative
Red wine	ω-hydroxylauric acid ↑	MI ↓	-	negative
Alcohol	ω-hydroxylauric acid ↑	MI ↓	-	-
Salmon	Unk1 ↑	MI ↓	negative	negative
Eicosapentaenoic acid C20:5	Unk1 ↑	MI ↓	negative	-
Docosapentaenoic acid C22:5	Unk1 ↑	MI ↓	negative	-
Docosahexaenoic acid C22:6	Unk1 ↑	MI ↓	negative	negative
PCB118	Unk1 ↑	MI ↓	-	-
PFNA	PE(P-16:0/22:6) ↑	MI ↓	-	-
PFDA	PE(P-16:0/22:6) ↑	MI ↓	-	-
PFUnDA	PE(P-16:0/22:6) ↑	MI ↓	-	-

Mechanistic investigations were undertaken for pipecolate, ω-hydroxylauric acid, PE(P-16:0/22:6) and Unk1. Pipecolate, derived from dietary sources including dried fruits and plant-based food[368], demonstrated an inverse association with MI. This protective effect may occur through its contribution to the biosynthesis of rapamycin [369,370], which exerts anti-aging effects through the inhibition of the mammalian target of rapamycin (mTOR) [371]. ω-Hydroxylauric acid also had an inverse association with MI, possibly through its role in myocardial energy metabolism under hypoxic conditions [372], although its correlation with red wine and alcohol has not been previously reported. Unk1, which correlated with salmon and ω-3-fatty acids and was inversely associated with MI, may act as a mediator of the cardioprotective effects of ω-3-fatty acids [373–375]. PE(P-16:6/22:6) associated inversely with MI, possibly though mechanisms involving phosphatidylcholine synthesis that supports membrane stability and exerts anti-inflammatory effects[376]. However, mediation analysis showed that its correlation with PFAS may reflect confounding rather than a direct causal relationship. Additional mediation analysis highlights that PE(P-16:0/22:6) may instead mediate the association of salmon and ω-3-fatty acids with MI

risk, suggesting that the observed cardioprotective effect is primarily attributable to the beneficial nutrients from salmon, while PFAS likely present as a co-exposure resulting from salmon consumption.

In triplot visualization of the 33 replicated outcome-related features, only component 3 exhibited correlations with salmon, ω-3 fatty acids and PFAS (|ρ|>0.2) and associated inversely with MI (**Figure 21**). Similar to the analysis using individual metabolite features as mediators, counterfactual mediation analysis also indicated that component 3 may mediate the association between salmon, ω-3 fatty acids and MI. However, component 3 did not appear to mediate the PFAS-MI association.

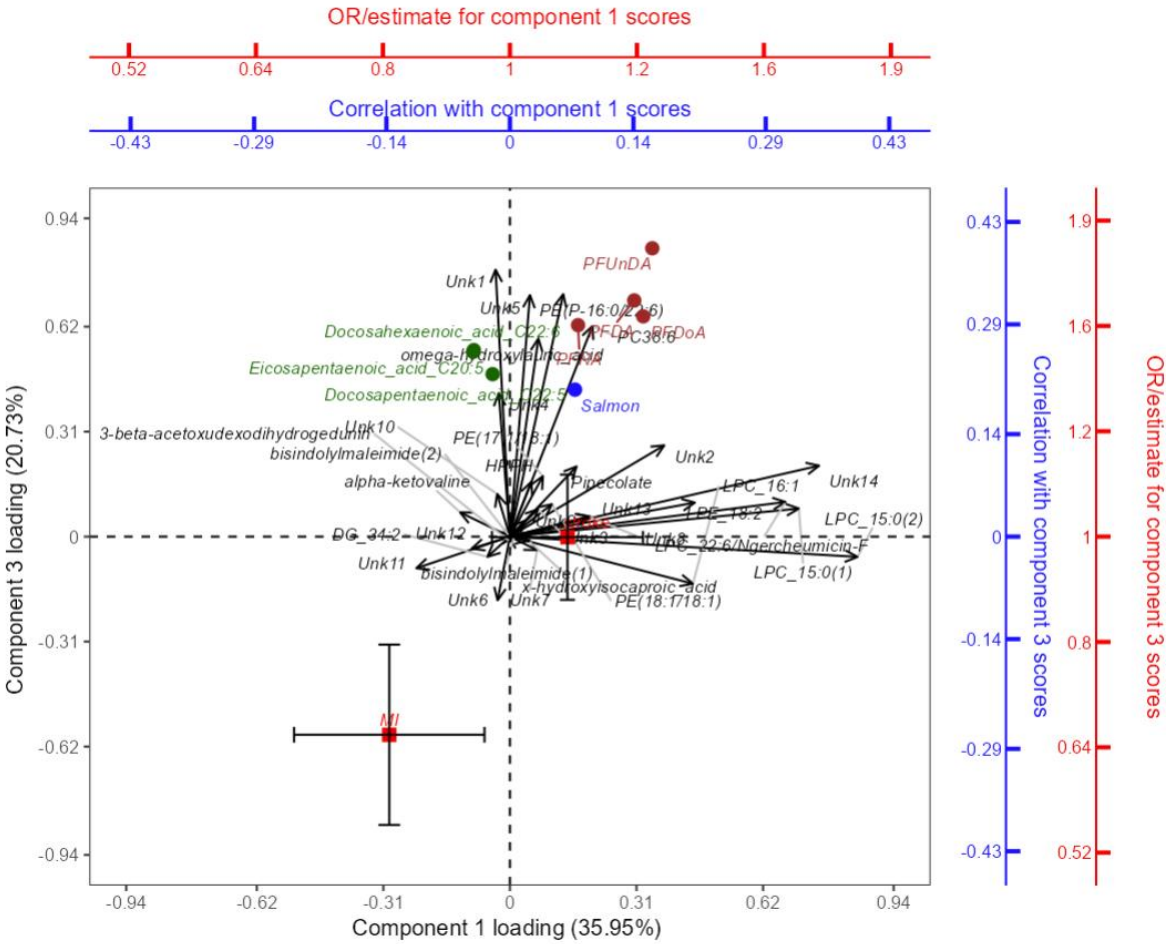


Figure 21. Triplot visualization of the 33 replicated outcome-related metabolite features with exposures and outcomes in SIMP-UC. Principal component analysis (n=4, varimax rotation) was performed on the features. Metabolite loadings (black arrows) are co-visualized with Spearman correlation of component scores with food items (blue dots), nutrients (dark green points), PFAS (brown points) and component associations with MI and stroke risks (red squares). Confounder adjustment for the different analyses is presented in **Paper VI.** Correlations are presented as coefficients and associations are presented as odds ratio with 95% confidence intervals. Features, exposures and outcomes pulling in the same direction are associated, whilst opposite directions indicate inverse associations. The triplot shows components 1 and 3 on the x and y axes. OCs, organochlorine compounds; PFAS, per- and polyfluoroalkyl substances.

Additionally, in triplot visualization of the 8 replicated outcome-related features that also associated with exposures in the SIMP-UC discovery cohort, several mediations were supported: Component 2 showed indirect effects linking salmon, ω-3-fatty acids, red wine and alcohol with

MI; Component 4 showed indirect effects linking nut/almonds, apricots/other dried fruit, ascorbic acid/vitamin C, salmon, red wine and alcohol to MI. Components 1 and 3 showed indirect effects between nut/almonds and MI. Importantly, mediation was not supported between POPs (OCs and PFAS) and MI through any of the components (**Figure 22**).

Component 3 was primarily dominated by the stroke-related Unk12 while component 4 was characterized by pipecolate originating from dietary sources. Metabolite features pulling in the direction of components 1 and 2 seemingly represent an MI-related pattern. These metabolites may reflect a metabolic pattern linking salmon intake, ω-3 fatty acids level, red wine and alcohol to reduced MI risk.

The triplot visualizations revealed distinct metabolite patterns associated with MI: Pipecolate, despite its inverse association with MI, did not appear to coincide with the main MI-related metabolic pattern. This effectively suggests multiple exposure-related molecular mechanisms relevant to MI risk and that pipecolate may reflect a different perturbation compared to the group of metabolites that showed similar loading pattern, indicating evidence of concerted perturbations. Notably, while PE(P-16:0/22:6) contributed to the MI-related metabolic pattern, mediation of PFAS-MI association was neither supported through PE(P-16:0/22:6) nor the MI-related metabolic pattern. This suggests that the observed correlation between PFAS and PE(P-16:6/22:6) likely stems from confounding rather than a direct causal relationship. Post-hoc mediation analysis instead supported the role of PE(P-16:0/22:6) as a potential mediator of salmon intake and ω-3 fatty acids with MI risk. These findings highlight the importance of distinguishing between biomarkers driven by common exposure pathways and those reflecting direct mechanistic links.

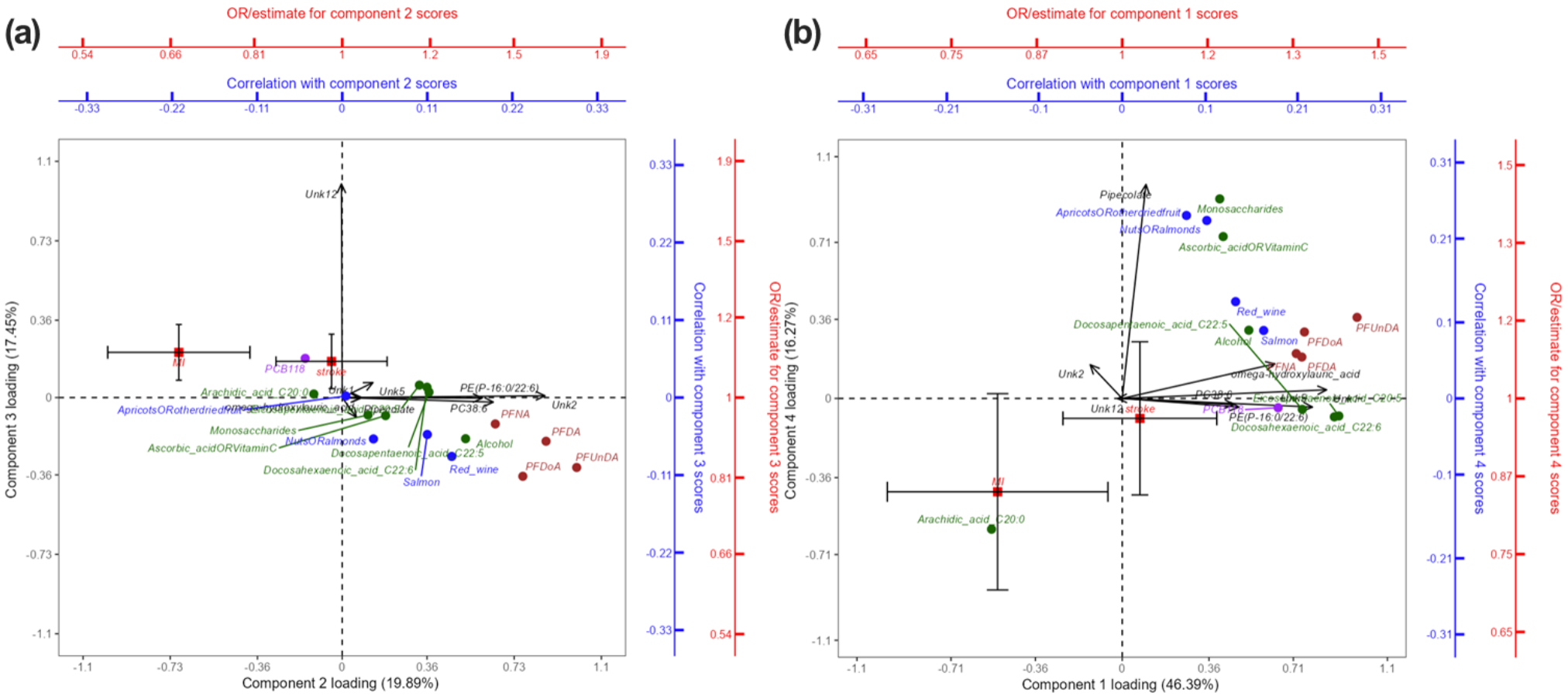


Figure 22. Triplot visualization of the 8 replicated outcome-related features that also associated with exposures in SIMP-UC. Principal component analysis (n=4, varimax rotation) was performed on the features. Metabolite loadings (black arrows) are co-visualized with Spearman correlation of component scores with food items (blue dots), nutrients (dark green points), PFAS (brown points) and component associations with MI and stroke risks (red squares). Confounder adjustment for the different analyses is presented in **Paper VI.** Correlations are presented as coefficients and associations are presented as odds ratio with 95% confidence intervals. Features, exposures and outcomes pulling in the same direction are associated, whilst opposite directions indicate inverse associations. The triplots show **(a)** Components 2 and 3, **(b)** Components 1 and 4 on the x and y axes. OCs, organochlorine compounds; PFAS, per- and polyfluoroalkyl substances. OCs, organochlorine compounds; PFAS, per- and polyfluoroalkyl substances.

7. General discussion

The methodological development and cohort studies presented in this thesis were designed to investigate how combined exposures influence metabolic health through omics approaches. In **Paper III**, the MUVR algorithm was employed for feature selection, while the triplot tool facilitated visualization and interpretation. With the subsequent development of MUVR2 (**Paper I**) and TriplotGUI (**Paper II**), these tools became instrumental to the analytical procedures in **Paper IV**.

Although MUVR2 and TriplotGUI are independent tools applicable to omics research, their functions are highly complementary. Omics data typically consist of thousands of features, many of which may correlate, introduce noise or be irrelevant to the exposure or outcome of interest. MUVR2 is particularly effective at narrowing down these vast datasets to a focused set of features of interest. These selected features can then be further analyzed and visualized using TriplotGUI, which condenses their information through data reduction techniques and relates them to exposures and/or outcomes. In essence, MUVR2 serves as a discovery tool for identifying meaningful information, while TriplotGUI excels at summarizing and visualizing this information.

In cohort studies, **Paper III** focused on the associations between POP components (two exposures) with MI, stroke and composite CVD (three outcomes) through genetics, proteomics and metabolomics as potential effect modifiers or mediators within a nested case-control design. In contrast, **Paper IV** expanded the investigation to encompass more complex exposures, including diet, POPs and gut microbiota (multiple exposure variables) and their associations with MI and stroke (two outcomes) specifically through metabolomics under the hypothesis of mediation. The study in **Paper IV** also adopted a nested case-control design but incorporated a cross-cohort discovery and replication approach, made possible by the recent availability of metabolomics and gut microbiota data in SIMP-VC. The approaches applied in **Papers III** and **IV** are similar yet with important distinctions: **Paper III** represents an exposure-centric approach (downstream from exposure to outcome), investigating how POPs may affect metabolic patterns that in turn may explain the mechanisms underlying CVD risks. On the other hand, **Paper IV** represents an outcome-centric approach (upstream from outcome to exposure), investigating how MI- and stroke-related metabolite features relate to exposures and potentially mediate exposure-outcome associations.

Paper III identified metabolic patterns linking OCs and PFAS to MI and stroke and investigated possible mechanisms: One subpattern associated with disrupted lipid and inflammatory pathways, elevated OCs, lower PFAS and higher risk of MI; The other subpattern related to carnitines and possible mitochondrial dysfunction and associated with OCs and stroke. In **Paper IV**, specific metabolite features were discovered that linked the exposome to CVD outcomes and mediation analysis was used to further probe causal relationships: Pipecolate from nuts and dried fruits appeared to reduce MI risk, possibly through the mTOR pathway and anti-inflammatory mechanisms. PE(P-16:0/22:6) associated with lower MI risk and PFAS exposure. However, mediation analysis did not support a causal association, instead indicating that salmon intake is the primary exposure underlying the observed PE-MI association, with PFAS exposure likely occurring concurrently as a result of salmon consumption.

Both **Papers III** and **IV** discovered inverse associations between PFAS and MI, through metabolic patterns or individual metabolite features. While **Paper III** explored possible mechanisms using meet-in-the-middle analysis, **Paper IV** provided further evidence from mediation analysis that the association is likely not directly causal, but possibly confounded by shared dietary sources,

particularly salmon. Additionally, **Paper III** relied on data from a single cohort and identified several omics features linking POPs to CVDs. However, the cross-cohort discovery and replication approach applied in **Paper IV** revealed that only a minority of the exposure-metabolite-outcome associations discovered in single cohorts could be consistently reproduced. The results from this thesis thus strongly support the notion that cross-cohort validation represents a critical methodological necessity and replication should be an absolute requirement for exploring exposure-omics-outcome associations in molecular epidemiology.

As shown in this thesis, the integration of omics approaches in epidemiological research, combined with the development of specialized methodological tools, offers substantial potential for elucidating metabolic perturbations underlying exposure-outcome associations. However, the complexity inherent in exposures and high-dimensional omics data necessitates rigorous attention to study design and interpretation. The following discussion addresses key considerations relevant in particular to epidemiological methodology, POP exposures, and the application of omics technologies.

Epidemiological considerations

Causal assumptions and reverse causation: A central causal assumption in this thesis, and in the entire triplot/TriplotGUI methodology, is that omics features or patterns act as molecular intermediates linking exposures to outcomes (**Paper II, III and IV**). While this assumption is less emphasized in the meet-in-the-middle analysis, it is methodologically encoded in the mediation analysis framework. However, this assumption can be challenged e.g., by reverse causation. For instance, host metabolites, such as bile acid, may modulate microbial gene expression upon reaching the gut[377]. Additionally, changes in protein and metabolite concentrations may reflect disease symptoms rather than underlying pathway alterations.

Moreover, the cross-sectional nature of exposure and omics assessment in **Papers III** and **IV** further complicates causal interpretation since temporal precedence is not established. TriplotGUI is also currently designed for cross-sectional associations between exposures and omics rather than prospective analyses (**Paper II**). However, in the cohort studies, although exposure information and blood samples were collected simultaneously, we assume that the information reflects past exposures, such as long-term dietary habits, accumulated POP levels, or stable gut microbiota composition. Nevertheless, even if the assumption of omics acting as intermediates holds, the associations observed in our studies cannot be interpreted as evidence of causality. These limitations represent a fundamental constraint in observational epidemiology, necessitating careful interpretation of mediation analyses and explicit acknowledgment of assumptions. To further establish causal relationships, other confirmatory studies are necessary, as well as mechanistic studies to further elucidate and establish the pathways underlying these associations.

Covariate adjustment: In the statistical analysis, both cohort studies began by filtering out a large bulk of apparently irrelevant features using machine learning. In **Paper III**, while the MUVR algorithm can select features relevant to POPs, these associations might be confounded, which the employed RF procedure cannot accurately address. A partial correlation was therefore conducted to adjust for potential confounders. However, this univariate approach fails to evaluate features jointly and does not account for collinearity or the joint effects of multiple variables. With the development of MUVR2, covariate adjustment was integrated through elastic net modeling. While this streamlined the analytical process in **Paper IV**, enabling simultaneous variable selection and covariate adjustment in linear models, this approach remains limited by

its assumptions of linearity and of Gaussian distributions. Further methodological development is therefore needed for flexible, non-linear frameworks, capable of confounder adjustment and applicable for holistic analysis of high-dimensional omics data.

While tailored covariate adjustment can improve the validity of estimated associations, a more conservative set of potential covariates was selected for adjustment in **Paper IV** compared to **Paper III**, for three primary reasons: (1) Avoiding overadjustment and risks of controlling for colliders or potential mediators. (2) By selecting a limited set of covariates, we achieved consistent analyses for all exposures investigated in **Paper IV**, thereby enhancing comparability. (3) Some variables, although potential confounders, were treated as exposures in our study and thus were not adjusted for. However, the conservative covariate adjustment may also lead to residual confounding, potentially biasing estimates or obscuring true associations. For instance, diet can serve as a source of POPs through the consumption of contaminated foods, particularly fish and meat, and also affects metabolite levels in blood. As a result, diet could introduce spurious associations between POPs and metabolite features. This underscores the need for careful study design and sensitivity analysis in future work.

Matching: While both **Papers III** and **IV** employed nested case-control designs to minimize selection and information bias, their approaches for selecting controls differed, with important implications for our findings. In **Paper III**, cases and controls were matched on age and sample date to control confounding. However, this may also carry the risk of overmatching, where the control group is artificially made to be so similar to cases that it no longer accurately represents the underlying reference population. Consequently, the effect estimates are potentially biased towards the null, thereby hindering the detection of true associations. In contrast, **Paper IV** employed random matching with repeated down sampling to create balanced groups for ML analysis. Although this approach may be less effective at reducing bias by design, it reduces stochastic variability and may yield a more representative control group. This trade-off highlights the importance of carefully considering the study context to achieve an appropriate balance between confounder control, bias reduction, and generalizability when selecting matching and sampling strategies for nested case-control studies.

Multiple testing: Notably, neither **Paper III** nor **Paper IV** applied adjustments for multiple comparisons when interpreting p-values, for the following reasons: (1) Both studies were exploratory, using p-values to filter relevant features for hypothesis generation rather than formal hypothesis testing. (2) Machine learning algorithms were applied to all features simultaneously to select the best predictors, rather than testing each feature individually. (3) Several exposures and metabolite features were correlated and thus not entirely independent. (4) Perhaps most importantly, the number of selected features differed in relation to exposure (**Paper III**), to outcome (**Paper IV**) and to discovery cohort (**Paper IV**), which would have led to inconsistent penalization of p-values when applying multiple testing adjustments. However, not adjusting for multiple comparisons requires careful considerations, since it may increase the risk of false discoveries, particularly given the high-dimensional nature of omics data. In **Paper III**, we performed permutation tests to assess model performance. In addition, in **Paper IV** we mitigated the risk of false positives by seeking replication of findings across multiple independent cohorts, Together, these strategies reduced the likelihood that our findings were driven by chance, enhancing the reliability of the results in the context of high-dimensional omics analyses.

External validity: Finally, both **Papers III** and **IV** focused on older Nordic populations, with **Paper III** specifically examining Swedish women exclusively. Even assuming no systematic differences between participants and non-participants, effect modification by age and sex may limit the

generalizability of our results to the broader Nordic population. Additionally, our findings may not be generalizable to other ethnic or geographic groups. This limitation in population diversity represents a significant constraint on the broader applicability of findings, particularly given known differences in metabolism, dietary patterns, and pollutant exposure profiles across populations.

Exposure considerations

Whereas **Paper III** aggregated POPs into PCA-derived components representing OC and PFAS subpatterns, **Paper IV** analyzed individual POPs to capture all potential exposure-metabolite associations. This highlights a fundamental tradeoff: Aggregating exposures can reveal synergistic, pattern-level effects of co-occurring contaminants, which may better reflect cumulative and real-world exposure scenarios; Analysis at the individual compound level allows for identifying specific biological effects posed by particular contaminants, enabling direct interpretation. Integrating individual and aggregated exposure analysis may be necessary to fully elucidate the health impact of POPs. In future research, data reduction techniques could be applied to different exposure groups to capture shared variance, or clustering methods or network analysis could be used to detect communities of related exposures [378–380]. Approaches such as mixture analysis and the development of polyexposure scores (similar to polygenic risk scores[381]) could help quantify the cumulative biological impact of exposure mixtures and enable dose-response analyses for complex exposomes[382].

BMI was adjusted for in relation to OCs in **Papers III** and **IV**, given that BMI is associated with CVD and OC levels in blood may fluctuate based on BMI. However, managing lipophilic contaminants in relation to lipids when assessing health outcomes remains controversial. Elevated blood lipids are associated with CVD risk[383] and individuals with higher blood lipid concentrations tend to have proportionally higher levels of lipophilic OCs in their circulation[384]. For this reason, the total lipid level may serve as a confounder to be adjusted for, as in **Paper IV** [384]. However, POPs may also increase blood lipids, which would place lipids in the causal pathway between POPs and CVD and therefore adjustment should be avoided, as in **Paper III**[384]. To investigate these different perspectives further, we conducted a sensitivity analysis in **Paper IV** without adjusting for lipid levels and found that the results remained largely unchanged. This suggests that within our study population, the potential bias introduced by either lipid adjustment strategy was minimal.

Omics considerations

In **Paper III**, proteins and metabolites were selected over genetic polymorphisms relating POPs to CVD. This raises important considerations regarding both biological relevance and methodological design: Proteins and metabolites have inherently closer proximity to phenotypic expression, which may result in stronger signal-to-noise ratios in predicting POPs. On the other hand, the lack of genetic contributions in our findings may be influenced by analytical choices: The use of single-block data integration[385], where all omics layers are concatenated into a single predictor matrix, may not adequately account for the distinct scales and distributions in PLS and EN analyses, as well as biological roles of genetic, proteomic, and metabolomic data. Furthermore, limitations in sample size or ML methods may also have constrained the ability to detect associations involving genetic polymorphisms. Future research may benefit from a deeper investigation of the genetic contributions to POPs-CVD associations, employing other multi-omics integration strategies, such as multi-block or network-based methods[385].

While the multi-cohort discovery and replication in **Paper IV** is arguably one of the greatest strengths in this thesis, replication was severely hindered by technical variability: A significant proportion of metabolite features discovered could not be matched in other cohorts, highlighting the non-stable nature of untargeted LC-MS analysis. Metabolite annotation represents another bottleneck. Standardized protocols, expanded reference libraries and improved annotation protocols (including algorithms) are critical to address these challenges to enhance the potential for replication and annotation.

8. Conclusions

This thesis presents novel advanced tools for omics data analysis and investigates molecular associations linking diet, persistent organic pollutants, and microbiota to cardiovascular outcomes. The thesis comprises four papers, progressing from methodological development (**Papers I** and **II**) to cross-cohort applications (**Papers III** and **IV**), reflecting the four objectives:

MUVR2 facilitates the selection of omics variables that contribute to the prediction of specific exposures or outcomes and demonstrates state-of-the-art performance while minimizing overfitting. Through the incorporation of elastic net modeling, MUVR2 also supports covariate adjustment. In addition, it provides a framework for assessing prediction performance and overfitting (**Paper I**).

The TriplotGUI toolbox provides advanced yet accessible analysis of metabolic perturbations linking exposures to outcomes. This is achieved through data reduction, meet-in-the-middle and mediation analyses, providing visualization and interpretation capabilities for molecular epidemiology (**Paper II**).

In a cohort of older Swedish women, we identified two distinct molecular patterns of proteins and metabolites that link POPs to CVD risk. The first pattern represented perturbed lipid metabolism and inflammation and linked higher OC and lower PFAS levels to increased MI risk. The second pattern represented carnitines and mitochondrial dysfunction and linked higher OC levels to increased stroke risk. These findings imply that OCs and PFAS associate with distinct omics signatures that may exert opposing effects on CVD risk. However, the observed associations may be influenced by additional factors such as age, adiposity, or co-exposures, underscoring the need for further research to clarify the interplay between POPs, metabolic perturbations and cardiovascular outcomes (**Paper III**).

In an ambitious multi-cohort discovery and replication scheme, we identified specific metabolites that link complex exposures to CVD risk. Pipecolate from nuts and dried fruit may reduce MI risk, possibly through the mTOR pathway and anti-inflammatory mechanisms. Salmon intake and ω-3 fatty acids were inversely associated with MI, possibly through PE(P-16:0/22:6) and an unknown metabolite mediating the cardioprotective effect. Additionally, although it remains to be elucidated, associations were observed for ω-hydroxylauric acid with both red wine and reduced MI risk. In contrast to **Paper III**, PFAS was not associated with reduced MI risk, likely related both to the stringent replication limiting false discovery, and the more advanced analytical approach employing mediation analysis. Only a limited number of exposure-metabolite-outcome associations could be fully reproduced, stressing the importance of replication for generalized conclusions (**Paper IV**).

In summary, this thesis advances the field of molecular epidemiology by providing robust, accessible tools: MUVR2 for unbiased variable selection and covariate adjustment, and TriplotGUI for intuitive visualization and interpretation of complex exposure-omics-outcome relationships. Their application in large-scale cohort studies highlighted the molecular support for the contribution of OCs to CVD risk and for certain healthy foods to reduced MI risk. Importantly, our findings revealed that many associations observed in single cohorts may not be replicated across populations, necessitating stringent methodology and thorough replication to ensure robust findings.

9. Future perspectives

This thesis provides a steppingstone in the development towards robust tools and analytical frameworks to link exposures to outcomes through metabolic perturbations. However, substantial methodological and empirical advances remain necessary. This chapter outlines future directions to refine the developed tools.

Computational scalability and performance optimization: Both tools face scalability challenges with large datasets. MUVR2 often requires high-performance computing environments, especially for the larger datasets investigated in epidemiological settings, while the graphical interface of TriplotGUI may exhibit slow responsiveness as the dimensionality of the input data increases. Future development could address these issues by refactoring code for enhanced memory management, implementing core functions in compiled languages like C++ and GPU processing for increased speed[344].

Temporal modeling**:** While MUVR2 can select predictors for a prospective outcome, the tool lacks support for time-to-event outcomes, which are critical for longitudinal studies. Implementing random survival forest[386] and penalized Cox regression model[387] within MUVR2's framework would enable survival analysis with high-dimensional predictors.

In addition, TriploGUI uses correlation analysis to evaluate exposure-mediator association cross-sectionally. Future updates should introduce linear models that can better account for temporal order. Additionally, it could be expanded to accommodate survival analysis for both exposure-omics and omics-outcome associations, better aligning with mediation frameworks for survival data.

Enhanced functionality and method integration**:** The MUVR2 algorithm could be expanded by incorporating additional machine learning methods, while TriplotGUI could include more data reduction techniques and mediation analysis approaches (e.g., involving time-varying exposures or multiple mediators[388–390]), as well as sensitivity analysis to test mediation assumptions. However, balancing advanced options with usability remains paramount to avoid overwhelming non-technical users.

Validation and benchmarking**:** MUVR2 and TriplotGUI have primarily been applied to omics data and health-related exposures and outcomes. Future research could explore their use with other types of data and domains, such as linking policy to health outcomes through omics-based mediators. Additionally, benchmarking MUVR2 and TriplotGUI's core functionalities against established frameworks (e.g., tidyomics ecosystem[391], Qlucore Omics Explorer[392]) is essential to validate versatility and robustness.

10. Acknowledgements

The 4-year journey is approaching an end. I am 28 now. I remember clearly, 4 years ago on a Friday around 6 pm, Calle called me to tell me I got the PhD position. And that was the start. With a niche in data analysis and R, I have worked from simple univariate analysis to machine learning, to interfaces and websites and to complex molecular epidemiology studies that link everything. It is an okay journey.

First, I would like to thank my main supervisor, **Carl Brunius**. You have guided and helped me in many aspects during the four years. I like discussing problems with you and finding solutions for them one by one. You also give me a lot of freedom to develop on my own. When I am up for a course or conference, you always support me. I appreciate your effort for details in reviewing and editing my manuscripts, especially when there are so many versions. Compared to you, I am a bit too serious about everything and often worry about what if something goes wrong. I like your optimistic attitude and enthusiasm, which encourages me also to deal with chaos and uncertainties more calmly and not panic in front of tight deadlines:). And not least, I am grateful that you always listen to and understand my difficulties, provide help for my future and career, and answer all the phone calls from recruiters as my reference.

My co-supervisors **Agneta Åkesson** and **Ingegerd Johansson**, thank you for the help and support throughout my project. Our project meeting discussions regarding cohort data and epidemiological considerations are often insightful. Your domain-specific knowledge is invaluable for my projects. I am also very grateful for your thoughtful feedback on my manuscript and thesis.

My co-supervisor **Anton Ribbenstedt**, thank you for the fast response and practical help all the time. You are always on Teams :). When I have a technical problem with shiny app or any questions regarding instruments and chemical compounds, I know where I can turn to.

Tessa Schillemans, thanks for sharing your knowledge of POPs and epidemiology and providing help in TriplotGUI. I read your thesis thoroughly and it gives me a lot of ideas of how to write mine! Seeing your academic growth and finishing your PhD is inspiring.

Ingvar Bergdahl, I appreciate your help regarding POPs and cohort data. When I ask you questions or for feedback, I always get very detailed and clear answers/comments in a short time. I appreciate all the long emails explaining things to the clearest and our meetings focusing on practical details. I value the efficiency and precision in our communications.

Rikard Landberg, thank you for being my examiner. I am grateful for your supportive attitude when I applied for many postdoc positions and for always being willing to provide help for my future and career.

Anna-Karin Kolseth. Thank you for all the detailed replies regarding SIMPLER data! **Gael Toubon**, thank you for your suggestions in microbiota data analysis. Special thanks to **Thérése Hjorth** who has answered me so many questions regarding thesis and defense. **Clemens Wittenbecher** and **Nathalie Scheers**, thank you for the thorough revision of my thesis.

I would like to express my deepest gratitude to my friends who have supported, comforted, and encouraged me during my most difficult times. I wish you all the best in your lives and careers. I would like to thank my parents and all my colleagues at Chalmers, Epihub and the LongiTools EU

project (longitools.org) for insightful conversations, engaging discussions, and unwavering support throughout this journey.

I am grateful to SciLifeLab for organizing courses that have significantly enhanced my technical skills and proficiency in R programming. I would also like to acknowledge the Swedish Research Council Formas (grant 2020-01653) for their financial support, which made this work possible. I would like to thank all study participants in all cohorts who donated their biological samples to advance research and contribute to a healthier future, as well as to the infrastructures to manage and deliver data that enabled this work. I am thankful to the Chalmers Mass Spectrometry Infrastructure (CMSI) for the delivery of metabolomics data and to Uppmax/Bianca for providing a secure computational environment that supported my research. I would also like to acknowledge the use of Perplexity pro for text editing during thesis preparation and take full responsibility for the accuracy and integrity of the thesis.

11. Reference

1. NCD Countdown 2030 collaborators. NCD Countdown 2030: pathways to achieving Sustainable Development Goal target 3.4. 2020;(January):19–21.
2. Mendis S. Global progress in prevention of cardiovascular disease. *Cardiovasc Diagn Ther*. 2017;**7**(Suppl 1):S32–S38.
3. Silander K, Alanne M, Kristiansson K, et al. Gender differences in genetic risk profiles for cardiovascular disease. *PLoS One*. 2008;**3**(10).
4. Huma S, Tariq R, Amin Dr. F, Mahmood Dr. KT. Modifiable and non-modifiable predisposing risk factors of myocardial infarction -A review. *Journal of Pharmaceutical Sciences and Research*. 2012;**4**(1):1649–1653.
5. World Health Organization. Cardiovascular diseases (CVDs) [Internet]. 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
6. Bays HE, Kulkarni A, German C, et al. Ten things to know about ten cardiovascular disease risk factors – 2022. *Am J Prev Cardiol* [Internet]. Elsevier B.V.; 2022;**10**(December 2021):100342. Available from: https://doi.org/10.1016/j.ajpc.2022.100342
7. Lind PM, Salihovic S, Stubleski J, Kärrman A, Lind L. Association of Exposure to Persistent Organic Pollutants With Mortality Risk: An Analysis of Data From the Prospective Investigation of Vasculature in Uppsala Seniors (PIVUS) Study. *JAMA Netw Open*. 2019;**2**(4):e193070.
8. Nesci A, Carnuccio C, Ruggieri V, et al. Gut Microbiota and Cardiovascular Disease: Evidence on the Metabolic and Inflammatory Background of a Complex Relationship. *Int J Mol Sci*. 2023;**24**(10).
9. Khatib O. Noncommunicable diseases: Risk factors and regional strategies for prevention and care. *Eastern Mediterranean Health Journal*. 2004;**10**(6):778–788.
10. Norman RE, Carpenter DO, Scott J, Brunea MN, Sly PD. Environmental exposures: An underrecognized contribution to noncommunicable diseases. *Rev Environ Health*. 2013;**28**(1):59–65.
11. Rippe JM. Lifestyle Strategies for Risk Factor Reduction, Prevention, and Treatment of Cardiovascular Disease. *Am J Lifestyle Med*. 2019;**13**(2):204–212.
12. Nestel PJ, Mori TA. Dietary patterns, dietary nutrients and cardiovascular disease. *Rev Cardiovasc Med*. 2022;**23**(1).
13. Schillemans T, Yan Y, Ribbenstedt A, et al. OMICs Signatures Linking Persistent Organic Pollutants to Cardiovascular Disease in the Swedish Mammography Cohort. *Environ Sci Technol*. 2024;**58**(2):1036–1047.
14. Newgard CB. Metabolomics and Metabolic Diseases: Where Do We Stand? *Cell Metab* [Internet]. Elsevier Inc.; 2017;**25**(1):43–56. Available from: http://dx.doi.org/10.1016/j.cmet.2016.09.018
15. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*. Nature Publishing Group; 2012;**13**(4):263–269.
16. Schneider M V., Orchard S. Omics Technologies, Data and Bioinformatics Principles. *Methods in Molecular Biology* 2011.
17. Richmond RC, Hemani G, Tilling K, Davey Smith G, Relton CL. Challenges and novel approaches for investigating molecular mediation. *Hum Mol Genet*. 2016;**25**(R2):R149–R156.
18. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. *Annu Rev Psychol*. 2007;**58**(February):593–614.
19. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R package for causal mediation analysis. *J Stat Softw*. 2014;**59**(5).

20. Chadeau-Hyam M, Athersuch TJ, Keun HC, et al. Meeting-in-the-middle using metabolic profiling-a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers*. 2011;**16**(1):83–88.
21. Akbaraly T, Würtz P, Singh-Manoux A, et al. Association of circulating metabolites with healthy diet and risk of cardiovascular disease: Analysis of two cohort studies. *Sci Rep*. 2018;**8**(1):1–14.
22. Li J, Guasch-Ferré M, Chung W, et al. The Mediterranean diet, plasma metabolome, and cardiovascular disease risk. *Eur Heart J*. 2020;**41**(28):2645–2656.
23. Singh P, Meenatchi R, Ahmed ZHT, et al. Implications of the gut microbiome in cardiovascular diseases: Association of gut microbiome with cardiovascular diseases, therapeutic interventions and multi-omics approach for precision medicine. *Medicine in Microecology* [Internet]. Elsevier B.V.; 2024;**19**(November 2023):100096. Available from: https://doi.org/10.1016/j.medmic.2023.100096
24. Perng W, Aslibekyan S. Find the needle in the haystack, then find it again: Replication and validation in the 'omics era. *Metabolites*. 2020;**10**(7):1–13.
25. Perakakis N, Yazdani A, Karniadakis GE, Mantzoros C. Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism*. 2018;**87**:A1–A9.
26. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. *Annu Rev Public Health*. 2019;**41**:21–36.
27. Hawkins DM. The Problem of Overfitting. *J Chem Inf Comput Sci*. 2004;**44**(1):1–12.
28. Posma JM, Garcia-Perez I, Ebbels TMD, et al. Optimized Phenotypic Biomarker Discovery and Confounder Elimination via Covariate-Adjusted Projection to Latent Structures from Metabolic Spectroscopy Data. *J Proteome Res*. 2018;**17**(4):1586–1595.
29. Mensah GA, Roth GA, Fuster V. The Global Burden of Cardiovascular Diseases and Risk Factors: 2020 and Beyond. *J Am Coll Cardiol*. 2019;**74**(20):2529–2532.
30. Roth GA, Abate D, Abate KH, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*. 2018;**392**(10159):1736–1788.
31. Li S, Peng Y, Wang X, et al. Cardiovascular events and death after myocardial infarction or ischemic stroke in an older Medicare population. *Clin Cardiol*. 2019;**42**(3):391–399.
32. Thygesen K, Alpert JS, White HD. Universal definition of myocardial infarction. *Circulation*. 2007;**116**(22):2634–2653.
33. Niranjan Ojha; Amit S. Dhamoon. Myocardial Infarction [Internet]. *StatPearls Publishing* 2025. Available from: https://www.ncbi.nlm.nih.gov/books/NBK537076/
34. Reed GW, Rossi JE, Cannon CP. Acute myocardial infarction. *The Lancet*. 2017;**389**(10065):197–210.
35. Lu L, Liu M, Sun RR, Zheng Y, Zhang P. Myocardial Infarction: Symptoms and Treatments. *Cell Biochem Biophys* [Internet]. Springer US; 2015;**72**(3):865–867. Available from: http://dx.doi.org/10.1007/s12013-015-0553-4
36. Krishnamurthi R V., Feigin VL, Forouzanfar MH, et al. Global and regional burden of first-ever ischaemic and haemorrhagic stroke during 1990-2010: Findings from the Global Burden of Disease Study 2010. *Lancet Glob Health* [Internet]. Krishnamurthi et al. Open Access article distributed under the terms of CC BY-NC-ND; 2013;**1**(5):e259–e281. Available from: http://dx.doi.org/10.1016/S2214-109X(13)70089-5
37. Krafft PR, Bailey EL, Lekic T, et al. Etiology of stroke and choice of models. *International Journal of Stroke*. 2012;**7**(5):398–406.
38. Turner GM, McMullan C, Aiyegbusi OL, et al. Stroke risk following traumatic brain injury: Systematic review and meta-analysis. *International Journal of Stroke*. 2021;**16**(4):370–384.

39. National Institue of Neurological Disorders and Stroke. Brain Basics: Preventing Stroke [Internet]. 2024. Available from: https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-preventing-stroke
40. Powers WJ, Rabinstein AA, Ackerson T, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke a guideline for healthcare professionals from the American Heart Association/American Stroke A. *Stroke* 2019.
41. Ferns GAA. New and emerging risk factors for CVD. *Proceedings of the Nutrition Society*. 2008;**67**(2):223–231.
42. Kilkenny MF, Dunstan L, Busingye D, et al. Knowledge of risk factors for diabetes or cardiovascular disease (CVD) is poor among individuals with risk factors for CVD. *PLoS One*. 2017;**12**(2):1–11.
43. Yusuf S, Joseph P, Rangarajan S, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *The Lancet*. 2020;**395**(10226):795–808.
44. Brauer M, Casadei B, Harrington RA, et al. Taking a Stand Against Air Pollution—The Impact on Cardiovascular Disease: A Joint Opinion from the World Heart Federation, American College of Cardiology, American Heart Association, and the European Society of Cardiology. *J Am Coll Cardiol* [Internet]. Elsevier; 2021;**77**(13):1684–1688. Available from: http://dx.doi.org/10.1016/j.jacc.2020.12.003
45. Liu M, Lv J, Pan Z, Wang D, Zhao L, Guo X. Mitochondrial dysfunction in heart failure and its therapeutic implications. *Front Cardiovasc Med*. 2022;**9**(August):1–12.
46. Deprince A, Haas JT, Staels B. Dysregulated lipid metabolism links NAFLD to cardiovascular disease. *Mol Metab* [Internet]. Elsevier GmbH; 2020;**42**(October):101092. Available from: https://doi.org/10.1016/j.molmet.2020.101092
47. Li K, Wan B, Li S, et al. Mitochondrial dysfunction in cardiovascular disease: Towards exercise regulation of mitochondrial function. *Front Physiol*. 2023;**14**(January):1–9.
48. Ketola E, Sipilä R, Mäkelä M. Effectiveness of individual lifestyle interventions in reducing cardiovascular disease and risk factors. *Ann Med*. 2000;**32**(4):239–251.
49. Fukumoto Y. Lifestyle intervention for primary prevention of cardiovascular diseases. *Eur J Prev Cardiol* [Internet]. Oxford University Press; 2022;**29**(17):2250–2251. Available from: https://doi.org/10.1093/eurjpc/zwac245
50. Gupta R, Wood DA. Primary prevention of ischaemic heart disease: populations, individuals, and health professionals. *The Lancet* [Internet]. Elsevier Ltd; 2019;**394**(10199):685–696. Available from: http://dx.doi.org/10.1016/S0140-6736(19)31893-8
51. Yusuf S, Joseph P, Rangarajan S, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *The Lancet*. 2020;**395**(10226):795–808.
52. Sisti LG, Dajko M, Campanella P, Shkurti E, Ricciardi W, Waure C de. The effect of multifactorial lifestyle interventions on cardiovascular risk factors: a systematic review and meta-analysis of trials conducted in the general population and high risk groups. *Prev Med (Baltim)* [Internet]. Elsevier; 2018;**109**(December 2017):82–97. Available from: https://doi.org/10.1016/j.ypmed.2017.12.027
53. Lloyd-Jones DM, Wilson PWF, Larson MG, et al. Framingham risk score and prediction of lifetime risk for coronary heart disease. *American Journal of Cardiology*. 2004;**94**(1):20–24.
54. Ioannidis JPA. Exposure-wide epidemiology: Revisiting Bradford Hill. *Stat Med*. 2016;**35**(11):1749–1762.

55. Vanderweele TJ. Outcome-wide epidemiology. *Epidemiology*. 2017;**28**(3):399–402.
56. Wild CP. Complementing the genome with an 'exposome': The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers and Prevention*. 2005;**14**(8):1847–1850.
57. Bibbò S, Ianiro G, Giorgio V, et al. The role of diet on gut microbiota composition. *Eur Rev Med Pharmacol Sci*. 2016;**20**(22):4742–4749.
58. Odland JO. The importance of diet on exposure and effects of persistent organic pollutants on human health in the arctic. *Annual Review of Biomedical Sciences*. 2005;**7**:161–181.
59. Tian Y, Gui W, Rimal B, et al. Metabolic impact of persistent organic pollutants on gut microbiota. *Gut Microbes* [Internet]. Taylor & Francis; 2020;**12**(1):1–16. Available from: https://doi.org/10.1080/19490976.2020.1848209
60. Popli S, Badgujar PC, Agarwal T, Bhushan B, Mishra V. Persistent organic pollutants in foods, their interplay with gut microbiota and resultant toxicity. *Science of the Total Environment* [Internet]. Elsevier B.V.; 2022;**832**(March):155084. Available from: https://doi.org/10.1016/j.scitotenv.2022.155084
61. Sunderland EM, Hu XC, Dassuncao C, Tokranov AK, Wagner CC, Allen JG. A review of the pathways of human exposure to poly- and perfluoroalkyl substances (PFASs) and present understanding of health effects. *J Expo Sci Environ Epidemiol* [Internet]. Springer US; 2019;**29**(2):131–147. Available from: http://dx.doi.org/10.1038/s41370-018-0094-1
62. Forsthuber M, Kaiser AM, Granitzer S, et al. Albumin is the major carrier protein for PFOS, PFOA, PFHxS, PFNA and PFDA in human plasma. *Environ Int* [Internet]. Elsevier; 2020;**137**(November 2019):105324. Available from: https://doi.org/10.1016/j.envint.2019.105324
63. Li QQ, Loganath A, Chong YS, Tan J, Obbard JP. Persistent organic pollutants and adverse health effects in humans. *Journal of Toxicology and Environmental Health - Part A: Current Issues*. 2006;**69**(21):1987–2005.
64. Münzel T, Miller MR, Sørensen M, Lelieveld J, Daiber A, Rajagopalan S. Reduction of environmental pollutants for prevention of cardiovascular disease: it's time to act. *Eur Heart J*. 2020;**41**(41):3989–3997.
65. Li QQ, Loganath A, Chong YS, Tan J, Obbard JP. Persistent organic pollutants and adverse health effects in humans. *Journal of Toxicology and Environmental Health - Part A: Current Issues*. 2006;**69**(21):1987–2005.
66. Forsthuber M, Kaiser AM, Granitzer S, et al. Albumin is the major carrier protein for PFOS, PFOA, PFHxS, PFNA and PFDA in human plasma. *Environ Int* [Internet]. Elsevier; 2020;**137**(February):105324. Available from: https://doi.org/10.1016/j.envint.2019.105324
67. Palencia M, Lerma TA, Garcés V, Mora MA, Martínez JM, Palencia SL. Passive sampler of organochloride compounds in water and air. *Eco-friendly Functional Polymers*. Elsevier; 2021 Jan 1;297–325.
68. Singh S, Sharma S, Sarma SJ, Misra K, Brar SK. Pesticides in water. *Handbook of Water Purity and Quality*. Academic Press; 2021 Jan 1;231–253.
69. Martyniuk CJ, Mehinto AC, Denslow ND. Organochlorine pesticides: Agrochemicals with potent endocrine-disrupting properties in fish. *Mol Cell Endocrinol*. 2020;**507**:1–24.
70. Jayaraj R, Megha P, Sreedev P. Review Article. Organochlorine pesticides, their toxic effects on living organisms and their fate in the environment. *Interdiscip Toxicol*. 2016;**9**(3–4):90–100.
71. Paasivirta J. Organochlorine compounds in the environment. *Water Science and Technology*. 1988;**20**(2):119–129.

72. Valdespino C, Guillen-Guillen ZG, Albino-Miranda S, Osten JR von, Vázquez G. Spatiotemporal distribution of organochlorine pesticides in the upper La Antigua watershed, Veracruz Mexico. *Sci Rep*. 2024;**14**(1):1–10.
73. Donat-Vargas C, Moreno-Franco B, Laclaustra M, Sandoval-Insausti H, Jarauta E, Guallar-Castillon P. Exposure to dietary polychlorinated biphenyls and dioxins, and its relationship with subclinical coronary atherosclerosis: The Aragon Workers' Health Study. *Environ Int* [Internet]. Elsevier; 2020;**136**(January):105433. Available from: https://doi.org/10.1016/j.envint.2019.105433
74. Monica Lind P, Bavel B van, Salihovic S, Lind L. Circulating levels of persistent organic pollutants (POPs) and carotid atherosclerosis in the elderly. *Environ Health Perspect*. 2012;**120**(1):38–43.
75. Park SH, Lim J eun, Park H, Jee SH. Body burden of persistent organic pollutants on hypertension: a meta-analysis. *Environmental Science and Pollution Research* [Internet]. Environmental Science and Pollution Research; 2016;**23**(14):14284–14293. Available from: http://dx.doi.org/10.1007/s11356-016-6568-6
76. Donat-Vargas C, Akesson A, Tornevi A, et al. Persistent organochlorine pollutants in plasma, blood pressure, and hypertension in a longitudinal study. *Hypertension*. 2018;**71**(6):1258–1268.
77. Raffetti E, Donat-Vargas C, Mentasti S, Chinotti A, Donato F. Association between exposure to polychlorinated biphenyls and risk of hypertension: A systematic review and meta-analysis. *Chemosphere* [Internet]. Elsevier Ltd; 2020;**255**:126984. Available from: https://doi.org/10.1016/j.chemosphere.2020.126984
78. Donat-Vargas C, Schillemans T, Kiviranta H, et al. Blood Levels of Organochlorine Contaminants Mixtures and Cardiovascular Disease. *JAMA Netw Open*. 2023;**6**(9):e2333347.
79. Ali Mohammadkhani M, Shahrzad S, Haghighi M, Ghanbari R, Mohamadkhani A. Insights into Organochlorine Pesticides Exposure in the Development of Cardiovascular Diseases: A Systematic Review. *Arch Iran Med* [Internet]. 2023;**26**(10):592–599. Available from: https://doi.org/10.34172/aim.2023.86
80. Monica Lind P, Lind L. Are persistent organic pollutants linked to lipid abnormalities, atherosclerosis and cardiovascular disease? A review. *J Lipid Atheroscler*. 2020;**9**(3):334–348.
81. Ghosh R, Siddharth M, Singh N, et al. Organochlorine pesticide-mediated induction of NADPH oxidase and nitric-oxide synthase in endothelial cell. *Journal of Clinical and Diagnostic Research*. 2017;**11**(1):BC09-BC12.
82. Perkins JT, Petriello MC, Newsome BJ, Hennig B. Polychlorinated biphenyls and links to cardiovascular disease. *Environmental Science and Pollution Research*. 2016;**23**(3):2160–2172.
83. Rubini E, Paglia G, Cannella D, et al. β-Hexachlorocyclohexane: A Small Molecule with a Big Impact on Human Cellular Biochemistry. *Biomedicines*. 2020;**8**(11):1–17.
84. Lee H, Ko E, Shin S, Choi M, Kim KT. Differential mitochondrial dysregulation by exposure to individual organochlorine pesticides (OCPs) and their mixture in zebrafish embryos. *Environmental Pollution* [Internet]. Elsevier Ltd; 2021;**277**:115904. Available from: https://doi.org/10.1016/j.envpol.2020.115904
85. Park CM, Kim KT, Rhyu DY. Exposure to a low concentration of mixed organochlorine pesticides impairs glucose metabolism and mitochondrial function in L6 myotubes and zebrafish. *J Hazard Mater* [Internet]. Elsevier B.V.; 2021;**414**(February):125437. Available from: https://doi.org/10.1016/j.jhazmat.2021.125437
86. Salihovic S, Ganna A, Fall T, et al. The metabolic fingerprint of p,p'-DDE and HCB exposure in humans. *Environ Int*. 2016;**88**:60–66.

87. OECD. Reconciling Terminology of the Universe of Per- and Polyfluoroalkyl Substances: Recommendations and Practical Guidance Series on Risk Management No.61. *Organisation for Economic Co-operation and Development*. 2021;(61):1–44.
88. Glüge J, Scheringer M, Cousins IT, et al. An overview of the uses of per- And polyfluoroalkyl substances (PFAS). *Environ Sci Process Impacts*. 2020;**22**(12):2345–2373.
89. OECD. Toward a new comprehensive global database of per- and polyfluoroalkyl substances (PFASs): Summary report on updating the OECD 2007 list of per- and polyfluoroalkyl substances (PFASs). *Organisation for Economic Co-operation and Development*. 2018;(39):1–24.
90. Williams AJ, Gaines LGT, Grulke CM, et al. Assembly and Curation of Lists of Per- and Polyfluoroalkyl Substances (PFAS) to Support Environmental Science Research. *Front Environ Sci*. 2022;**10**(1):1–28.
91. Giesy JP, Kannan K. Global distribution of perfluorooctane sulfonate in wildlife. *Environ Sci Technol*. 2001;**35**(7):1339–1342.
92. Lau C, Anitole K, Hodes C, Lai D, Pfahles-Hutchens A, Seed J. Perfluoroalkyl acids: A review of monitoring and toxicological findings. *Toxicological Sciences*. 2007;**99**(2):366–394.
93. Pérez F, Nadal M, Navarro-Ortega A, et al. Accumulation of perfluoroalkyl substances in human tissues. *Environ Int*. 2013;**59**:354–362.
94. Steenland K, Fletcher T, Stein CR, et al. Review: Evolution of evidence on PFOA and health following the assessments of the C8 Science Panel. *Environ Int*. 2020;**145**.
95. Radke EG, Wright JM, Christensen K, et al. Epidemiology Evidence for Health Effects of 150 per-and Polyfluoroalkyl Substances: A Systematic Evidence Map. *Environ Health Perspect*. 2022;**130**(9):1–10.
96. Mastrantonio M, Bai E, Uccelli R, Cordiano V, Screpanti A, Crosignani P. Drinking water contamination from perfluoroalkyl substances (PFAS): An ecological mortality study in the Veneto Region, Italy. *Eur J Public Health*. 2018;**28**(1):180–185.
97. Huang M, Jiao J, Zhuang P, Chen X, Wang J, Zhang Y. Serum polyfluoroalkyl chemicals are associated with risk of cardiovascular diseases in national US population. *Environ Int* [Internet]. Elsevier; 2018;**119**(March):37–46. Available from: https://doi.org/10.1016/j.envint.2018.05.051
98. Feng X, Long G, Zeng G, Zhang Q, Song B, Wu KH. Association of increased risk of cardiovascular diseases with higher levels of perfluoroalkylated substances in the serum of adults. *Environmental Science and Pollution Research* [Internet]. Springer Berlin Heidelberg; 2022;**29**(59):89081–89092. Available from: https://doi.org/10.1007/s11356-022-22021-z
99. Dong J, Yang S, Zhuang Q, et al. The Associations of Lipid Profiles With Cardiovascular Diseases and Death in a 10-Year Prospective Cohort Study. *Front Cardiovasc Med*. 2021;**8**(November):1–8.
100. Angelantonio E Di, Sarwar N, Perry P, et al. Major lipids, apolipoproteins, and risk of vascular disease. *JAMA*. 2009;**302**(18):1993–2000.
101. Wang P, Liu D, Yan S, Cui J, Liang Y, Ren S. Adverse Effects of Perfluorooctane Sulfonate on the Liver and Relevant Mechanisms. *Toxics*. 2022;**10**(5).
102. Ory DS. Nuclear receptor signaling in the control of cholesterol homeostasis: Have the orphans found a home? *Circ Res*. 2004;**95**(7):660–670.
103. Behr AC, Plinsch C, Braeuning A, Buhrke T. Activation of human nuclear receptors by perfluoroalkylated substances (PFAS). *Toxicology in Vitro* [Internet]. Elsevier; 2020;**62**(August 2019):104700. Available from: https://doi.org/10.1016/j.tiv.2019.104700
104. Fragki S, Dirven H, Fletcher T, et al. Systemic PFOS and PFOA exposure and disturbed lipid homeostasis in humans: what do we know and what not? *Crit Rev Toxicol* [Internet].

Taylor & Francis; 2021;**51**(2):141–164. Available from: https://doi.org/10.1080/10408444.2021.1888073
105. United States Environmental Protection Agency. Health Effects Support Document for Perfluorooctanoic Acid (PFOA) [Internet]. *Toxics* Taylor & Francis; 2016. Available from: https://doi.org/10.1080/10408444.2021.1888073
106. Bonato M, Corrà F, Bellio M, et al. Pfas environmental pollution and antioxidant responses: An overview of the impact on human field. *Int J Environ Res Public Health*. 2020;**17**(21):1–45.
107. Mattsson K, Rignell-Hydbom A, Holmberg S, et al. Levels of perfluoroalkyl substances and risk of coronary heart disease: Findings from a population-based longitudinal study. *Environ Res* [Internet]. Elsevier; 2015;**142**:148–154. Available from: http://dx.doi.org/10.1016/j.envres.2015.06.033
108. Simpson C, Winquist A, Lally C, Steenland K. Relation between perfluorooctanoic acid exposure and strokes in a large cohort living near a chemical plant. *Environ Res* [Internet]. Elsevier; 2013;**127**:22–28. Available from: http://dx.doi.org/10.1016/j.envres.2013.10.002
109. Hutcheson R, Innes K, Conway B. Perfluoroalkyl substances and likelihood of stroke in persons with and without diabetes. *Diab Vasc Dis Res*. 2020;**17**(1).
110. Schillemans T, Donat-Vargas C, Lindh CH, et al. Per-and Polyfluoroalkyl Substances and Risk of Myocardial Infarction and Stroke: A Nested Case–Control Study in Sweden. *Environ Health Perspect*. Public Health Services, US Dept of Health and Human Services; 2022 Mar 1;**130**(3).
111. Andersen ME, Hagenbuch B, Apte U, et al. Why is elevation of serum cholesterol associated with exposure to perfluoroalkyl substances (PFAS) in humans? A workshop report on potential mechanisms. *Toxicology*. 2021;**459**(2021).
112. Jamie C. DeWitt, Alexander Shnyra, Mostafa Z. Badr, Scott E. Loveless, Denise Hoban, Steven R. Frame, Robyn Cunard, Stacey E. Anderson, B. Jean Meade, Margie M. Peden-Adams RWL &Michael I a, Jamie C. DeWitt, Alexander Shnyra, Mostafa Z. Badr, Scott E. Loveless, Denise Hoban, Steven R. Frame, Robyn Cunard, Stacey E. Anderson, B. Jean Meade, Margie M. Peden-Adams RWL &Michael IL. Immunotoxicity of Perfluorooctanoic Acid and Perfluorooctane Sulfonate and the Role of Peroxisome Proliferator-Activated Receptor Alpha. *Crit Rev Toxicol*. 2009;**39**(1):76–94.
113. Martínez-Valenzuela MDC, Waliszewski SM, Gómez-Arroyo S, et al. Comparison of organochlorine pesticide levels between human blood serum and adipose tissue. *Revista Internacional de Contaminacion Ambiental*. 2017;**33**(3):393–401.
114. Minelli EV, Ribeiro ML. Quantitative method for the determination of organochlorine pesticides in serum. *J Anal Toxicol*. 1996;**20**(1):23–26.
115. Miller V, Micha R, Choi E, Karageorgou D, Webb P, Mozaffarian D. Evaluation of the Quality of Evidence of the Association of Foods and Nutrients with Cardiovascular Disease and Diabetes: A Systematic Review. *JAMA Netw Open*. 2022;**5**(2):1–15.
116. Yu E, Malik VS, Hu FB. Cardiovascular Disease Prevention by Diet Modification: JACC Health Promotion Series. *J Am Coll Cardiol*. 2018;**72**(8):914–926.
117. Mozaffarian D, Rosenberg I, Uauy R. History of modern nutrition science-implications for current research, dietary guidelines, and food policy. *BMJ (Online)*. 2018;**361**:1–6.
118. Marques-Vidal P, Tsampasian V, Cassidy A, et al. Diet and nutrition in cardiovascular disease prevention: a scientific statement of the European Association of Preventive Cardiology and the Association of Cardiovascular Nursing & Allied Professions of the European Society of Cardiology. *Eur J Prev Cardiol* [Internet]. Oxford University Press; 2025;1–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/40504596

119. Slavin JU of M, Lloyd Beate PNGR. Health benefits of fruits and vegetables. *Advances in Nutrition* [Internet]. 2012;**3**(4):506–516. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3649719/
120. Slavin J. Whole grains and human health. *Nutr Res Rev*. 2004;**17**(1):99–110.
121. Ros E. Nuts and CVD. *British Journal of Nutrition*. 2015;**113**(S2):S111–S120.
122. Ramel A, Nwaru BI, Lamberg-Allardt C, et al. White meat consumption and risk of cardiovascular disease and type 2 diabetes: a systematic review and meta-analysis. *Food Nutr Res*. 2023;**67**:1–33.
123. Giosuè A, Calabrese I, Lupoli R, Riccardi G, Vaccaro O, Vitale M. Relations between the Consumption of Fatty or Lean Fish and Risk of Cardiovascular Disease and All-Cause Mortality: A Systematic Review and Meta-Analysis. *Advances in Nutrition*. 2022;**13**(5):1554–1565.
124. Brasky TM, Neuhouser ML, Cohn DE, White E. Associations of long-chain ω-3 fatty acids and fish intake with endometrial cancer risk in the VITamins and Lifestyle cohort. *American Journal of Clinical Nutrition* [Internet]. American Society for Nutrition.; 2014;**99**(3):599–608. Available from: https://doi.org/10.3945/ajcn.113.070524
125. Skåre J, Brantsæter A, Frøyland L, et al. Benefit-risk Assessment of Fish and Fish Products in the Norwegian Diet – An Update. *Eur J Nutr Food Saf* 2015.
126. Li JJ, Dou KF, Zhou ZG, et al. Role of omega-3 fatty acids in the prevention and treatment of cardiovascular Diseases: A consensus statement from the Experts' Committee Of National Society Of Cardiometabolic Medicine. *Front Pharmacol*. 2022;**13**(December):1–19.
127. Fernández-Solà J. The effects of ethanol on the heart: Alcoholic cardiomyopathy. *Nutrients*. 2020;**12**(2):1–18.
128. Eshak ES, Yamagishi K, Iso H. Dietary Fat and Risk of Cardiovascular Disease. *Encyclopedia of Cardiovascular Research and Medicine*. 2017;**1–4**:60–89.
129. Iqbal MP. Trans fatty acids - A risk factor for cardiovascular disease. *Pak J Med Sci*. 2014;**30**(1):194–197.
130. Tu Z, Yang J, Fan C. The role of different nutrients in the prevention and treatment of cardiovascular diseases. *Front Immunol*. 2024;**15**(May):1–15.
131. Reguant-Closa A, Pedolin D, Herrmann M, Nemecek T. Review of Diet Quality Indices that can be Applied to the Environmental Assessment of Foods and Diets. *Curr Nutr Rep* [Internet]. Springer US; 2024;**13**(2):351–362. Available from: https://doi.org/10.1007/s13668-024-00540-0
132. Williams CM, Lovegrove JA, Griffin BA. Dietary patterns and cardiovascular disease. *Proceedings of the Nutrition Society*. 2013;**72**(4):407–411.
133. Nestel PJ, Mori TA. Dietary patterns, dietary nutrients and cardiovascular disease. *Rev Cardiovasc Med*. 2022;**23**(1).
134. Zampelas A, Magriplis E. Dietary patterns and risk of cardiovascular diseases: A review of the evidence. *Proceedings of the Nutrition Society*. 2020;**79**(1):68–75.
135. Bailey RL. Overview of dietary assessment methods for measuring intakes of foods, beverages, and dietary supplements in research studies. *Curr Opin Biotechnol* [Internet]. Elsevier Ltd; 2021;**70**:91–96. Available from: https://doi.org/10.1016/j.copbio.2021.02.007
136. Burrows TL, Ho YY, Rollo ME, Collins CE. Validity of Dietary Assessment Methods When Compared to the Method of Doubly Labeled Water: A Systematic Review in Adults. *Front Endocrinol (Lausanne)*. 2019;**10**(December).
137. Matijašić M, Meštrović T, Paljetak HČ, Perić M, Barešić A, Verbanac D. Gut microbiota beyond bacteria-mycobiome, virome, archaeome, and eukaryotic parasites in IBD. *Int J Mol Sci*. 2020;**21**(8):1–21.

138. Hou K, Wu ZX, Chen XY, et al. Microbiota in health and diseases. *Signal Transduct Target Ther*. Springer US; 2022;**7**(1).
139. Leonard JM, Toro D Del. Defining the Microbiome Components (Bacteria, Viruses, Fungi) and Microbiome Geodiversity. *Surg Infect (Larchmt)*. 2023;**24**(3):208–212.
140. Jandhyala SM, Talukdar R, Subramanyam C, Vuyyuru H, Sasikala M, Reddy DN. Role of the normal gut microbiota. *World J Gastroenterol*. 2015;**21**(29):8836–8847.
141. Rowland I, Gibson G, Heinken A, et al. Gut microbiota functions: metabolism of nutrients and other food components. *Eur J Nutr*. Springer Berlin Heidelberg; 2018;**57**(1):1–24.
142. Martin AM, Sun EW, Rogers GB, Keating DJ. The influence of the gut microbiome on host metabolism through the regulation of gut hormone release. *Front Physiol*. 2019;**10**(MAR):1–11.
143. Uebanso T, Shimohata T, Mawatari K, Takahashi A. Functional Roles of B-Vitamins in the Gut and Gut Microbiome. *Mol Nutr Food Res*. 2020;**64**(18):1–10.
144. Wu HJ, Wu E. The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes*. 2012;**3**(1):4–14.
145. Witkowski M, Weeks TL, Hazen SL. Gut Microbiota and Cardiovascular Disease. *Circ Res*. 2020;**127**(4):553–570.
146. Williams CE, Hammer TJ, Williams CL. Diversity alone does not reliably indicate the healthiness of an animal microbiome. *ISME Journal*. 2024;**18**(1):1–5.
147. Drapkina OM, Ashniev GA, Zlobovskaya OA, et al. Diversities in the Gut Microbial Patterns in Patients with Atherosclerotic Cardiovascular Diseases and Certain Heart Failure Phenotypes. *Biomedicines*. 2022;**10**(11).
148. Jin M, Qian Z, Yin J, Xu W, Zhou X. The role of intestinal microbiota in cardiovascular disease. *J Cell Mol Med*. 2019;**23**(4):2343–2350.
149. Tang WHW, Hazen SL. The contributory role of gut microbiota in cardiovascular disease. *Journal of Clinical Investigation*. 2014;**124**(10):4204–4211.
150. Zhou W, Cheng Y, Zhu P, Nasser MI, Zhang X, Zhao M. Implication of Gut Microbiota in Cardiovascular Diseases. *Oxid Med Cell Longev*. 2020;**2020**.
151. Hu T, Wu Q, Yao Q, Jiang K, Yu J, Tang Q. Short-chain fatty acid metabolism and multiple effects on cardiovascular diseases. *Ageing Res Rev* [Internet]. Elsevier B.V.; 2022;**81**(March):101706. Available from: https://doi.org/10.1016/j.arr.2022.101706
152. Geng J, Sui Z, Dou W, et al. 16S rRNA Gene Sequencing Reveals Specific Gut Microbes Common to Medicinal Insects. *Front Microbiol*. 2022;**13**(May):1–16.
153. Min YW, Rezaie A, Pimentel M. Bile Acid and Gut Microbiota in Irritable Bowel Syndrome. *J Neurogastroenterol Motil*. 2022;**28**(4):549–561.
154. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol*. 2017;**35**(9):833–844.
155. Durazzi F, Sala C, Castellani G, Manfreda G, Remondini D, Cesare A De. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Sci Rep* [Internet]. Nature Publishing Group UK; 2021;**11**(1):1–10. Available from: https://doi.org/10.1038/s41598-021-82726-y
156. Moraes ACF de, Fernandes GR, Silva IT da, et al. Enterotype may drive the dietary-associated cardiometabolic risk factors. *Front Cell Infect Microbiol*. 2017;**7**(FEB):1–9.
157. Wu GD, Chen J, Hoffmann C, et al. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science (1979)* [Internet]. 2011;**334**(October):105–109. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3368382&tool=pmcentrez&rendertype=abstract
158. Noh H, Jang HH, Kim G, et al. Taxonomic composition and diversity of the gut microbiota in relation to habitual dietary intake in korean adults. *Nutrients*. 2021;**13**(2):1–16.
159. Lim MY, Rho M, Song YM, Lee K, Sung J, Ko G. Stability of gut enterotypes in Korean monozygotic twins and their association with biomarkers and diet. *Sci Rep*. 2014;**4**:1–7.

160. Gasull M, Bosch de Basea M, Puigdomènech E, Pumarega J, Porta M. Empirical analyses of the influence of diet on human concentrations of persistent organic pollutants: A systematic review of all studies conducted in Spain. *Environ Int* [Internet]. Elsevier Ltd; 2011;**37**(7):1226–1235. Available from: http://dx.doi.org/10.1016/j.envint.2011.05.008
161. Zhang L, Nichols RG, Correll J, et al. Persistent organic pollutants modify gut microbiota–host metabolic homeostasis in mice through aryl hydrocarbon receptor activation. *Environ Health Perspect*. 2015;**123**(7):679–688.
162. Roth K, Imran Z, Liu W, Petriello MC. Diet as an Exposure Source and Mediator of Per- and Polyfluoroalkyl Substance (PFAS) Toxicity. *Frontiers in Toxicology*. 2020;**2**(December).
163. Seshasayee SM, Rifas-Shiman SL, Chavarro JE, et al. Dietary patterns and PFAS plasma concentrations in childhood: Project Viva, USA. *Environ Int* [Internet]. Elsevier Ltd; 2021;**151**:106415. Available from: https://doi.org/10.1016/j.envint.2021.106415
164. Sultan H, Buckley JP, Kalkwarf HJ, et al. Dietary per- and polyfluoroalkyl substance (PFAS) exposure in adolescents: The HOME study. *Environ Res* [Internet]. Elsevier Inc.; 2023;**231**(P1):115953. Available from: https://doi.org/10.1016/j.envres.2023.115953
165. Zhang J, Hu L, Xu H. Dietary exposure to per- and polyfluoroalkyl substances: Potential health impacts on human liver. *Science of the Total Environment* [Internet]. Elsevier B.V.; 2024;**907**(August 2023):167945. Available from: https://doi.org/10.1016/j.scitotenv.2023.167945
166. Yu HY, Guo Y, Zeng EY. Dietary intake of persistent organic pollutants and potential health risks via consumption of global aquatic products. *Environ Toxicol Chem*. 2010;**29**(10):2135–2142.
167. Fan X, Wang Z, Li Y, Wang H, Fan W, Dong Z. Estimating the dietary exposure and risk of persistent organic pollutants in China: A national analysis. *Environmental Pollution* [Internet]. Elsevier Ltd; 2021;**288**(July):117764. Available from: https://doi.org/10.1016/j.envpol.2021.117764
168. Odediran A, Obeng-Gyasi E. Association between Combined Metals and PFAS Exposure with Dietary Patterns: A Preliminary Study. *Environments - MDPI*. 2024;**11**(6):1–17.
169. Odland JO. The importance of diet on exposure and effects of persistent organic pollutants on human health in the arctic. *Annual Review of Biomedical Sciences*. 2005;**7**:161–181.
170. Levitan EB, Wolk A, Mittleman MA. Fish consumption, marine omega-3 fatty acids, and incidence of heart failure: A population-based prospective study of middle-aged and elderly men. *Eur Heart J*. 2009;**30**(12):1495–1500.
171. Lai KP, Ng AHM, Wan HT, et al. Dietary exposure to the environmental chemical, PFOS on the diversity of gut microbiota, associated with the development of metabolic syndrome. *Front Microbiol*. 2018;**9**(OCT):1–11.
172. Chiu K, Warner G, Nowak RA, Flaws JA, Mei W. The impact of environmental chemicals on the gut microbiome. *Toxicological Sciences*. 2020;**176**(2):253–284.
173. Richmond RC, Hemani G, Tilling K, Davey Smith G, Relton CL. Challenges and novel approaches for investigating molecular mediation. *Hum Mol Genet*. 2016;**25**(R2):R149–R156.
174. Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet* [Internet]. Nature Publishing Group; 2018;**19**(5):299–310. Available from: http://dx.doi.org/10.1038/nrg.2018.4
175. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. Genome Biology; 2017;**18**(1):1–15.
176. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. *Nat Rev Genet*. Nature Publishing Group; 2011;**12**(1):56–68.
177. Nossal GJV. The digital code of DNA. *Nature*. 2003;**421**:444–448.

178. Manzoni C, Kia DA, Vandrovcova J, et al. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Brief Bioinform*. 2018;**19**(2):286–302.
179. Dai X, Shen L. Advances and Trends in Omics Technology Development. *Front Med (Lausanne)*. 2022;**9**(July).
180. Manera J. Transcriptomics : Decoding the language of gene expression . 2023;**7**(5):1–2.
181. Cao B. Cellular Components: Proteins and Their Crucial Role in Cell Function. *J Biochem*. 2023;**6**(3):55–57.
182. Antohe F. Mass spectrometry based proteomics. *Acta Endocrinol (Copenh)*. 2015;**11**(2):139–142.
183. Scalbert A, Brennan L, Manach C, et al. The food metabolome: A window over dietary exposure. *American Journal of Clinical Nutrition* [Internet]. American Society for Nutrition.; 2014;**99**(6):1286–1308. Available from: https://doi.org/10.3945/ajcn.113.076133
184. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: Beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol*. 2016;**17**(7):451–459.
185. Gorrochategui E, Jaumot J, Lacorte S, Tauler R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *TrAC - Trends in Analytical Chemistry* [Internet]. Elsevier B.V.; 2016;**82**:425–442. Available from: http://dx.doi.org/10.1016/j.trac.2016.07.004
186. Wishart DS. Metabolomics for investigating physiological and pathophysiological processes. *Physiol Rev*. 2019;**99**(4):1819–1875.
187. Khoomrung S, Wanichthanarak K, Nookaew I, et al. Metabolomics and integrative omics for the development of Thai traditional medicine. *Front Pharmacol*. 2017;**8**(JUL):1–11.
188. Vandeputte D, Joossens M. Effects of low and high FODMAP diets on human gastrointestinal microbiota composition in adults with intestinal diseases: A systematic review. *Microorganisms*. 2020;**8**(11):1–15.
189. Gibson PR, Halmos EP, Muir JG. Review article: FODMAPS, prebiotics and gut health-the FODMAP hypothesis revisited. *Aliment Pharmacol Ther*. 2020;**52**(2):233–246.
190. Hansen LBS, Roager HM, Søndertoft NB, et al. A low-gluten diet induces changes in the intestinal microbiome of healthy Danish adults. *Nat Commun*. 2018;**9**(1).
191. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov*. Nature Publishing Group; 2016;**15**(7):473–484.
192. Krautkramer KA, Fan J, Bäckhed F. Gut microbial metabolites as multi-kingdom intermediates. *Nat Rev Microbiol* [Internet]. Springer US; 2021;**19**(2):77–94. Available from: http://dx.doi.org/10.1038/s41579-020-0438-4
193. Guijas C, Montenegro-Burke JR, Warth B, Spilker ME, Siuzdak G. Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nat Biotechnol*. 2018;**36**(4):316–320.
194. Johnson C, Ivanisevic J, Siuzdak G. Beyond Biomarkers and towards Mechanisms. *Nat Rev Mol Cell Biol*. 2016;**17**(July):451–459.
195. Walker DI, Valvi D, Rothman N, Lan Q, Miller GW, Jones DP. The metabolome: A key measure for exposome research in epidemiology. *Curr Epidemiol Rep*. 2019;**6**:93–103.
196. Newgard CB. Metabolomics and Metabolic Diseases: Where Do We Stand? *Cell Metab* [Internet]. Elsevier Inc.; 2017;**25**(1):43–56. Available from: http://dx.doi.org/10.1016/j.cmet.2016.09.018
197. Ramautar R, Berger R, Greef J van der, Hankemeier T. Human metabolomics: Strategies to understand biology. *Curr Opin Chem Biol* [Internet]. Elsevier Ltd; 2013;**17**(5):841–846. Available from: http://dx.doi.org/10.1016/j.cbpa.2013.06.015

198. Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *J Am Soc Mass Spectrom*. 2016;**27**(12):1897–1905.
199. Fraga-Corral M, Carpena M, Garcia-Oliveira P, Pereira AG, Prieto MA, Simal-Gandara J. Analytical Metabolomics and Applications in Health, Environmental and Food Science. *Crit Rev Anal Chem* [Internet]. Taylor & Francis; 2022;**52**(4):712–734. Available from: https://doi.org/10.1080/10408347.2020.1823811
200. Mlynárik V. Introduction to nuclear magnetic resonance. *Anal Biochem*. 2017;**529**:4–9.
201. Zhang H, Ouyang Z, Zhang W. Advances in mass spectrometry for clinical analysis: Data acquisition, interpretation and information integration. *TrAC - Trends in Analytical Chemistry* [Internet]. Elsevier B.V.; 2023;**169**(September):117380. Available from: https://doi.org/10.1016/j.trac.2023.117380
202. Haggarty J, Burgess KE. Recent advances in liquid and gas chromatography methodology for extending coverage of the metabolome. *Curr Opin Biotechnol* [Internet]. Elsevier Ltd; 2017;**43**:77–85. Available from: http://dx.doi.org/10.1016/j.copbio.2016.09.006
203. Segers K, Declerck S, Mangelings D, Heyden Y Vander, Eeckhaut A Van. Analytical techniques for metabolomic studies: A review. *Bioanalysis*. 2019;**11**(24):2297–2318.
204. Spicer RA, Salek R, Steinbeck C. Comment: A decade after the metabolomics standards initiative it's time for a revision. *Sci Data*. 2017;**4**:2–4.
205. Schymanski EL, Jeon J, Gulde R, et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environ Sci Technol*. 2014;**48**(4):2097–2098.
206. Chaleckis R, Meister I, Zhang P, Wheelock CE. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Curr Opin Biotechnol* [Internet]. Elsevier Ltd; 2019;**55**:44–50. Available from: https://doi.org/10.1016/j.copbio.2018.07.010
207. Deja S, Porebska I, Kowal A, et al. Metabolomics provide new insights on lung cancer staging and discrimination from chronic obstructive pulmonary disease. *J Pharm Biomed Anal* [Internet]. Elsevier B.V.; 2014;**100**:369–380. Available from: http://dx.doi.org/10.1016/j.jpba.2014.08.020
208. Peng J, Chen YT, Chen CL, Li L. Development of a universal metabolome-standard method for long-term LC-MS metabolome profiling and its application for bladder cancer urine-metabolite- biomarker discovery. *Anal Chem*. 2014;**86**(13):6540–6547.
209. Qiu Y, Cai G, Zhou B, et al. A distinct metabolic signature of human colorectal cancer with prognostic potential. *Clinical Cancer Research*. 2014;**20**(8):2136–2146.
210. Torre D La, Seppänen-Laakso T, Larsson HE, et al. Decreased cord-blood phospholipids in young age-at-onset type 1 diabetes. *Diabetes*. 2013;**62**(11):3951–3956.
211. Wang-Sattler R, Yu Z, Herder C, et al. Novel biomarkers for pre-diabetes identified by metabolomics. *Mol Syst Biol*. 2012;**8**(615).
212. Zordoky BN, Sung MM, Ezekowitz J, et al. Metabolomic fingerprint of heart failure with preserved ejection fraction. *PLoS One*. 2015;**10**(5):1–19.
213. Wang J, Li Z, Chen J, et al. Metabolomic identification of diagnostic plasma biomarkers in humans with chronic heart failure. *Mol Biosyst*. 2013;**9**(11):2618–2626.
214. Lind L, Fall T, Ärnlöv J, Elmståhl S, Sundström J. Large-Scale Metabolomics and the Incidence of Cardiovascular Disease. *J Am Heart Assoc*. 2023;**12**(2).
215. Lv J, Pan C, Cai Y, et al. Plasma metabolomics reveals the shared and distinct metabolic disturbances associated with cardiovascular events in coronary artery disease. *Nat Commun* [Internet]. Springer US; 2024;**15**(1). Available from: http://dx.doi.org/10.1038/s41467-024-50125-2
216. Iliou A, Mikros E, Karaman I, et al. Metabolic phenotyping and cardiovascular disease: An overview of evidence from epidemiological settings. *Heart*. 2021;**107**(14):1123–1129.

217. Bordoni A, Capozzi F. The foodomics approach for discovering biomarkers of food consumption in nutrition studies. *Curr Opin Food Sci* [Internet]. Elsevier Ltd; 2015;**4**:124–128. Available from: http://dx.doi.org/10.1016/j.cofs.2015.07.005
218. Brennan L, Hu FB. Metabolomics-Based Dietary Biomarkers in Nutritional Epidemiology—Current Status and Future Opportunities. *Mol Nutr Food Res*. 2019;**63**(1):1–5.
219. Garcia-Perez I, Posma JM, Gibson R, et al. Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial. *Lancet Diabetes Endocrinol* [Internet]. The Author(s). Published by Elsevier Ltd. This is an Open Access Article under the CC BY license; 2017;**5**(3):184–195. Available from: http://dx.doi.org/10.1016/S2213-8587(16)30419-3
220. Playdon MC, Moore SC, Derkach A, et al. Identifying biomarkers of dietary patterns by using metabolomics. *American Journal of Clinical Nutrition*. 2017;**105**(2):450–465.
221. Esko T, Hirschhorn JN, Feldman HA, et al. Metabolomic profiles as reliable biomarkers of dietary composition. *American Journal of Clinical Nutrition*. 2017;**105**(3):547–554.
222. O'Gorman A, Brennan L. Metabolomic applications in nutritional research: A perspective. *J Sci Food Agric*. 2015;**95**(13):2567–2570.
223. Savolainen O, Lind MV, Bergström G, Fagerberg B, Sandberg AS, Ross A. Biomarkers of food intake and nutrient status are associated with glucose tolerance status and development of type 2 diabetes in older Swedish women. *American Journal of Clinical Nutrition*. 2017;**106**(5):1302–1310.
224. Playdon MC, Ziegler RG, Sampson JN, et al. Nutritional metabolomics and breast cancer risk in a prospective study. *American Journal of Clinical Nutrition* [Internet]. American Society for Nutrition.; 2017;**106**(2):637–649. Available from: https://doi.org/10.3945/ajcn.116.150912
225. India-Aldana S, Yao M, Midya V, et al. PFAS Exposures and the Human Metabolome: A Systematic Review of Epidemiological Studies. *Curr Pollut Rep*. 2023;**9**(3):510–568.
226. Ma X, Cai D, Chen Q, et al. Hunting Metabolic Biomarkers for Exposure to Per- and Polyfluoroalkyl Substances: A Review. *Metabolites* 2024.
227. Yan Q, Paul KC, Walker DI, et al. High-Resolution Metabolomic Assessment of Pesticide Exposure in Central Valley, California. *Chem Res Toxicol*. 2021;**34**(5):1337–1347.
228. Liu Q, Wang Q, Xu C, et al. Organochloride pesticides impaired mitochondrial function in hepatocytes and aggravated disorders of fatty acid metabolism. *Sci Rep*. Nature Publishing Group; 2017;**7**(October 2016):1–11.
229. Qi C, He G, Qian K, et al. gutMGene v2.0: an updated comprehensive database for target genes of gut microbes and microbial metabolites. *Nucleic Acids Res*. Oxford University Press; 2025;**53**(D1):D783–D788.
230. Dekkers KF, Sayols-Baixeras S, Baldanzi G, et al. An online atlas of human plasma metabolite signatures of gut microbiome composition. *Nat Commun*. Springer US; 2022;**13**(1):1–12.
231. Agus A, Clément K, Sokol H. Gut microbiota-derived metabolites as central regulators in metabolic disorders. *Gut*. 2021;**70**(6):1174–1182.
232. Parker A, Fonseca S, Carding SR. Gut microbes and metabolites as modulators of blood-brain barrier integrity and brain health. *Gut Microbes* [Internet]. Taylor & Francis; 2020;**11**(2):135–157. Available from: https://doi.org/10.1080/19490976.2019.1638722
233. Salihovic S, Lind L, Larsson A, Lind PM. Plasma perfluoroalkyls are associated with decreased levels of proteomic inflammatory markers in a cross-sectional study of an elderly population. *Environ Int* [Internet]. Elsevier; 2020;**145**(September):106099. Available from: https://doi.org/10.1016/j.envint.2020.106099
234. Li J, Guasch-Ferré M, Chung W, et al. The Mediterranean diet, plasma metabolome, and cardiovascular disease risk. *Eur Heart J*. 2020;**41**(28):2645–2656.

235. Jansen VL, Gerdes VE, Middeldorp S, Mens TE van. Gut microbiota and their metabolites in cardiovascular disease. *Best Pract Res Clin Endocrinol Metab*. 2021;**35**(3).
236. Deng P, Li X, Petriello MC, Wang C, Morris AJ, Hennig B. Application of metabolomics to characterize environmental pollutant toxicity and disease risks. *Rev Environ Health*. 2019;**34**(3):251–259.
237. Yan Y, Schillemans T, Skantze V, Brunius C. Adjusting for covariates and assessing modeling fitness in machine learning using MUVR2. *Bioinformatics Advances*. 2024;**4**(1).
238. Li R, Li L, Xu Y, Yang J. Machine learning meets omics: Applications and perspectives. *Brief Bioinform*. 2022;**23**(1):1–22.
239. Filzmoser P, Liebmann B, Varmuza K. Repeated double cross validation. *J Chemom*. 2009;**23**(4):160–171.
240. Shi L, Westerhuis JA, Rosén J, Landberg R, Brunius C. Variable selection and validation in multivariate modelling. *Bioinformatics*. 2019;**35**(6):972–980.
241. Maharana K, Mondal S, Nemade B. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*. Elsevier B.V.; 2022;**3**(1):91–99.
242. Amorim LBV de, Cavalcanti GDC, Cruz RMO. The choice of scaling technique matters for classification performance. *Appl Soft Comput* [Internet]. Elsevier B.V.; 2023;**133**:109924. Available from: https://doi.org/10.1016/j.asoc.2022.109924
243. Feng C, Wang H, Lu N, et al. Log-transformation and its implications for data analysis. *Shanghai Arch Psychiatry*. 2014;**26**(2):105–109.
244. Kwak SK, Kim JH. Statistical data preparation: Management of missing values and outliers. *Korean J Anesthesiol*. 2017;**70**(4):407–411.
245. Berg RA van den, Hoefsloot HCJ, Westerhuis JA, Smilde AK, Werf MJ van der. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*. 2006;**7**:1–15.
246. Eckhardt CM, Madjarova SJ, Williams RJ, et al. Unsupervised machine learning methods and emerging applications in healthcare. *Knee Surgery, Sports Traumatology, Arthroscopy* [Internet]. Springer Berlin Heidelberg; 2023;**31**(2):376–381. Available from: https://doi.org/10.1007/s00167-022-07233-7
247. Wong PC. Unsupervised Machine Learning. *Methodology of Educational Measurement and Assessment* 2021.
248. Naeem S, Ali A, Anam S, Ahmed MM. An Unsupervised Machine Learning Algorithms: Comprehensive Review. *International Journal of Computing and Digital Systems*. 2023;**13**(1):911–921.
249. Wold S. Principal Component Analysis. *Chemometrics and intelligent laboratory systems* [Internet]. 1987;**2**(1–3):37–52. Available from: http://files.isec.pt/DOCUMENTOS/SERVICOS/BIBLIO/Documentos de acesso remoto/Principal components analysis.pdf
250. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;**2**(4):433–459.
251. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;**9**.
252. DiLeo M V., Strahan GD, Bakker M den, Hoekenga OA. Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. *PLoS One*. 2011;**6**(10).
253. Li J, Zhou D, Qiu W, et al. Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design. *Sci Rep*. 2018;**8**(1):1–8.
254. Jiang T, Gradus JL, Rosellini AJ. Supervised Machine Learning: A Brief Primer. *Behav Ther* [Internet]. Elsevier Ltd; 2020;**51**(5):675–687. Available from: https://doi.org/10.1016/j.beth.2020.05.002

255. Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. *iScience* [Internet]. The Author(s); 2022;**25**(2):103798. Available from: https://doi.org/10.1016/j.isci.2022.103798
256. Nasteski V. An overview of the supervised machine learning methods. *HorizonsB*. 2017;**4**(December):51–62.
257. Nikolaos Perakakis, Alireza Yazdani, George E. Karniadakis CM. Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism*. 2018;1–21.
258. Grissa D, Pétéra M, Brandolini M, Napoli A, Comte B, Pujos-Guillot E. Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data. *Front Mol Biosci*. Frontiers Media S.A.; 2016 Jul 8;**3**(JUL):1–15.
259. Iguyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003;**3**:1157–1182.
260. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* [Internet]. Elsevier B.V.; 2020;**415**:295–316. Available from: https://doi.org/10.1016/j.neucom.2020.07.061
261. Valdenegro-Toro M, Sabatelli M. Machine Learning Students Overfit to Overfitting. *Proc Mach Learn Res*. 2022;**207**:46–51.
262. Dietterich T. Overfitting and Undercomputing in Machine Learning. *ACM Computing Surveys (CSUR)*. 1995;**27**(3):326–327.
263. Alturayeif N, Hassine J. Data leakage detection in machine learning code: transfer learning, active learning, or low-shot prompting? *PeerJ Comput Sci*. 2025;**11**:1–29.
264. Apicella A, Isgrò F, Prevete R. Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning. 2024; Available from: http://arxiv.org/abs/2401.13796
265. Faber NM. A closer look at the bias-variance trade-off in multivariate calibration. *J Chemom*. 1999;**13**(2):185–192.
266. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc Natl Acad Sci U S A*. 2019;**116**(32):15849–15854.
267. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*. 2010;**11**:2079–2107.
268. Tampu IE, Eklund A, Haj-Hosseini N. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Sci Data*. Springer US; 2022;**9**(1):1–8.
269. Lee HT, Cheon HR, Lee SH, Shim M, Hwang HJ. Risk of data leakage in estimating the diagnostic performance of a deep-learning-based computer-aided system for psychiatric disorders. *Sci Rep* [Internet]. Nature Publishing Group UK; 2023;**13**(1):1–9. Available from: https://doi.org/10.1038/s41598-023-43542-8
270. Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The 'K' in K-fold cross validation. *ESANN 2012 proceedings, 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 2012;(April):441–446.
271. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit* [Internet]. Elsevier; 2015;**48**(9):2839–2846. Available from: http://dx.doi.org/10.1016/j.patcog.2015.03.009
272. T R M, V VK, V DK, Geman O, Margala M, Guduri M. The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. *Healthcare Analytics* [Internet]. Elsevier Inc.; 2023;**4**(August):100247. Available from: https://doi.org/10.1016/j.health.2023.100247
273. Sadeghi-Bazargani H, Banani A, Mohammadi S. Using SIMCA statistical software package to apply orthogonal projections to latent structures modeling. *2010 World Automation Congress*. 2010. p. 1–9.

274. Westerhuis JA, Hoefsloot HCJ, Smit S, et al. Assessment of PLSDA cross validation. *Metabolomics*. 2008 Mar;**4**(1):81–89.
275. Wainer J, Cawley G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Syst Appl* [Internet]. Elsevier Ltd; 2021;**182**(December 2019):115222. Available from: https://doi.org/10.1016/j.eswa.2021.115222
276. Moreno-Torres JG, Saez JA, Herrera F. Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Trans Neural Netw Learn Syst*. 2012;**23**(8):1304–1312.
277. Kim JH. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput Stat Data Anal* [Internet]. Elsevier B.V.; 2009;**53**(11):3735–3745. Available from: http://dx.doi.org/10.1016/j.csda.2009.04.009
278. Ishibuchi H, Nojima Y. Repeated double cross-validation for choosing a single solution in evolutionary multi-objective fuzzy classifier design. *Knowl Based Syst* [Internet]. Elsevier B.V.; 2013;**54**:22–31. Available from: http://dx.doi.org/10.1016/j.knosys.2013.09.023
279. Liebal UW, Phan ANT, Sudhakar M, Raman K, Blank LM. Machine learning applications for mass spectrometry-based metabolomics. *Metabolites* MDPI AG; 2020. p. 1–23.
280. Mendez KM, Reinke SN, Broadhurst DI. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics*. Springer New York LLC; 2019 Dec 1;**15**(12):1–15.
281. Rubingh CM, Bijlsma S, Derks EPPA, et al. Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics*. 2006 Jun;**2**(2):53–61.
282. Westerhuis JA, Velzen EJJ van, Hoefsloot HCJ, Smilde AK. Multivariate paired data analysis: Multilevel PLSDA versus OPLSDA. *Metabolomics*. 2010 Mar;**6**(1):119–128.
283. S. Wold, A. Ruhe, H. Wold WJD. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *Society for Industrial and Applied Mathematics Journal on Scientific and Statistical Computing*. 1984;**5**(3):735–743.
284. Pokhrel DR, Sirisomboon P, Khurnpoon L, Posom J, Saechua W. Comparing Machine Learning and PLSDA Algorithms for Durian Pulp Classification Using Inline NIR Spectra. *Sensors*. 2023;**23**(11):1–22.
285. Wold S, Sjöström M, Eriksson L. PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*. 2001;**58**(2):109–130.
286. Gromski PS, Xu Y, Correa E, Ellis DI, Turner ML, Goodacre R. A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. *Anal Chim Acta* [Internet]. Elsevier B.V.; 2014;**829**:1–8. Available from: http://dx.doi.org/10.1016/j.aca.2014.03.039
287. Steyrl D, Scherer R, Faller J, Müller-Putz GR. Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: A practical and convenient non-linear classifier. *Biomedizinische Technik*. 2016;**61**(1):77–86.
288. Zhu T. Analysis on the applicability of the random forest. *J Phys Conf Ser*. 2020;**1607**(1).
289. Biau G, Devroye L, Lugosi G. Consistency of Random Forests and Other Averaging Classifier. *Journal of Machine Learning Research*. 2008;**9**:2015–2033.
290. Antonelli J, Claggett BL, Henglin M, et al. Statistical workflow for feature selection in human metabolomics data. *Metabolites* MDPI AG; 2019. p. 143.
291. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;**67**(2):301–320.
292. Fei Z, Li Y. Estimation and inference for high dimensional generalized linear models: A splitting and smoothing approach. *Journal of Machine Learning Research*. 2021;**22**:1–32.
293. Amaya-Tejera N, Gamarra M, Vélez JI, Zurek E. A distance-based kernel for classification via Support Vector Machines. *Front Artif Intell*. 2024;**7**.

294. Armaghani DJ, Ming YY, Mohammed AS, Momeni E, Maizir H. Effect of SVM Kernel Functions on Bearing Capacity Assessment of Deep Foundations. *Journal of Soft Computing in Civil Engineering*. 2023;**7**(3):111–128.
295. Afifi A, Zanaty EA, Ghoniemy S. Improving the Classification Accuracy Using Support Vector Machines (Svms) With New Kernel. *Journal of Global Research in Computer Science* [Internet]. 2013;**4**(2):1–7. Available from: www.jgrcs.info
296. Eftekhary M, Gholami P, Safari S, Shojaee M. Ranking normalization methods for improving the accuracy of SVM algorithm by DEA method. *Mod Appl Sci*. 2012;**6**(10):26–36.
297. Max A, Wing J, Weston S, et al. Package ' caret '. 2024.
298. Marcilio WE, Eler DM. From explanations to feature selection: Assessing SHAP values as feature selection mechanism. *Proceedings - 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2020*. 2020;340–347.
299. Sharma S, Sharma S, Anidhya A. Understanding Activation Functions in Neural Networks. *International Journal of Engineering Applied Sciences and Technology*. 2020;**4**(12):310–316.
300. Amari S ichi. Backpropagation and stochastic gradient descent method. *Neurocomputing*. 1993;**5**(4–5):185–196.
301. Nemade M, Wadhe V, Shah P, et al. A Comprehensive Study of Using Artificial Neural Networks for Natural Language Processing. *International Journal of Science and Research (IJSR)*. 2023;**12**(9):2042–2048.
302. Kim Y-S, Kim MK, Fu N, Liu J, Wang J, Srebric J. Investigating the Impact of Data Normalization Methods on Predicting Electricity Consumption in a Building Using different Artificial Neural Network Models. *Sustain Cities Soc* [Internet]. Elsevier Ltd; 2024;**118**(October 2023):105570. Available from: https://doi.org/10.1016/j.scs.2024.105570
303. Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecol Modell*. 2004;**178**(3–4):389–397.
304. Hartvigsson O, Barman M, Rabe H, et al. Associations of maternal and infant metabolomes with immune maturation and allergy development at 12 months in the Swedish NICE-cohort. *Sci Rep* [Internet]. Nature Publishing Group UK; 2021;**11**(1):1–12. Available from: https://doi.org/10.1038/s41598-021-92239-3
305. Nordin E, Hellström PM, Dicksved J, Pelve E, Landberg R, Brunius C. Effects of FODMAPs and Gluten on Gut Microbiota and Their Association with the Metabolome in Irritable Bowel Syndrome: A Double-Blind, Randomized, Cross-Over Intervention Study. *Nutrients*. 2023;**15**(13).
306. Schillemans T, Shi L, Liu X, Åkesson A, Landberg R, Brunius C. Visualization and interpretation of multivariate associations with disease risk markers and disease risk—The triplot. *Metabolites*. 2019;**9**(7):1–12.
307. Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: The new in light of the old. *Cancer Epidemiology Biomarkers and Prevention*. 2007;**16**(10):1954–1965.
308. Schillemans T, Shi L, Donat-Vargas C, et al. Plasma metabolites associated with exposure to perfluoroalkyl substances and risk of type 2 diabetes – A nested case-control study. *Environ Int*. Elsevier Ltd; 2021 Jan 1;**146**.
309. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: Methods, interpretation and bias. *Int J Epidemiol*. 2013;**42**(5):1511–1519.
310. Pearl J. Direct and Indirect Effects. *Probabilistic and Causal Inference*. 2022;373–392.
311. VanderWeele TJ. Mediation Analysis: A Practitioner's Guide. *Annu Rev Public Health*. 2016;**37**:17–32.

312. Rijnhart JJM, Lamp SJ, Valente MJ, MacKinnon DP, Twisk JWR, Heymans MW. Mediation analysis methods used in observational research: a scoping review and recommendations. *BMC Med Res Methodol* [Internet]. BioMed Central; 2021;**21**(1):1–17. Available from: https://doi.org/10.1186/s12874-021-01426-3
313. Vansteelandt S. Estimating direct effects in cohort and case-control studies. *Epidemiology*. 2009;**20**(6):851–860.
314. Agler R, Boeck P De. On the interpretation and use of mediation: Multiple perspectives on mediation analysis. *Front Psychol*. 2017;**8**(NOV):1–11.
315. Memon MA, Cheah JH, Ramayah T, Ting H, Chuah F. Mediation analysis: issues and recommendations. *Journal of Applied Structural Equation Modeling*. 2018;**2**(1):I–IX.
316. Shrout PE, Bolger N. Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychol Methods*. 2002;**7**(4):422–445.
317. Lee H, Cashin AG, Lamb SE, et al. A Guideline for Reporting Mediation Analyses of Randomized Trials and Observational Studies: The AGReMA Statement. *JAMA - Journal of the American Medical Association*. 2021;**326**(11):1045–1056.
318. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;**51**(6):1173–1182.
319. Petersen ML, Sinisi SE, Laan MJ Van Der. Estimation of direct causal effects. *Epidemiology*. 2006;**17**(3):276–284.
320. Chi WE, Huang S, Jeon M, Park ES, Melguizo T, Kezar A. A Practical Guide to Causal Mediation Analysis: Illustration With a Comprehensive College Transition Program and Nonprogram Peer and Faculty Interactions. *Front Educ (Lausanne)*. 2022;**7**(August):1–14.
321. Mackinnon DP. Introduction to statistical mediation analysis. *Introduction to Statistical Mediation Analysis*. 2012;(November):1–477.
322. Gelfand L, Mensinger J, Tenhave T. Mediation analysis: A retrospective snapshot of practice and more recent directions. *Journal of General Psychology*. 2009;**136**(2):153–178.
323. Byeon S, Lee W. An introduction to causal mediation analysis with a comparison of 2 r packages. *Journal of Preventive Medicine and Public Health*. 2023;**56**(4):303–311.
324. Valente MJ, Rijnhart JJM, Smyth HL, Muniz FB, MacKinnon DP. Causal Mediation Programs in R, Mplus, SAS, SPSS, and Stata. *Structural Equation Modeling* [Internet]. Routledge; 2020;**27**(6):975–984. Available from: https://doi.org/10.1080/10705511.2020.1777133
325. Shi L, Brunius C, Johansson I, et al. Plasma metabolites associated with type 2 diabetes in a Swedish population: a case–control study nested in a prospective cohort. *American Journal of Clinical Nutrition*. Diabetologia; 2018;**108**(3):564–575.
326. Harris H, Håkansson N, Olofsson C, et al. The Swedish mammography cohort and the cohort of Swedish men : study design and characteristics of two population- based longitudinal cohorts. 2013;**1**(2):1–8.
327. Norberg M, Wall S, Boman K, Weinehall L. The Västerbotten Intervention Programme: background, design and implications. *Glob Health Action*. 2010;**3**(1):4643.
328. Lanuza F, Bondonno NP, Zamora-Ros R, et al. Comparison of Flavonoid Intake Assessment Methods Using USDA and Phenol Explorer Databases: Subcohort Diet, Cancer and Health-Next Generations—MAX Study. *Front Nutr*. 2022;**9**(April):1–11.
329. Unión-Caballero A, Meroño T, Zamora-Ros R, et al. Metabolome biomarkers linking dietary fibre intake with cardiometabolic effects: results from the Danish Diet, Cancer and Health-Next Generations MAX study. *Food Funct*. Royal Society of Chemistry; 2024;**15**(3):1643–1654.
330. Lanuza F, Zamora-Ros R, Rostgaard-Hansen AL, et al. Descriptive analysis of dietary (poly)phenol intake in the subcohort MAX from DCH-NG: “Diet, Cancer and Health—

Next Generations cohort". *Eur J Nutr* [Internet]. Springer Berlin Heidelberg; 2023;**62**(1):337–350. Available from: https://doi.org/10.1007/s00394-022-02977-x
331. Zheng R, Brunius C, Shi L, et al. Prediction and evaluation of the effect of pre-centrifugation sample management on the measurable untargeted LC-MS plasma metabolome. *Anal Chim Acta*. 2021;**1182**.
332. Schillemans T, Yan Y, Ribbenstedt A, et al. OMICs Signatures Linking Persistent Organic Pollutants to Cardiovascular Disease in the Swedish Mammography Cohort. *Environ Sci Technol*. 2024;**58**(2):1036–1047.
333. Michaëlsson K, Baron JA, Byberg L, et al. Combined associations of body mass index and adherence to a Mediterranean-like diet with all-cause and cardiovascular mortality: A cohort study. *PLoS Med*. 2020;**17**(9):1–18.
334. Ibsen DB, Chiu YH, Gémes K, Wolk A. Hypothetical 22-Year Intervention With the Dietary Approaches to Stop Hypertension and Risk of Heart Failure in a General Population. *Am J Epidemiol*. 2024;**193**(1):96–106.
335. Koponen J, Rantakokko P, Airaksinen R, Kiviranta H. Determination of selected perfluorinated alkyl acids and persistent organic pollutants from a small volume human serum sample relevant for epidemiological studies. *J Chromatogr A* [Internet]. Elsevier B.V.; 2013;**1309**:48–55. Available from: http://dx.doi.org/10.1016/j.chroma.2013.07.064
336. Norén E, Lindh C, Glynn A, Rylander L, Pineda D, Nielsen C. Temporal trends, 2000–2017, of perfluoroalkyl acid (PFAA) concentrations in serum of Swedish adolescents. *Environ Int*. 2021;**155**(January).
337. Bernert JT, Turner WE, Patterson DG, Needham LL. Calculation of serum 'total lipid' concentrations for the adjustment of persistent organohalogen toxicant measurements in human samples. *Chemosphere*. 2007;**68**(5):824–831.
338. Boulund F. StaG Metagenomic Workflow Collaboration [Internet]. 2023. Available from: https://github.com/ctmrbio/stag-mwc
339. Boulund F. boulund/stag-mwc: StaG-mwc v0.3.0-beta [Internet]. Zenodo; 2018. Available from: https://zenodo.org/records/1483891
340. Manghi P, Blanco-Míguez A, Manara S, et al. MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice. *Cell Rep*. 2023;**42**(5).
341. Gorelick R. Combining richness and abundance into a single diversity index using matrix analogues of Shannon's and Simpson's indices. *Ecography*. 2006;**29**(4):525–530.
342. Rostgaard-Hansen AL, Esberg A, Dicksved J, et al. Temporal gut microbiota variability and association with dietary patterns: From the one-year observational Diet, Cancer, and Health - Next Generations MAX study. *American Journal of Clinical Nutrition*. 2024;**119**(4):1015–1026.
343. Szymańska E, Saccenti E, Smilde AK, Westerhuis JA. Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*. 2012 Jun;**8**:3–16.
344. Krasnovidov A V., Khomonenko AD. Integration of MatLab and R with high-level languages using C# and Microsoft Visual Studio as an example. *J Phys Conf Ser*. 2021;**2131**(2).
345. Escriba-Montagut X, Basagaña X, Vrijheid M, Gonzalez JR. Software Application Profile: ExposomeShiny - A toolbox for exposome data analysis. *Int J Epidemiol*. 2022;**51**(1):18–26.
346. Darzé BC, Lima ICA, Pinto L, Luna AS. Chemometrics web app part 1: Data handling. *Chemometrics and Intelligent Laboratory Systems*. 2022;**231**(July).
347. Bolker BM, Brooks ME, Clark CJ, et al. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol*. 2009;**24**(3):127–135.

348. Dixon P. Models of accuracy in repeated-measures designs. *J Mem Lang* [Internet]. Elsevier Inc.; 2008;**59**(4):447–456. Available from: http://dx.doi.org/10.1016/j.jml.2007.11.004
349. Tay JK, Narasimhan B, Hastie T. Elastic Net Regularization Paths for All Generalized Linear Models. *J Stat Softw*. 2023;**106**:1–24.
350. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform*. 2016;**17**(4):628–641.
351. Steen J, Loeys T, Moerkerke B, Vansteelandt S. Medflex: An R package for flexible mediation analysis using natural effect models. *J Stat Softw*. 2017;**76**(1).
352. Chen M, Chernozhukov V, Fernández-Val I, Melly B. Counterfactual: An R package for counterfactual analysis. *R Journal*. 2017;**9**(1):370–384.
353. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol*. 1989;**18**(1):269–274.
354. Hutcheson R, Innes K, Conway B. Perfluoroalkyl substances and likelihood of stroke in persons with and without diabetes. *Diab Vasc Dis Res*. 2020;**17**(1).
355. Mattsson K, Rignell-Hydbom A, Holmberg S, et al. Levels of perfluoroalkyl substances and risk of coronary heart disease: Findings from a population-based longitudinal study. *Environ Res* [Internet]. Elsevier; 2015;**142**:148–154. Available from: http://dx.doi.org/10.1016/j.envres.2015.06.033
356. Winquist A, Steenland K. Modeled PFOA Exposure and Coronary Artery Disease, Hypertension, and High Cholesterol in Community and Worker Cohorts. *Environ Health Perspect*. 2014;**122**(12):1299–1305.
357. Carrizo D, Chevallier OP, Woodside J V., et al. Untargeted metabolomic analysis of human serum samples associated with exposure levels of Persistent organic pollutants indicate important perturbations in Sphingolipids and Glycerophospholipids levels. *Chemosphere* [Internet]. Elsevier Ltd; 2017;**168**:731–738. Available from: http://dx.doi.org/10.1016/j.chemosphere.2016.11.001
358. Walker DI, Marder ME, Yano Y, et al. Multigenerational metabolic profiling in the Michigan PBB registry. *Environ Res*. 2019;**172**(September 2018):182–193.
359. Kersten S. Peroxisome proliferator activated receptors and lipoprotein metabolism. *PPAR Res*. 2008;**2008**.
360. Salihovic S, Ganna A, Fall T, et al. The metabolic fingerprint of p,p'-DDE and HCB exposure in humans. *Environ Int*. 2016;**88**:60–66.
361. Zheng Y, Li Y, Rimm EB, et al. Dietary phosphatidylcholine and risk of all-cause and cardiovascular-specific mortality among US women and men. *American Journal of Clinical Nutrition*. 2016;**104**(1):173–180.
362. Nordestgaard BG, Benn M, Schnohr P, Tybjærg-Hansen A. Nonfasting triglycerides and risk of myocardial infarction, ischemic heart disease, and death in men and women. *JAMA*. 2007;**298**(3):299–308.
363. Yu GW, Laseter J, Mylander C. Persistent organic pollutants in serum and several different fat compartments in humans. *J Environ Public Health*. 2011;**2011**.
364. Dambrova M, Makrecka-Kuka M, Kuka J, et al. Acylcarnitines: Nomenclature, Biomarkers, Therapeutic Potential, Drug Targets, and Clinical Trials. *Pharmacol Rev*. 2022;**74**(3):506–551.
365. Quinn CL, Wania F. Understanding differences in the body burden-age relationships of bioaccumulating contaminants based on population cross sections versus individuals. *Environ Health Perspect*. 2012;**120**(4):554–559.
366. Hardell E, Carlberg M, Nordström M, Bavel B van. Time trends of persistent organic pollutants in Sweden during 1993-2007 and relation to age, gender, body mass index, breast-feeding and parity. *Science of the Total Environment* [Internet]. Elsevier B.V.;

2010;**408**(20):4412–4419. Available from: http://dx.doi.org/10.1016/j.scitotenv.2010.06.029
367. Fängström B, Hovander L, Bignert A, et al. Concentrations of polybrominated diphenyl ethers, polychlorinated biphenyls, and polychlorobiphenylols in serum from pregnant faroese women and their children 7 years later. *Environ Sci Technol*. 2005;**39**(24):9457–9463.
368. Stråvik M, Hartvigsson O, Noerman S, et al. Biomarker Candidates of Habitual Food Intake in a Swedish Cohort of Pregnant and Lactating Women and Their Infants. *Metabolites*. 2024;**14**(5).
369. Ritacco F V., Graziani EI, Summers MY, et al. Production of novel rapamycin analogs by precursor-directed biosynthesis. *Appl Environ Microbiol*. 2005;**71**(4):1971–1976.
370. König A, Schwecke T, Molnár I, et al. The pipecolate incorporating enzyme for the biosynthesis of the immunosuppressant rapamycin. *Eur J Biochem* [Internet]. 1997;**247**:526–534. Available from: http://hdl.handle.net/10993/12416
371. Zhang N, Wang X, Feng M, et al. Early-life exercise induces immunometabolic epigenetic modification enhancing anti-inflammatory immunity in middle-aged male mice. *Nat Commun*. Springer US; 2024;**15**(1):1–15.
372. Qiang Z. Physiological Adaptability of Yak in Extreme High-Altitude Habitat Using Serum Metabolite Analysis. *Pakistan J Zool* [Internet]. 2023;1–11. Available from: https://dx.doi.org/10.17582/journal.pjz/20230119050142
373. Singer P. Omega-3 Fatty Acids. *MAGGIE B COVINGTON*. 2004;**70**(1):133–139.
374. Chang CL, Deckelbaum RJ. Omega-3 fatty acids: mechanisms underlying "protective effects" in atherosclerosis. *Curr Opin Lipidol*. 2013;**24**(4):345–350.
375. Adkins Y, Kelley DS. Mechanisms underlying the cardioprotective effects of omega-3 polyunsaturated fatty acids. *Journal of Nutritional Biochemistry* [Internet]. Elsevier B.V.; 2010;**21**(9):781–792. Available from: http://dx.doi.org/10.1016/j.jnutbio.2009.12.004
376. Zhang Y, Ma J, Wei F, et al. Polyenylphosphatidylcholine alleviates cardiorenal fibrosis, injury and dysfunction in spontaneously hypertensive rats by regulating Plpp3 signaling. *Front Cardiovasc Med* [Internet]. 2024;**11**(September). Available from: https://doi.org/10.3389/fcvm.2024.1458173
377. Hou J, Xiang J, Li D, Liu X, Pan W. Gut microbial response to host metabolic phenotypes. *Front Nutr*. 2022;**9**(November):1–16.
378. Epskamp S, Borsboom D, Fried EI. Estimating psychological networks and their accuracy: A tutorial paper. *Behav Res Methods*. Behavior Research Methods; 2018;**50**(1):195–212.
379. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2004;**69**(2 2):1–15.
380. Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2006;**74**(1):1–14.
381. Maher BS. Polygenic Scores in Epidemiology: Risk Prediction, Etiology, and Clinical Utility. *Curr Epidemiol Rep*. 2015;**2**(4):239–244.
382. Chung MK, House JS, Akhtari FS, et al. Decoding the exposome : data science methodologies and implications in exposome-wide association studies (ExWASs). 2024;**4**(1).
383. Ballantyne C, Arroll B, Shepherd J. Lipids and CVD management: Towards a global consensus. *Eur Heart J*. 2005;**26**(21):2224–2231.
384. Donat-Vargas C, Akesson A, Tornevi A, et al. Persistent organochlorine pollutants in plasma, blood pressure, and hypertension in a longitudinal study. *Hypertension*. 2018;**71**(6):1258–1268.
385. Wörheide MA, Krumsiek J, Kastenmüller G, Arnold M. Multi-omics integration in biomedical research – A metabolomics-centric review. *Anal Chim Acta*. 2021;**1141**:144–162.

386. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Annals of Applied Statistics*. 2008;**2**(3):841–860.
387. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*. 2005;**21**(13):3001–3008.
388. Clark-Boucher D, Zhou X, Du J, et al. Methods for mediation analysis with highdimensional DNA methylation data: Possible choices and comparisons. *PLoS Genet* [Internet]. 2023;**19**(11 November):1–26. Available from: http://dx.doi.org/10.1371/journal.pgen.1011022
389. VanderWeele T, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiol Methods*. 2013;**2**(1):95–115.
390. Dai JY, Stanford JL, LeBlanc M. A Multiple-Testing Procedure for High-Dimensional Mediation Hypotheses. *J Am Stat Assoc* [Internet]. Taylor & Francis; 2022;**117**(537):198–213. Available from: https://doi.org/10.1080/01621459.2020.1765785
391. Hutchison WJ, Keyes TJ, Crowell HL, et al. The tidyomics ecosystem: enhancing omic data analyses. *Nat Methods* [Internet]. Springer US; 2024;**21**(7):1166–1170. Available from: http://dx.doi.org/10.1038/s41592-024-02299-2
392. Kirik U, Hansson K, Krogh M, et al. Discovery-based protein expression profiling identifies distinct subgroups and pathways in leiomyosarcomas. *Molecular Cancer Research*. 2014;**12**(12):1729–1739.