Resolving Inflammatory Networks in Atherosclerosis Using Proteomics and Bioinformatics

PhD Thesis

Hasman Maria

School of Cardiovascular Medicine & Sciences

British Heart Foundation Centre of Research Excellence

2022

Submitted for the degree of Doctor of Philosophy from

King's College London

Supervisors: Dr Konstantinos Theofilatos & Professor Manuel Mayr

Abstract

Network reconstruction is a crucial component of quantitative omics data analysis, allowing the study of molecular interactions and regulatory associations among genes, transcripts, and proteins. Biological networks have been widely used in atherosclerosis and cardiovascular diseases to identify causal genes and potential biomarkers and to design drug targets.

In this Thesis, we first describe the main computational techniques for reconstructing and analyzing different types of protein networks and summarize the previous applications of such techniques in cardiovascular diseases, particularly atherosclerosis. Existing tools were critically compared, discussing when each method is preferred and presenting examples of reconstructing protein networks of different types (regulatory, co-expression, and protein-protein interaction networks). We demonstrate the necessity to reconstruct networks separately for each tissue type and disease entity, with different cardiovascular diseases serving as examples. We then demonstrate and discuss how the findings of protein networks could be interpreted using single-cell RNA-sequencing data.

Networks of protein interactions in atherosclerosis and relevant phenotype, sex, and vasculature-specific mechanisms, have not been fully elucidated. As existing network analysis techniques typically use the same association threshold for all proteins when inferring positive interactions, they may not be able to identify and incorporate the study of negative associations and interactions among genes and proteins, as well as lacking directionality in the reconstructed networks. In the present Thesis, we also introduce a pipeline of directional regulatory network reconstruction with adaptive partitioning (DiRec-AP), which overcomes existing problems of reconstructing regulatory and co-expression networks. DiRec-AP was benchmarked against three representative network reconstruction methods, using golden standard datasets and networks, and significantly outperformed all of them. Then, it was applied to the reconstruction of atherosclerotic plaque extracellular matrisome networks and phenotype-, sex- and vasculature-specific networks. The reconstructed

atherosclerotic-directed networks can be used to formulate new hypotheses for atherosclerosis mechanisms and to identify potential novel drug targets.

Mass spectrometry-based proteomics offers balanced identification and quantification performance compared to antibody and aptamer-based techniques. However, the complexity of data produced, high missingness, and low reproducibility restrict the application in clinical practice. In the third chapter of this Thesis, we introduce a novel multi-objective optimization framework, HOpTar-omics, based on a Pareto-based Evolutionary Optimization Algorithm, which is designed to streamline and optimize the preprocessing pipeline of three different types of targeted mass spectrometry techniques. The proposed optimization framework outperformed benchmark methods and tools when applied to three large-scale tissue and plasma samples datasets. The application of this pipeline to apolipoprotein measurements in the matrisome of atherosclerotic plaques enabled the largest absolute quantification study of apolipoproteins in the human plaques matrisome and facilitated insight into how apolipoproteins accumulate in human plaques and adjacent tissue.

In the final chapter of this Thesis, using proteomics, we try to reveal novel molecular subtypes of human atherosclerotic lesions, study their associations with histology and imaging and relate them to long-term cardiovascular outcomes. We use discovery and targeted proteomics analysis on plaque samples from 120 patients undergoing carotid endarterectomy, along with a combination of statistical, bioinformatics, and machine learning methods to perform differential expression, network, enrichment analysis, and train and evaluate prognostic models. This proteomics analysis doubles the coverage of the plaque proteome compared to the largest proteomics study on atherosclerosis so far. We identify proteomic signatures for plaque calcification, inflammation and sex, validate them using different types of RNA-sequencing data, and compare them with proteomic signatures of plaque ultrasound and histology measurements. Through proteomics, we define plaque subgroups. A signature of four proteins predicts cardiovascular endpoints with an Area Under the Curve of 75% in the discovery and 65% in the validation cohort, improving the prognostic performance of imaging, and histology. Finally, we develop a prototype of the relational database,

the PlaqueMS database, to store our findings and provide easy access to atherosclerotic plaque proteomics datasets and results.

Acknowledgements

First, I would like to thank all my colleagues in the Vascular Proteomics lab for their collaboration, advice, and valuable conversations throughout my PhD journey. More specifically, thank you Javier Barallobre-Barreiro and Kaloyan Takov for your advice and guidance. Thank you, Marika Fava, Elisa Duregotti and Ferheen Baig for your continuous support, for always helping me with everything I need and for the long conversations where you let me share my thoughts with you. Ferheen Baig and Lukas Schmidt thank you very much for your help with Mass Spectrometry-related issues and for training me to use Proteomics Software, and Xiaoke Yin thank you for answering all sorts of proteomics questions I had. Bhawana, thank you for sharing with me valuable conversations on bioinformatics-related issues. Ferheen Baig, Lukas Schmidt, Elisa Duregotti, Marika Fava, Kaloyan Takov, Siqi Yin, Clemens Gutmann, Isabella Ragone, Sean Burnap, Ella Reed and Panos Vorkas thank you also for the enjoyable times and laughters we had. Thank you, Magda Swiecicka, for always being very helpful with all “organizational” issues.

A special thank you to the Vascular Proteomics group members who performed all proteomics and wet lab-related experiments whose data I used in my thesis since I have not participated in any of the proteomics or/and wet lab experiments: Ferheen Baig, Lukas Schmidt, Sean Burnap, Javier Barallobre-Barreiro and Xiaoke Yin. I would also like to thank Ammar Amjad Shahabuddin for performing the Maxquant analysis.

I would also like to thank all our external colleagues for their contribution to this PhD work. Dr Tsoka and Dr Mavroudi for their help, suggestions, and advice on the directional regulatory networks manuscript I prepared and Dr Stefan Stojkovic and Prof Johann Wojta at the Department of Internal Medicine II, Medical University Vienna, Vienna, AT, for providing the carotid endarterectomy samples and their helpful insights on atherosclerotic plaques biology. I would also like to thank Prof Mayr and Dr Theofilatos for giving me access to the carotid endarterectomy data interpretation and analysis, respectively.

I would like to thank my thesis committee for their time, guidance and helpful feedback on my PhD, Dr Tsoka and Dr Mardakheh. I would also like to thank Dr Alex Ivetic for being a very helpful and supportive PhD coordinator, always there for any questions I might have.

I would also like to thank the BHF organization for funding this project and my studentship.

I would not have been able to finish this PhD without the love and support of my friends and family. Fani, thank you for always being there for me to listen to my fears and support me. Christina and Lina, thank you for sharing with me the ups and downs of our PhDs. Thanasis and Theofilos, thank you for always being there to cheer me up. Daniel, thank you for believing in me. Most importantly, thank you very much mum for always believing in me, for motivating and supporting me and for hearing me out and loving me unconditionally. Without you, I would not have been able to be there doing a PhD.

Finally, I would like to express my gratitude to my supervisors, Prof Mayr and Dr Theofilatos. Prof Mayr, I am sincerely grateful for giving me the chance to do a PhD and work in a group full of talented scientists and participate in cutting-edge research. I have learned very much next to you and gained valuable experience. Dr Theofilatos, I am sincerely grateful for your continuous help and support throughout all these years. You've been a real mentor for me.

For my grandparents, A. Mitropoulos and M. Mitropoulou

Table Of Contents

List of Figures

List of Tables

Abbreviations

A1BG	Alpha 1-B glycoprotein
ABI3BP	Target of Nesh-SH3
ACTA	aortic smooth muscle cell actin
AEBP1	Adipocyte enhancer-binding protein 1
AHA	American Heart Association
ALB	Albumin
APID	Agile Protein Interaction DataAnalyzer
APOA/LPA	apolipoprotein(a)
APOA1	apolipoprotein A1
APOA4	apolipoprotein A-IV
APOB	apolipoprotein B
APOC1	Apolipoprotein C-I
APOC3	apolipoprotein C-III
APOJ	apolipoprotein J
ARACNe-AP	Algorithm for the Reconstruction of Accurate Cellular Networks with adaptive partitioning
ASCOT	Anglo-Scandinavian Cardiac Outcomes Trial
AUPR	area under the precision-recall
AUROC	area under the ROC curve
BC	Betweenness Centrality
BMI	body mass index
BNW	Bayesian Network Webserver
C163A	scavenger receptor cysteine-rich type 1 protein M130
CALD1	Caldesmon
CD14	monocyte differentiation antigen CD14
CERU	Ceruloplasmin
CF	Complement Component C7
CFHR1	Complement Factor H Related 1
CHF/CFAH	Complement Factor H
CLU	Clusterin
CMI2	Conditional Mutual Inclusive Information
CMPK1	UMP-CMD kinase
CNN1	calponin-1
COPD	chronic obstructive pulmonary disease
COVID-19	coronavirus disease
CPXM2	Carboxypeptidase X, M14 Family Member 2
CPXM2	Inactive carboxypeptidase-like protein X2
CRP	C-reactive protein
CSPG2	Versican
CTHRC1	Collagen Triple Helix Repeat Containing 1
CTSD	Cathepsin D
CVD	cardiovascular disease
DEF1	neutrophil defensin 1
DERM	dermatopontin

DIA	Data Independent Acquisition
DiRec-AP	directional regulatory network reconstruction with adaptive partitioning
DM	type 2 diabetes mellitus
DPT	Dermatopontin
EC	Endothelial cells
ECM	extracellular matrix
EMC	Enhanced Markov clustering
F13A1	Coagulation factor XIII A chain
F2	Coagulation Factor II / Thrombin
FB	Fibroblast cells
FETUA	fetuin-A
FN1	Fibronectin
FRIH	ferritin heavy chain
FTL	ferritin light chain
GAS6	growth-arrest specific 6
GENA	Gradually expanding neighborhoods with adjustment
GENIE3	GEne Network Inference with Ensemble of trees
GRN	gene regulatory network
GSN	Gelsolin
GuHCl	guanidine extract
HDL	High-Density Lipoprotein
HipMCL	High-performance MCL
HOpTar-omics	Heuristically Optimized Targeted Proteomics
HSPG2	Basement membrane-specific heparan sulfate proteoglycan core protein
HTN	Hypertension
HTRA1	Serine protease HTRA1
IGHA1	Immunoglobulin Heavy Constant Alpha 1
IGHG3	Immunoglobulin Heavy Constant Gamma 3
IGHM	Immunoglobulin heavy constant mu
IGKC	Immunoglobulin Kappa Constant
IID	Integrated Interactions Database
iRT	indexed Retention Time
ITGB1	Integrin beta-1
L/H	Light to Heavy peptide ratio
LDL	low-density lipoprotein
LEG3	galectin-3
LGALS3BP	galectin-3-binding protein
LRP1	Prolow-density lipoprotein receptor-related protein 1
LTBP1	Latent-transforming growth factor beta-binding protein 1
LTBP1	Latent-transforming growth factor beta-binding protein 1
MANOVA	multiple analyses of variant
MCL	Markov Clustering
mg/L	milligrams per litre
MGP	matrix Gla protein
MI	mutual information
MIIC	Multivariate Information-based Inductive Causation

MMP	matrix metalloproteinase
MMP12	macrophage metalloelastase
MMP9	matrix metalloproteinase 9
Mo	Monocytes
MOEA	multi-objective evolutionary algorithm
MRC1	macrophage mannose receptor 1
MRM	Multiple Reaction Monitoring
MS	mass spectrometry
NaCl	salt extract
NID2	nidogen 2
NK	Natural Killer cells
OSTP / OPN	Osteopontin
oxLDL	oxidized LDL
PERM	myeloperoxidase
PGCA	Aggrecan
PICKLE	Protein InteraCtion KnowLegdE
PODN	Podocan
POSTN	Periostin
PPI	protein-protein interaction
PRDX1	Peroxiredoxin 1
PRELP	Proline and Arginine Rich End Leucine Rich Repeat Protein
PRM	Parallel Reaction Monitoring
PROC	vitamin K-dependent protein C
PROS1	Protein S
PTN	Pleiotropin
PTPRC / CD45	receptor-type tyrosine-protein phosphatase C
RNSC	Restricted Neighborhood Search Clustering Algorithm
ROS	reactive oxygen species
S/N	signal-to-noise ratio
scRNA-seq	Single-cell RNA-sequencing
SDF1	stromal cell-derived factor 1
SDS	sodium dodecyl sulfate
SEC	Sparse Estimation of the Correlation matrix
SERPH	serpin H1
SERPINA1	serpin family A member 1
SERPINA1	Alpha-1-antitrypsin
SERPINF2	Alpha-2-antiplasmin
SERPING1	Serpin Family G Member 1
SFRP3	Secreted frizzled-related protein 3
SMC	smooth muscle cells
SRM	Selected Reaction Monitoring
SUMO3	small ubiquitin-related modifier 3
SVM	Support Vector Machines
TAGL	Transgelin
TF	Tissue Factor
THBS1	Thrombospondin-1
THRB	Prothrombin
TIA	transient ischemic attack

TII	Total Ion Intensity
TIMP1	metalloproteinase inhibitor 1
TMT	tandem mass tags
TNC	Tenascin
TTR	Transthyretin
UMAP	Uniform Manifold Approximation
usCRP	ultra-sensitive C-reactive protein
VCL	Vinculin
VWA1	von Willebrand Factor A domain-containing 1
VWF	von Willebrand factor
WGCNA	Weighted gene correlation network analysis

1. Introduction

1.1 Introduction to Atherosclerosis

Atherosclerotic cardiovascular disease (CVD) is the leading cause of vascular disease worldwide. Atherosclerosis is a chronic, multifactorial disease that develops in the intima of the arteries due to lipid deposition. A normal artery consists of the intima (inner layer with endothelial cells and some smooth muscle cells), the media (middle layer consisting of smooth muscle cells), and the adventitia (outer layer with mast cells, capillaries, and nerve endings) (1, 2). After an endothelial injury, the cells lose their proper functionality resulting in lipoprotein accumulation to the arterial wall (especially low-density lipoprotein or LDL, which carries cholesterol through the blood), which binds to proteoglycans in the extracellular matrix (ECM) (3) and get oxidized by reactive oxygen species (ROS) (4). Oxidized LDL (oxLDL) can promote atherogenesis (5). As a response to oxLDL, circulating leukocytes infiltrate the subendothelial space, initiating a chronic inflammatory process. Monocytes differentiate into macrophages, which in turn phagocytose oxLDL and turn into foam cells, with this mechanism being the hallmark of early atherosclerosis. This initial stage of atherosclerosis is known as the “fatty streak” (6). Research has shown that apart from differentiated macrophages, smooth muscle cell (SMC) metaplasia can lead to foam cell formation (7). Foam cells, mast cells, and T-lymphocytes secrete cytokines, which, along with ROS, activate SMC migration to the intima and stimulate the secretion of ECM proteins. As a result, a fibrous cap is created on top of the lipid-filled core of the atherosclerotic plaque (1, 7), which is developed by the programmed cell death of macrophages and SMCs (8). Inefficient clearance of dead cells can also contribute to the formation of the core (9). Moreover, LDL particles aggregated with proteoglycans can enter SMCs, which accumulate cholesterol (10). Lipids can be contained in these SMCs and macrophages, contributing thus to atherosclerosis progression. Accumulation of foam cells initiates inflammatory processes that drive the progression of the disease and can cause plaque rupture, which can in turn cause myocardial infarction and stroke (6). Plaque rupture leads to the exposure of the contents of the plaques to the blood. This material can lead to thrombosis, the biggest complication of atherosclerosis (8). Atherosclerotic plaques have been shown to also

contain T lymphocytes, with a contradictory role in the diseases. Some T-cell subtypes such as T regulatory cells moderate atherosclerosis, whereas other subtypes such as type 1 T helper cells can further promote it (11). A schematic overview of atherosclerosis progression is shown in Figure 1.1.

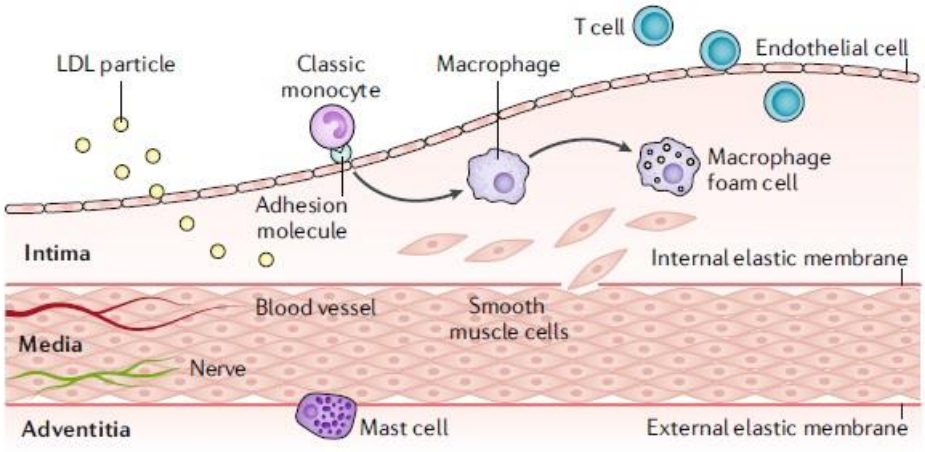


Figure 1.1 Atherosclerosis initiation and progression. Figure adapted from Libby, P. et al. 2019 (8).

Atherosclerosis is progressing slowly, so most cases remain asymptomatic for a long time. The clinical manifestation of the disease depends on the characteristics of the vascular lesion it is occurring. When affecting the peripheral arteries, atherosclerosis can cause gangrene that can affect limb function. Atherosclerosis affecting heart circulation can lead to ischaemic strokes, aneurysms, or acute coronary syndromes like myocardial infarction. Heart disease due to atherosclerosis and strokes are the number one cause of death worldwide. Even patients that survive acute coronary syndromes can be left with defective cardiac function which can lead to heart failure. Advanced plaques can also obstruct blood flow and cause tissue ischaemia (8).

1.1.1 Role of inflammation in atherosclerosis

There are several risk factors for atherosclerosis, with the most important being age, male sex, smoking, hypercholesterolemia, hypertension, vascular stiffening, and remodelling, as well as coagulation and diabetes (12, 13). The exact mechanisms of action of these risk factors and their processes for activating atherogenesis are not fully understood, but many of these activate inflammatory pathways (8).

Inflammation in the vascular wall is the response of our immune system to oxLDL. Inflammation can alter the function of cells of the arterial wall. Inflammatory cells can be found in abundance in atherosclerotic lesions but not in a normal vessel wall. The inflammatory cells contribute to all stages of the disease, as well as to the progression of the atherosclerotic plaque and its rupture due to the release of ECM-degrading proteases (14, 15). Moreover, the inflammasome (the protein complexes governing the inflammatory response of the organism) contributes to the initial steps of inflammatory diseases and has been linked to atherosclerosis (13). Inflammation has also been linked to aging. “Inflammaging”, a mild form of chronic inflammation, is observed in older people and is known to be one of the most important causes of age-related diseases, such as atherosclerosis (16). Additionally, inflammation is linked to vascular calcification, which is another hallmark of late-stage atherosclerosis. Inflammation activates the initiation of calcification so inflammatory signals trigger the calcification response in atherosclerosis (16). Finally, inflammation is related to thrombosis as it leads to a procoagulant state and predisposes to blood clot formation (14, 15). Inflammation is one of the key features that distinguishes unstable from stable plaques. Many prior studies have attempted to identify the involved proteins of inflammation in blood vessels and plaques with the work of Langley et al. (17) revealing a highly connected in the PPI network neutrophil-related biosignature, which can be detected in both blood and carotid endarterectomies and is predictive of outcomes and atherosclerosis stage.

1.1.2 Role of calcification in atherosclerosis

Intimal vascular calcification is another feature that has been linked to atherosclerosis. Vascular calcification is associated with a higher cardiovascular risk, as it reflects the level of ongoing inflammation in the vascular tree (18). Vascular calcification shares some mechanisms with bone formation, but its precise molecular mechanisms are not clear (19). Calcification is thought to occur due to an imbalance of proteins that promote or inhibit mineralization in the vascular wall, leading to arterial stiffening (20). The role of calcification in atherosclerotic plaques depends on the type of calcification. It has been suggested that microscopic or spotty calcification, “microcalcification”, a nodular and spotty form of calcification, is related to plaque

instability, with a high risk for rupture and thrombosis (21, 22). On the other hand, “macrocalcification”, a widespread form of calcification mostly in the necrotic core, has a protective role in plaques. Macrocalcification is thought to reflect a “healed” plaque, in which the inflammatory response has subsided (23). It has also been shown that the incidence of preoperative neurological symptoms was similar in patients with calcified and non-calcified plaques (24), therefore calcified lesions cannot be considered necessarily as stable. Despite the importance of calcification in plaque stability and outcomes, the driver mechanisms have not been revealed and the reconstruction of directed protein networks can contribute significantly to this field.

1.1.3 Role of vascular ECM in atherosclerosis

The ECM is a three-dimensional structure, present in all tissues, but different for each organ. It is composed of proteins such as proteoglycans, collagens, elastin, and glycoproteins (25). Many of these ECM proteins can trap lipoproteins and promote the accumulation of lipids in the intima. ECM remodelling in atherosclerosis plays an important role in plaque destabilization and progression, as it triggers the migration of SMCs into the intima (26). ECM proteins control ECM remodelling via signalling with matrix metalloproteinases (MMPs), which degrade the constituents of ECM. Production of MMPs can develop an inflammatory response and contribute to different stages of atherosclerosis (27). Inflammatory processes can prevent the synthesis of collagens by SMCs in the plaques, destroying their ability to maintain the shape of fibrous cap (28). As mentioned above, the secretion of ECM proteins from vascular SMCs in the intima covers the atherosclerotic lesions with a fibrous cap, which encloses a lipid-rich necrotic core. Proteoglycans that are secreted into the intima can further promote the retention of LDL and therefore the development of atherosclerosis (29).

1.1.4 Atherosclerotic plaque classification and treatment

According to the atherosclerotic plaque histology (AHA classification), atherosclerotic lesions are classified into eight different types: intimal thickening, fibroatheroma, late fibroatheroma, healed plaque rupture, fibrocalcific plaques, erosions, thin-capped fibroatheroma, and ruptured plaques (16, 30). According to ultrasound

measurements, atherosclerotic plaques are classified as echolucent, mixed and echogenic plaques. Echolucent plaques are soft, unstable plaques with a thin fibrous cap, a necrotic lipid core and continuous inflammation. These plaques are more prone to rupture and can cause clinical overt diseases such as heart attacks or strokes due to thromboembolism. Echogenic are calcified plaques with less likelihood of clinical symptoms, possibly due to reduced inflammation and increased ECM deposition (31).

The diagnosis of atherosclerosis mostly depends on either invasive (e.g. intravascular ultrasonography mostly for coronary arteries) or non-invasive (e.g. CT angiography for almost all vascular territories, MRI mostly for carotid artery and aorta, carotid ultrasonography) imaging techniques (8). Carotid plaque vulnerability can also be assessed by ultrasound center frequency shifts, a non-invasive ultrasound method (32).

Apart from the alterations in the lifestyle that an individual can do to avoid developing risk factors that induce atherogenesis, such as smoking or a high-fat diet, there are additional medications that help in the treatment of atherosclerosis. Such medications are lipid-lowering drugs, such as statins, which target LDL cholesterol, non-statin lipid-lowering drugs, anti-platelet drugs, which are though avoided due to the high risk of bleeding, and anti-inflammatory drugs such as colchicine (8). Despite the use of existing medications, such as statins, atherosclerosis until now remains partially untreated. Thus, the use of other approaches to identify new therapeutic targets is of high importance.

1.2 Proteomics methods to study atherosclerosis-related protein changes

Since atherosclerosis is a multifactorial disease, a systems-biology approach including measurements from multiple platforms would facilitate its research. In parallel to other omics approaches, such as transcriptomics, proteomics is a useful tool for the analysis of proteins of interest and the extraction of useful conclusions (33). Mass spectrometry (MS) based proteomics is widely used as a discovery technique to identify and quantify proteins in biosamples. Discovery proteomics is non-biased since it does not require prior protein information, can quantify a wide range of high-

abundance peptides, can derive useful conclusions about protein expression in different tissues, and facilitate the discovery of biomarkers and possible drug targets (33). MS proteomics though is prone to errors, shows a high false discovery rate, favours high-abundant proteins and does not fully cover the proteome (34, 35). These limitations can be faced with the depletion of high-abundant proteins and the use of other methods, such as targeted proteomics, which only analyze proteins of interest, or proximity extension assays, such as Olink antibody-based commercial platform, which combines antibody techniques with DNA amplification (33).

1.2.1 Mass Spectrometry

A mass spectrometer is a tool that measures the mass-to-charge ratio of one or more molecules in a biosample. It generates mass spectra from peptide fragments and detects the ions to measure molecule abundance. A typical liquid chromatography MS method (LC-MS/MS) requires the digestion of proteins into peptides usually by trypsin (Figure 1.3), the peptide separation in the liquid chromatography column, the peptide ionization and the detection of fragment ions (mass spectrum) (34). The spectra generated are searched against databases and identify the existing proteins of a biosample (Figure 1.2) (33).

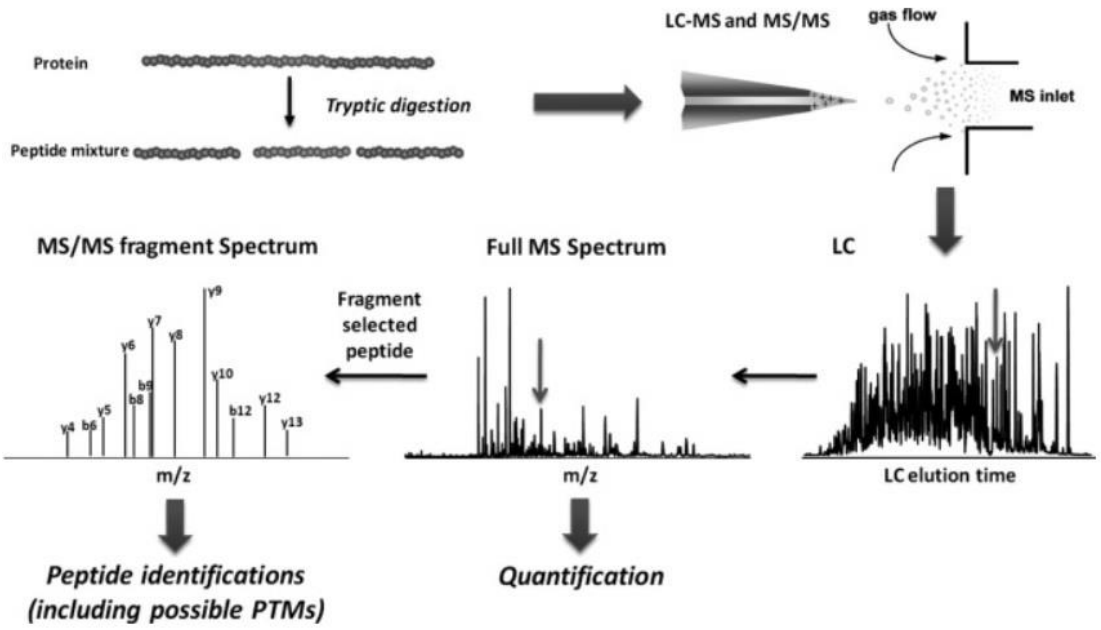


Figure 1.2 Classic MS/MS workflow. Figure adapted from Xie et al. 2011 (36).

In a non-typical MS workflow, enzymatic digestion with trypsin can be limited to only one terminal (semi-tryptic search) or not conducted at all (non-tryptic search). In this way, different peptides would be identified, and this alternative identification could lead to the identification of novel peptides and proteins (37). What is more, non-specific searches can lead to the identification of more peptides for an already identified protein and thus improve protein quantification allowing also to study fragmentation events. However, these types of searches increase the search space and can be time-consuming, especially in large-scale proteomics from complex biosamples, while the increase of the search space can also lead to fewer identified peptides when maintaining a constant false discovery rate (38).

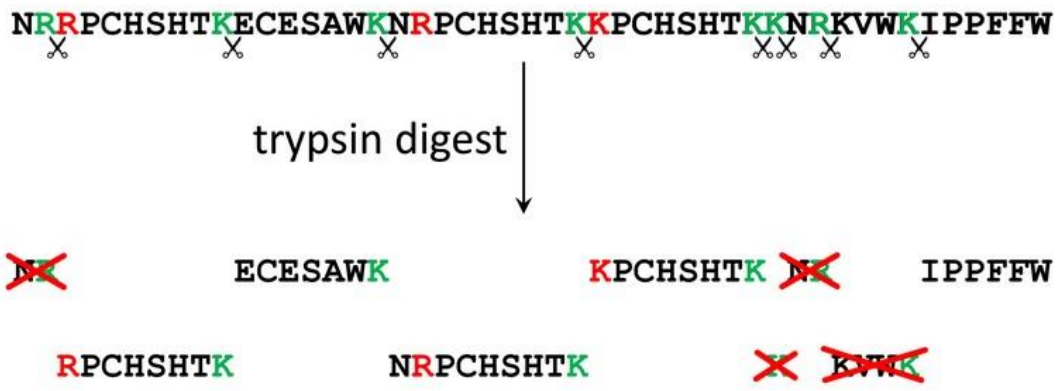


Figure 1.3 Trypsin digestion. Trypsin cleaves the protein at the C-terminal, in the lysine (K) or arginine (R) amino acids. Figure adapted from (39).

1.2.2 Discovery Mass Spectrometry

Discovery or untargeted proteomics are used to identify and quantify proteins, without a prior hypothesis. Relative quantification of proteins using discovery MS is either label-free or labelling techniques. The latter offer higher accuracy and precision in quantification, whereas the label-free approaches have broader proteome coverage (36). Label-free quantification can be based on MS ion intensity or peak area of the peptides and this approach is used in a typical LC-MS/MS workflow (36). With this approach, only the top abundant peptides are selected for fragmentation (usually top 15 or 20 peptides), a stochastic process that can lead to high missingness, and the confidence of the identification depends on the accuracy of the mass instrument (40). The use of isobaric labels, such as tandem mass tags (TMT), results in more accurate

quantification compared to label-free MS, as they produce a higher signal-to-noise ratio and are more robust (41). TMT are the most used chemical isobaric tags for MS quantification and can analyze two to twenty samples simultaneously.

A more recent approach is the Data Independent Acquisition (DIA), which analyses a bigger number of proteins as it isolates and fragments together all peptide ions within a defined mass-to-charge window (42). It combines the wide protein coverage of data-dependent acquisition and the accuracy, reproducibility, and consistency of targeted analysis (43). Its major limitation is the computational complexity of the fragmented spectra acquired and, thus, the complex data analysis.

1.2.3 Targeted Mass Spectrometry

Targeted proteomics is a hypothesis-driven MS acquisition method, which focuses on the quantification of proteins of interest. This method can identify low-abundant proteins efficiently compared to the discovery proteomics (33). The basic targeted proteomics techniques are Selected Reaction Monitoring (SRM) or Multiple Reaction Monitoring (MRM) and Parallel Reaction Monitoring (PRM). In SRM, a specific peptide is selected and fragmented, and specific fragment ions (typically three product ions) are selected for detection. In PRM the first two steps remain the same, but in the last step, all fragment ions are analyzed using a high-resolution accurate mass instrument (Figure 1.4) (42).

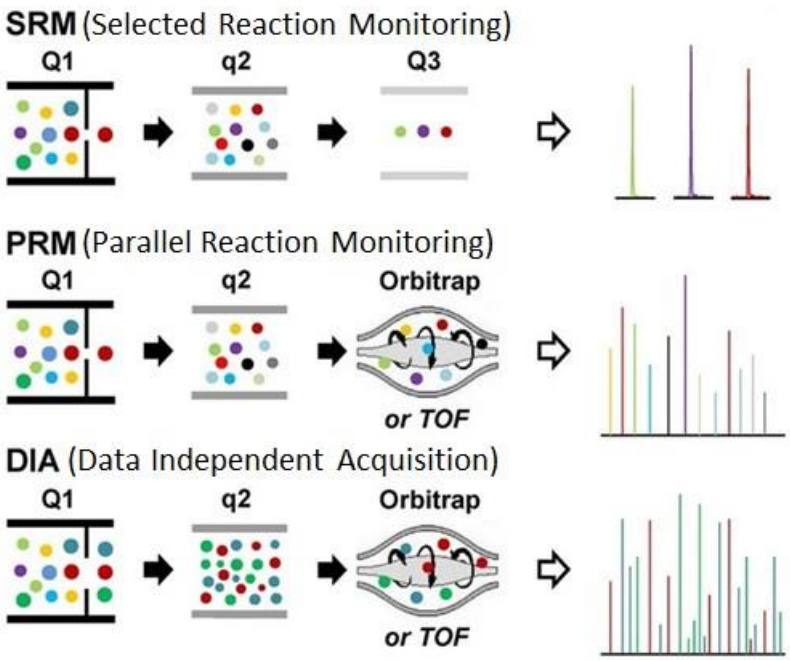


Figure 1.4 Targeted MS proteomics methods. Figure adapted from Shi et al. 2016 (42).

Targeted proteomics has been shown to overcome further limitations of discovery proteomics, such as reproducibility issues or false identifications (36). The specificity of detection in targeted proteomics can be further improved with the use of labelled internal (heavy) standards (36), which offer accurate quantification, and the use of multiple proteotypic peptides across the whole amino acid sequence of a protein. Their major disadvantages are the complex method development (as in PRM the accurate precursor masses of the target peptides need to be optimized while developing the method and in SRM both the peptides and the fragment ions / specified transitions must be optimized) and the analysis time, as they have the disadvantage of producing a complex output, which requires elaborate signal pre-processing and quality control analysis to achieve accurate quantification values.

1.3 Networks in biology

Networks are computational models that have been widely used in biology to model molecule interactions in systems biology (44). Biological networks are scale-free networks, whose nodes can be genes, proteins, or other biomolecules. Their edges represent the relationship of these molecules and can be either directed (regulatory networks) or undirected (co-expression networks). These networks are sparse and contain few highly connected hubs, central for each network (33).

The basic methods for reconstructing biological networks are based on correlation metrics, mutual information metrics and mathematical modelling. Correlation-based methods cannot model non-linear associations. Thus, information theory-based techniques have been introduced to address this issue. The problem of reconstructing co-expression networks is a mathematical optimization problem, and this is the reason why mathematical and probabilistic modelling techniques, such as the Bayesian ones, have been widely used. However, these techniques suffer from a high number of assumptions and do not apply to large-scale networks.

Most existing network approaches reconstruct static networks, whereas cardiovascular diseases are dynamic processes. As mentioned above, some of the most important mechanisms involved in atherosclerosis are inflammation, calcification, vascular ageing, and degradation of the extracellular matrix, which are interconnected. Network approaches can contribute to resolving these mechanisms and their synergistic function, which leads to the development and progression of atherosclerosis. Despite the use of existing medications, such as statins, atherosclerosis until now remains partially untreated. Thus, the use of network approaches to identify new therapeutic targets is of high importance. In chapter 2 of this Thesis, we will thoroughly describe and compare the existing methods for network reconstruction, their application to cardiovascular diseases and especially atherosclerosis so far, as well as provide some examples that highlight the need to reconstruct networks separately for each tissue type and disease entity.

1.4 Hypothesis

Atherosclerotic CVD is the leading cause of vascular disease worldwide, but the role of vascular inflammation and its association with calcification, sex and other risk factors, and with the histological and imaging characterization of the plaques, has not been fully elucidated. Expanding our understanding of vascular inflammation is key to allow us to move a step closer towards developing novel therapeutic targets, with the potential to inhibit vascular rather than systemic inflammation. Also, novel biomarkers are required to stratify cardiovascular patients who are likely to benefit from anti-inflammatory and other atherosclerosis-related therapies.

Atherosclerosis and its inflammation-related mechanisms are multifactorial conditions affected by multiple types of cellular and extracellular mechanisms as well as demographics, lifestyle and medications of the patients. MS proteomics technologies can quantify the abundancies of thousands of proteins in parallel, but their data analysis processes are not optimal, while to promote understanding of their measurements and use them to state and confirm molecular biology hypotheses, proper network reconstruction and analysis techniques are required.

In the present Thesis, we hypothesized that developing optimized network reconstruction and analysis tools, methods for optimizing the processing of targeted MS datasets and applying them to process the largest known collection of proteomics data of atherosclerotic plaques, will allow for the identification of novel proteomic biosignatures of inflammation, associate them with other phenotypes, develop new prognostic models for atherosclerosis and identify potential therapeutic targets for specific patient groups.

1.5 Aims

The aims of this study are as follows:

1. **Novel proteomic network reconstruction method**
    a. Design and develop a new method for reconstructing networks from proteomic datasets overcoming the limitations of existing techniques, such as the lack of directionality in networks, limited accuracy, and biased selection of thresholds.
    b. Evaluate the newly introduced network reconstruction method by comparing it with state-of-the-art techniques on golden standard datasets.

2. **Novel method and tool for the optimization of targeted and DIA MS data.**
    a. Design and implement a new method for optimizing the processing of targeted and DIA MS datasets using external proteomic measurements (clinical, discovery or other MS protein measurements).
    b. Benchmark the designed technique on multiple tissue and blood

cardiovascular research-related datasets against commercial and open-source software solutions.

3. **Apply the novel data analysis techniques to multiple proteomics datasets to resolve inflammation and other molecular mechanisms of atherosclerosis.**
    a. Combine multiple proteomics techniques to characterize the proteome of atherosclerotic plaques.
    b. Identify and validate the inflammation biosignatures of carotid plaques and associate them with calcification, sex, ultrasound and histology signatures.
    c. Reconstruct cellular and extracellular protein networks of atherosclerotic plaques and identify clusters related to inflammatory cells and mechanisms. Explain these clusters using bulk, single-cell and spatial RNA-sequencing data.
    d. Identify prognostic proteomic signatures to predict outcomes of atherosclerosis patients and group patients into clinically meaningful clusters.
    e. Identify potential therapeutic targets by reconstructing and comparing networks in different phenotypes of atherosclerosis (eg. Symptomatic vs Asymptomatic plaques and plaques from male and female patients).

2. Uncovering protein networks in cardiovascular diseases and atherosclerosis using proteomics

2.1 Introduction

Cardiovascular diseases (CVDs), the first cause of morbidity and mortality worldwide, are disorders that affect the heart, vasculature and blood (45). CVDs include coronary artery disease, cardiomyopathy, heart failure, myocardial infarction, stroke, and atherosclerosis (46). The pathogenesis of cardiovascular diseases is multifactorial. Some of the most important risk factors are hypertension, hypercholesterolemia, diabetes, central obesity, lack of physical activity (47), and smoking (48). Different kinds of measures to control the development and the progression of the risk factors have been developed, such as lipid-lowering drugs (eg statins), beta-blockers, ACE inhibitors, (49) antihypertensive treatments, insulin-sensitizing agents, or powerful antismoking campaigns, and have contributed to the reduction of incidences of cardiovascular diseases (50).

In this chapter, emphasis is given to atherosclerosis, which is a chronic and progressive inflammatory disease of the large and medium-sized arteries, and its association with other cardiovascular diseases. Atherosclerosis develops in the intima of the arteries due to lipid deposition and results in the formation of plaques in arteries (17, 51). Atherosclerosis is affected by sex and its progression depends on age (16, 18).

Biological networks have been widely used in many different diseases to identify potential biomarkers, causal genes and drug targets (52). The basic types of protein networks are the experimentally or in silico reconstructed protein-protein interaction (PPI) networks and the functional networks, which could show similar protein coregulation, expression or function. The latter can be split into regulatory and co-expression networks, according to whether their edges are directed or not.

Networks can reveal useful biological and molecular information by inspecting two different and complementary types of network properties, the topological and the functional ones (53). Topological characteristics are used to represent the structural

features of the network and are associated with biological properties and certain parameters, such as the betweenness centrality, which is used to reveal critical nodes. The functional approach clusters the nodes based on their functional information, such as co-localization (cell compartments and molecular functions) (54). Basic network analysis steps involve network clustering and identification of modules, which could represent biological processes found by enrichment analysis or be related to specific disease phenotypes found by differential expression analysis, identification of hub genes, which are more relevant to the functionality of the network compared to other nodes and are the most likely nodes where the network could be perturbed in pathophysiological conditions, and comparison between modules of a network or different networks (55, 56).

Recently, a systems-level approach, which involves focusing on a group of genes or proteins rather than on individual molecules, has been introduced to unravel mechanisms of complex diseases, which involve groups of genes or proteins (57). In this context, network analysis has facilitated atherosclerosis research importantly. Xi et al. (58) identified 48 hub genes for atherosclerosis using network analysis, Koplev et al. (59) used gene regulatory networks and gene modules to identify causal mechanisms for cardiovascular and cardiometabolic disease phenotypes and Banik et al. (60) used network analysis to identify causal genes and miRNAs that regulate them for the pathogenesis of atherosclerosis. Herrington et al. (61) used proteomic network analyses to identify biomarkers for subclinical atherosclerosis, but their work was based on samples from biopsies, reducing the ability to conclude because of unknown effects of protein degradation. In a recent article (62), Bandaru et al. used a protein-protein interaction network approach and identified filamin A as a novel target that can reduce the activity of macrophages and therefore be used as a treatment for atherosclerosis. Moreover, Abe et al. (63) deployed the IPA tool to confirm the mechanism of action of the membrane-associated guanylate kinase with inverted domain structure-1 (MAGI1), which was associated with endothelial activation and ER stress and recognized as a driver of atherosclerosis.

Several reviews are describing the different types of biological networks and analyses. Hu et al. (64) outlined different computational methods for identifying PPI networks,

Vella et al. (56) described PPI and co-expression network reconstruction and analysis methods as well as studies involving the use of proteomics co-expression networks, Liu et al. (65) described different ways to identify critical nodes and Meng et al. (66) thoroughly presented the different topological properties of PPI networks. Furthermore, several reviews (46, 55, 67–69) have described the application of network analysis in several diseases as well as systems biology approaches, which include proteomics integration with other -omics technologies, network analysis and their application in cardiovascular diseases (17, 33) but, to our knowledge, none of them has focused on the different network types, the technical aspects of the network reconstruction methods and their applications on atherosclerosis.

In this chapter, we discuss the role and types of protein networks, the different network analysis techniques and tools, and focus on their application to studying the underlined mechanisms of cardiovascular diseases and atherosclerosis, discussing the effects of medications in the networks and the applications of network analysis to identify common and disease-specific mechanisms for various cardiovascular diseases.

2.2 Main Body

2.2.1 Existing methods for protein networks reconstruction and analysis

The basic steps of network analysis involve the reconstruction of static PPI networks, undirected protein co-expression networks and directed protein regulatory networks, their clustering to identify significant modules and their analysis and visualization to reveal key hub proteins that can serve as diagnostic, prognostic, or therapeutic biomarkers.

However, each one of these analysis steps can be implemented with different tools and methods, with each one of them having specific limitations and strengths. **Error! Reference source not found.** presents the basic categories for protein network reconstruction and analysis, some indicative tools, and their advantages and disadvantages.

Analysis Type	**Subcategory**	**Indicative Methods**	**Advantages**	**Disadvantages**	**Implementation**
Protein-Protein Interaction Network Reconstruction	Computationally and experimentally verified	STRING (70), Metascape (71)	Interactions for many species Interactions from many sources User-friendly network visualization and analysis	High false positive rate Not full coverage of interactomes	Webtool / Standalone (R, Bioconductor Package)
Protein co-expression networks	Correlation-based	WGCNA (72)	Sparse Networks Optimized Threshold	One threshold for all nodes Linear Associations Only Undirected Networks	Standalone (R / Python)
	Mutual Information Based	MIDER (73)	Distinguishes between direct and indirect links	Undirected Networks	Standalone (Matlab)
	Probabilistic	SEC (74)	Creates sparse matrices	The threshold needs to be decided with trial and error Undirected Networks	Standalone (Matlab)
Protein Regulatory Networks	Mutual Information Based	ARACNe-AP (75)	Sparse Networks Directed Networks Non-linear associations Removes indirect links	Cannot discriminate between positive and negative associations Requires a given set of transcription factors	Standalone (JAVA)
	Probabilistic	BNW (76)	Handles noise and uncertainty Directed Networks	Does not support large networks Feedback loops are not allowed	Webtool
	Machine-learning based	GENIE3 (77)	Directed Networks	Requires TF information	Standalone (Python / Matlab / R)
Network Clustering	Hard-clustering	HipMCL (78)	Fast Clustering in Large Networks Supports edge-weighted graphs clustering	Does not allow unclustered nodes Does not allow overlapping clusters	Standalone (C++)
	Soft-clustering	ClusterOne (79)	Allows overlapping clusters	Lower Interpretability	Standalone (JAVA) / Cytoscape plug-in

			Allows for unclustered nodes Does not allow small clusters		

Table 2.1 Existing methods for the reconstruction and analysis of proteomic networks, indicative tools, and their implementations

2.2.1.1 Physical protein-protein interaction networks

Protein-protein interactions can be divided into physical (direct) interactions and functional (indirect) interactions and both types of networks can be either reconstructed with experimental methods or predicted using in silico machine learning, mathematical modelling or other methods (53, 67). There are various types of databases for protein-protein interaction information, including primary databases such as IntAct (80) and BioGRID (81), which include only experimentally validated curated interactions, secondary databases or meta-databases, such as iRefIndex (82) and Protein InteraCtion KnowLegdE (PICKLE) (83). The latter group of databases combines and integrates information from primary databases, databases which include computational predictions such as STRING (70), Integrated Interactions Database (IID) (84) and Metascape (71), or subject-specific databases such as VirHostNet (85) which explores virus-host protein-protein interactions.

iRefIndex (82) is one of the most widely used meta-databases. It contains non-redundant binary and complex participation interaction data from 10 databases including BIND, BioGRID, CORUM, DIP, HPRD, IntAct, MINT, MPact, MPPI and OPHID, more than 1000 organisms and also provides a web interface, iRefWeb (86). Agile Protein Interaction DataAnalyzer (APID (87)) contains experimentally verified physical protein interactions for more than 400 organisms, mined from 5 primary databases, including BioGRID, DIP, HPRD, IntAct and MINT, and from experimentally resolved 3D structures from PDB when more than two distinct proteins have been identified. APID groups interaction detection methods in binary (information coming from techniques that detect direct interactions between protein pairs) and indirect (information coming from co-complex methods, identifying interaction in protein complexes) and curates interactions by removing duplicate and incomplete records. APID also includes a network visualization web tool which displays the produced network with color mapping of the nodes according to functional enrichment and the interactions according to selected properties such as the reliability of the edges. PICKLE 3.0 (83) consists of interactions between humans and mice, inferred from 4 primary databases, including IntAct, BioGRID, HPRD and DIP. PICKLE uses the complete human proteome

from UniProtKB as a reference, creates heterogeneous networks containing edges from genes, mRNAs and proteins and integrates the protein-protein interaction networks at either protein or gene level, filtering interactions based on the probability to be direct (88). PICKLE 3.0 enables the extension of mouse interactions through homology and the comparison between mouse and human networks.

Apart from the experimentally curated, protein-protein interactions can also be computationally predicted in many different ways such as protein sequences (89) or protein neighborhoods (90). Along the databases that include computationally predicted links, STRING (70) is the most popular one. It contains more than 20 billion interactions for more than 14000 organisms, collecting information from primary databases, text mining, computational predictions from co-expression, protein homology and more showing the different interaction resources to the network via different colored edges. STRING is available in Cytoscape (91) via a plugin, from R language through the Bioconductor package and its website. Interactions are scored using a “combined score” based on whether they are biologically meaningful and, in the case that there is evidence that they form a complex, also using a physical interaction score. Apart from protein-protein interaction network generation and visualization, STRING also offers network statistics, such as average node degree and protein-protein interaction enrichment p-value, network functional enrichment, and clustering using either k-means or Markov Clustering algorithm (MCL). IID (84) contains interactions for 18 organisms inferred from 7 primary databases, text mining, computational predictions (for humans and yeast only) and protein orthology. The resulting interactions can be filtered or enriched for tissue, cellular location, disease, and drug annotation and analyzed for topology, based on node degree and betweenness. Metascape (71) is a web tool that integrates information from more than 40 knowledge bases and offers an integrated OMICs data analysis. Metascape produces protein networks and analyses them, using the MCODE algorithm for network clustering and enrichment analysis for each cluster. Additionally, Metascape also offers gene annotation information, including gene summary, genomic variants and cellular localization, membership search, interactome analysis and functional

enrichment. Such databases have been extensively used for the reconstruction of PPI networks in several diseases (66, 92).

2.2.1.2 Protein co-expression networks reconstruction

The existing methods for co-expression network reconstruction were grouped into four basic categories: correlation-based, information theory-based, mathematical modelling and other methods.

2.2.1.2.1 Correlation-based Reconstruction Techniques

Weighted gene correlation network analysis (WGCNA) (72) is one of the most widely used correlation-based approaches to construct co-expression networks. During its procedure, WGCNA quantifies the interaction between individual pairs of genes and the degree to which these genes have identical neighbors. Then the topological matrix constructed is converted into a dissimilarity measure for hierarchical clustering, which creates a dendrogram. Finally, there is the possibility for the singular branches to be clustered in separate modules. WGCNA provides tools to construct networks, identify modules, determine topological attributes, simulate and visualise data and gives the possibility to the user to select between different correlation metrics and weighted (soft-thresholding) or unweighted (hard-thresholding) network construction.

RMTGeneNet combines Pearson correlation and random matrix theory to construct networks. After the similarity matrix has been constructed RMT is used to threshold it and separate biologically relevant correlations from noise. More specifically, after finding a significant threshold RMTGeneNet continues to iterate through lower thresholds until a Chi-square of 200 is found. This way an adjacency matrix is constructed, where values below the threshold are set to zero and non-zero values represent the edges of the network (93). MPICorMat (94) is a parallelization of RMTGeneNet. It constructs the similarity matrix using Pearson correlation by taking advantage of the computational capabilities of multicore clusters.

2.2.1.2.2 Information Theory-based Reconstruction Techniques

MIDER (73) takes as input time-series data related to quantitative features of the network nodes and combines mutual information-based distances with entropy reduction to infer the network structure. It consists of two steps. Firstly, the network

is represented in a way that the statistical closeness is indicated by the distance among the nodes. Secondly, the algorithm refines the prediction of existing edges to distinguish between direct and indirect links and directionality is assigned using transfer entropies.

2.2.1.2.3 Mathematical Modelling Reconstruction Techniques

Reverse phase protein array (RPPA) (95) is a computational method which constructs protein networks and detects complex patterns in protein signalling by using statistical analysis of protein pairs. Protein pairs are analyzed using multiple analysis of variant (MANOVA), significant protein pairs are selected comparing each knockdown gene versus negative controls and each protein pair is scored based on its significance level in the group comparisons, correlation coefficient, and the number of times each protein is significant in a pair in the MANOVA model. Final networks for each comparison are constructed based on a threshold value to select the high-scoring protein pairs.

A Bayesian network is a representation of a mutual causal probability distribution of a set of random variables. Nodes representing the random variables edges between pairs of nodes representing the causal relationship of these nodes, and a conditional probability distribution in each of the nodes are all parts of the network, which also uses two major mathematical principles, the one of the directed acyclic graph and the one of probability (96). Their strength is their ability to estimate model parameters even in the presence of incomplete data, which makes them perfect for modelling protein-protein interaction networks, and their major limitation is that they cannot support large-size networks (97). Depending on whether Bayesian networks are generated using data with multiple timepoints or not, they can be used to create protein regulatory networks too. A representative example of a tool generating this kind of network is the Bayesian Network Webserver (BNW) (76). It can infer the network structure, perform parameter learning and display the network model.

The Sparse Estimation of the Correlation matrix (SEC) (74) is an approach that estimates a sparse correlation matrix and penalizes the correlations according to the empirical ones (larger amount of penalization to smaller empirical correlations). This

estimation is based on the Accelerated Proximal Gradient algorithm and achieves a fast rate of convergence.

2.2.1.2.4 Other Network Techniques

Motif Analysis GeneNet (98) expresses an approach for the detection of gene expression modules which are regulated by known promoter motifs. The methods used during the procedure are based on motif enrichment and motif position bias.

Boolean networks model and analyze complex systems with switch-like causal interactions based on binary logic. This means that in the system, each node logical function defines the node state of the following step, related to the input values the other nodes present. Probabilistic Boolean networks are an extension of Boolean networks, as more than one transition is applied and is strengthened concerning uncertainty since at any time point the gene state vector transitions occur following the laws of the associated network (96). Like Bayesian networks, Boolean networks can also be used to generate protein regulatory-directed networks, depending on the input data. A representative tool for constructing Boolean networks is CABeRNET (99), which is a standalone tool and a Cytoscape app plug-in, which can generate and analyse Boolean networks, focusing also on their augmentation in case partial functional or topological information is provided.

2.2.1.3 Protein regulatory networks reconstruction

The existing methods for regulatory network reconstruction were grouped into three basic categories: information theory-based, mathematical modelling and machine learning-based methods.

2.2.1.3.1 Information Theory-based Reconstruction Techniques

The Algorithm for the Reconstruction of Accurate Cellular Networks using adaptive partitioning strategy (ARACNe-AP) (75), uses an information theoretic context based on the data processing inequality theorem to infer direct regulatory relations among transcriptional regulator proteins and target genes. ARACNe-AP estimates the mutual information threshold, uses bootstraps to reconstruct networks and finally constructs a consensus network. The application of the algorithm based on an adaptive partitioning strategy for estimating the mutual information presents more sensitivity

includes the Data Processing Inequality (DPI) implementation and supports multithreading.

Conditional Mutual Inclusive Information (CMI2)-based Network Inference (CMI2NI) (100) algorithm constructs gene regulatory networks by combining CMI2 with path consistency algorithm. CMI2 is used to quantify the mutual information between two genes given a third including and excluding the edge between the two genes. Firstly, a complete connected graph is generated. Secondly, the mutual information between an adjacent gene pair is computed and in case it is low or equal to zero, the edge between the genes is deleted. Then, CMI2 is computed for an adjacent gene pair, given a third neighboring gene. The higher order CMI2 is calculated until there is no change in network topology.

The multivariate Information-based Inductive Causation (MIIC) algorithm is based on a method which combines a constraint-based learning approach and maximum likelihood framework to construct networks. More specifically, it starts from a fully connected graph and iteratively removes edges based on their contributions in indirect paths and orients the remaining edges by causality. MIIC has also an online interface where, apart from the dataset, the user can also give as an optional input a defined network layout, can exclude specific edges based on prior knowledge or provide a threshold to filter retained edges based on their confidence (101).

2.2.1.3.2 Mathematical Modelling / Probabilistic Reconstruction Techniques

Yan et al. (102) introduced a method for directed network construction, taking into consideration degree heterogeneity and homophily simultaneously. Degree heterogeneity is denoted in the model via node-specific parameterization. They parameterize the degree of each node by an incomingness parameter characterizing how attractive the node is, and an outgoingness parameter characterizing how attracted the node is to other nodes. Homophily is taken into consideration by incorporating node covariates.

Dynamic Bayesian networks enriched the standard Bayesian networks by also involving the concept of time. This offers the ability to operate with time series, expressing different conditions due to different contexts over time (96). A tool that

uses dynamic Bayesian networks is FALCON (103), which creates large regulatory networks from noisy data using a large number of perturbation experiments.

Ordinary differential equation constructs a dynamic system of gene regulatory networks because of the use of continuous variables. They are considered to be the best-applied methods for non-linear systems (96). An indicative example tool is HiDi (104), a standalone Matlab tool which uses a linear differential equation model to identify large regulatory networks from time-course data.

2.2.1.3.3 Machine-learning based Reconstruction Techniques

Artificial Neural Networks (ANNs) can recognize any input pattern entered and create models of the data structure and the connections that occurred during the procedure. The recurrent neural network topology is widely presented as the most effective neural network-based model for gene regulatory network construction because of its ability to represent and model feedback and memory mechanisms (96). Rubiolo et al. (105) have used the Extreme Learning Machine (ELM) supervised neural model to reconstruct regulatory networks from time series.

GEne Network Inference with an Ensemble of trees (GENIE3) (77) is an unsupervised method for network inference based on regression trees. This method is scalable, suitable for non-linear data and efficient in the case of a large number of features. GENIE3 reconstructs a regulatory network from x molecules by decomposing the problem into x different regression subproblems, trying to find the subset of molecules that can best predict the target.

As the methods for reconstructing regulatory networks increase, Netbooks (106) have recently been developed and are based on Jupyter Notebook (107), which is an open-source web application that supports over a hundred of programming languages and allows the user to create and share text, code or other related documents of a project. Netbooks are Jupyter notebooks, that provide the user with step-by-step guidance on case studies involving regulatory networks and thus facilitate reproducibility.

2.2.1.4 Protein networks clustering methods

The first step after constructing and visualizing the network (with the most popular and widely used tool for visualization being Cytoscape (91)), is often to divide the network into separate clusters. Network clustering is used to group proteins with similar expression profile and results in groups of nodes that often correspond to different functional groups. Clustering methods can be separated into two categories, the hard clustering techniques, where all proteins are clustered and each protein can only belong to one cluster, and the soft clustering technique, where clusters can overlap, some proteins could remain unclustered and a protein can exist in more than one clusters with a certain probability or degree of membership.

Most existing network clustering techniques perform hard clustering. As mentioned above, WGCNA is widely used for the identification of modules through hierarchical clustering. There are algorithms though, that are clustering-specific. Markov Clustering (MCL) algorithm (108) is one of the most widely used hard clustering algorithms in the field. It uses random walks, expansion and inflation operators and mathematical bootstrapping to identify the cluster structure. It was first used for clustering of protein sequences to detect protein families, but later its use was expanded in co-expression networks clustering. Because of its high memory usage and computational running time, High-performance MCL (HipMCL) (78) was introduced. HipMCL is MCL's scalable parallel implementation, which uses MPI and OpenMP to make clustering detection faster and more efficient on large-scale. Because of certain drawbacks of MCL, such as its biased process for tuning and selection of the inflation parameter by the user, which affects its performance, there have been expansions of the algorithm trying to overcome these. Enhanced Markov clustering (EMC) (109) finds protein complexes by first performing MCL clustering and then filtering the results with one or a combination of 4 filters, namely the density, haircut operation, best neighbor and cutting edge. Theofilatos et al. (110) combined EMC with an evolutionary algorithm to predict functional protein complexes from weighted protein-protein interaction networks. This methodology, the evolutionary enhanced Markov clustering (EE-MC), performs MCL clustering, uses the EMC's 4 parameters adjusted for weighted networks in a certain order (best neighbor, haircut, density, cutting edge)

to enhance MCL clustering and applies a genetic algorithm to optimize EMC's parameters and MCL's inflation rate in parallel by maximizing a fitness function based on a cost function, producing possibly overlapping protein complexes. Restricted Neighborhood Search Clustering Algorithm (RNSC) (111) is a cost-based algorithm for protein-protein interaction networks clustering. RNSC forms clusters by trying to minimize a cost function, which is calculated based on the number of intra- and inter-cluster edges. Each time a node is moved from one cluster to another if the cost function is lower and this process finishes when a certain number of moves have happened without the cost function being reduced. Gradually expanding neighborhoods with adjustment (GENA) (112) is an unsupervised clustering algorithm for the prediction of protein complexes from weighted protein-protein interaction networks. It functions in two steps. The first step is the initialization step, which starts from some seed nodes and expands the clusters, including neighboring nodes. More specifically, the seed nodes are selected based on the highest clustering coefficient and each seed node uses an l-forward r-backward searching algorithm to expand its cluster, minimizing an evaluation function based on the cluster connectivity. The second step is the adjustment step, where GENA uses local search based on the RNSC method and expands its moves. More specifically, GENA uses 3 moves to define the cluster nodes, the removal of a node from one cluster to another with high probability and the copy of a node to a cluster or its deletion from a cluster with low probability. This step terminates when the maximum number of moves has been performed or the evaluation function did not improve for the several (1000) last steps. Finally, clusters with one node are removed and highly overlapping clusters are merged. This algorithm allows for overlapping clusters and the processing of weighted protein-protein networks.

Molecular Complex Detection (MCODE) (113) algorithm is widely used in biological networks and allows for the detection of overlapping clusters in large protein-protein interaction networks. It performs clustering in 3 steps; node weighting based on the clustering coefficient which measures the cliquishness of the neighborhood of the nodes, complex prediction based on the weighted nodes, and an optional post-

processing step to filter out or add proteins in the resulting clusters by certain connectivity criteria.

ClusterONE (79) is a local-cluster-quality-based soft clustering method for detecting potentially overlapping clusters in protein-protein interaction networks. It depends on overlapping neighborhood expansion and consists of 3 steps. First, it starts from a single vertex and adds or removes vertices using a greedy procedure. This process is repeated from different seeds and forms many possibly overlapping clusters. In the second step the clusters which have high overlap scores are merged and in the final step clusters with less than three proteins or with a density below a certain threshold are removed. Finally, some algorithms that have been applied to complex networks, such as fuzzy c-means, could be also used for protein networks.

2.2.1.5. Network analysis

Network analysis is used to identify several characteristics of the networks and answer certain biological questions of interest. In Table 2.2 we present the series of the most commonly used local and global network characteristics, their mathematical definition, and their biological meaning (114).

	Metric	Definition	Mathematical Type	Biological Meaning
Global	**Degree Distribution**	The distribution of the percentage of nodes of degree k, P(k), over all k ie. the probability that a node has degree k	P(k) ~ k^(-γ), with often 2<=γ<3	Shows the connectivity of the network
	Clustering Coefficient	The number of edges between a node's neighbors divided by the number of possible connections between these neighbors	$C_i = \frac{2e}{k(k-1)}$, C_i: clustering coefficient of node I, with $0 \leq C_i \leq 1$ k: degree e: number of edges between the k neighbors of node i	Shows whether the network tends to form tightly connected communities

Shortest Path Length	The minimal distance (number of edges) that needs to be traversed to reach from one node to another		Shows how close two proteins are in the network
Betweenness Centrality	Number of shortest paths from all nodes to all others, that pass through a node	$C_{bet}(i)=\frac{\sigma_{xy}(i)}{\sigma_{xy}}$ σ_{xy}: total number of shortest paths from node x to node y σ_{xy} (i): total number of these paths that pass through node i	Shows how influential (thus important) a protein is to the network
Closeness Centrality	Is inversely proportional to the average shortest distance of a node to all other nodes in the network	$C_{clo}=\frac{1}{\sum dist_{ij}}=\frac{N-1}{\sum dist_{ij}}$ $Dist_{ij}$: the distance between nodes I and j N: number of nodes	Detects important proteins that can communicate with others quickly in the network

Local	**Degree**	Number of edges adjacent to a node	(In case of directed network) $deg_i=deg_i^{in}+deg_i^{out}$	Shows the number of interactions a protein is engaged in
	Network Motifs	Repeated subgraphs that appear in any frequency and show certain patterns of interactions between nodes		Can carry functional information and reflect the underlying evolutionary processes that generated the network (are evolutionary conserved, especially in PPI networks)
	Eccentricity	The eccentricity of a node v is the reciprocal of the longest shortest path between node v and all other nodes in the network	$Ecc=\frac{1}{dist(v,K)}$, with K being the most distant node from node v	How easily a protein can be functionally reached by all other proteins in the network

Table 2.2. Global and local network characteristics

The step after clustering the network is often the identification of essential nodes, and hubs. Hub nodes are central nodes to the network and could be separated into intra- and inter-modular hubs. Intra-modular hubs are central to each network cluster they belong to, whereas inter-modular hubs are central to the whole network (55). The most widely used methods for identifying hubs in a network are based on topological characteristics. The measures used for this are the centrality measures, such as degree centrality which is measuring the importance of a node based on the number of neighbors it has, closeness centrality which calculates the mean of the average shortest paths from one node to another, neighborhood centrality which is based on edge clustering coefficients and betweenness centrality which is the most typical centrality measure and defines a node as important by the number of shortest paths that pass through it (65). CytoHubba (115) is an example tool, which uses topological features to evaluate network nodes. It is a Cytoscape plug-in and contains 4 different local-based topological analysis methods, such as node degree and maximal clique centrality, and 7 global-based methods, such as closeness and betweenness centrality. When applied to a yeast protein-protein interaction network, CytoHubba's best-performing method (maximal clique centrality) was able to predict essential proteins with 90% precision in the top 20 ranked nodes. Guimera and Amaral (116) defined the role of nodes in complex networks, such as an E Coli metabolic network. Their methodology first identifies network modules and then determines the nodes by their intra-modular degree and their participation coefficient, which measures how well are their edges distributed across different modules. Han et al. (117) categorized hubs based on their expression with their interactors. That way, "party hubs" are co-expressed and have a high averaged Pearson correlation coefficient with their interactors, and "date hubs" have a low averaged Pearson correlation coefficient with their interactors. They suggested based on protein-protein interaction networks of yeast that party function inside modules whereas date hubs connect groups of nodes active in different time points or processes. Paci et al. (118) applied the date/party hub classification in gene co-expression networks from human cancer data, using the so-called clusterphobic coefficient as a measure of global connectivity, which measures the ratio of internal to external connections of the nodes, and the global within-module degree as a measure of local connectivity, which measures the rates of

intra-cluster connectivity over their global connectivity. For this purpose, they implemented SWItchMiner (SWIM) algorithm and later on its open-source version SWIMmeR (119) and found two more types of nodes, “fight-club hubs” with high clusterphobic coefficient and negative correlation with their first neighbors and “switch genes”, a part of fight-club hubs which also have a low within-module degree. Switch genes were shown to have high clinical and biological relevance in cancer and more recently in other diseases, including ischemic and non-ischemic cardiomyopathy (52). Hubs can also be found in external user-defined data. Contextual Hub Analysis Tool (CHAT) (120) is a Cytoscape plug-in, which constructs interaction networks from a list of genes or proteins of interest, integrates user-provided contextual data, such as differentially expressed molecules, identifies hubs by examining whether they are connected to contextual nodes more than expected by chance and can compare hub nodes to degree-based hubs. In a case study of a viral infection, hub nodes were found to be more functionally relevant compared to degree-based hubs.

Comparing networks can reveal biologically relevant characteristics of certain proteins related to network topology or identify proteins with different expression profiles under different conditions, such as diseases or tissue types. Several tools exist for network comparison or differential co-expression analysis. GraphCrunch 2 (121) is a software used for modelling biological, undirected networks according to 7 well-known network models, for network comparison between two networks based on either network topological characteristics, such as average pathlength and diameter, or small connected subgraphs of the network, the graphlets, for network alignment using the GRAAL algorithm (122), and clustering based on topological characteristics of the network. TopNet (123) is a web tool for the analysis of protein interaction networks, which creates subnetworks and compares them based on average degree, clustering coefficient, characteristic path length and diameter. Network Analysis Tools (NEAT) (124) clusters interaction networks and compares them mostly based on the intersection and union of nodes, whereas the more recent web tool NetConfer (125) includes different network comparison methods and separates them in different analysis workflows, each one designed to achieve different analysis objectives such as identification and comparison of key nodes, inference and comparison of community

structures, community clustering and creation of superclusters or comparative analysis of network cliques. BioNetStat (126) can compare two or more correlation networks of predefined variables, based on the probability distribution of graph or node features. It returns the compared subnetworks or significant nodes, topological or node properties, the statistics of the test and the p-values. DiffCoEx (127), CoXpress (128) and CoDiNA (129) build modules and then detect molecules with differential correlation among different biological states. DICER (130) finds molecules with differential co-expression in a state of interest compared to all other states, identifying either groups of molecules differentially correlated in one state or pairs of molecule sets where the molecules within each set are correlated across all states, but the sets have a different correlation between states. These sets are called meta-modules.

2.2.2 Major network reconstruction and analyses applications in atherosclerosis

Recent studies have used protein networks to infer conclusions about complex diseases. Emilsson et al. (131) used WGCNA to construct a protein regulatory network of more than 4000 serum proteins from a large cohort of elderly individuals. They identified structurally preserved network modules, some of which are associated with several disease conditions, including coronary heart disease, heart failure, type 2 diabetes, and metabolic syndrome. Certain modules associated with incident disease and all-cause or post-coronary heart disease mortality indicating that the protein network was predictive of future events and disease outcomes. They also identified reproducible protein hubs which showed a strong association with disease-related outcomes, such as UMP-CMD kinase (CMPK1) for example, which was positively associated with type 2 diabetes. Another example was the small ubiquitin-related modifier 3 (SUMO3) which was positively associated with prevalent heart failure, and high levels of this protein could predict reduced survival post-incident coronary heart disease. These results indicated that serum proteins might be possible biomarkers of disease.

Langley et al. (17) initially performed proteomics analysis of the extracellular matrix and associated proteins from 12 samples of human carotid endarterectomy atherosclerotic plaque specimens. Selected findings were integrated with

corresponding transcriptomes from 121 carotid endarterectomies and 9 candidate biomarkers were identified. Biomarkers were validated in two independent prospective studies with a 10-year follow-up and a final 4 biomarker signature that significantly improved risk prediction for cardiovascular diseases was identified. The identified biomarkers were revealed and linked in a correlation-based co-expression network constructed from the proteomic data. Herrington et al. (61) performed proteomic analysis of 99 human coronary artery and 99 human aortic samples and used WGCNA to construct reproducible protein co-expression modules for left anterior descending samples and abdominal aorta samples. The authors performed an enrichment analysis of the modules showing that they were enriched in different cellular functions, such as mitochondrial proteins with cellular respiration. They observed that mitochondrial proteins were more abundant in the left anterior descending samples than in the aortic ones and that both sample types were highly enriched in fibrous plaques and normal intima. When comparing the fibrous plaque- and normal intima-enriched left anterior descending samples, they identified 89 significant proteins and used knowledge-fused differential dependency networks to characterize significant network rewiring between normal and fibrous plaque samples. 26 rewiring hub proteins were revealed with significant enrichment of tricarboxylic acid proteins, whose analysis showed a reduction in fibrous plaque-enriched samples, demonstrating that network rewiring could reveal biological results not easily evident from other means.

Kamal et al. (132) did a proteomics analysis of murine macrophage cells treated with lipopolysaccharide and statin to investigate the mechanism of proteins during inflammation and the effect of statin and lipopolysaccharide on the protein level, consistently quantifying 344 proteins. They used Ingenuity Pathways Analysis (IPA) to create 24 protein-protein interaction sub-networks. Proteins plectin and prohibitin 2 had a high ranking In the IPA analysis and were also found highly expressed in treated cells compared to controls, thus they were selected for targeted interactome analysis. Their targeted networks revealed direct and indirect interactions with other proteins. They generated a disease and function-based protein network with the identified proteins, which revealed certain inflammatory diseases and their functions, such as

cell immune response, inflammation of absolute anatomical regions and organ inflammation. Moreover, they found some identified proteins in the treated cells to be associated with cardiac inflammation, such as tubulin family proteins with pericarditis and carditis.

In a more recent study, Hartman et al. (133) used sex-specific gene regulatory networks to infer a mechanism of differences in atherosclerosis between men and women. More specifically, they used RNA-seq data from the aortic root of 160 males and 160 females from the STARNET study and created sex-specific co-expression networks using WGCNA. They performed enrichment analysis on the 32 female modules identified and found that 24 were enriched for biological processes ranging from immunity to metabolism and muscle tissue development. They prioritized female modules for their importance in female atherosclerosis, performed processes enrichment for the top 3 most significant ones and showed that the cell type was determinant for each module. They then used Bayesian network analysis to infer directionality and generate gene regulatory networks from the 3 significant co-expression modules, on which key driver analysis was performed. The key driver genes of each module were further analyzed using scRNA-seq data from 37 human carotid endarterectomy atherosclerotic plaques to determine their cellular expression, confirming the cell specificity of the female modules. Key drivers were differentially expressed between the sexes in plaque smooth muscle cells, uncovering female atherosclerotic biology in these cells.

Koplev et al. (59) used transcriptomic data from seven tissues from 600 patients with coronary artery disease and 250 controls from the STARNET study to identify mechanisms that cause cardiometabolic and cardiovascular disease phenotypes. They used WGCNA and inferred 135 co-expression tissue-specific modules and 89 cross-tissue ones, which were found highly associated with cardiometabolic phenotypes, coronary artery disease scores and differentially expressed genes in coronary artery disease after enrichment analysis, highlighting co-expression modules' ability to provide functional and clinical information to study cardiometabolic and cardiovascular diseases. They then transformed the co-expression modules to gene regulatory networks using GENIE3, carried out key driver analysis in each subnetwork

and found that the regulatory interactions contain more than 40000 independent expression regulatory SNPs whose contribution to coronary artery disease heritability was almost 60%. Finally, using Bayesian network modelling of the eigengene values of each module they created a supernetwork representing the intra- and inter-organ organization of the regulatory networks and the connectivity between the modules and found that cross-tissue networks are critical for the development of coronary artery disease.

2.2.3 Comparison of network approaches for atherosclerotic tissue proteomics

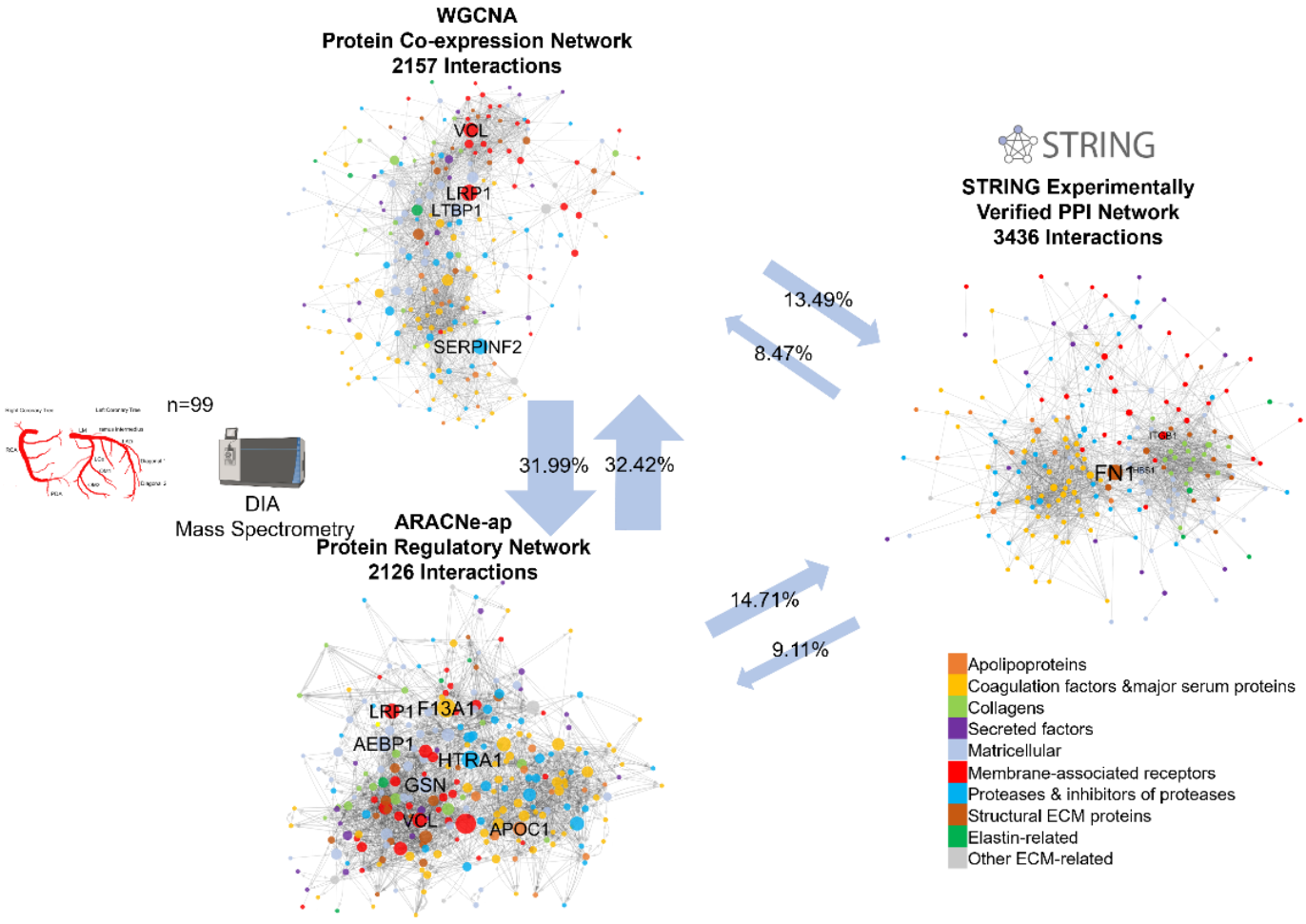


Figure 2.1. Different types of networks capture different aspects of the matrisome network of atherosclerotic plaques. Example of different types of networks for the matrisome of human left anterior descending coronary artery. DIA MS data of 99 samples were used from the Parker et al. study (134) and filtered to keep only matrisome-related proteins, according to a custom matrisome database composed of extracellular proteins from the MatrisomeDB (135), apolipoproteins and other secreted proteins, that are consistently quantified (less than 20% missing values). The

experimentally verified PPI network was created by mining matrisome interactions from the STRING web tool (70). The protein co-expression network was reconstructed using the WGCNA pipeline (72), with Pearson's Correlation as the interaction metric and 0.5 with a soft power of 10 as a threshold to infer interactions. ARACNe-AP (75) with default parameters was used to reconstruct the protein regulatory network with the same data. All networks were visualized using Cytoscape (91), proteins were colored based on the matrisome group they belong to, and node size was set to be proportional to the betweenness centrality of the node in the network. Only the nodes with betweenness centrality above 0.5 were labelled. The confirmed interactions of a network against another type of network are depicted with arrows connecting the different networks and the percentages (%) of common interactions between the two types of networks are also depicted. VCL: Vinculin; LRP1: Prolow-density lipoprotein receptor-related protein 1; LTBP1: Latent-transforming growth factor beta-binding protein 1; SERPINF2: Alpha-2-antiplasmin; F13A1: Coagulation factor XIII A chain; AEBP1: Adipocyte enhancer-binding protein 1; HTRA1: Serine protease HTRA1; GSN: Gelsolin; APOC1: Apolipoprotein C-I; FN1: Fibronectin; THBS1: Thrombospondin-1; ITGB1: Integrin beta-1

Figure 2.1 demonstrates a visual example of reconstructing the basic types of networks presented in **Error! Reference source not found.** (PPI, protein co-expression and protein regulatory networks). We have chosen a proteomics dataset for the matrisome of the left anterior descending coronary human arteries, using the most widely used method from each category. The matrisome was defined as the ensemble of ECM and ECM-associated proteins. Bayesian methods and other mathematical modelling methods were not used since the examined tools (**Error! Reference source not found.**) were not supporting datasets of this sample and protein markers size. Limited overlap was found among the static PPI network and the reconstructed networks, as the percentages of confirmed interactions of the PPI network against both the protein co-expression (8.47%) and the protein regulatory (9.11%) network was less than 10%. PPI networks, even the ones based on experimental evidence, are being created based on evidence from different types of tissues and conditions, and most of them are likely not relevant to a particular tissue. As shown in Figure 2.1, significant proteins in the PPI networks, such as FN1, were not returned as significant in the other networks. Protein co-expression and protein regulatory networks presented higher overlap, with more than 30% of the interactions of one network confirmed in the other (31.99% and 32.42%, respectively). Moreover, significant overlap was observed in the hub proteins, with the membrane-associated proteins VCL and LRP1 being hubs (betweenness centrality over 0.05) for both protein co-

expression and regulatory networks. Thus, instead of just exploring the static PPI networks, this analysis should be complemented with the reconstruction and analysis of co-expression and regulatory networks in the specific tissues of interest to probe real interactions and disease mechanisms.

2.2.4 The effect of medications on cardiovascular diseases networks

The pleiotropic effects of several drugs in various cardiovascular diseases (136–138), including statins (139), have been previously discussed and reported in the literature. However, little is known about the effect of drugs on the network level and co-expression networks can be particularly useful to promote our understanding of the pleiotropic effect of drugs.

To better demonstrate the above concept, we conducted co-expression network reconstruction using MS-based apolipoprotein measurements on plasma samples before and after the use of statins. Statins are among the most commonly used medications to treat cardiovascular diseases, such as atherosclerosis, coronary artery disease, and heart failure (140).

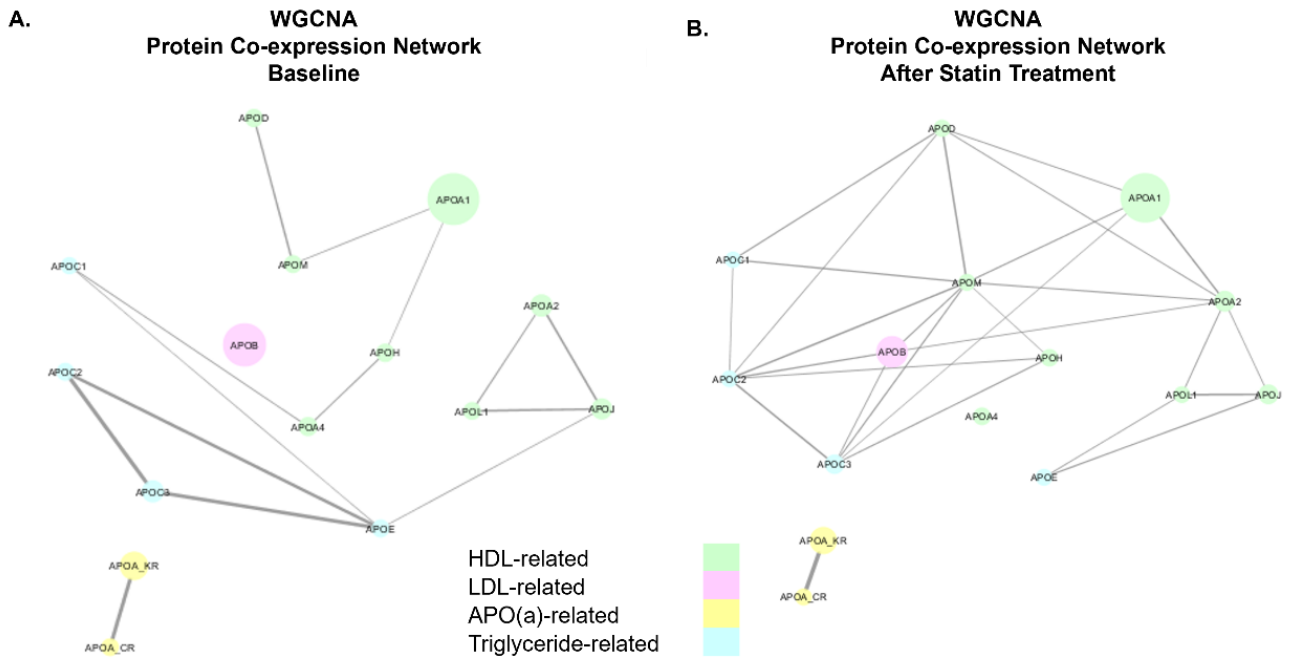


Figure 2.2 Apolipoproteins' co-expression network in plasma samples before and after the use of statins. Apolipoprotein data from 20 patients **A.** before and **B.** 1 year

after statin treatment from the ASCOT trial (141) were used to build co-expression networks. WGCNA pipeline (72) with Pearson correlation as the interaction metric and 0.5 with a soft power of 4 as a threshold to infer interactions was used to create the networks. Networks were visualized using Cytoscape (91), apolipoproteins were colored based on the functional category they belong to, edge width was set to be proportional to correlation metric and node size was set to be proportional to the average absolute apolipoprotein abundance (milligrams per litre, mg/L).

Figure 2.2 shows an example of the effect of statin usage in a co-expression network of apolipoproteins from plasma samples, taken from the Anglo-Scandinavian Cardiac Outcomes Trial (ASCOT) (141). The ASCOT trial was a double-blind randomised 2 × 2 factorial trial of blood pressure lowering and lipid-lowering treatment. Among the 14,412 participants, apolipoprotein measurements using MS (142) were conducted for a random sample of 40 individuals (20 allocated to atorvastatin and 20 allocated to placebo) from an age and sex-matched subset of treated and controls in the ASCOT trial with available plasma samples.

Apolipoprotein co-expression network after statin usage differed significantly, showing higher connectivity. LDL-related apolipoprotein B (APOB) was not significantly associated with any other apolipoprotein in the baseline network but showed a high correlation with certain HDL- and triglyceride-related proteins after statin therapy. Statins significantly lowered the LDL and triglyceride-related apolipoproteins, with APOB being the most affected protein as its levels decreased 5 times more compared to apolipoprotein C-III (APOC3), the protein with the second highest decrease in the network. The new connections introduced between APOB and triglyceride-related apolipoproteins after statin treatment could be potentially explained by the substantial reduction of LDL-related APOB, which ended up leaving in the data a higher proportion of triglycerides-related APOB. Only apolipoprotein(a) (APOA) levels remained the same after statin usage. Thus, apolipoprotein network comparison confirmed that statins target LDL-related as well as triglyceride-related proteins and that this has a massive effect on the network suggesting that proper corrections should be conducted on the data for medications and other cofounding factors before reconstructing co-expression networks (143).

2.2.5 Examples of cardiovascular disease-specific networks

Network analysis has been applied to various cardiovascular diseases, dissecting molecular mechanisms underlying each disease and identifying novel therapeutic targets. Several biological network-based methods have been used to explain coronary artery disease. Iwata et al. (144) linked PARP9-PARP14 with coronary artery disease through network analysis, Lempiäinen et al. (145) integrated gene co-expression networks with PPI networks and GWAS studies to identify subnetwork modules associated with coronary artery disease and selected candidate drug targets via network analysis whereas Huan et al. (146) identified lipid metabolism and inflammation processes to be involved in coronary artery disease pathogenesis, via differentially expressed gene networks and their analysis. Network medicine has also been used in different cardiovascular diseases. Nakano et al. (147) analyzed protein networks from mice macrophages and linked certain protein clusters with atherosclerosis and myocardial infarction, whereas a recent study (148) integrated gene expression profiles and PPI networks to form condition-specific co-expression networks and identify modules from different processes driving the progression of heart failure. Schlotter et al. (149) used a protein network of highly expressed proteins in aortic valves and identified specific pathways critical to the network topology which could be possible therapeutic targets for preventing valvular calcification. Network analysis has also aided in finding possible drug targets or driving mechanisms for further diseases, such as carotid artery stenosis, stroke, hypertension, or cardiac hypertrophy (150, 151). Finally, a recent study (152) introduced PPI patient-specific networks in hypertrophic cardiomyopathy utilizing individual patients' network topology and suggested that unique protein interactions could be important to understand the biological variability across different phenotypes of cardiomyopathy.

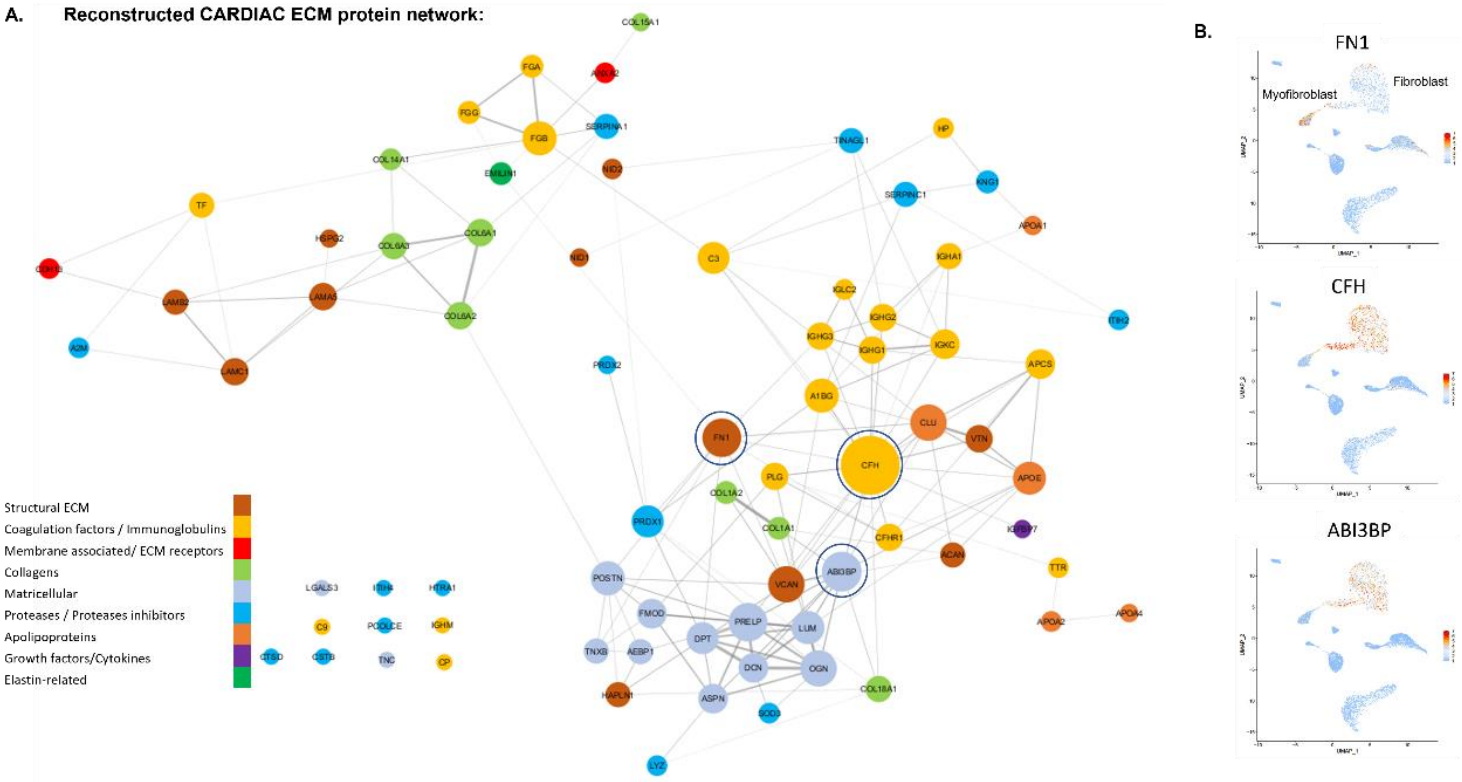


Figure 2.3 Heart tissue matrisome network and explanation of top changes using single-cell RNA-sequencing data. A. Label-free discovery MS data of 65 ischemic heart tissue samples were used from the Barallobre-Barreiro et al. study (153) and filtered to keep only matrisome-related proteins, according to a custom matrisome database composed of extracellular proteins from the MatrisomeDB (135), apolipoproteins and other secreted proteins, that are consistently quantified (less than 30% missing values). ARACNe-AP (75) with default parameters was used to reconstruct the regulatory networks filtering out negative associations using the SIREN algorithm (154). The network was visualized using Cytoscape(91), matrisome proteins were colored based on the functional category they belong to, edge width was set to be proportional to mutual information metric, node size was set to be proportional to its degree centrality and hub proteins are highlighted in blue. **B.** Uniform Manifold Approximation (UMAP) feature plot of the expression of the top 3 central nodes (hubs) of the network, using the Hocker et al. (155) scRNA-seq dataset of human heart tissue samples (8993 cells from two healthy donors) from the ExpressHeart web portal (156).

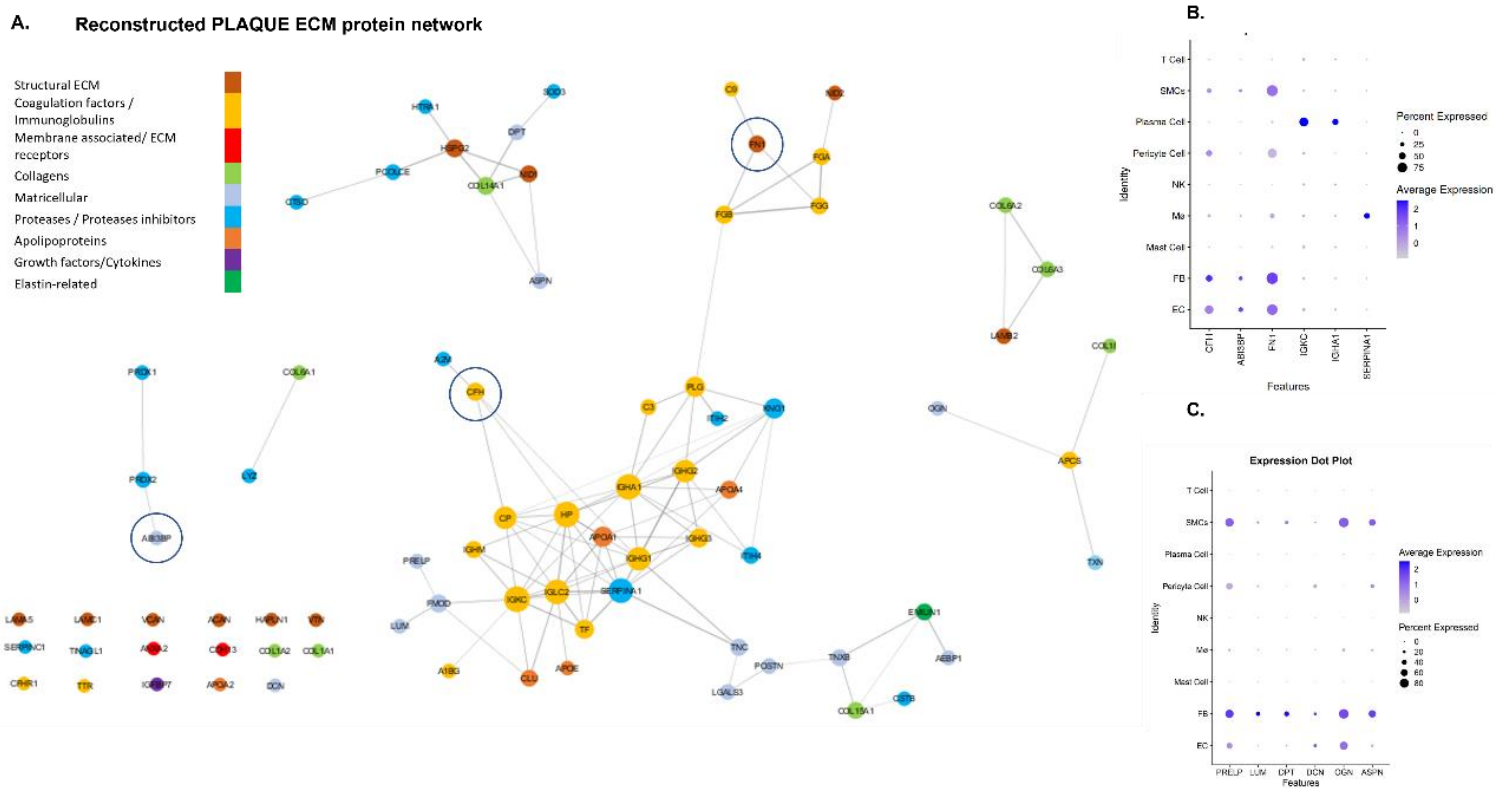


Figure 2.4 Carotid plaques matrisome network and explanation of top changes using single-cell RNA-sequencing data. **A.** Label-free discovery MS data of 12 carotid endarterectomy samples were used from the study by Langley et al. (17) and filtered to keep only matrisome-related proteins, according to a custom matrisome database composed of extracellular proteins from the MatrisomeDB (135), apolipoproteins and other secreted proteins, that are consistently quantified (less than 30% missing values). ARACNe-AP (75) with default parameters was used to reconstruct the regulatory network, filtering out negative associations using the SIREN algorithm (154). The network was visualized using Cytoscape (91), matrisome proteins were colored based on the functional category they belong to, edge width was set to be proportional to mutual information metric, node size was set to be proportional to its degree centrality and hub proteins of the heart tissue network (Figure 2.3) are highlighted in blue. **B.** Expression dot plot of top 3 central proteins (hubs) in the carotid plaques network and the heart tissue network (Figure 2.3), using the PlaqView web tool (157), scRNA-seq data from carotid plaques (Pan et al. (158) dataset, n=3) and Aran et al. (159) method to label cell clusters. **C.** Expression dot plot of selected proteoglycans in different cells using the PlaqView web tool (157), scRNA-seq data from carotid plaques (Pan et al. dataset (158), n=3) and Aran et al. (159) method to label cell clusters. EC: Endothelial cells, FB: Fibroblast cells, SMCs: Smooth Muscle Cells, NK: Natural Killer cells, Mo: Monocytes.

To illustrate the tissue and disease specificity of the reconstructed networks, we reconstructed two networks of the extracellular matrisome using label-free proteomics data from samples of previously published studies on ischemic heart

failure (153) (Figure 2.3.A) and atherosclerotic carotid plaques (17) (Figure 2.4.A). For the reconstruction of these networks, we used only extracellular matrisome proteins identified and consistently quantified in both datasets. With the network reconstruction, we were able to confirm known interactions in the heart tissue network, such as the ones between collagens (*COL6A1, COL6A2, COL6A3*) and the ones between laminins (*LAMA5, LAMB2, LAMC1*). As expected, the two disease and tissue-specific networks do not show many similarities, with the heart tissue network having higher connectivity and a smaller number of sub-networks than the carotid plaques one. The major differences between the two networks are in the decreased connectivity among the matricellular proteins and Complement Factor H (*CHF*) in the carotid plaques versus the ischemic heart tissue network. In contrast, serum proteins and proteases are more central in the carotid plaque matrisome network since carotid endarterectomy lesions are expected to have more inflammation than heart tissue.

The most significant difference is observed in the top hub proteins using the degree centrality of each network. CHF, FN1 and Target of Nesh-SH3 (ABI3BP) are the top 3 interconnected proteins of the heart tissue network (with their degree centrality being 40, 23 and 22 respectively). Publicly available scRNA-seq data from heart tissue samples (Figure 2.3.B) showed that these matricellular and structural ECM proteins are mostly expressed in fibroblast and myofibroblast cells in the heart. On the contrary, these proteins lost their central role in the carotid plaque network and stop being hubs (Figure 2.4.A), having a very smaller degree, belonging to different subnetworks and being also expressed in smooth muscle and endothelial cells (Figure 2.4.B). Immunoglobulins (Immunoglobulin Kappa Constant: *IGKC*, Immunoglobulin Heavy Constant Alpha 1: *IGHA1*) and serpin family A member 1 (*SERPINA1*) were the top 3 hub proteins in plaques (with a degree of 11 for immunoglobulins and 10 for SERPINA1 respectively) and were highly expressed in plasma cells and monocytes, respectively (Figure 2.4.B). Moreover, proteins with similar functionality were more closely connected in the heart tissue network than the carotid plaque. One such example is the matricellular proteins and especially the proteoglycans (*DPT, PRELP, LUM, DCN, OGN, ASPN),* which form a highly connected network component in the heart tissue network (Figure 2.3) whereas in plaques showed less interconnectivity

and belonged to different subnetworks. In opposite to their cellular expression in the heart tissue (myofibroblasts and fibroblasts), in plaques these proteins were not only expressed in fibroblasts but also endothelial and smooth muscle cells (Figure 2.4.C), reflecting the higher cell heterogeneity of carotid plaques. Thus, the combination of network analysis and scRNA-seq data verified that the known cell composition differences between the two tissues and diseases are reflected in the reconstructed networks and identified different hub proteins for each matrisome network.

2.3 Conclusion

The reconstruction and study of different types of atherosclerotic networks as demonstrated above, can shed light on the mechanisms involved in atherosclerosis pathogenesis, development, and progression, identify new therapeutic and drug targets and assist in the development of more accurate diagnostic and prognostic biosignatures and models.

However, the current prevalent approach is mostly based on the reconstruction of co-expression networks using established pipelines such as WGCNA, or the use of static commercial (e.g. IPA) or publicly available (e.g. STRING) PPI networks. Despite their simplicity and ease of use, they cannot be used for causality analysis and thus fails to provide a stable hypothesis for the identification of new biosignatures and drug targets. Many researchers try to overcome this issue by combining proteomics with genomics in an attempt to identify pQTLs (160, 161), but these approaches are limited to genomic variants which alter protein abundances and thus fail to capture pathogenic mechanisms related to post-translational modifications, protein degradation and protein-protein interactions. To overcome these limitations, more robust pipelines and methods are needed to reconstruct and study directed regulatory networks, which are more suitable for generating and validating causality hypotheses and performing simulations. Moreover, multi-omics networks (33, 162) combining RNA and protein networks and the integration of single-cell RNA-sequencing and proteomics data (163), further allow the parallel analysis of transcriptional and translational mechanisms, but this analysis requires the development of new methods such as consensus clustering (164), which has recently been introduced.

Lately, a lot of emphasis has been given to uncovering the sex-specific mechanisms involved in atherosclerosis, as an attempt to develop new medications and therapies for the overlooked atherosclerosis in females. Some sex-specific therapeutic approaches have been recently introduced (165), but the promotion of the understanding of the differences in the basic mechanisms involved in atherosclerosis among the two sexes, as attempted in the Hartman et al. (133) manuscript, can substantially speed up advances in this field.

As demonstrated in this chapter, the network reconstruction process is highly affected by covariates and particularly by medications, such as statins, which not only target specific proteins but present pleiotropic effects. Thus, medications should be taken into consideration when reconstructing networks and when interpreting findings based on them. Finally, protein interaction networks are significantly different among different diseases and sample types, with cell composition being one of the most important factors. Thus, it is paramount that networks should be reconstructed specifically for each tissue type and disease entity. Then, the comparisons of these networks become more meaningful in identifying potential pathophysiological mechanisms that can then be further validated experimentally.

3. Resolving atherosclerotic networks with directional regulatory network reconstruction

3.1 Introduction

Data from various studies, such as PPI and gene expression measurements, have been employed and integrated for network representation and analysis (166). Network modelling has contributed significantly to insights into biological procedures and disease mechanisms. Identification of disease biomarkers provides valuable means of understanding pathophysiological processes and acquiring information on pharmacological responses to therapeutic interventions (167, 168). Due to great scientific achievements in measurement technologies and the availability of high-throughput profiling data, the pathogenesis of human diseases has been elucidated and has become accessible for treatment. Emphasis is placed on interactions in biological systems, rather than on individual components of biological procedures (169). The importance of network construction in biological research relies on its capability to offer crucial information for the discovery of new drugs and generally for medical treatment fields, where working models, based on network modelling, can help researchers pose hypotheses and create experimental designs (96).

Advances in computational methods, reliable statistical methods, and the availability of a wide variety of transcriptomics datasets have offered a premium context for the identification of gene functions in unresolved or not fully understood conditions (170). Useful outcomes from the application of gene regulatory networks (GRNs) include understanding information flow in a biological system, the determination of circuits inside the system that represent a specific function, and modelling perturbations in gene expression expressed in various situations (171).

As mentioned before (chapter 2), the basic types of biological networks are PPI and functional networks, which can be split into regulatory and co-expression networks according to whether their edges are directed or not. PPI networks can be divided into physical (direct) and functional (indirect) interactions and can be either reconstructed from experimental methods or predicted using in silico machine learning,

mathematical modelling, or other methods (53, 67). Databases of PPI networks usually include interactions from many sources and for many species but have a high false positive rate and do not fully cover the interactomes (64). Existing approaches for co-expression and regulatory network reconstruction vary and can be grouped into four basic categories: correlation-based reconstruction, information theory-based reconstruction, mathematical modelling or probabilistic reconstruction, and other techniques, such as machine learning-based reconstruction. Importantly, each method presents specific limitations. Most correlation-based reconstructed networks lack directionality in the interactions and can represent linear associations only, without distinguishing direct from indirect links (72). Although many methods based on information theory can remove indirect links, they cannot discriminate between positive and negative associations of the interacting molecules (75). Some probabilistic methods on the other hand, such as Bayesian networks, do not support large networks and most of them need additional user-defined information, such as the determination of certain thresholds which are usually obtained via trial-and-error (97).

Biological networks have been widely used in atherosclerosis and cardiovascular diseases to identify causal genes and potential biomarkers, and design drug targets (33, 52). In particular, as described in chapter 2, recently a holistic approach is being used to find mechanisms of complex diseases, involving more than one group of genes or proteins and this approach has been widely applied for atherosclerosis applications. Despite the use of existing medications, such as statins, atherosclerosis remains partially untreated and thus, the use of network approaches to identify novel therapeutic targets is of high importance.

Lately, emphasis has been given to uncovering the phenotype or sex-specific mechanisms in atherosclerosis, as an attempt to develop new medications and therapies in less well studied population subgroups, including female atherosclerosis patients. Some sex-specific therapeutic approaches have been recently introduced (165), but the understanding of differences in the basic mechanisms between the two sexes in atherosclerosis, as attempted in the Hartman et al. manuscript (133), can substantially speed up advances in this field.

Despite the emphasis given to sex and phenotype-specific changes in atherosclerosis, the majority is based on differential expression analysis exploring different genes or proteins separately, while network approaches have focused on reconstructing and analyzing general global networks, without studying and comparing phenotype-specific networks. Finally, existing methods possess specific limitations, such as the lack of directionality in the interactions and the high number of false positives or indirect interactions, limiting their applicability in clinical research.

In this chapter, we introduce a new methodology for reconstructing directed co-expression networks. The directional regulatory network reconstruction with adaptive partitioning (DiRec-AP) combines two existing, conditional mutual information-based methods, ARACNe-AP (75) and SIREN (154), expands them to get simplified views of the networks without multiple edges and allows a different threshold for each node, to reflect better the behaviour of some genes and proteins that may possess more interacting partners than others. The proposed method was benchmarked against the most widely used methods for reconstructing protein co-expression networks in three different publicly available datasets from the DREAM challenge, showing a significant increase in the accuracy of the reconstructed networks.

The DiRec-AP pipeline was then applied to proteomics datasets from carotid endarterectomy plaques to reconstruct matrisome networks in symptomatic, asymptomatic, male, and female samples. The reconstructed networks were compared first with each other to identify network-specific changes, and then against networks from coronary artery samples to explore which of the mechanisms are specific to carotid plaques and which can generally be observed in atherosclerosis. The reconstructed, atherosclerotic-directed networks can be used to formulate new hypotheses for mechanisms involved.

3.2 Methods

3.2.1 DiRec-AP Description

We developed a novel network reconstruction technique, extending the ARACNe-AP method from Lachmann et al. (75) described above (Chapter 2). The ARACNe-AP JAVA implementation (version 1.8.0_221) was first modified so that the bootstrapping

feature would be disabled for low sample-size datasets (sample threshold n=10 was used as the default threshold to define small sample size studies). For the rest of our method implementation, a Python (version 3.6.4) script was generated. The first step is the calculation of the significance threshold for the mutual information (MI) for each node of the network and the reconstruction of the initial network. The main difference between our method compared to ARACNe-AP implementation is that a threshold is calculated for each protein/gene separately and dynamically, to allow for some proteins to have more interactions similar to the "real" biological networks, where few proteins act as hubs while the rest possess fewer connections. This is achieved by initially resetting the MI threshold calculated by ARACNe-AP, the average and standard deviation of the MI values of every interaction per molecule are calculated. Then, a new threshold is calculated separately for each protein by assuming the MI values are following normal distribution and applying a user-defined confidence interval to calculate the threshold (default is 95%). In the case where the dynamic threshold is smaller than the one calculated by ARACNe-AP, the dynamic threshold is maintained, otherwise, the ARACNe-AP estimated one is kept. Subsequently, we use the Benjamini-Hochberg method (172) to correct the calculated by ARACNE-AP nominal p-values of the interactions for multiple testing, and the SIREN algorithm (154) to classify the edges into ones of activation or inhibition. Finally, self-loops are filtered out and the edge with the greatest weight is kept in cases where multiple edges among two proteins exist (either activation or inhibition) in both directions. All reconstructed networks during the process are also exported as intermediate steps.

The code takes as input 7 arguments and two optional ones: the path for the initial jar file, the path for the modified jar file, the expression file, the folder name for the output folder, the TFs file, the path for the SIREN folder, a float for the p-value threshold, an optional float argument standing for the confidence interval to calculate the threshold (the default values is 95%), and finally, as an optional argument, an int standing for the number of bootstraps. If the optional argument is not given as input, the default number of bootstraps is 100, as suggested in the ARACNe-AP manuscript. The parameters used in the code for this analysis were the 95% interval of trust used for the dynamic threshold calculation, 0.05 as a p-value threshold, and 100 bootstraps.

The code is available at the following link: https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis. The network reconstruction flowchart and the suggested pipeline for network analysis are displayed in Figure 3.1.

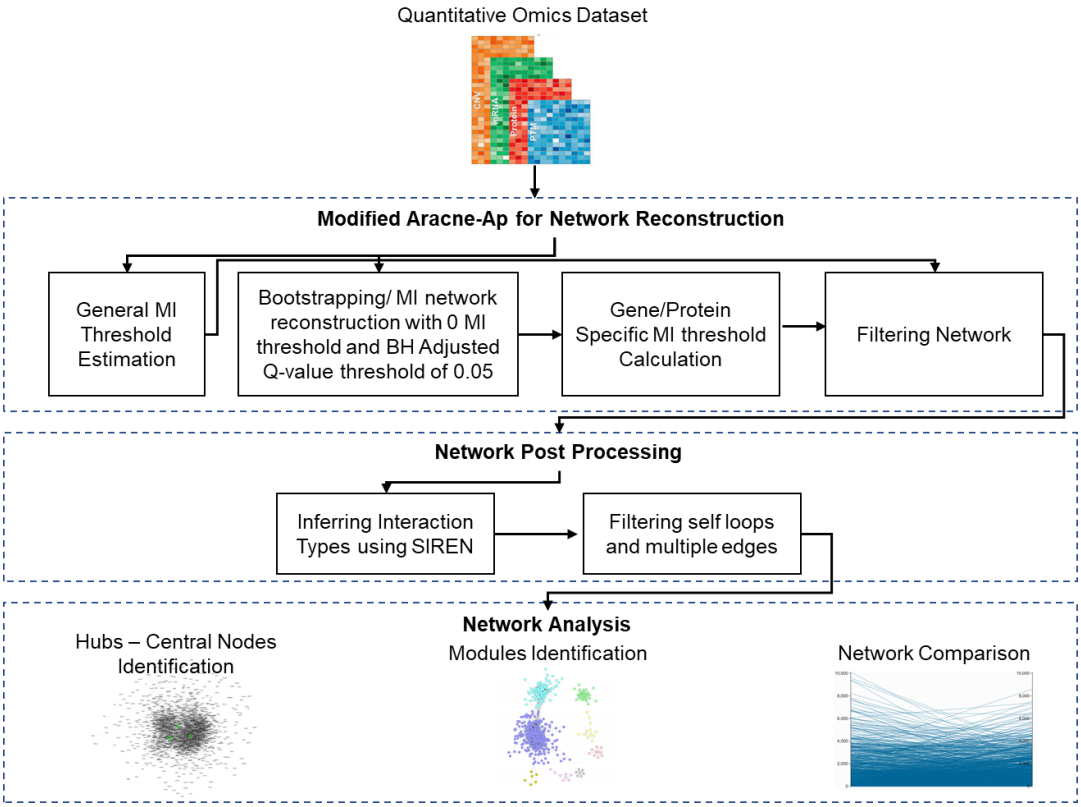


Figure 3.1 DiRec-AP network reconstruction pipeline and network analysis example.

3.2.2 Benchmark Techniques

To estimate the performance of the proposed method in reconstructing co-expression networks, we benchmarked its performance against the most widely used network reconstruction techniques, including the ones most frequently applied in cardiovascular research, namely WGCNA (72) and ARACNe-AP (75).

WGCNA (Weighted Gene Correlation Network Analysis) is one of the most widely used correlation-based methods for co-expression network reconstruction. It quantifies the interaction between pairs of genes and the degree to which these genes have identical neighbours and offers the possibility to the user to select between different correlation metrics and weighted (soft-thresholding) or unweighted (hard-thresholding) network construction (72). We run WGCNA in its default mode using

Pearson correlation and the soft thresholding mode, with soft power being identified for each dataset as the lowest power for which the scale-free topology index reached 0.90.

ARACNe-AP (75) is an information theory-based regulatory network reconstruction technique, which uses conditional mutual information and the data processing inequality theorem to infer direct regulatory relations among transcriptional regulator proteins and target genes. We run ARACNe-AP 10 times using different seeds and their default parameters, p-value threshold equal to 1E-8, and 100 reproducible bootstraps.

We complemented these two methods with a representative mathematical modelling method, namely the Sparse Estimation of high-dimensional Correlation Matrices (SEC) (74). SEC estimates a sparse correlation matrix and penalizes the correlations according to the empirical ones (larger amount of penalization to smaller empirical correlations). We run SEC with its default parameters and the correlation coefficient rho testing 4 different values, 0.2, 0.3, 0.4, and 0.5 respectively. We chose the value 0.2 for rho, for which a scale-free topology of the networks was achieved. To obtain the final networks, we transformed the correlation matrices produced as output from the SEC method into networks, by keeping the interactions of the molecules for which the correlation was not 0.

We used the networks and the evaluation method from the DREAM5 contest to evaluate the performance of the proposed method, compared to the benchmark methods. The evaluation was a binary classification task with the predicted edges being either present or absent. All interactions that were not part of the gold standard networks were considered negatives. The performance metrics used for evaluation were the area under the ROC curve (AUROC, which plots the true positive rate versus the false positive rate), the area under the precision-recall curve (AUPR), their transformation into p-values P_AUROC and P_AUPR, respectively, and an overall score. The AUROC and AUPR scores were transformed into p-values, P_AUROC and P_AUPR respectively, for the areas under the curve to gain statistical significance. More specifically, recall is defined as the number of true positives in the predicted network divided by the number of positives in the gold standard and is equivalent to

the true-positive rate, precision is defined as the number of true positives in the predicted network divided by the number of edges, and false-positive rate is defined as the number of false positives in the predicted network divided by the number of negatives in the gold standard. The AUROC is calculated by taking the integral of the ROC curve and the AUPR by taking the integral of the precision versus recall curve respectively. The AUROC and AUPR scores were transformed into p-values, P_AUROC and P_AUPR respectively, for the areas under the curve to gain statistical significance. This was achieved by sampling a random list of edges from the submitted networks of the participants and simulating a null distribution for 25000 random networks. AUROC and AUPR values were computed for the random predictions and the joint probability density function (pdf) of obtaining at the same time an (AUROC, AUPR) pair would be the probability of AUROC and AUPR being higher than the estimated ones respectively (AUROC>α_{ROC} and AUPR>α_{PR}). In the case that the actual performance of the algorithms is more accurate than random chance, the randomly obtained AUROC and AUPR values were fitted with stretched exponentials, as described in detail in Stolovitzky et al. manuscript (173). The AUROC and AUPR scores across all three gold standard networks are computed as follows:

AURC_score= $\frac{1}{3}\sum_{i=1}^{3} -log_{10}$P_AUROC(i) (1)

and

APR_score=$\frac{1}{3}\sum_{i=1}^{3} -log_{10}$P_AUPR(i). (2)

The overall score is computed by taking the mean of (1) and (2).

3.2.3 Network Analysis Techniques

The proposed method was applied to networks in carotid endarterectomy atherosclerotic samples from three different extracts, SDS, GuHCl, and NaCl. The produced networks were visualized in Cytoscape (91) (version 3.7.1) and were then analyzed with Cytoscape's built-in plugin, NetworkAnalyzer. Network clustering was performed using the MCL algorithm (108) (Cytoscape plugin v1.0) and enrichment analysis of the statistically significant revealed clusters with the DAVID web tool (174).

NetConfer web tool (125) was used to compute network characteristics and compare networks from different phenotypes. Finally, to evaluate the performance of the method in identifying physical protein-protein interaction networks and transcription factors, we used the human protein interactome of StringDB (70).

3.2.4 Data Preparation

3.2.2.1. DREAM 5 Evaluation Datasets

Three datasets and networks from the DREAM5 Network Inference Challenge (175) were used as gold standard datasets, to evaluate the performance of the proposed method. Gold standard transcriptional networks were compiled in datasets from three different microarray experiments, comprising an in silico network with gene expression profiles derived from GeneNetWeaver (176) (1643 genes over 805 measurements), a network from *Escherichia coli* expression data (4511 genes over 805 chip measurements with raw data downloaded from Gene Expression Omnibus database (177) - Platform ID: GPL199) and a network from *Saccharomyces cerevisiae* expression data (5950 genes over 536 chip measurements with raw data downloaded from Gene Expression Omnibus database - Platform ID: GPL). Microarrays were normalized to remove batch effects using Robust Multichip Averaging (178). Microarray data were also background adjusted, quantile normalized, and probe sets were summarized using median polish. The normalized expression data were also logarithmised. Control probe sets and probe sets that did not map unambiguously to one gene were removed. Moreover, when multiple probe sets were mapped to a single gene, then expression values were averaged. Known transcription factors for each organism were used to infer interactions and assess method performance. For each dataset, certain genes were picked as potential transcription factors. For E. coli the transcription factor list comprised genes defined by RegulonDB (179) (Release 6.8), and genes were identified using Gene Ontology terms. More specifically, selected genes were annotated with a biological process related to transcription (mRNA transcription or transcription, DNA dependent) and a molecular function of DNA binding or any child terms. Through these processes, a total of 334 genes were selected as potential transcription factors. For S. cerevisiae the transcription factor list comprised of genes defined by Zhu et al. (180) and genes identified using Gene

Ontology terms as described above. Through these processes, a total of 333 genes were selected as potential transcription factors. For network assessment, organism-specific gold standards were compiled, including the known transcription factor to gene interactions. For the in-silico network, the known network structure was used as the gold standard, resulting in 4012 interactions/edges. The gold standard for E. coli was compiled from RegulonDB (Release 6.8), including interactions with at least one strong evidence, resulting in 2066 interactions/edges. For S. cerevisiae gold standards from 3 different sources were tested, using the one based on the strictest thresholds from MacIsaac et al. publication (181), resulting in 3940 interactions/edges. A detailed description of the data and evaluation method, as well as the evaluation scripts and datasets, are provided in Marbach et al. manuscript (175).

3.2.2.2. Carotid Plaques Dataset

Data from 12 age and sex-matched patients undergoing carotid endarterectomies were collected. Six samples were obtained after an acute cerebrovascular event (symptomatic) and the rest six were collected during an elective surgery (asymptomatic). This dataset was generated using the three-extract strategy (182) with GuHCl (guanidine) and NaCl (salt) extracts being sent for label-free tandem mass spectrometry (MS/MS), to identify and quantify the core and soluble matrisome of carotid atherosclerotic plaques, respectively. For the GuHCl extract, 12 additional technical replicates per sample were used, resulting in a total of 24 samples, 12 symptomatic and 12 asymptomatic respectively. Label-free tandem MS/MS was used on the top 6 ions, with a mass-to-charge range of 450-1600. Data were matched to the UniProtKB database (RELEASE 2022_03) and Scaffold software (version 4.3.2, Proteome Software Inc.) was used for quantification. Identified proteins with less than 95% probability, less than two independent peptides, or precursor ion mass accuracy greater than 10ppm were not used in the analysis. 512 unique proteins were identified in the GuHCl extract and 878 unique proteins in the NaCl extract. We further filtered the identified proteins to keep only matrisome-related proteins according to a custom matrisome database, composed of extracellular proteins from the MatrisomeDB (135), apolipoproteins, and other secreted proteins, resulting in 119 consistently quantified (with less than 30% of missing values) proteins in the core matrisome

(GuHCl extract) and 204 in the soluble matrisome (NaCl extract), respectively. A detailed description of the cohort and the data is provided in Langley et al. (17), but for this manuscript, the raw data were searched again using the Proteome Discoverer software and an up-to-date UniprotKB (RELEASE 2022_03) human proteome database.

3.2.2.3. Coronary Arteries Dataset

Publicly available data were used from a cohort of 100 relatively young (age <50 years), male and female individuals of any race, without a prior diagnosis of cardiovascular disease within 48 hours of death. More specifically, this dataset was generated by running DIA MS on human tissues from the left anterior descending (LAD) coronary artery, resulting in 99 LAD tissue specimens. 12371 peptides from 2055 proteins were quantified, 1582 proteins of which were identified with less than 50% of missing values and included in the available dataset. A detailed description of the cohort, method, and dataset used is provided in the Parker et al. manuscript (134). We further filtered the protein list to keep only proteins with more than 1 peptide count for identification, resulting in a final list of 1100 proteins.

3.3 Results

3.3.1 DiRec-AP Overcomes Benchmark Methods

The efficiency of the proposed method was benchmarked against 3 state-of-the-art methods for co-expression network reconstruction described above (section 4.2.2), using 3 gold standard datasets from the DREAM5 challenge (section 4.2.2.1). Figure 3.2 displays the overall score for the performance of the 4 methods in all 3 networks. The proposed method performs better that the other methods, having an overall score of 0.83. The next best-performing method was ARACNe-AP, with an overall score of 0.26.

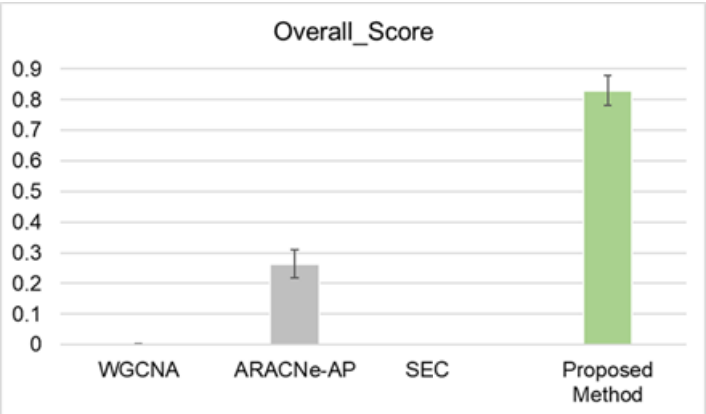


Figure 3.2 Benchmarking proposed network reconstruction method. The overall score from the three gold standard networks was computed based on the integral of the ROC and precision-recall curves for the three gold standard networks. The proposed method was compared against 3 other methods, the WGCNA, the ARACNe-AP method and the SEC method. Because of the stochastic nature of ARACNe-AP and the proposed method, we run both methods 10 times with different seeds and used the summary of all runs for computing the performance metrics.

The detailed performance of the methods in each network is shown in Table 3.1, Table 3.2, Table 3.3 and Table 3.4 below. Because of the stochastic nature of ARACNe-AP and the proposed method, we ran both methods 10 times with different seeds and used the combination of all runs (average +- standard deviation) for computing the performance metrics. SEC method was not able to compute a correlation matrix for the S. cerevisiae network, probably because of the large number of input genes and limitations of the provided implementation in Matlab. Its performance for this network was considered as 0 for the computation of the overall scores.

WGCNA

Network	Overall_Score	AUROC_Score	AUPR_Score	
	0.003080	0.000160	0.006000	
	P_AUPR	**P_AUROC**	**AUPR**	**AUROC**
In-silico	0.9999	0.9994	0.049	0.527
E. coli	0.9597	0.9996	0.039	0.531
S. cerevisiae	0.9997	0.9999	0.018	0.501

Table 3.1 Weighted Gene Coexpression Network Analysis performance metrics for each network. AUROC: Area under the ROC curve; AUPR: Area under the precision-recall; P_AUPR: Probability of AUPR; P_AUROC: Probability for AUROC.

ARACNe-AP (All Runs, Mean±SD)

Network	Overall_Score	AUROC_Score	AUPR_Score	
	0.263667±0.045236	0.472667±0.077184	0.054667±0.013317	
	P_AUPR	**P_AUROC**	**AUPR**	**AUROC**
In-silico	0.9999±1.3597E-16	0.8539±0.0151	0.0643±0.0006	0.6777±0.0006

E. coli	0.9992±0.0000	0.9978±0.0015	0.0240±4.2492E-18	0.5653±0.0015
S. cerevisiae	0.6864±0.0644	0.0494±0.0282	0.0200±0.0000	0.5163±0.0006

Table 3.2 ARACNe-AP's performance metrics for each network. Because of the stochastic nature of ARACNe-AP, we ran it 10 times with different seeds and used the summary of all runs for computing the performance metrics. SD: Standard Deviation; AUROC: Area under the ROC curve; AUPR: Area under the precision-recall; P_AUPR: Probability of AUPR; P_AUROC: Probability for AUROC.

SEC

Network	**Overall_Score**	**AUROC_Score**	**AUPR_Score**	
	0.000206	0.000217	0.000196	
	P_AUPR	**P_AUROC**	**AUPR**	**AUROC**
In-silico	0.9999	0.9994	0.0160	0.5010
E. coli	0.9992	0.9996	0.0140	0.4990
S. cerevisiae	-	-	-	-

Table 3.3 Sparse Estimation of High-dimensional Correlation Matrices' performance metrics for each network. This method was not able to compute a correlation matrix for the S. cerevisiae network, thus its performance for this network was considered as 0 for the computation of the overall scores. AUROC: Area under the ROC curve; AUPR: Area under the precision-recall; P_AUPR: Probability of AUPR; P_AUROC: Probability for AUROC.

DiRec-AP (All Runs, Mean±SD)

Network	**Overall_Score**	**AUROC_Score**	**AUPR_Score**	
	0.830000±0.047791	1.656333±0.093853	0.003333±0.001527	
	P_AUPR	**P_AUROC**	**AUPR**	**AUROC**
In-silico	0.9999±1.3597E-16	0.0760±0.0213	0.0390±0.0000	0.6930±0.0010
E. coli	0.9992±0.0000	0.1874±0.0760	0.0220±0.0010	0.5853±0.0011

S. cerevisiae	0.9794±0.0087	0.0010±0.0007	0.0190±0.0000	0.5197±0.0006

Table 3.4 DiRec-AP's performance metrics for each network. Because of the stochastic nature of DiRec-AP, we ran it 10 times with different seeds and used the summary of all runs for computing the performance metrics. SD: Standard Deviation; AUROC: Area under the ROC curve; AUPR: Area under the precision-recall; P_AUPR: Probability of AUPR; P_AUROC: Probability for AUROC.

3.3.2 Reconstructing atherosclerotic plaque networks

Since DiRec-AP outperformed the previously applied methods for reconstructing co-expression and regulatory networks for atherosclerotic samples, we then applied it to reconstruct directional networks using a previously studied carotid endarterectomy dataset (17).

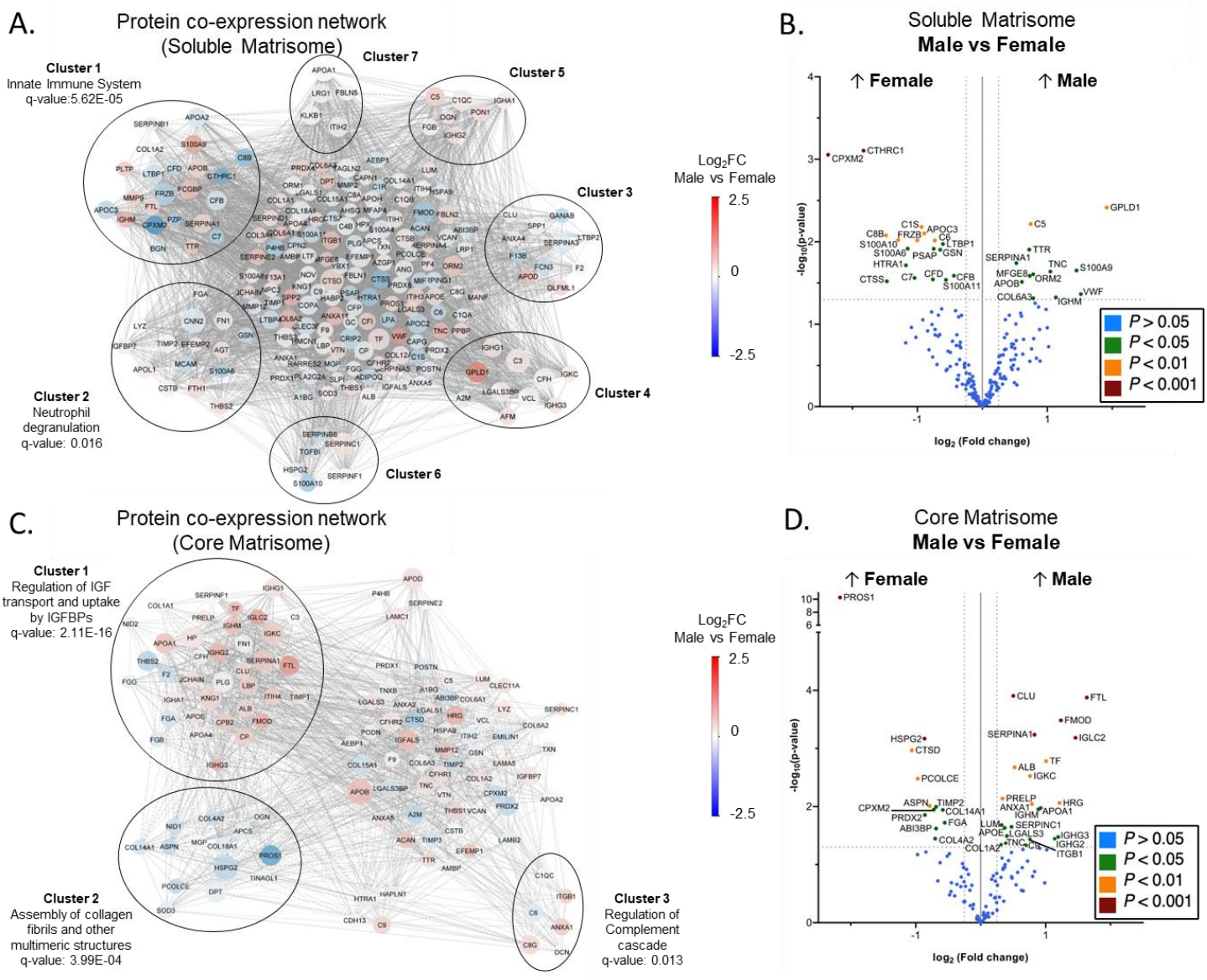


Figure 3.3 Reconstructed proteomic networks and differential expression analysis of soluble (NaCl extract) and core (GuHCl extract) matrisome of symptomatic and asymptomatic atherosclerotic carotid plaques. **A.** Protein coexpression network from the soluble matrisome of the plaques, displaying log2 fold changes for the comparisons between symptomatic vs asymptomatic carotid plaques. Edges with negative SIREN (154) scores (inhibitions) were reported with dashed lines while the width of the edge was relevant to the calculated mutual information score and the node size was relevant to the betweenness centrality. Clustering was conducted using the MCL algorithm (108) with an inflation value parameter of 3.0, using only interactions with a positive SIREN score, and a minimum number of 5 proteins per cluster. Unclustered proteins are reported without a circle. The top enriched pathway term is reported for each cluster using the DAVID web tool (174) and Benjamini Hochberg method (172) to correct for multiple testing. **B.** Volcano plot displaying the results of differential expression analysis in soluble matrisome (NaCl extract) between symptomatic (n=6) and asymptomatic carotid plaques (n=6). The Ebayes method of the limma package (183) was considered for differential expression analysis, correcting for age and sex. **C.** Protein coexpression network from the core matrisome of the plaques, displaying log2 fold changes for the comparisons between symptomatic vs asymptomatic carotid plaques. Network reconstruction was done in the same way as the **A.** panel. **D.** Volcano plot displaying the results of differential expression analysis in core matrisome (GuHCl extract) between symptomatic (n=12, 2 technical replicates for each of the 6 biological replicates) and asymptomatic carotid plaques (n=12, 2 technical replicates for each of the 6 biological replicates). The Ebayes method of the limma package (183) was considered for differential expression analysis, correcting for age.

The DiRec-AP pipeline was applied separately to reconstruct the networks of the soluble (Figure 3.3.A) and the core (Figure 3.3.B) matrisome using data from the NaCl and the GuHCl extracts respectively. Since this dataset was initially analyzed with a previous version of the human proteome definition, we re-searched the data using the latest version of the UNIPROT database (RELEASE 2022_03) and performed comparisons using the Ebayes method of the limma package, correcting for age and sex when comparing among symptomatic and asymptomatic plaques (Figure 3.3.C) and for age when comparing between plaques from males and females (Figure 3.3.D). Two of the most significantly upregulated proteins in symptomatic plaques were EGF Containing Fibulin Extracellular Matrix Protein 2 (EFEMP2) and Apolipoprotein L1 (APOL1), which were clustered together with at a cluster that was enriched in proteins involved in neutrophil degranulation (q-value: 0.016). Another important cluster was Cluster 1 including the Matrix metalloproteinase-9 (MMP9) and S100 Calcium Binding

Protein A9 (S100A9) proteins, which were part of the validated symptomatic plaques biosignature introduced by Langley et al. (17), with this cluster being significantly enriched in the Innate Immune System Reactome pathway (q-value: 5.62E-05). From the analysis of the core matrisome, two major clusters were revealed with one being enriched in the regulation of IGF transport and uptake by IGFBPs (q-value: 2.11E-16), including proteins upregulated in symptomatic plaques, and the other being downregulated in symptomatic plaques and enriched in proteins involved in the assembly of collagen fibrils and other multimeric structures (q-value: 3.99E-04). Among significant findings of our differential expression analysis, apolipoproteins B (APOB) and A-I (APOA1) were also significantly upregulated in symptomatic patients in the work of Langley et al. and asporin (ASPN) was significantly upregulated in asymptomatic patients. Network analysis agreed with differential expression analysis, with apolipoprotein B (APOB) being the most significantly upregulated protein in symptomatic plaques and among the ones with the highest betweenness centralities.

Then, comparisons between female and male plaques were conducted (Figure 3.4.B and Figure 3.4.D) and projected on the reconstructed protein networks of the soluble and the core matrisome (Figure 3.4.A and Figure 3.4.C). From the soluble matrisome, it was revealed that a signature related to inflammation (including S100A9 and APOB) was upregulated in males, while the most upregulated proteins were Collagen Triple Helix Repeat Containing 1 (CTHRC1) and Carboxypeptidase X, M14 Family Member 2 (CPXM2), that were clustered in the Innate Immune System enriched cluster. The core matrisome analysis revealed that cluster 1, which is enriched in proteins involved in the regulation of IGF transport and uptake by IGFBPs, was upregulated in male plaques, while cluster 2, which is enriched in proteins involved in the assembly of collagen and other multimeric structures, was upregulated in female plaques, with Protein S (PROS1) being both significantly upregulated in female plaques and central in the network.

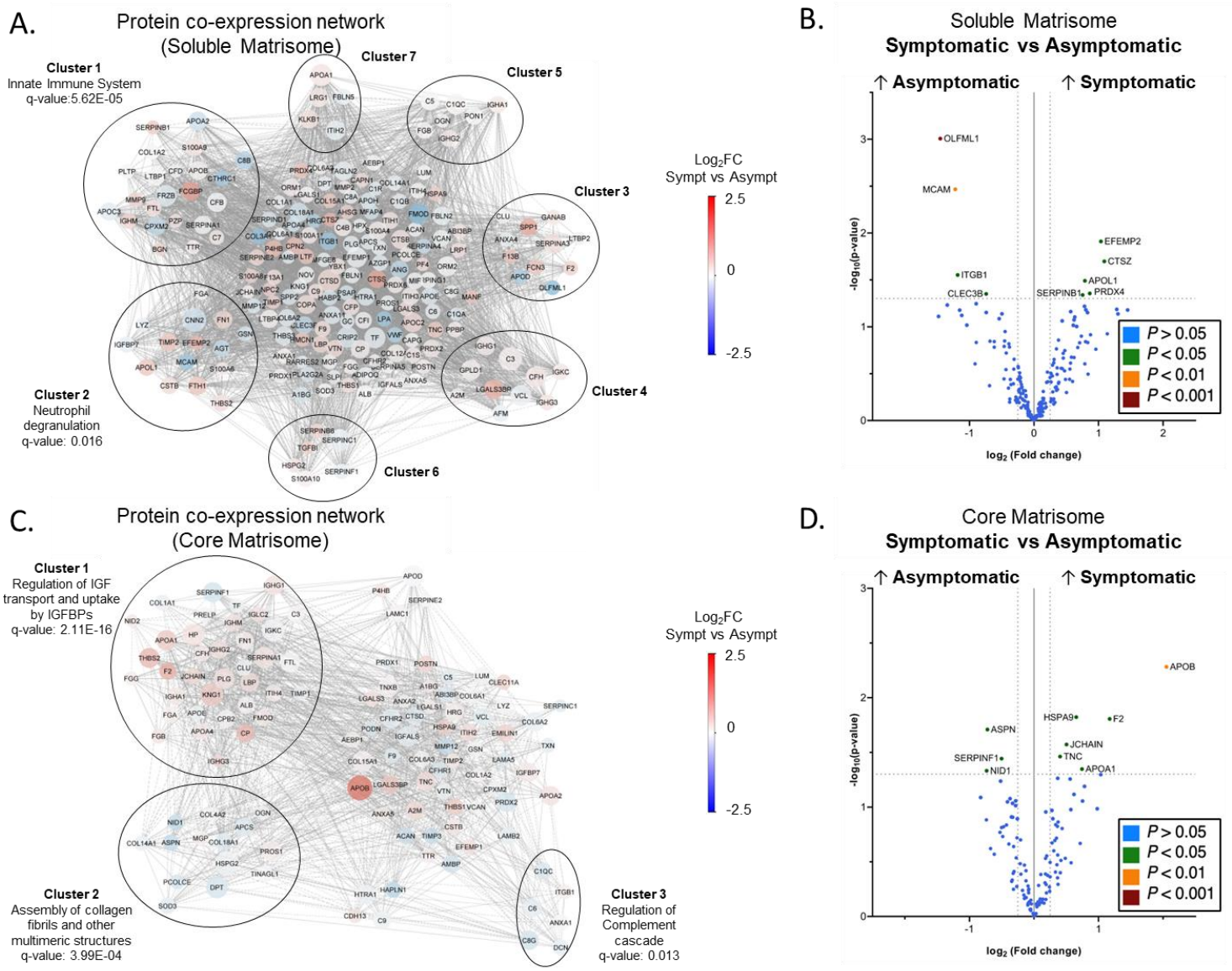


Figure 3.4 Reconstructed proteomic networks and differential expression analysis of soluble (NaCl extract) and core (GuHCl extract) matrisome of male and female atherosclerotic carotid plaques. **A.** Protein coexpression network from the soluble matrisome of the plaques, displaying log2 fold changes for the comparisons between male vs female carotid plaques. Edges with negative SIREN (154) scores (inhibitions) were reported with dashed lines while the width of the edge was relevant to the calculated mutual information score and the node size was relevant to the betweenness centrality. Clustering was conducted using the MCL algorithm (108) with an inflation value parameter of 3.0, using only interactions with a positive SIREN score, and a minimum number of 5 proteins per cluster. Unclustered proteins are reported without a circle. The top enriched pathway term is reported for each cluster using the DAVID web tool (174) and Benjamini Hochberg method (172) to correct for multiple testing. **B.** Volcano plot displaying the results of differential expression analysis in soluble matrisome (NaCl extract) between male (n=6) and female carotid plaques (n=6). The Ebayes method of the limma package (183) was considered for differential expression analysis, correcting for age and sex. **C.** Protein coexpression network from

the core matrisome of the plaques, displaying log2 fold changes for the comparisons between symptomatic vs asymptomatic carotid plaques. Network reconstruction was done in the same way as the **A.** panel. **D.** Volcano plot displaying the results of differential expression analysis in core matrisome (GuHCl extract) between male (n=12, 2 technical replicates for each of the 6 biological replicates) and female carotid plaques (n=12, 2 technical replicates for each of the 6 biological replicates). The Ebayes method of the limma package (183) was considered for differential expression analysis, correcting for age.

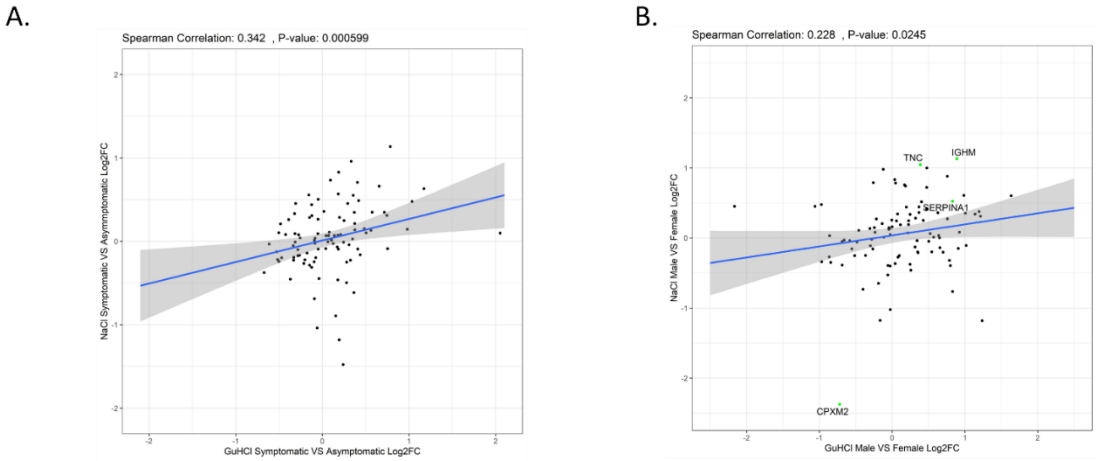
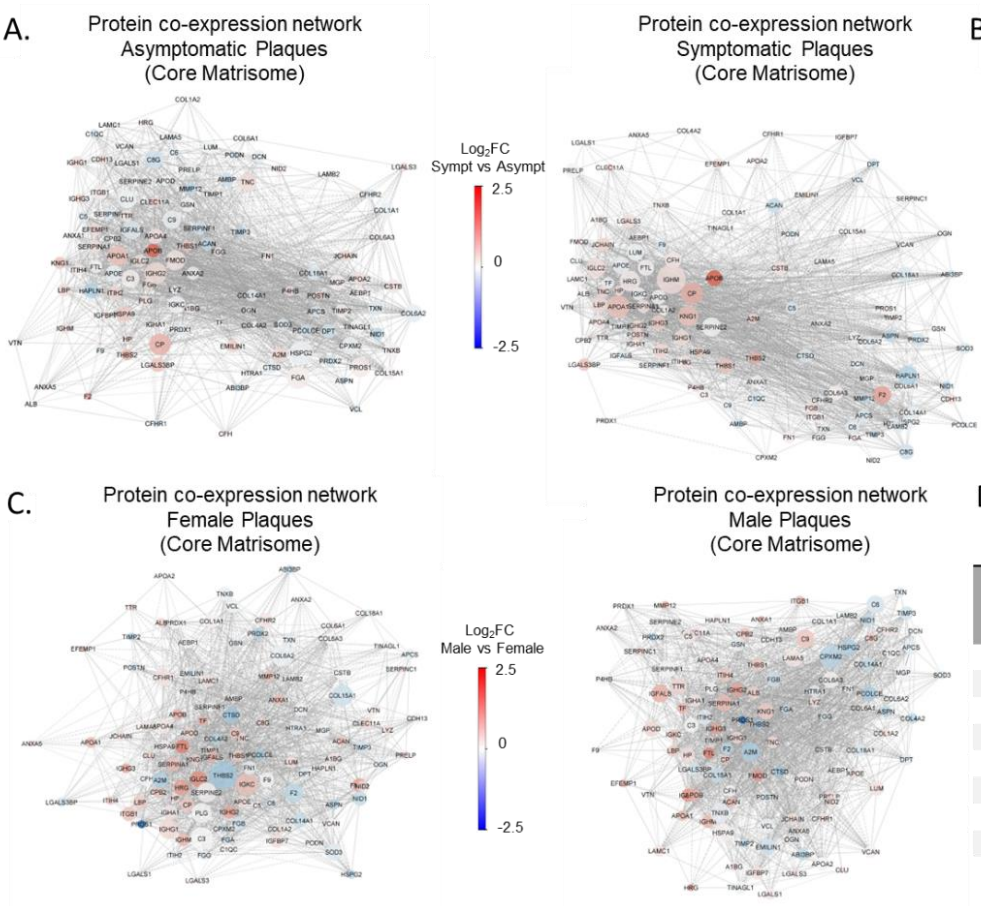


Figure 3.5 Correlation of fold changes between A. symptomatic vs asymptomatic and B. male vs female carotid plaques in core (GuHCl) and soluble (NaCl) matrisome. Scatterplots depicting the log2 fold changes, Spearman correlation and the corresponding p-value. Significant proteins in both comparisons are colored green and labelled while the rest are colored black. Linear regression plots are plotted to depict their 95% confidence intervals. Ebayes method of the limma package (183) was considered for differential expression analysis, correcting for **A.** age and sex and **B.** age. Q-values were calculated correcting for multiple testing using the Benjamini-Hochberg method (172). TNC: Tenascin; IGHM: Immunoglobulin heavy constant mu; SERPINA1: Alpha-1-antitrypsin; CPXM2: Inactive carboxypeptidase-like protein X2.

The correlation of the fold changes in symptomatic versus asymptomatic and male versus female comparisons between core and soluble matrisome was assessed using Spearman's correlation analyses (Figure 3.5) and a significant but weak correlation in both comparisons between the two extracts were revealed. The fact that the correlation was not as high as expected is possible due to the small sample size (n=12) of the soluble matrisome, but it is also partially expected because the two extracts are complementary measuring proteins with different solubility of the matrisome.



Gene Name	BC Asymptomatic Plaques	BC Symptomatic Plaques	Log2FC of BC in Symptomatic vs Asymptomatic Plaques
F2	0.157	1.200	2.933
COL6A1	0.089	0.656	2.890
CFH	0.183	1.080	2.563
LAMB2	0.137	0.760	2.470
GSN	0.826	0.134	-2.623
CPXM2	1.066	0.146	-2.872
TXN	0.884	0.092	-3.269
PCOLCE	0.606	0.055	-3.452
PRDX1	0.529	0.044	-3.581

Gene Name	BC Female Plaques	BC Male Plaques	Log2FC of BC in Male vs Female Plaques
APOA2	0.061	0.773	3.660
CPXM2	0.158	1.606	3.345
P4HB	0.102	0.729	2.842
LGALS1	0.103	0.584	2.500
TTR	0.196	1.082	2.468
HRG	1.028	0.146	-2.819
CLU	0.271	0.014	-4.320
F9	0.928	0.024	-5.258

Figure 3.6 Network comparison of symptomatic/asymptomatic and male/female core (GuHCl extract) matrisome networks. A. Reconstructed core matrisome (GuHCl extract) networks of symptomatic and asymptomatic patients. **B.** Proteins with statistically significant (p-value<0.05) change in the betweenness centrality in the asymptomatic and symptomatic plaque networks. **C.** Reconstructed core matrisome (GuHCl extract) networks of male vs female patients. **D.** Proteins with statistically significant (p-value<0.05) change in the betweenness centrality in the female and the male plaque networks. Edges with negative SIREN scores (inhibitions) were reported with dashed lines while the edges' width was relevant to the calculated mutual information score and the node size was relevant to the betweenness centrality. Networks were visualized using the edge-weighted spring-embedded layout of Cytoscape, based on the calculated SIREN scores.

For core matrisome, 24 samples were available since the analysis was conducted in two technical replicates for each biological replicate. The analysis on the soluble matrisome was based on 12 samples. Thus, phenotype-specific networks were then reconstructed for asymptomatic and symptomatic (Figure 3.6.A), female, and male (Figure 3.6.C) plaques for the core matrisome, as the number of samples was acceptable for network reconstruction. From the comparison of these networks using the approach suggested in Nagpal et al. (125) and Theofilatos et al. (166) manuscripts, it was found that the role of 9 proteins, measured by the betweenness centrality, changed significantly when comparing female and male plaques (Figure 3.6.B). Among

these proteins, Coagulation Factor II (F2) was the one with the most increased centrality in the symptomatic plaque network (log2 fold change: 2.933) and it was also significantly upregulated in symptomatic plaques. From the comparison of male vs female reconstructed networks, 8 proteins were revealed to have a statistically significantly different role in the sex-specific networks (Figure 3.6.D). Among these proteins were CPXM2 and Transthyretin (TTR), which were also significantly upregulated and downregulated in male plaques, respectively.

3.3.3 Differences between symptomatic carotid plaque and CAD plaque protein-protein interaction networks

To explore which differentially expressed and network biosignatures are carotid plaque-specific or reproducible in other types of atherosclerotic samples, we compared the reconstructed core matrisome network of carotid plaques (Figure 3.7.A) with a reconstructed network using the LAD coronary artery proteomic dataset (Figure 3.7.B). Both networks revealed a close association of major serum proteins with other secreted factors, apolipoproteins, and matricellular proteins, while collagens were mostly interacting with structural ECM proteins, and membrane-associated ECM receptors and proteases seemed to be associated in both groups of proteins. This separation was also confirmed by a physical protein-protein interaction network (Figure 3.7.C). Comparison of the carotid plaques and the LAD coronary artery protein networks revealed 8 proteins, IGKC, SERPINA1, Dermatopontin (DPT), Complement Factor H Related 1 (CFHR1), Podocan (PODN), Proline and Arginine Rich End Leucine Rich Repeat Protein (PRELP), Peroxiredoxin 1 (PRDX1) and Immunoglobulin Heavy Constant Gamma 3 (IGHG3), whose role in the network changed between the two networks (statistically significantly changing betweenness centralities).

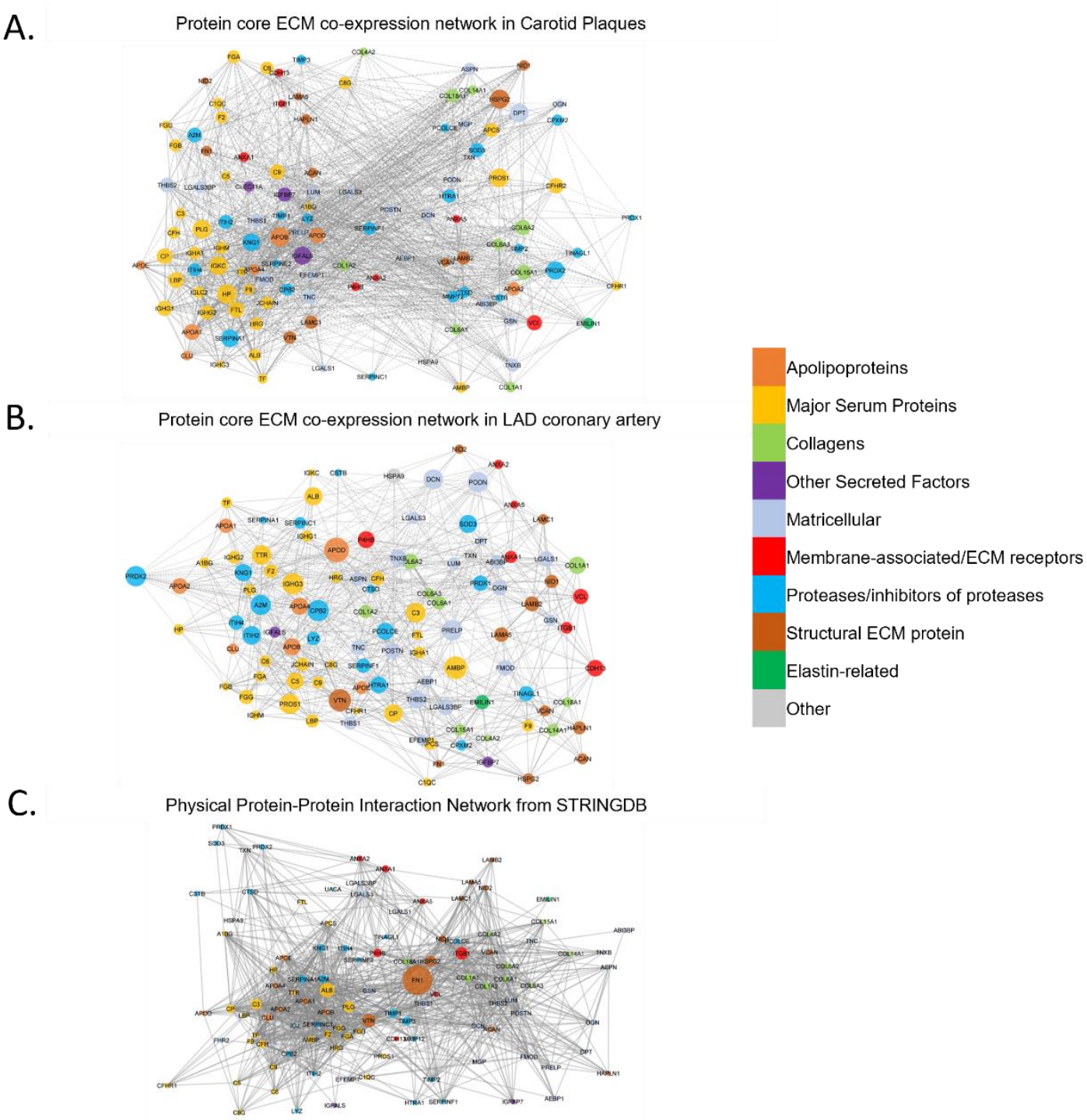


Figure 3.7 Reconstructed networks for carotid plaques, LAD coronary artery samples, and physical protein-protein interactions. **A.** Reconstructed core matrisome (GuHCl extract) networks for carotid plaques. **B.** Reconstructed core matrisome (GuHCl extract) networks for LAD coronary arteries. **C.** Physical PPI network for core matrisome proteins mined from StringDB (70). For all networks, edges with negative SIREN (154) scores (inhibitions) were reported with dashed lines while the width of the edge was relevant to the calculated mutual information score and the node size was relevant to the betweenness centrality. Networks were visualized using the edge-weighted spring-embedded layout of Cytoscape based on the calculated SIREN scores

for the reconstructed co-expression networks and based on the summary StringDB interaction score for the PPI network. ECM proteins were colored based on their functional category.

When assessing the correlation of fold changes between carotid plaques and coronary arteries in the sex comparison (Figure 3.8), we observed that there was no agreement in sex changes, suggesting that postmortem changes in autopsy samples may have erased sex changes.

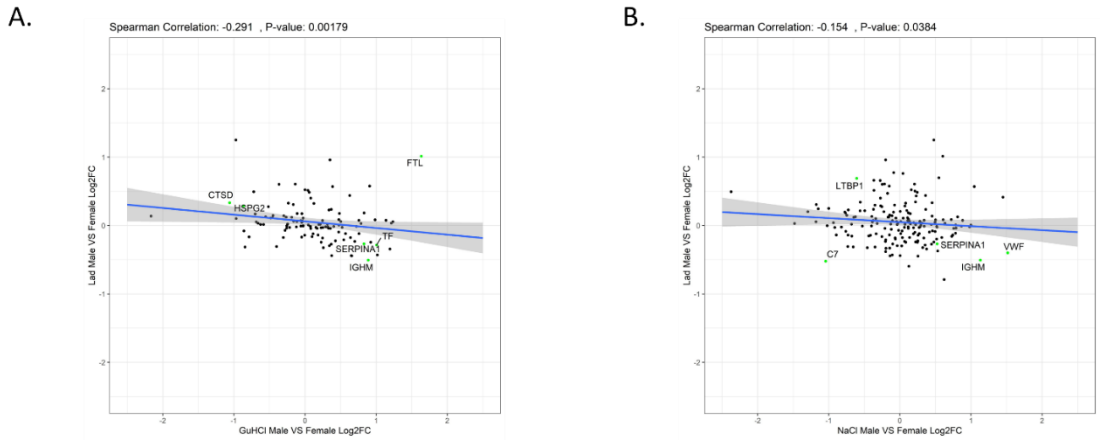


Figure 3.8 Correlation of fold changes between male vs female comparison in A. core (GuHCl) matrisome of carotid plaques and coronary arteries and B. soluble (NaCl) matrisome of carotid plaques and coronary arteries. Scatterplots depicting the log2 fold changes, Spearman correlation and the corresponding p-value. Significant proteins in both comparisons are colored green and labelled while the rest are colored black. Linear regression plots are plotted to depict their 95% confidence intervals. Ebayes method of the limma package (183) was considered for differential expression analysis, correcting for age. Q-values were calculated correcting for multiple testing using the Benjamini-Hochberg method (172). FTL: Ferritin light chain; CTSD: Cathepsin D; HSPG2: Basement membrane-specific heparan sulfate proteoglycan core protein; SERPINA1: Alpha-1-antitrypsin; TF: Tissue Factor; IGHM: Immunoglobulin heavy constant mu; LTBP1: Latent-transforming growth factor beta-binding protein 1; C7: Complement Component C7; VWF: von Willebrand factor.

3.4 Discussion

Network reconstruction is a critical component of quantitative omics data analysis, as it allows studying the interactions and regulatory associations among genes, transcripts, and proteins. In the context of the present manuscript, we introduced the DiRec-AP pipeline, to overcome the inherent limitations of existing solutions, such as the use of the same association threshold for all proteins in inferring positive interactions, the lack of directionality in the reconstructed networks, and the lack of

ability to locate and incorporate negative associations and interactions among genes and proteins. DiRec-AP was compared against representative network reconstruction methods, including the most widely used pipeline (WGCNA), the best-performing conditional mutual information-based method (ARACNe-AP), and a recently reported mathematical modelling-based method (SEC). Comparisons using the DREAM 5 challenge golden-standard datasets confirmed that the technical advantages of the DiRec-AP pipeline allowed the reconstruction of the networks with fewer false positives and false negatives, improving the produced metrics substantially.

The ability of the method to reconstruct directed protein regulatory networks allowed the reconstruction of directed protein networks for the extracellular matrisome of atherosclerotic plaques. To our knowledge, this is the first time that such a network is reconstructed with such large coverage since previous attempts either put the emphasis on serum proteome (162) or focus on transcriptomic networks (59). The reconstructed proteomic networks for the soluble and core matrisome, as well as for sex- and phenotype-specific networks, can be utilized to generate new hypotheses on regulatory relationships between matrisome proteins and to screen for potential novel drug targets through the identification of network-specific proteins (33, 184).

The analyses of the reconstructed networks and the comparison of the symptomatic and asymptomatic networks confirmed that the blood coagulation pathway is significantly altered in symptomatic compared to asymptomatic plaques, with the centrality of thrombin being significantly increased. The role of thrombin in atherogenesis has been extensively studied (185), and this protein is among the proteins that have been targeted by drugs for atherosclerosis treatment (186, 187).

The role of the innate immune system, neutrophil degranulation, regulation of IGF transport and uptake by IGFBPs, and assembly of collagen fibrils and other multimeric structures pathways in symptoms as well as in sex-specific changes, was uncovered from the analyses of the reconstructed soluble and core matrisome networks. As opposed to the role of the innate immune system in atherosclerosis and plaque symptoms, which has been extensively studied and reported (17), the role of the other reported pathways was neglected in recent research.

The comparison between male and female-specific networks revealed that CPXM2 has a substantially central role in female networks and is upregulated in female plaques, while Transthyretin is substantially more central in male plaques and upregulated in male plaques. Despite TTR being recently introduced as a new biomarker for insulin resistance (188) and found to be negatively associated with intima thickening in carotid plaques (189), more research is required to completely understand its exact mechanisms of actions in atherosclerosis, and how these mechanisms are differentiated between males and females.

The comparison between the reconstructed matrisome protein networks from carotid endarterectomy samples and LAD coronary artery samples showed agreement in the global structure of the network, but three proteins (IGKC, SERPINA1, DPT) were specific to the carotid plaque network, while five proteins were specific to the LAD coronary artery network (CFHR1, PODN, PRELP, PRDX1, IGHG3), something that was also observed in the correlation of protein changes between the two sample types. However, LAD samples were from autopsies. Generalizing conclusions for atherosclerosis from autopsy samples may lead to erroneous conclusions due to postmortem changes in plaque proteome.

4. Automatic optimization of targeted MS proteomics data processing pipeline, using a multi-objective evolutionary algorithm

4.1 Introduction

MS-based proteomics is widely used as a discovery technique to identify and quantify proteins of interest in biosamples. Discovery approaches for MS have the potential to detect a large number of proteins but their quantification measurements are prone to errors (35) and require further validation. For this reason, various targeted proteomics methods have been developed for the reproducible quantitative analysis of the proteome in biosamples (190).

Targeted proteomics techniques, such as SRM and PRM, have the disadvantage of producing a complex output, which requires elaborate signal pre-processing and quality control analysis to achieve accurate quantification values. Several software solutions have been developed including Skyline (191), Picky (192), TRIC (193), and TargetedMSQC (194) for performing data preprocessing and optimizing parts of this analysis, such as alignment and quality control.

DIA combines the wide protein coverage of Data Dependent Acquisition (DDA) and the accuracy, reproducibility, and consistency of targeted analysis (43). Its major limitation is the computational complexity of the fragmented spectra acquired. The software tools developed for the analysis of these data can be split into two major categories about spectral library presence. Firstly, several tools have been introduced for data preprocessing of DIA data, including tools for generating libraries, such as MaxQuant (195), library-based analysis, such as Spectronaut (196), and spectral library-free analysis, such as DIA-Umpire (197). Additionally, tools have been introduced to optimize the processing of DIA data, including the cloud-based tool QCloud (198), which performs instrument performance quality assessment and evaluation, Avant-garde (199), which optimizes the transitions and peak boundaries and minimizes the FDR during quantification, optimization of conditions or MS acquisitions (200), and MSstatsQC (201) for statistical quality control of the data.

However, to the best of our knowledge, until now no method exists to optimize the complete processing pipeline in an unbiased manner.

Evolutionary optimization algorithms have been used on protein datasets for several different applications, one of the most widely used being the prediction of protein structure, from the secondary (202) to 3D (203, 204) (tertiary) protein structure. Further applications are the discovery of protein biomarkers (205, 206), disease prediction (207), optimization of processing of large-scale protein datasets (208) and functional module detection in protein-protein interaction networks (209). Optimization algorithms have also been used to combine targeted and DIA proteomics by trying to maximize the quantification agreement between the two methods (210). For DIA proteomics, genetic algorithms have been used for the selection and refinement of peak transitions (199).

In the present chapter, we introduce the Heuristically Optimized Targeted Proteomics (HOpTar-omics) tool, a novel multi-objective optimization framework based on a Pareto-based Evolutionary Optimization Algorithm, which is designed for the optimization of the preprocessing pipeline of three different types of targeted proteomics datasets: SRM, PRM, and DIA. The proposed optimization framework was applied to three different datasets and outperformed existing methods for processing targeted proteomics data.

4.2 Methods

4.2.1 HOpTAr-omics tool

The proposed tool is a multi-objective evolutionary algorithm (MOEA) method to optimize parameters and variables for each step of the targeted proteomics preprocessing pipelines. The utilized algorithm is a Pareto-optimization technique since its selection process is driven by organizing solutions to non-dominated fronts and assigning close fitness values to solutions belonging to the same front.

The iterative process of the optimization framework begins by initializing a set of solutions. The composition of every solution depends on the targeted proteomics data

type and is presented in Table 4.1 below. The first population of solutions is generated by randomizing values considering the normal distribution for each variable.

Method	**Range of Values**	**PRM**	**PRM heavy**	**DIA**	**DIA heavy**
S/N Threshold	3-7			**X**	**X**
Q-value Threshold	0.001-0.05			**X**	**X**
Missing Threshold	0.1-0.5		**X**	**X**	**X**
Correlation Method	Spearman/Pearson	**X**	**X**	**X**	**X**
Correlation threshold	0.3-0.9	**X**	**X**	**X**	**X**
Normalisation Method	TII/iRT normalisation	**X**		**X**	
Protein Quantification Method	Sum/mean/median of peptides	**X**	**X**	**X**	**X**
K-number	3-20	**X**	**X**	**X**	**X**
Truncated Peaks Threshold	3-7	**X**	**X**		

Table 4.1 Optimization variables, range of values, and their applicability depending on the deployed proteomics method. S/N: signal-to-noise ratio; TII: Total Ion Intensity; iRT: indexed Retention Time.

The optimization goals that were formulated as fitness functions are provided in Table 4.2. These goals were selected to maximize the reproducibility of the analysis in comparison with additional data from other protein quantification methods (e.g. clinical laboratory-based methods).

Fitness Function	**Definition**
Fitness Function 1	FF1 = 1/(1+#Missing_Values) #Missing_Values: total number of missing values
Fitness Function 2	FF2 = 1/(1+#Values_Below_Limit_Of_Detection)

	#Values_Below_Limit_Of_Detection: total number of quantification values
Fitness Function 3	FF3 = (Average_Values_Correlation_Coefficient +1)/2 Average_Values_Correlation_Coefficient: the average correlation of each protein quantification value with the ones measured with an additional method (clinical or another proteomics method)
Fitness Function 4 *(for more than 20 proteins)*	FF4 = (Average_Log2FC_Correlation_Coefficient +1)/2 Average_Log2FC_Correlation_Coefficient: the average correlation of the log2 fold change in the deployed statistical comparison using each protein's quantification values versus the log2 fold changes calculated using the values measured with an additional method (clinical or another proteomics method)
Fitness Function 4 *(for less than 20 proteins)*	FF4 = 1/(Log2FC_Average_Absolute_Error +1) Log2FC_Average_Absolute_Error: the average absolute error of the log2 fold change in the deployed statistical comparison using each protein's quantification values versus the log2 fold changes calculated using the values measured with an additional method (clinical or another proteomics method)
Fitness Function 5	FF5 = 1/(1+#Outliers) #Outliers: total number of outliers detected using the Bland-Altman method (211) to compare the quantification values against the values of an additional protein quantification method

Table 4.2 Fitness Functions of the proposed HOpTar-omics tool.

The utilized fitness functions aim to optimize data quality and their correlation with untargeted MS or clinical measurements.

After the evaluation of the population, the Pareto fronts of non-dominated solutions are calculated, and solutions are assigned a fitness value based on their Pareto front. The Roulette Wheel Selection method is applied to generate a new population of solutions which are then differentiated using the Genetic Algorithms two-point crossover and Gaussian Mutation Operators. The new population is evaluated, and this iterative process continues until the algorithm converges or reaches the maximum number of generations (200). A population size of 50 was used for all experiments. Regarding the convergence criterion, the population is considered to converge when the minimum distance of the weighted sum of the fitness values of the population is less than 5% from the weighted sum of the fitness values of the best individual solution. A two-point-crossover probability of 90% was used while a dynamic mutation operator was applied to promote better exploration of the search space in initial iterations and better exploitation of the search space as the optimization process continues.

4.2.2 Benchmark methods and tools for targeted and DIA MS analysis

Skyline (191) is the most widely used software for the analysis of targeted MS experiments. Its main objective is to generate MS methods and analyze the data collected from chromatography for quantitation, but it also gives users the ability to integrate with the analysis their custom analysis tools via an external tools framework (212). Skyline ecosystem offers additional software integrated with Skyline to facilitate the analysis. Repositories, such as Panorama (213) or Panorama Public (214) and CHORUS, for sharing Skyline experimental results and raw MS files (215) respectively, allow interlaboratory collaborations. Tools for retention time prediction, such as SSRCalc (216) or indexed retention time (iRT) peptides method (217), are also incorporated. Furthermore, prediction algorithms such as PREGO (218), which predicts peptides with the most intense MS signal using an artificial neural network, are implemented as a plug-in for peptide selection. Statistical tools, such as MSstats (201) and SProCop (219), assess certain metrics to ensure reproducibility across runs.

The result of Skyline's processing is a calculated peak area for each peptide ion, a report with several metrics and information for the proteins and peptides detected as well as a visualization of the data.

Spectronaut (196) is a commercial software designed by Biognosys to process and analyze DIA acquisition proteomics experiments. It processes the data using a similar strategy to targeted analysis, predicts retention time based on iRT (217) and uses mProphet (220) scoring. Spectronaut supports both library-driven and direct DIA analysis, post-translational modification analysis, post-analysis such as differential abundance and enrichment analysis, and visualization of the data. The analysis results can be exported in a reporting schema with metrics and information about the proteins and peptides detected. The analysis software of Biognosys for targeted proteomics data is SpectroDive (221). The methods and algorithms used for the processing and analysis of the data are similar to these of Spectronaut.

DIA-NN (222) is a tool for processing complex DIA data. It also includes both spectra library-based and library-free analysis modes and provides a graphical interface and a command-line tool. The retention time alignment is done using endogenous peptides and the retention time window is determined automatically. Each elution peak is scored based on its characteristics, the best peak per precursor is selected, potentially interfering peptides are detected and removed and the precursor q-values to distinguish real signals from noise are calculated based on deep neural networks. The protein inference and quantification results are reported in a text format report.

The outputs of Skyline or Spectrodive were used either as inputs to the HoPTar-omics tool or for benchmarking the algorithm. Skyline's and Spectronaut's exports were used as inputs for the HoPTar-omics tool for targeted and DIA proteomics data respectively. For DIA proteomics data Spectronaut's pipeline export and DIA-NN's export were used to compute the examined fitness values and compare the results with the proposed algorithm's results. For targeted proteomics (with or without heavy standards) Skyline's pipeline export and Spectrodive's export were used for the comparative results. The identity of a specific peptide was confirmed by the presence of multiple transitions at the same retention time. The total fragment peak areas were used for

quantification. Peptides for reported proteins were manually checked to ensure accurate peak integration across all samples. For the targeted workflow there were cases where the scheduled acquisition window did not capture the entire elution profile of a peptide peak. If any integrated peak had one of its boundaries at either terminal point of the chromatogram and the intensity at that end was greater than 1% of the peak height higher than the intensity at the other integration boundary, this peak was marked as truncated by Skyline. Truncated peaks were marked as missing values, and their imputed values were compared to the initial Skyline-exported value. In case the imputed value was smaller, the initial value was kept. In the case of DIA and targeted proteomics with heavy standards, final quantitative comparisons were conducted using the light/heavy peptide abundance ratio.

4.2.3 Targeted Proteomics PRM data: In-solution protein digestion

10 µL of inactivated serum or plasma were denatured by the addition of urea (final concentration 7.2 M) and reduced using dithiothreitol (final concentration 5 mM) for 1 h at 37 °C and shaking at 180 rpm. Reduced proteins were cooled to room temperature before being alkylated in the dark for 1 h using iodoacetamide (final concentration 25 mM). An aliquot equivalent to 40 µg of alkylated protein was added to a 0.1 M triethylammonium bicarbonate solution (pH 8.2) and digested for 18 h at 37 °C, shaking at 180 rpm using 1.6 µg of Trypsin/LysC (Promega, V5072). Digested peptide solutions were acidified using trifluoroacetic acid (TFA, final concentration 1%).

4.2.4 Targeted Proteomics PRM data: Peptide clean-up and stable isotope-labelled standard (SIS) spike-in

Peptide clean-up was achieved using a Bravo AssayMAP Liquid Handling Platform (Agilent). After conditioning and equilibration of the resin, acidified peptide solutions were loaded onto AssayMAP C18 Cartridges (Agilent, 5190-6532), washed using 1% acetonitrile (ACN), 0.1% TFA (aq), and eluted using 70% ACN, 0.1% TFA (aq). Eluted peptides were vacuum centrifuged (Thermo Scientific, Savant SPD131DDA) to dry and resuspended in 40 µL of 2% ACN, 0.05% TFA (aq). For clinical cohort analysis, 6 µL of cleaned peptide solution was added to two injection equivalents of PQ500 SIS mix

(Biognosys, Ki-3019-96) (223) using a Bravo Liquid Handling Platform (Agilent).4.2.5 DIA–MS analysis.

Peptides were analyzed using high-performance liquid chromatography (HPLC)–MS assembly consisting of an UltiMate 3000 HPLC system (Thermo Scientific) which was equipped with a capillary flow selector and coupled via an EASY-Spray NG Source (Thermo Scientific) to an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Scientific). To generate DIA data for serum samples (GSTT COVID-19 ICU cohort), peptides were injected onto a C18 trap cartridge (Thermo Scientific, 160454) at a flow rate of 25 µL/min for 1 min, using 0.1% formic acid (FA, aq). The initial capillary flow rate was reduced from 3 to 1.2 µL/min in 1 min at 1% B. Peptides were then eluted from the trap cartridge and separated on an analytical column (Thermo Scientific, ES806A, at 50 °C) using the following gradient: 1–11 min, 1–5% B; 11–32 min, 5–18% B; 32–52 min, 18–40% B; 52–52.1 min, 40–99% B; 52.1–58 min, 99% B. The flow rate was increased to 3 µL/min and the column was washed using the following gradient: 58–58.1 min, 99–1% B; 58.1–59.9 min, 1–99% B; 59.9–60 min, 99–1% B. Finally, the column was equilibrated at 1% B for 6 min. In all HPLC-DIA-MS analyses, mobile phase A was 0.1% FA (aq) and mobile phase B was 80% ACN, 0.1% FA (aq). Precursor MS1 spectra were acquired using Orbitrap detection (resolution 60000 at 200 m/z, scan range 329–1201 m/z). Quadrupole isolation was used to sequentially scan 30 precursor m/z windows of variable width. Per isolation window, semi-targeted Orbitrap MS2 spectra (resolution 30000 at 200 m/z) were collected following higher-energy C-trap dissociation.

4.2.6 MS database search for DIA–MS analysis

PQ500 (223) SIS-spiked DIA data from all serum samples of the GSTT COVID-19 ICU cohort were analyzed in Spectronaut v14 (Biognosys AG), using the provided PQ500 analysis plug-in. MS1 and MS2 mass tolerance strategies were set to dynamic. Retention time calibration was achieved using the spiked iRT peptides included in the PQ500 SIS mix. The precursor and protein *Q*-value cutoff was set to 0.01. Quantification was conducted at an MS2 level using peak areas and individual runs were normalized using the global strategy set to the median. Peptides for reported

proteins were manually checked to ensure accurate peak integration across selected samples.

4.3 Results

4.3.1 Materials and Datasets.

Three datasets were used for the evaluation of the proposed MS data preprocessing method, including, a PRM dataset from carotid endarterectomy plaque samples; a PRM dataset from the same samples but including heavy labelled peptide standards; and a DIA dataset for plasma samples for Covid-19 patients.

The first case study (Case Study 1: PRM Proteomics of Atherosclerotic Plaques) was based on PRM data from human carotid plaques. The raw dataset used in this report comes from carotid endarterectomies from 120 patients (chapter 5). Carotid endarterectomy specimens have been obtained from two areas of the plaque, area A (the core of the plaque) and area B (the periphery of the plaque). More than 500 extracellular proteins were identified using label-free discovery proteomics and 129 proteins were quantified from 190 samples using targeted PRM proteomics and the GuHCl extract, which contains the core matrisome, using the three-step protein extraction, as previously described (182). Proteotypic peptides of ECM proteins were selected using label-free data. Information on the peptides used for quantification for each one of these proteins is provided in Supplemental Table 1 (https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis).

The second case study (Case Study 2: PRM Proteomics in Atherosclerotic Plaques with injected heavy standards) involved PRM proteomics and heavy standards of the human carotid plaque samples also used for the first case study. Apolipoproteins (total of 22) and CFAH were quantified in this dataset. Information about peptides used for the quantification of each of these proteins is shown in Supplemental Table 2 (https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis).

The final case study (Case Study 3: DIA proteomics for Covid-19 patients serum samples) involved the analysis of a previously published raw DIA dataset from serum samples of Covid-19 patients (224). Measurements of 62 blood serum samples from

hospitalized covid patients were used for this case study. Information about the peptides used for quantification for each one of these proteins is presented in Supplemental Table 3 (https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis).

The raw MS proteomics data of all three deployed datasets have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository. In particular, the targeted proteomics data of carotid plaques are available with the identifiers PXD031052 and the DIA proteomics data of plasma samples from Covid-19 patients with the identifiers PXD024026 and PXD024089.

4.3.2 Benchmark models and methods for Targeted and DIA data processing.

Processing targeted or DIA MS data involves the use of software packages responsible to perform peak detection in a semi-automatic manner, peak alignment and quality control steps to identify and quantify peptides, and then combining them for protein quantification. A combination of established open-source and commercial software solutions was used to compare their performance with the proposed method in the three case studies. Skyline (191) is the most widely used software for the analysis of targeted MS experiments and therefore it was used for comparison against the proposed method in case studies 1 and 2. The result of Skyline's processing is a calculated peak area for each peptide ion, a report with several metrics and information for the proteins and peptides detected, as well as a visualization of the data. Spectronaut (196) is a commercial software designed by Biognosys to process and analyze DIA proteomics experiments, and version v14 of Spectronaut was used for processing DIA data in case study 3. Injections (total of 2) of heavy standards were conducted using the PQ500 SIS mix (Biognosys, Ki-3019-96) (223) to allow for absolute quantification of peptides and proteins. The analysis software of Biognosys for targeted proteomics data is SpectroDive (221), and this was used for processing the data of case studies 1 and 2 and comparison against the proposed method. The methods and algorithms used for the processing and analysis of the data are similar to the ones by Spectronaut. DIA-NN (222) is an open-source tool for processing complex DIA data which was also used for comparative analysis in case study 3.

4.3.3 HOpTar-omics Tool Outline.

Figure 4.1 describes the streamlined processing pipeline for targeted MS data which was used in the implementation of the HOpTar-omics tool. This pipeline starts by taking the result of Skyline software as input for PRM data and Spectronaut for DIA data, with an optional step of manually inspecting samples and peptides based on the initial q-values in the case of DIA. The next steps involve filtering peptide quantification values based on signal-to-noise (S/N) ratio, q-value and/or truncated peaks thresholds, marking missing values and filtering peptides based on their missingness frequency. Next, missing values are imputed using the KNN-impute method and a further peptide filtering step is applied to filter out peptides that do not highly correlate with other peptides of the same protein. In the case of a protein only having two peptides that do not correlate to each other, the one with no modifications or higher abundance is kept. Then, peptide quantification values are normalized to either iRT injected peptides (217), to the Total Ion Intensity (TII) or to the injected heavy standards in case they exist. Finally, protein quantities are inferred using the mean, median or sum of the peptides.

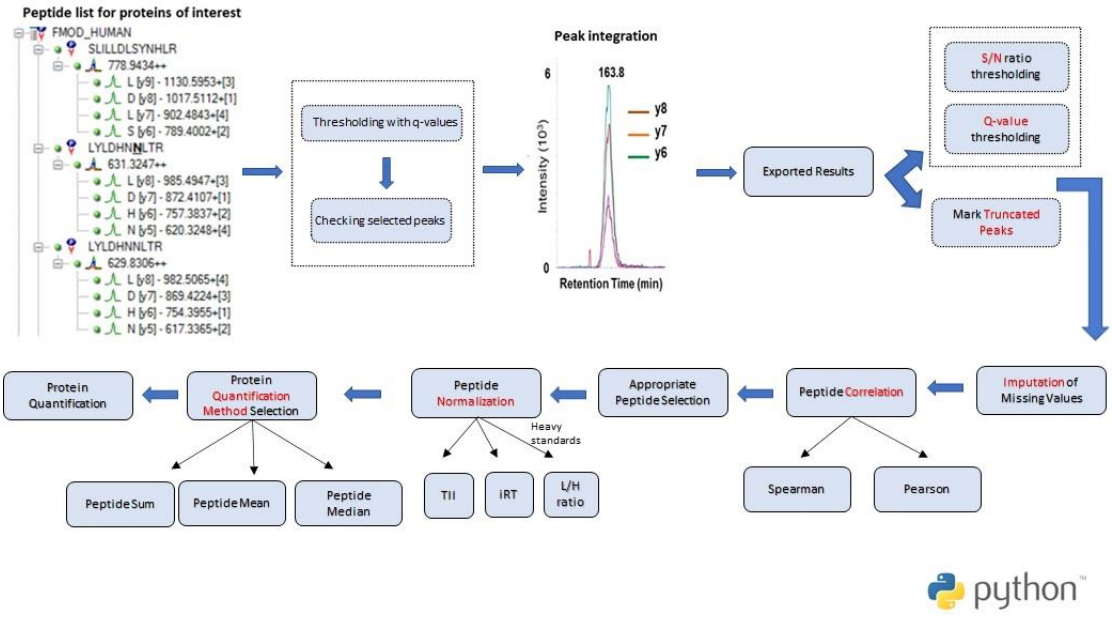


Figure 4.1 Streamlined Processing Pipeline for Targeted MS Data implemented in Python and used in the HOpTar-omics tool. The steps within dashed boxes denote optional steps, specific to either PRM MS data or DIA MS data. In red, steps/variables which require optimization are shown, which are: S/N ratio threshold, q-value, truncated peaks threshold, missing threshold (and the k for KNN-impute), the peptide

correlation method and threshold, the normalization method in the case of light peptides only and the protein quantification method. S/N: Signal to Noise, iRT: indexed retention time - empirically derived dimensionless peptide-specific values that are used for normalization across platforms and runs, TII: Total Ion Intensity, L/H: Light to Heavy peptide ratio.

This process involves many parameters and options, including an S/N ratio filter, q-value filter, truncated peaks filter, missing values frequency filter, k value of the KNN imputation method, peptide correlation method and correlation filtering threshold, normalization and protein quantification method with the most significant of them being highlighted in red in Figure 4.1. Since the values of these parameters are significantly affecting the final identification and quantification values, we have integrated this streamlined processing pipeline in the MOEA to optimize its parameters and method selection. The final pipeline which was implemented in the HOpTar-omics tool is presented in Figure 4.2. The iterative process of the optimization framework begins by initializing a set of solutions. The composition of every solution depending on the targeted proteomics data type is presented in Table 4.1. The first population of solutions is generated by randomizing values assuming a normal distribution for each variable. The optimization goals that were formulated as fitness functions are provided in Table 4.2. These goals were selected to maximize the reproducibility of the analysis in comparison with additional data from other protein quantification methods (e.g. clinical laboratory-based methods). The fitness functions aim to optimize data quality and their correlation with discovery or clinical measurements.

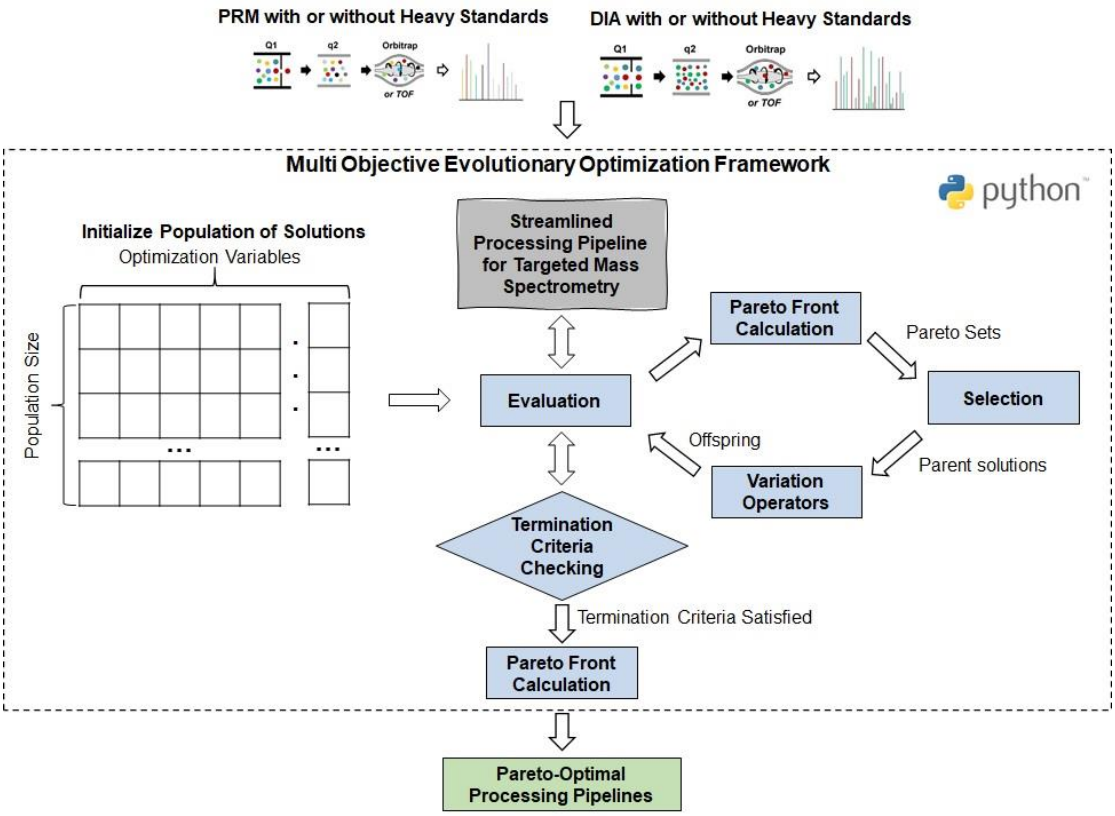


Figure 4.2 Workflow of HOpTar-omics tool. This tool can take as input the PRM or DIA MS Data in tab-delimited file format and returns as output a set of Pareto-optimal streamlined processing solutions as well as the best performing one based on the user-defined weights for the objectives of the optimization (Data Completeness, Consistency, Reproducibility and Discrimination Power). Solutions are represented as vectors of floating-point values representing the variables that should be optimized: S/N threshold, Q-value Threshold, Threshold for Truncated Peaks, Missing Threshold, k for Imputation with KNN-impute, Correlation Method for Peptide Filtering, Peptide Correlation Threshold, Normalization Method and Protein Quantification Method.

To guide the optimization process for the first and second cases, we have used discovery proteomics (TMT MS data) as an additional quantification method (Chapter 5). The quantified proteins in both cases were used to maximize the correlation of protein abundancies and the core versus periphery comparison in both cases was conducted, to maximize the agreement of fold changes. For the Covid-19 dataset, we have used the clinical measurements of albumin (ALB) and C-reactive protein (CRP) as well as the ELISA measurement of galectin-3-binding protein (LGALS3BP) to guide the optimization process calculating the defined fitness functions. The statistical comparison was the survival outcome 28 days after the baseline measurement of ICU-administered patients. The implementation of the HOpTar-omics tool is publicly

available at: https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis.

4.3.4 Software Benchmarking.

Existing tools and methods for processing PRM and DIA data were applied to analyze all datasets described in the Materials and Datasets section and be used as a benchmark to the proposed HOpTar-omics tool. Figure 4.3 provides comparative results for the 3 case studies deployed, using the fitness functions which are measuring data Completeness (absence of missing or zero quantification values), Consistency (absence of outliers calculated using the Bland-Altmann method (211)), Reproducibility (correlation of calculated quantification values with discovery proteomics or clinical measurements) and Discrimination Power (ability to reconstruct as close as possible the Fold Changes of discovery or clinical measurements of proteins). The same missing value imputation, missing value filtering and differential expression analysis methods were used between the benchmark and the proposed tool to ensure a fair comparison with the default parameters of the methods (30% missing values threshold, k=20 for KNN-impute) being selected for the benchmark methods. Because of the stochastic nature of the evolutionary algorithm used in the proposed tool, we have repeated the analysis of each case study 10 times and results are presented for the best and average performance.

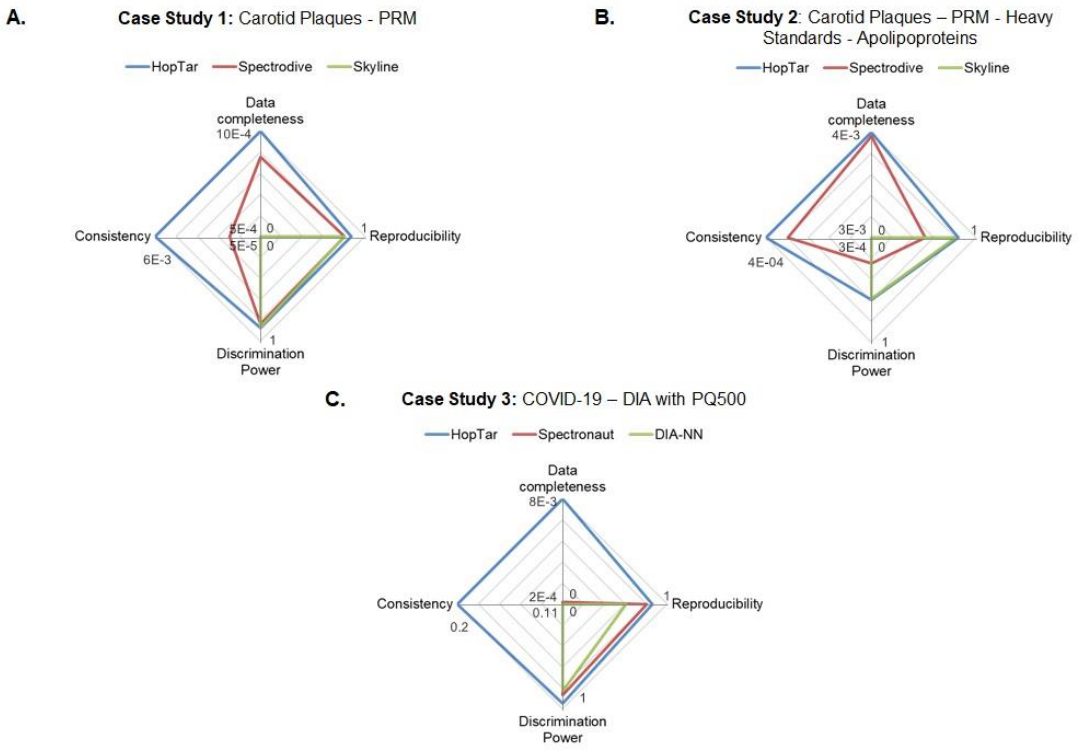


Figure 4.3 Comparative results of examined fitness values for the HopTar-omics against the other examined tools. **A.** Case study 1: Carotid plaques dataset using PRM targeted proteomics data (n=190 samples, n=129 proteins), **B.** Case study 2: Carotid plaque samples using PRM targeted proteomics with heavy standards injected for apolipoproteins (n=190 samples, n=23 proteins) and **C.** Case study 3: Covid-19 serum samples using DIA MS proteomics (n=62 samples, n=158 proteins). Data completeness (combination of Fitness Functions 1 and 2): 1/(1+number of missing values + several values below the limit of detection). Reproducibility: Fitness Function 3, Discrimination Power: Fitness Function 4, Consistency: Fitness Function 5.

Figure 4.4 below shows the non-linear regression curve and the 95% confidence interval from all runs of the best and average performance in each generation to demonstrate the convergence behaviour of the proposed algorithm.

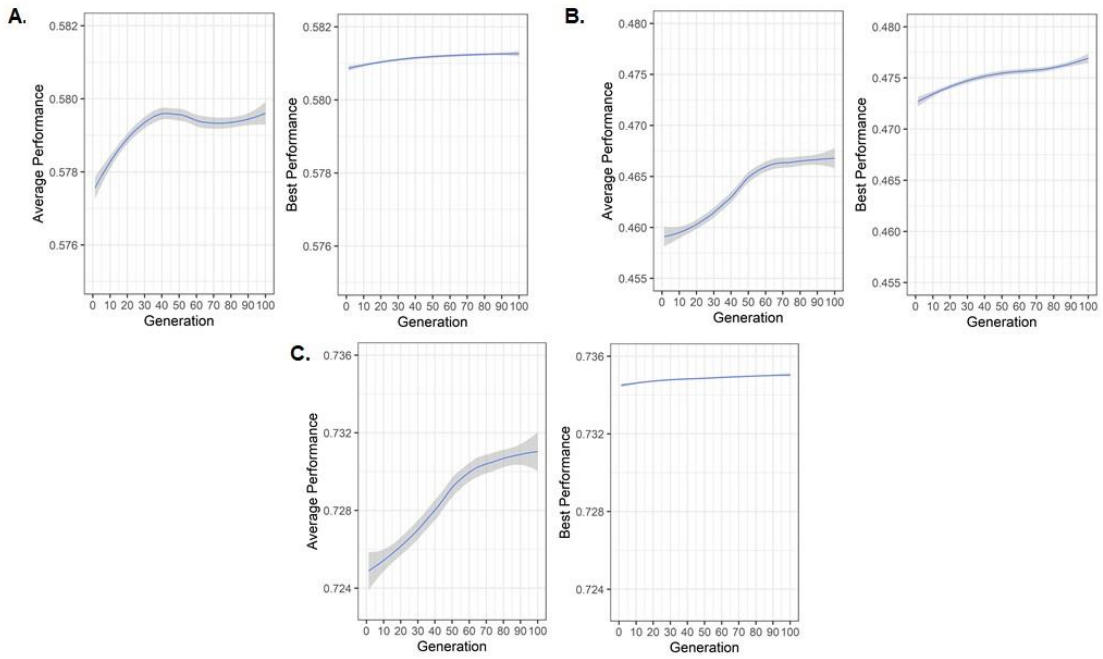


Figure 4.4 Average and best performance per generation when applying the HOpTar-omics tool on the three case studies. A. Case Study 1: Carotid Plaques – PRM, **B.** Case Study 2: Carotid Plaques – PRM - Heavy Standards – Apolipoproteins, **C.** Case Study 3: COVID-19 – DIA with PQ500. Performance is measured using the average of the fitness functions. Locally weighted smoothing regression curves are provided for the 10 executed runs and the 95% confidence intervals are visualized.

It is noted that the average performance has not increased substantially after approximately 50-60 generations, while most runs were prematurely terminated due to termination criteria satisfaction before 70 generations.

From the results of case study 1, it is readily observed that HOpTar was able to marginally improve reproducibility and discrimination power, while significantly improving data completeness and consistency goals compared to Spectrodive and Skyline. In case study 2, it is observed that, when used to analyze PRM data with heavy peptide standards injected, the HOpTar-omics tool was able to substantially increase the reproducibility and discrimination power of the produced targeted proteomics dataset compared to existing tools.

Regarding the DIA MS case study, HOpTar-omics improved all metrics and increased the consistency by 81.82% and data completeness metrics by 3900%. Figure 4.5**,** Figure 4.6 and Figure 4.7 depict the results of correlation analysis between the data

processed by benchmark tools and HOpTar-omics against their clinical (ALB, CRP) or Elisa (LGALS3BP) measurements.

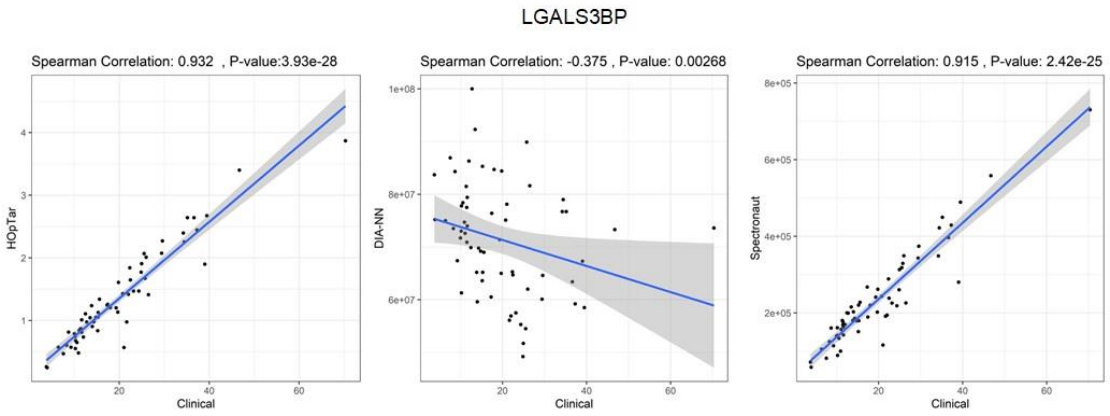


Figure 4.5 Correlation of proteomics vs Elisa measurements for LGALS3BP using all three tools. Scatterplots depicting the association of proteomics vs antibody measurements of LGALS3BP in serum samples of ICU-administered COVID-19 patients when DIA data are processed with HOpTar, Spectronaut and DIA-NN pipelines.

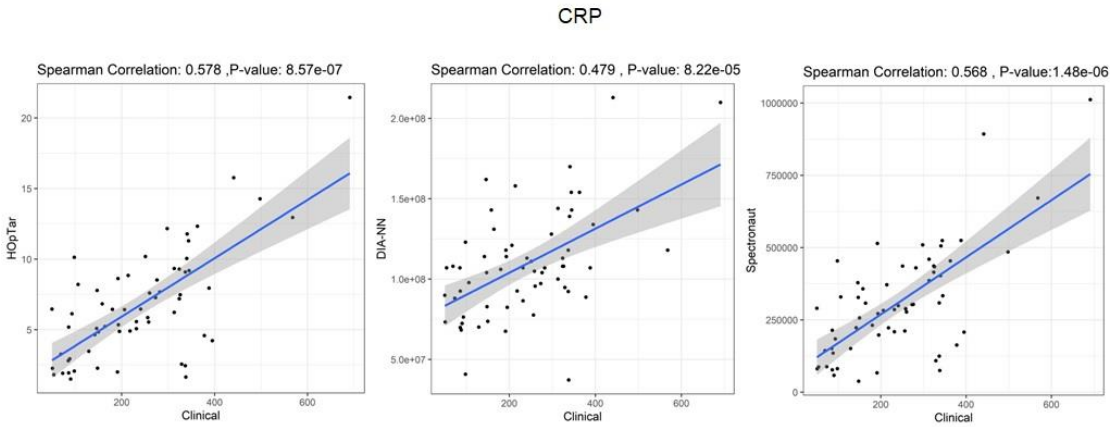


Figure 4.6 Correlation of proteomics vs clinical measurements for CRP using all three tools. Scatterplots depicting the association of proteomics vs clinical measurements of C-reactive protein in serum samples of ICU-administered COVID-19 patients when DIA data are processed with HOpTar, Spectronaut and DIA-NN pipelines.

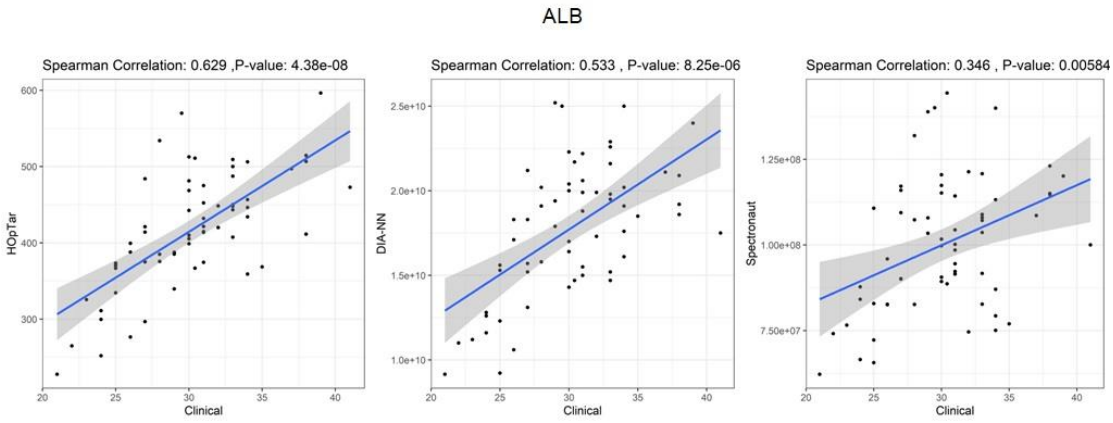


Figure 4.7 Correlation of proteomics vs clinical measurements for ALB using all three tools. Scatterplots depicting the association of proteomics vs clinical measurements of albumin in serum samples of ICU-administered COVID-19 patients when DIA data are processed with HOpTar, Spectronaut and DIA-NN pipelines.

The proposed tool outperformed benchmark methods for all three proteins. DIA-NN was not performing as well as other method. Table 4.3 list the performance metrics of all HOpTar-omics runs and the benchmark tools.

Case Study	***Fitness Function***	***Method***		
		Spectrodive	**Skyline**	**HopTar-omics**
1: Carotid Plaques - PRM	**Data completeness**	0.00085	0.00049	0.00096±0.00045
	Reproducibility	0.80258	0.80239	0.86563±0.0009
	Discrimination Power	0.83360	0.86103	0.87375±0.00162
	Consistency	0.00174	0.00005	0.00585±0.00015
		Spectrodive	**Skyline**	**HopTar-omics**
2: Carotid Plaques - Apolipoproteins - PRM with heavy Standards	**Data completeness**	0.00385	0.00300	0.00388±0.00006
	Reproducibility	0.51564	0.80256	0.8274±0.00112
	Discrimination Power	0.24920	0.58561	0.59655±0.00445
	Consistency	0.00309	0.00035	0.00381+0.00316

		DIA-NN	Spectronaut	HopTar-omics
3: DIA Covid-19 Serum Samples	**Data completeness**	0.00038	0.00022	0.0085±0.00084
	Reproducibility	0.80487	0.60617	0.85654±0.00001
	Discrimination Power	0.86448	0.82921	0.94579±0.00008
	Consistency	0.11111	0.11111	0.2±0.00001

Table 4.3 Detailed comparative results of proposed and benchmark methods in all three test cases. Because of the stochastic nature of the HOpTar-omics tool, results are provided as average ± standard deviation of the metrics in 10 independent runs.

Figure 4.8 below presents the optimized pipelines for the three case studies for the best-performing runs of the proposed tool.

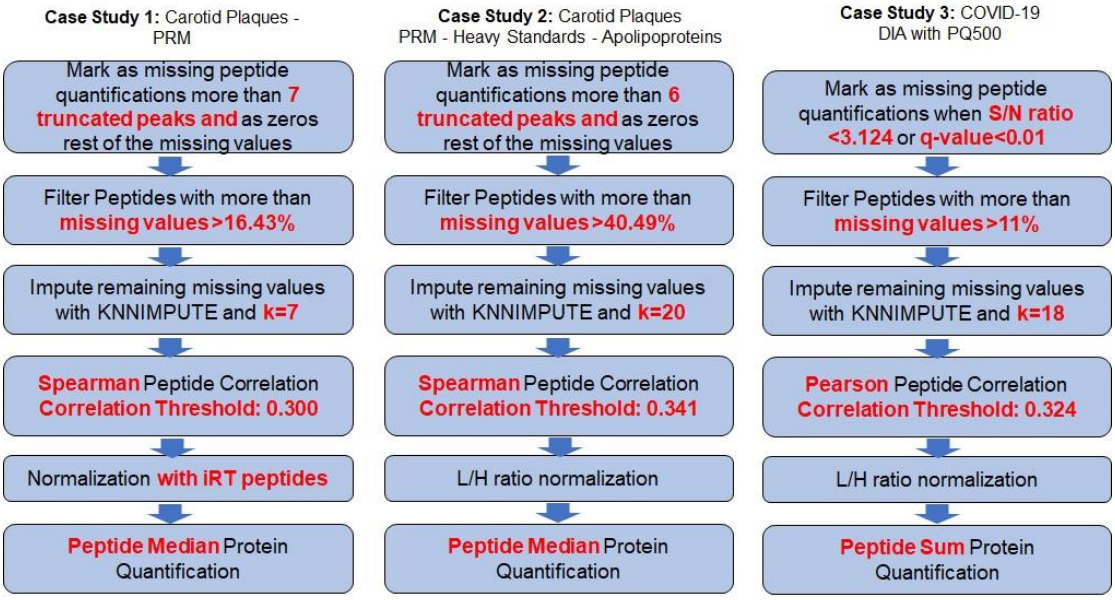


Figure 4.8 Optimized pipelines for each dataset. In red are shown the parameters that were optimized by the HOpTar-omics tool.

It is noteworthy that there are substantial differences among the parameters and the methods selected for PRM without heavy standards, PRM with heavy standards and DIA MS.

4.3.5 Accumulation of apolipoproteins in atherosclerotic plaques and adjacent tissue.

Figure 4.9 presents the Spearman correlation coefficient for apolipoproteins in the ECM (GuHCl extract) of carotid plaques quantified using PRM proteomics with heavy standards on atherosclerosis plaques in the core (n=95) and the periphery (n=95) of plaques. Data have been processed with the pipeline from the best-performing solution among all runs of the HOpTar-omics tool (Figure 4.8) to quantify the abundance of apolipoproteins. Hierarchical biclustering was applied to reveal clusters formed by apolipoproteins.

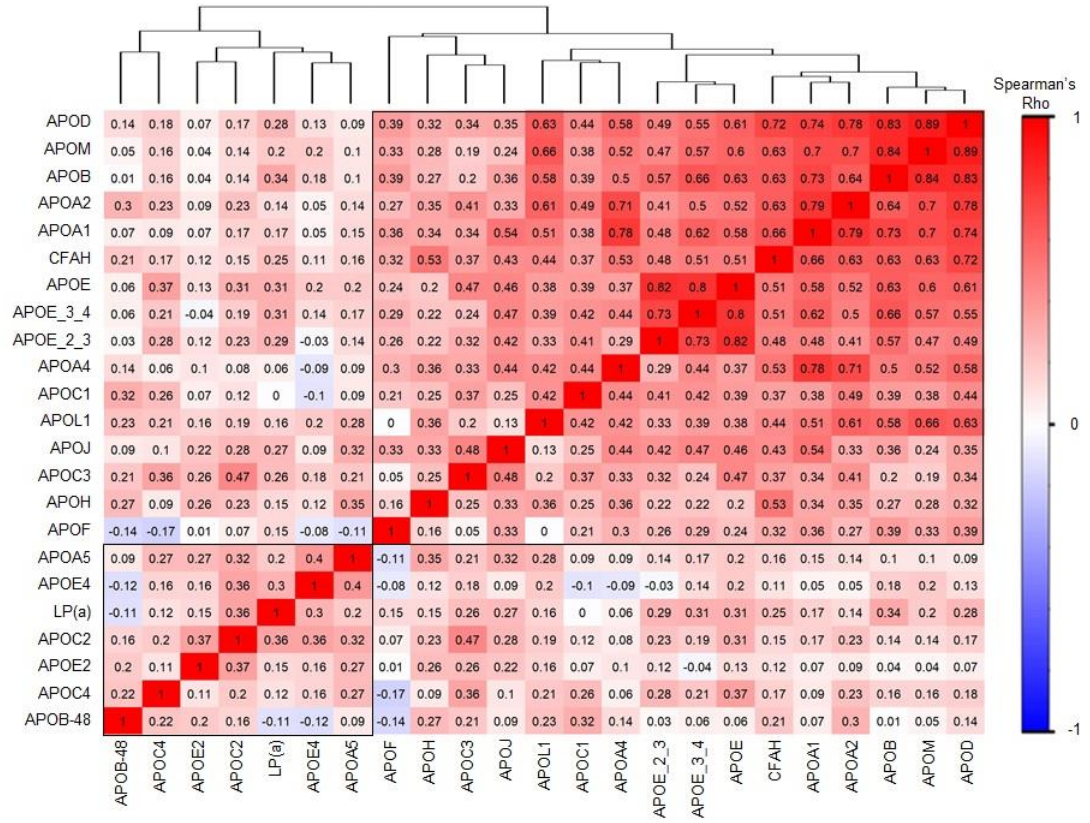


Figure 4.9 Correlation of apolipoproteins in the ECM. Heatmap showing Spearman correlation coefficient for apolipoproteins in the ECM (GuHCl extract) of carotid plaques quantified using PRM proteomics with heavy standards in the core (n=95) and the periphery (n=95) of atherosclerotic plaques.

Two clusters were identified with one composed of HDL, LDL and triglycerides-associated proteins, such as APOB, APOA1 and APOC3, and the other including LPA. A high correlation was observed among all apolipoproteins suggesting that the accumulation of lipids in the matrisome of carotid plaques is not limited to LDL, which only contains APOB.

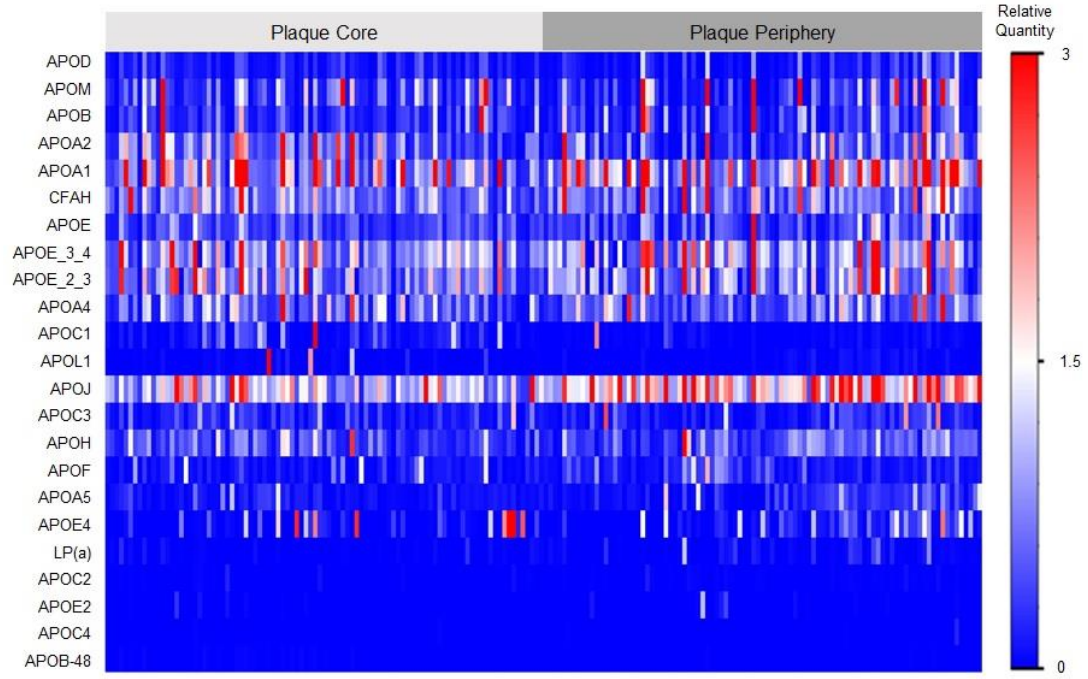


Figure 4.10 Relative quantification of apolipoproteins. Light-to-Heavy Ratio relative quantities of apolipoproteins measured with PRM using heavy standards in the ECM (GuHCl extract) of carotid plaques in the core and the periphery of the plaques. Proteins are ordered as revealed by hierarchical clustering of Figure 4.9.

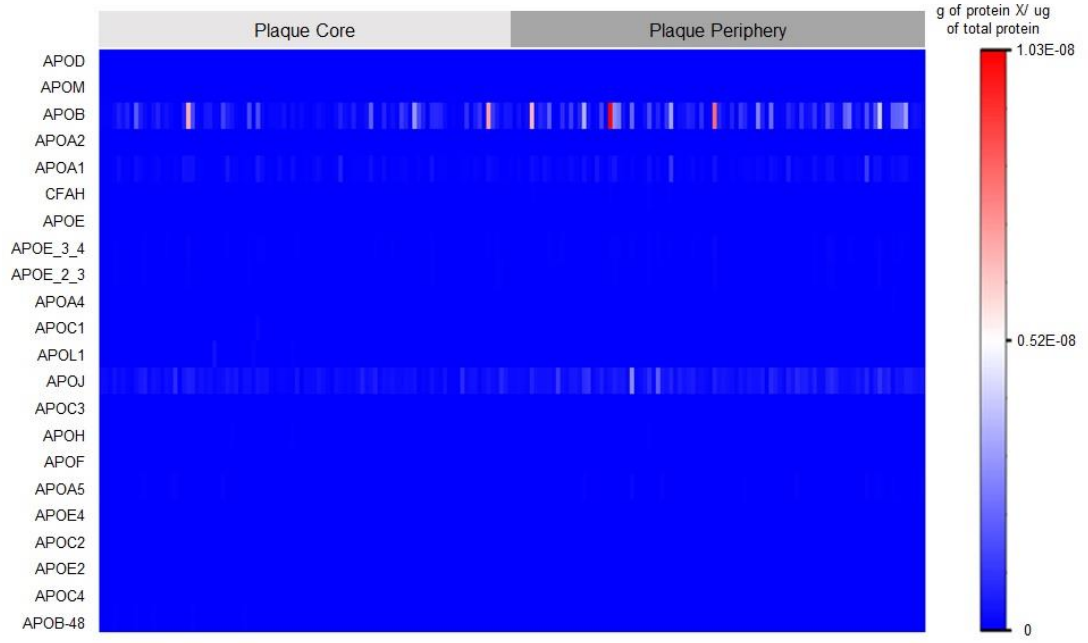


Figure 4.11 Absolute quantification of apolipoproteins. Absolute quantification values (g of protein X/ ug of total protein) of apolipoproteins were measured with PRM using heavy standards in the ECM (GuHCl extract) of carotid plaques in the core and the periphery of the plaques. Proteins are ordered as revealed by hierarchical clustering of Figure 4.9.

Figure 4.10 and Figure 4.11 above present the relative and absolute quantification values of apolipoproteins respectively. More specifically, Figure 4.10 presents the light-to-heavy normalized peptide ratio, and Figure 4.11 presents the absolute quantification values (g of protein X/ ug of total protein) of apolipoproteins measured with PRM using heavy standards in the core matrisome (GuHCl extract). Apolipoprotein(a) was not included in this calculation because of the presence of the kringle repeat peptide, which affects the total molecular weight of the protein across samples and individuals. The figures suggest that the most abundant apolipoproteins accumulated in carotid plaques are apolipoprotein A1 (APOA1) and APOB, thus further suggesting an accumulation of both LDL and High-Density Lipoprotein (HDL) lipoproteins, and apolipoprotein J (APOJ), mostly at the periphery of the plaque.

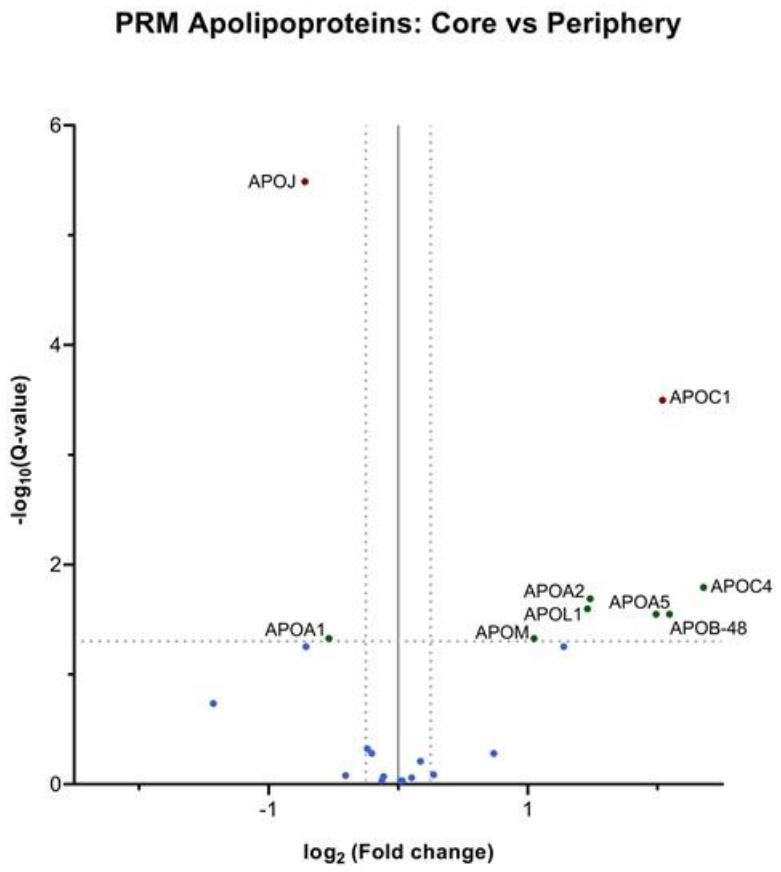


Figure 4.12 Differential expression analysis results for the core vs periphery comparison of apolipoproteins. Volcano plot for the comparison between core (n=95) vs periphery (n=95) of the apolipoproteins measured with PRM using heavy standards in the ECM (GuHCl extract) of the carotid plaque dataset. Ebayes method of the limma package (183) was considered for differential expression analysis, correcting for age,

sex and statins usages. Q-values were calculated correcting for multiple testing using the Benjamini-Hochberg method.

Differential expression analysis was conducted to reveal statistically significant changes in the accumulation of apolipoproteins in the core (n=95) and the periphery (n=95) of carotid plaques, demonstrating upregulation of most apolipoproteins in the core of the plaque while only the HDL-related protein APOA1 and APOJ were significantly upregulated in the periphery of the plaque (Figure 4.12).

4.4 Discussion

MS proteomics methods suffer from high data missingness, low reproducibility and low specificity, and these limitations restrict their applicability in large-scale applications (225). However, their unbiased approach and advantages compared to antibody or aptamer-based techniques (226) have made them a widely used tool for the identification and quantification of proteins in biosamples. Some of the most significant objectives of omics approaches are reproducibility and automation (33). Targeted and DIA MS proteomics have been able to increase reproducibility, while absolute quantification techniques such as using spiked-in peptides and calculating low-to-heavy peptide ratios have further contributed to this goal, while also improving specificity (227). However, the processing of targeted proteomics approaches is still not standardized, requires significant manual processing and several decisions to be made by the data analyst that might lead to biased results and calculation errors. Additionally, the substantial effort required to process such data prohibits their application to large samples and/or large protein-number studies. Moreover, a pipeline that is optimized for a specific type of biosample and a specific setup of targeted MS proteomics is not guaranteed to be optimal for other types of biosamples, MS setup and/or instrument settings. The proposed method was introduced to overcome such limitations by applying a multi-objective optimization framework to optimize both parameter and algorithm selection for processing targeted MS proteomics of different types.

The proposed method was benchmarked against widely used open-source and commercial solutions, demonstrating that it cannot only decrease data missingness and the presence of outliers in the dataset but can also improve substantially the

quantification accuracy of the MS techniques in all different types of MS data examined. To ensure a valid comparison between methods, we have applied the same missing value filtering, data imputation and statistical analysis pipelines on both the proposed and benchmark techniques. The proposed tool utilizes a data-driven evolutionary optimization approach to solve the optimization problem of reconstructing the optimal processing pipeline for targeted MS data. Evolutionary optimization was used because of its ability to perform well in big optimization problems, where a strictly mathematical optimization equation cannot be defined. Moreover, since the optimization problem to be solved has many contradictory objectives such as data completeness, reproducibility, consistency and discriminative power, a multi-objective Pareto-based technique was used to allow a good balance between exploration and exploitation properties of the algorithm, while also allowing the discovery of multiple non-dominated solutions (228). Experimental results showed that the proposed format of the HOpTar-omics can converge to near-optimal solutions after only 30-50 generations, making the computational optimization process computationally inexpensive and extremely efficient, considering the high workload of manual processing that would otherwise be required.

Clinical apolipoprotein measurements in blood are restricted to APOA-I and APOB, but recent evidence has demonstrated the potential of including additional apolipoproteins for a more efficient cardiovascular risk assessment (229–231). Despite increasing knowledge on the association of circulating apolipoproteins with cardiovascular risk, less is known so far about the accumulation of apolipoproteins in atherosclerotic plaques. Symptomatic or unstable atherosclerotic plaques are usually forming a lipid core that is enriched with apolipoproteins (232) but which apolipoproteins are accumulated in plaques has not yet fully elucidated, while the quantification of this accumulation requires the application of absolute quantification techniques. The ability of the HOpTar-omics tool to optimize and standardize the processing of targeted proteomics and the use of heavy standard peptides allowed absolute quantification in atherosclerotic plaques. Capitalizing on this unique generated dataset, we were able to confirm that Apolipoprotein A1 and Apolipoprotein B are mostly retained in the carotid plaques as expected because they

are among the most highly abundant apolipoproteins. However, it is noteworthy that the most abundant apolipoprotein found to be retained in the plaques was Apolipoprotein J, also known as clusterin (CLU), and it was particularly upregulated in samples from the periphery of the plaques. The role of clusterin in atherosclerosis has not been well studied with contradicting evidence available in the literature (233, 234).

In conclusion, the proposed HOpTar-omics tool outperforms benchmark methods for processing targeted and DIA MS data, as exemplified through several datasets of different types. This superior performance relates to a statistically significant improvement in data completeness, reproducibility and quantification accuracy goals. To facilitate the use of this tool in further proteomic applications, the code is stored in an open-source code repository to enable the proteomic and bioinformatic communities to test it and develop new methods and tools to further improve its performance.

5. A Proteomic Atlas of Atherosclerosis: Signatures of Plaque Inflammation, Calcification and Sex Differences and their Association with Outcomes

5.1 Introduction

Atherosclerosis is a multifactorial disease that develops in the intima of the arteries due to lipid deposition. Endothelial dysfunction, resulting in lipoprotein accumulation in the arterial wall (especially LDL), is thought to be the initiator of atherosclerosis (1, 235). Lipoproteins bind to proteoglycans in the ECM (3), get oxidized by ROS and become entrapped in the subendothelial space. As a response to oxLDL circulating leukocytes infiltrate the subendothelial space initiating a chronic inflammatory process. Monocytes differentiate into macrophages which phagocytose oxLDL and turn into foam cells. Accumulation of foam cells initiates inflammatory processes that drive the progression of atherosclerosis and can cause plaque rupture (6).

Atherosclerotic plaques are classified based on their histology (AHA classification) and ultrasound measurements. According to the AHA classification, there are 8 types of atherosclerotic lesions, intimal thickening, fibroatheroma, late fibroatheroma, healed plaque rupture, fibrocalcific plaques, erosions, thin-capped fibroatheroma, and ruptured plaques (16, 30). Calcification is a late feature of advanced atherosclerotic lesions (1). According to ultrasound measurements, atherosclerotic plaques are classified as echolucent, mixed and echogenic plaques. Echolucent plaques are soft, unstable plaques with a thin fibrous cap, a necrotic lipid core and continuous inflammation. These plaques are more prone to rupture and can cause clinical overt diseases such as heart attacks or strokes due to thromboembolism. Echogenic are calcified plaques with less likelihood of clinical symptoms, possibly due to reduced inflammation and increased ECM deposition (31).

The importance of the ECM and ECM-associated proteins in atherosclerotic plaque stability, vulnerability and progression have been highlighted in several studies, as ECM degradation processes have been linked with the development of vulnerable plaques (236). However, the differences in ECM composition across different plaque phenotypes have not been fully studied in atherosclerosis. Proteomics can

characterize the protein composition of human plaques and promote the current assessments by histology and imaging.

In this chapter, having established the network reconstruction and analysis pipelines and the computational frameworks to optimize the processing of MS proteomics, we applied them to the largest MS dataset of carotid plaques. With this analysis, we tried to reveal new molecular signatures of human atherosclerotic phenotypes, associate them with current classifications by imaging, and histology, relate them to cardiovascular outcomes, validate major findings using different proteomics techniques and measurements from an independent cohort and, finally, develop and store major findings to a relational database, the PlaqueMS database.

5.2 Methods

5.2.1 Description of Vienna plaque patient cohort

Data from 120 patients with carotid artery stenosis undergoing carotid endarterectomy were collected. All surgical procedures were performed at the Department of Vascular Surgery, Medical University of Vienna, Austria. The indication for surgery included symptomatic carotid artery stenosis or high-grade asymptomatic stenosis (>70%). Following carotid endarterectomy, carotid atherosclerotic plaques were excised and visually inspected and the plaque core was dissected from the plaque periphery. Patients were considered symptomatic if they had experienced a stroke, transient ischemic attack, or amaurosis fugax ipsilateral to the carotid artery stenosis within 6 months before carotid endarterectomy. Patients with hemorrhagic, lacunar, or cardioembolic stroke were excluded from the study. 9-year follow-up was collected for all patients and the primary cardiovascular endpoint was defined as the composite of cardiovascular death, myocardial infarction, transient ischemic attacks or stroke as well as atherosclerosis progression in the coronary or peripheral arteries requiring either interventional (percutaneous coronary intervention or peripheral balloon angioplasty with and without stenting) or surgical revascularization (aortocoronary bypass or peripheral bypass). A total of 219 human carotid endarterectomy samples (110 core and 109 periphery samples) were processed for

protein analysis. After quality control 207 samples were maintained (101 core and 106 from the periphery of the plaques).

5.2.2 Proteomics

Protein extraction was performed using a previously established three-step extraction protocol (182) from our lab, followed by two different proteomics methods: label-free (TopS) quantitation and quantitation using TMT multiplexing. Validation was performed by targeted (PRM) proteomics and by a Proximity Extension Assay platform of Olink proteomics (237). Significant findings were validated using label-free proteomics on carotid plaque samples from the independent Athero-Express cohort (238) (**Error! Reference source not found.**).

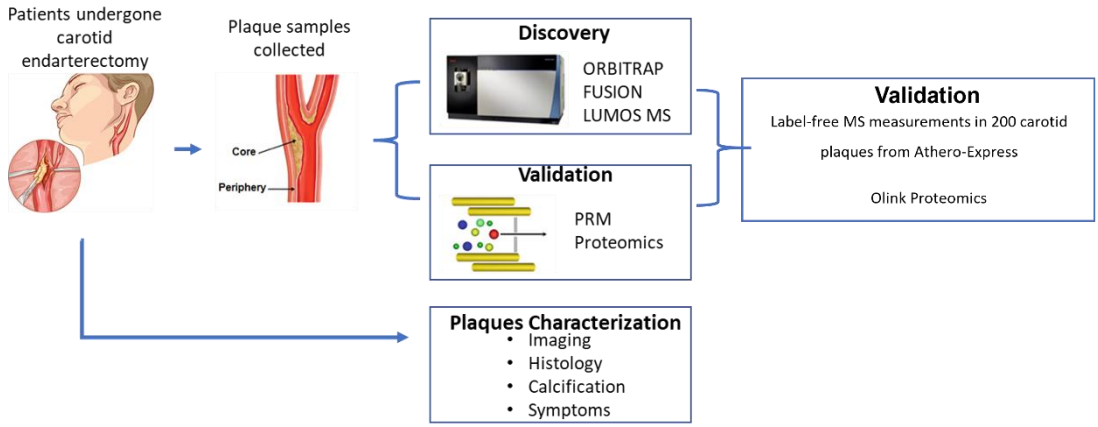


Figure 5.1. Sample extraction protocol and proteomic analysis.

5.2.2.1 Protein Extraction

Tissue sections were diced and weighed, and approximately 40-80 mg of tissue was taken for three-step protein extraction, as previously described (182). Briefly, samples were incubated in a NaCl extraction buffer (0.5 M NaCl, 25 mM EDTA, 10 mM Tris pH 7.5, plus protease inhibitors) with slow agitation for 1 h. The NaCl fraction (supernatant) was transferred to a new tube and stored at -80 ℃ for later use. Subsequently, samples were incubated in SDS buffer (0.1% SDS, 25 mM EDTA, and protease inhibitors) for 4 h to release cellular components. Finally, samples were incubated in a guanidine hydrochloride buffer (4 M GuHCl, 50 mM sodium acetate pH 5.8, and protease inhibitors) for 48 h to solubilize long-lived ECM proteins. Protein concentrations were estimated according to the 280 nm absorbance for NaCl and

GuHCl extracts, and SDS extracts were quantified using a Pierce BCA protein assay kit according to the manufacturer's instructions (Pierce BCA Protein Assay Kit, 23225, Thermo Scientific).

5.2.2.2 Deglycosylation

A two-step deglycosylation protocol was employed for GuHCl and NaCl extracts. First, sample pellets were resuspended in deglycosylation buffer (150 mM NaCl, 50 mM sodium acetate, 10 mM EDTA, pH 6.8, plus protease and phosphatase inhibitors) containing the following deglycosylation enzymes: Endo-α-N-acetylgalactosaminidase, α2-3,6,8,9-Neuraminidase, β-1,4-Galactosidase, β-N-Acetylglucosaminidase (all Merck-Millipore Glycoprotein Deglycosylation Kit, 362280), Chondroitinase ABC (Sigma-Aldrich, C3667), Heparinase II (Sigma-Aldrich, H6512), and Endo-β1,4-galactosidase (Sigma-Aldrich, G6920). Samples were incubated at 25 °C for 2 h, followed by 37 °C for 24 h. Second, samples were dried using a SpeedVac (Thermo Scientific, Savant SPD131DDA), reconstituted in 18O-labeled water (Taiyo Nippon Sanso, F03-0027) containing PNGase F (Merck-Millipore, 362280), and incubated at 37 °C for 48 h.

5.2.2.3 In-solution digestion

Proteins were denatured using 6 M urea and 2 M thiourea and reduced with 10 mM DTT at 37 °C for 1 h. The samples were then cooled to room temperature before being alkylated using 50 mM iodoacetamide followed by incubation in the dark for 45 min. Pre-chilled (-20 °C) acetone (10x volume) was used to precipitate the samples overnight at -20 °C. Samples were centrifuged at 14,000 x g for 40 min at 4 °C and the supernatant was subsequently discarded. Protein pellets were dried using a SpeedVac, resuspended in 0.1 M triethylammonium bicarbonate (TEAB) buffer, pH 8.2, containing MS grade trypsin (Thermo Scientific) (1:50 trypsin: protein), and digested overnight at 37 °C. Trypsin was inhibited by acidification of the samples with a final concentration of 1% trifluoroacetic acid (TFA). Peptide samples were then purified using a 96-well C18 spin plate according to the manufacturer's instructions (Harvard Apparatus). The dried peptide was reconstituted with 2% acetonitrile (ACN) and 0.05% TFA in water.

5.2.2.4 TMT Discovery Proteomics Workflow

All extracts were labelled using TMT and analysed on an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Scientific). TMT labelling was carried out according to the manufacturer's instructions (TMT10plex kit, QK226224, Thermo Scientific). Symptomatic and asymptomatic plaques were equally distributed into the batches of TMT labelling. Three technical replicates were injected for each sample. Thermo Scientific Proteome Discoverer software (version 2.1.0.81) was used to search raw data files against the human database, (UniProtKB/Swiss-Prot version 2017_01, 20,192 protein entries) using Mascot (version 2.3.01, Matrix Science). The mass tolerance was set at 10 ppm for precursor ions and 0.8 Da for fragment ions. Trypsin was used as the digesting enzyme with up to two missed cleavages being allowed. The in-built TMT10plex static modification was assigned for the detection of TMT labels. Carbamidomethylation of cysteine was chosen as a static modification and oxidation of methionine residues was chosen as a dynamic modification. MS/MS-based peptide and protein identifications were validated with the following filters, a peptide probability of greater than 95.0% (as specified by the Peptide Prophet algorithm (239)), a protein probability of greater than 99.0%, and at least two unique peptides per protein. Data were normalized to the total peptide amount to consider variation in abundances between samples. Scaling was conducted using a control pool sample, correcting for differences in reporter ion abundances due to different numbers of observed peptides. Data are available from the ProteomeXchange repository, under the PXD030975 identifier.

5.2.2.5 Label-free (TopS) Discovery Proteomics Workflow

The 0.5 M NaCl and 4 M GuHCl extracts were denatured and reduced in sample buffer containing 100 mM Tris, pH 6.8, 40% glycerol, 0.2% SDS, 2% β-mercaptoethanol, and 0.02% bromophenol blue and boiled at 96°C for 10 minutes (182). 35 µg of protein per sample was loaded and separated on Bis-Tris discontinuous 4%–12% polyacrylamide gradient gels (NuPage, Invitrogen) alongside protein standards (prestained All Blue, Precision Plus, Bio-Rad). Gels were stained using the PlusOne Silver staining Kit (GE Healthcare). Silver staining was used for band staining to avoid cross-contamination with fainter gel bands (240). All gel bands were excised in identical parallel positions

across lanes, and no empty gel pieces were left behind. Subsequently, all gel bands were subjected to in-gel tryptic digestion using an Investigator ProGest (Genomic Solutions) robotic digestion system. Tryptic peptides were separated on a nanoflow LC system (ThermoFisher Scientific UltiMate 3000) and eluted with eluent A (2% acetonitrile, 0.1% formic acid in H_2O) and B (90% acetonitrile, 0.1% formic acid in H_2O) using a 70-minute gradient (10%–25% B in 35 minutes, 25%–40% B in 5 minutes, 90% B in 10 minutes, and 2% B in 20 minutes). The column (ThermoFisher Scientific PepMap C18, 25-cm length, 75-µm internal diameter, 3-µm particle size) was coupled to a nanospray source (Picoview). During the liquid chromatography-mass spectrometry (LC-MS) run, spectra were collected from a high-mass accuracy analyzer (LTQ Orbitrap XL, ThermoFisher Scientific) using full ion scan mode over the mass-to-charge (*m/z*) range 450–1,600. Tandem MS (MS/MS) was performed on the top 6 ions in each MS scan using the data-dependent acquisition mode with dynamic exclusion enabled. MS/MS peak lists were generated by extract_msn.exe and matched to the human database (UniProtKB version 2013_8, 88,378 protein entries) using Mascot (version 2.3.01, Matrix Science). Carboxyamidomethylation of cysteine was chosen as a fixed modification, and oxidation of methionine, lysine, and proline was chosen as variable modifications. The variable modifications of lysine and proline were included due to the large abundance of collagens in the samples. The mass tolerance was set at 1.5 AMU for the precursor ions and at 1.0 AMU for fragment ions. Two missed cleavages were allowed. Proteome Discoverer software (version 2.1.0.81) was used for peptide identification. Peptide identifications were accepted if they could be established at greater than 95.0% probability as specified by the Peptide Prophet algorithm (239). Only tryptic peptides were included in the analysis. Protein identifications were accepted if they could be established at greater than 95.0% probability (241) with at least two independent peptides and a mass accuracy of ≤10 ppm of the precursor ion.

5.2.2.6 Targeted Proteomics Workflow

A PRM targeted proteomics workflow was developed using a Q Exactive HF mass spectrometer (ThermoFisher) and Skyline software (191) (version 4.1, MacCoss Lab Software), to quantify a selection of extracellular proteins of interest as described in

chapter 4 (section 4.2). Proteotypic peptides were selected using data from the untargeted label-free analysis. Precursor ions for ECM proteins of interest that were not detected in the untargeted analysis were predicted in-silico using SRM Atlas (242). Skyline was used to optimize retention times. Proteotypic peptides were scheduled using the retention time obtained from test experiments with the same LC configuration and eluting gradient; retention time windows were set at +/- 4 min. After initial testing, samples were quantified for a total of 205 peptides for 119 proteins of interest. Peptides with poor chromatography were excluded. A minimum of three fragment ions per peptide and a signal-to-noise level of 3 to 1 were required for each peptide to qualify as quantifiable. For proteins with more than one peptide, the peptide Spearman correlation was checked. In case more than two peptides per protein were detected, the ones with a correlation less than R=0.5 with the remaining peptides were filtered. In case two peptides per protein were detected, the most abundant peptide was kept when correlation was less than R = 0.5. 11 iRT synthetic peptides were spiked in during sample preparation and 10 were used for normalization (1 iRT peptide was excluded due to poor chromatography). Final protein abundance was calculated by summing up the quantified peptide abundances. The isolation list is shown in Supplemental Table 1 (https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis).

5.2.2.7 Olink Proteomics

To validate the findings of the MS discovery analysis, we used four Olink panels (237) (CARDIOMETABOLIC, CARDIOVASCULAR III, DEVELOPMENT, METABOLISM). Olink platform uses Proximity Extension Assay technology, which combines antibody- and DNA-based methodologies to quantify and detect protein biomarkers with high sensitivity and specificity (243). The cardiometabolic panel includes 92 proteins involved in biological processes such as cell adhesion and immune response, the Cardiovascular III panel includes 92 inflammatory and CVD protein markers, the Development panel includes 92 developmental-related biomarkers and the Metabolism panel 92 metabolism-related biomarkers.

5.2.2.8 Label-free Proteomics in Validation – Athero-Express Cohort

The validation cohort consisted of carotid endarterectomy samples from 200 patients with carotid artery stenosis undergoing carotid endarterectomy. A similar proteomics workflow as the one applied for the discovery was followed, using label-free quantification. Moreover, a 2-step protein extraction, similar to the previously described (182), but skipping the SDS extraction step was used. After the protein extraction, 20ug of proteins in GuHCl extracts from each sample were precipitated using 10x volume of ethanol overnight. Next, samples were deglycosylated. After in-solution digestion, the peptide samples were purified using C18 cartridges on a Bravo AssayMAP robotic system (Agilent) according to the manufacturer's instructions. Regarding the LC-MS/MS label-free analysis protocol, peptides were separated by a nanoflow LC system (Dionex UltiMate 3000 RSLC nano). Samples were injected onto a nano-trap column (Acclaim PepMap100 C18 Trap, inner diameter 300 µm x length 5 mm, particle size 5 µm, pore size 100 Å), at a flow rate of 25 µL/min for 3 min, using 0.1% formic acid (FA) in water. The following nano-LC gradient was then run at 0.25 µL/min to separate the peptides: 0−1 min, 1% B; 1−6 min, 1–6% B; 6–40 min, 6–18% B; 40-70 min, 18-35% B; 70–80 min, 35–45% B; 80–81 min, 45–99% B; 81–89.8 min, 99% B; 89.8–90 min, 99–1% B; 90–120 min, 1% B; where A = 0.1% FA in water, B = 80% ACN, 0.1% FA in water. The nano column (EASY-Spray PepMap RSLC C18, 75 µm x 500 mm, 2 µm, 100 Å), set at 45 °C, was connected to an EASY-Spray ion source (Thermo Scientific). Spectra were collected from an Orbitrap mass analyzer (Q Exactive HF, Thermo Scientific) using full MS mode over the m/z range 350–1600 with a resolution of 60,000 at 200 m/z. Data-dependent MS2 scan was performed using the Top15 method with HCD activation and Orbitrap detection with a resolution of 15,000 in each full MS scan with dynamic exclusion enabled. Thermo Scientific Proteome Discoverer software (version 2.4.1.15) was used to search raw data files against the human database, (UniProtKB/Swiss-Prot version 2021_01, 20,396 protein entries) using Mascot (version 2.6.0, Matrix Science). The mass tolerance was set at 10 ppm for precursor ions and 20 milli mass units (mmu) for fragment ions. Trypsin was used as the digesting enzyme with up to two missed cleavages being allowed. Carbamidomethylation of cysteine was chosen as a static modification. Oxidation of methionine, proline and lysine, and deglycosylation with the presence of O18-water

on asparagine was chosen as dynamic modifications. Protein identification FDR confidence was set to high and the minimum number of unique peptides per protein was 2. Only master proteins were shown. The precursor peak area was used for quantification and normalized to the total peptide peak area of each sample.

5.2.2.9 Semi-tryptic Search

The search was conducted using Thermo Scientific Proteome Discoverer (244) software (version 2.1.0.81). “No enzyme” was selected as the search parameter for the digestion enzyme. The rest of the search parameters were the same as the ones used for the standard discovery searches.

5.2.2.10 Gamma-carboxylation search

The search was conducted using Byonic^{TM}software (245) from the Protein Metrics platform, through its integration in Proteome Discoverer software. Byonic is a package which can identify proteins and peptides but is mostly used for PTM identification. It uses a combination of de novo sequencing and database search for identification, so it is more sensitive and accurate for non-standard PTMs (246) than for example Mascot (Matrix Science), which we used in all other Proteome Discoverer searches. For our search, we used a strict FDR (for hits with high confidence) of 0.01, a relaxed FDR (for hits with moderate confidence) of 0.05 for both PSMs and peptides, peptide confidence “LOW” and the rest parameters as described above. Finally, we manually filtered the identified peptides with known sites for this PTM from the literature and found that 19.44% of the identified peptides were including an amino acid that could be gamma-carboxylated according to the literature.

5.2.2.11 MaxQuant algorithm search

MaxQuant algorithm (195) was used to search raw data files against the human database UniProtKB/Swiss-Prot (version 2017_01, 20192 protein entries). There were three replicates for 24 sets of LC-MS runs: leading to a total of 72 isobaric channels. All extracts were labelled using tandem mass tags (TMT10plex) and analysed on an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Scientific). The reporter mass tolerance was set to 0.003 (Da) and the output was normalised on a “weighted ratio to reference channel” basis. Carbamidomethylation of cysteine was chosen as a

fixed modification and oxidation of methionine residues was chosen as a variable modification, with only a maximum of 5 modifications being allowed. Peptide and protein confidence was set at q-value< 0.01. Upon completion of the analysis, a small percentage of peptides identified were manually removed if they were marked as “reverse”, “potential contaminant”, or “only identified at the site” to improve peptide identification accuracy.

5.2.3 scRNAseq

UMAP of scRNA-seq data from carotid endarterectomies (n = 38) and dot-plot for the corresponding transcripts to the revealed cellular, inflammation and calcification biosignatures were reconstructed using the PlaqView tool (157) and the Aran et al. (159) reference-based algorithms to deduce cell identity of individual cells. The dataset used was the one of Slenders et al. (247) with 5633 cells from 38 patients.

5.2.4 Spatial RNAseq

Visium for formalin-fixed paraffin-embedded (FFPE) spatial gene expression slide and reagent kit was used according to manufacturer instructions (PN: 1000338; 10X Genomics). Each capture area (6.5 x 6.5 mm2) contains 5,000 barcoded spots that are 55 µm in diameter (100 µm centre-to-centre distance) providing an average resolution of 1 to 10 cells. A 5µm tissue section from one carotid plaque sample was placed onto one capture area of a Visium Spatial Gene Expression slide. The sample used for spatial RNA-seq was selected from a male, asymptomatic, 70 years old patient with 80-90% stenosis. Deparaffinization, dapi immunofluorescence staining, image acquisition and decrosslinking were performed as specified in the Visium Spatial Gene Expression for FFPE – Deparaffinization, Decrosslinking, Immunofluorescence Staining & Imaging protocol (CG000410) and performed at the Core Facility Imaging, Medical University of Vienna. Fluorescent images were acquired with an Olympus IX83 microscope equipped with a Hamamathsu Orca FLash camera for fluorescence image capture and parameters following those specified in the Visium Spatial Gene Expression Reagent Kits for FFPE User Guide sequencing instructions (read 1: 28 cycles; i7 index read: 10 cycles; i5 index read: 10 cycles; and read 2: 50 cycles) yielding 149 million sequenced reads. The FASTQ file and manually aligned histology image were analyzed with Space

Ranger (248) 1.3.1 and the human probeset provided by 10X genomics (Visium Human Transcriptome Probe Set v1.0 GRCh38-2020-A). Visualization of spatial cluster output from Space Ranger using the kmeans algorithm as well as differential gene expression analysis to determine cluster-specific up- or downregulated genes was performed on Loupe Browser (249) 6.0.0 (10X Genomics).

5.2.5 Statistical and Bioinformatics Analysis

5.2.5.1 Differential Expression Analysis

The proteins retrieved using the GuHCl and NaCl extracts were filtered to keep the ECM and related proteins using ECM protein annotation from the MatrisomeDB (135) and adding extracellular proteins, such as apolipoproteins, that were deemed important in the context of atherosclerosis, as stated above (Chapter 4). The dataset was further filtered using a method to discriminate between random missing values and values that are consistently missing because of abundances below the limit of detection. In particular, consistent missing values were identified and imputed with zeros when more than 90% missing values were observed in one phenotype and less than 10% in the other. Otherwise, proteins with more than 30% missing values were filtered out. All remaining missing values were imputed with the KNN-Impute method with k equal to 20 (default value). The relative quantities of the proteins were scaled using log2 transformation. The limma package (183) was used to compare different phenotypes using the Ebayes algorithm and correcting for selected covariates (age, sex and statins). The p-values were adjusted for multiple testing using the Benjamini-Hochberg method (172). Volcano plots were constructed using GraphPad Prism (250) (version 9.2.0).

Fisher's exact non-parametric testing was used to compare categorical variables. Scatterplots and heatmaps were constructed using the R environment (version 3.5.2).

5.2.5.2 Network Analysis

For the reconstruction of correlation networks from the MS proteomics data, we applied a simplified version of the DiRec-AP algorithm (chapter 3), which combined the ARACNe-AP (75) software to infer directed weighted correlation links between proteins and SIREN (154) algorithm, to characterize the edges of the network as

inhibition or activation. 100 bootstraps were used for the reconstruction of the networks and multiple edges and self-loops were filtered. Network clustering was conducted using the soft clustering method algorithm ClusterOne (79). The matrisome network was constructed by combining NaCl and GuHCl extract data. In particular, for proteins quantified in both extracts, the extract on which the protein had a higher number of matched spectra was used. Network visualizations were conducted using the Cytoscape tool (91) (version 3.7.1).

5.2.5.3 Bioinformatics Analysis

Spearman correlation was conducted using the scipy Python library (version 1.3.1) (251) and was used to correlate the relative expression levels of proteins with clinical variables, imaging, and histology measurements. Enrichment analysis was conducted using the David tool (174). This analysis included pathway terms from Reactome (252) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (253) and molecular function annotation from Gene Ontology (254). Survival analysis was conducted through the survival library of the R environment (version 3.5.2,) using Kaplan-Meier analysis. Principal component analysis (PCA) was conducted using the scikit-learn (255) Python library (version 0.19.2). Scree test was used to retain an adequate number of principal components to maintain at least the 90% variability of the data set. Logistic regression models were created and evaluated using the Hmisc package of the R environment (version 3.5.2). KMEANS algorithm from the Python library scikit-learn (version 0.19.2) was used for clustering. Inferring the optimal number of clusters was accomplished by experimenting with values from 2 to 20 measuring the Calinski-Harabasz score.

5.2.5.4 CVD Risk Prognostic Models

Proteins belonging to the revealed proteomics biosignatures were considered as potential inputs to classification models to predict the primary endpoint in the 9-year follow-up. An ensemble dimensionality reduction technique was used deploying a multi-objective evolutionary algorithm (208) to first identify the optimal feature subset to be used as input to Support Vector Machines (SVM) Classification models (256) and then optimize the SVM's parameters values (regularization parameter C and Radial Basis Function parameter gamma). The iterative process of the optimization framework begins by initializing a set of solutions. Each solution consists of values for

deciding if a feature/proteomic marker will be used as input (values greater than 0.5 force its use), and of two values for optimizing the gamma parameter of Radial Basis Functions Kernel and the regularization parameter C of SVM models. The first population of solutions is generated by randomizing values. The optimization goals that were formulated as Fitness Functions were the following:

- **Fitness Function 1:** Area Under The Receiver Operating Characteristic (ROC) Curve - AUC
- **Fitness Function 2:** 1/(1+number of support vectors)
- **Fitness Function 3:** 1/(1+number of selected features)

After the evaluation of the population, the Pareto fronts of non-dominated solutions are calculated, and solutions are assigned a fitness value based on their Pareto front. The Roulette Wheel Selection method is applied to generate a new population of solutions which are then differentiated using the Genetic Algorithms two-point crossover and Gaussian mutation operators. The new population is evaluated again, and this iterative process continues until it converges (the best solution's performance is less than 5% away from the mean performance of the population for 5 consecutive generations) or reaches the maximum number of generations.

5.2.5 PlaqueMS Knowledge Base

All datasets used, statistical analysis and network results produced along with relative datasets and analysis results of the Plaqomics Leducq consortium (257), were stored in the PlaqueMS relational database to ease access. WampServer (version 3.2.3) was used to create PlaqueMS and more specifically, phpMyAdmin (version 5.0.2) to create the project and MySQL (version 5.7.31) to create the database.

5.3 Results

5.3.1 Discovery cohort and patient characteristics

219 human carotid plaques from 120 patients were analyzed with the discovery MS proteomics. The clinical characteristics of the cohort are shown in **Error! Reference source not found.**. Symptomatic patients had suffered from TIA and stroke and had higher levels of high-sensitivity CRP and LDL and lower HDL cholesterol compared to asymptomatic patients. Based on the ultrasound and histology characterization of the

plaque, no significant changes were observed. Regarding the medications only the use of statins was significantly higher in symptomatic patients, so we corrected for statins usage in all statistical comparisons we conducted.

	Overall (n=120)	Asymptomatic (n=78)	Symptomatic (n=42)	P-value
Demographics				
Age	70 (64-74)	70 (63-73)	71 (66-75)	0.21
Sex (male)	88 (73.3)	54 (69.0)	34 (80.1)	0.20
Characteristics of carotid artery stenosis				
Grade of stenosis	90 (85-90)	90 (90-95)	90 (80-90)	**0.03**
Stenosis grade ≥90%	67 (55.8)	47 (60.3)	20 (47.6)	0.25
Contralateral stenosis	34 (28.3)	19 (24.4)	15 (35.7)	0.21
Peak systolic flow velocity, m/s	3.8 (3 – 4.5)	3.9 (2.9 -4.5)	3.6 (3 – 4.4)	0.61
Plaque morphology				
Echogenic	60 (50)	43 (55.1)	17 (40.5)	
Mixed	26 (21.7)	15 (19.2)	11 (26.2)	0.31
Echolucent	34 (28.3)	20 (25.6)	14 (33.3)	
Histological AHA classification				
Type V fibroatheroma	32 (26.7)	21 (26.9)	11 (26.2)	
Type VI complex lesion	52 (43.3)	34 (43.6)	18 (42.9)	0.96
Type VII calcified lesion	20 (16.7)	12 (15.4)	8 (19.1)	
Type VIII fibrotic lesion	16 (13.3)	11 (14.1)	5 (11.9)	
Calcified (clinical†)	70 (58.3)	46 (59.0)	24 (57.1)	0.85
Comorbidities and risk factors				
Transient ischemic attack	23 (19.2)	0 (0)	23 (54.7)	**<0.001**
Stroke	24 (20.0)	0 (0)	24 (57.1)	**<0.001**
History of stroke/TIA	23 (19.2)	14 (18.0)	9 (21.4)	0.64
Acute myocardial infarction	25 (20.8)	17 (22.0)	8 (19.1)	0.82
Coronary artery disease	42 (35.0)	24 (30.8)	18 (42.9)	0.23

Peripheral artery disease	49 (40.8)	37 (47.4)	12 (28.6)	0.05
Arterial hypertension	108 (90.0)	70 (90.0)	38 (90.1)	1
Diabetes mellitus type 2	36 (30.0)	26 (33.0)	10 (23.8)	0.30
Adipositas (BMI>30)	28 (23.3)	20 (25.6)	8 (19.1)	0.50
Smoking active	29 (24.2)	19 (24.4)	10 (23.8)	1
Past smoker	44 (36.7)	31 (39.7)	13 (31.0)	0.43
Pack-years	20 (0-45)	20 (0-45)	20 (0-40)	0.47
COPD	28 (23.3)	19 (24.4)	9 (21.4)	0.82
Medications				
Ace inhibitors	48 (40.0)	32 (41.03)	16 (38.10)	0.76
Angiotensin receptor blockers	41 (34.17)	24 (30.77)	17 (40.48)	0.29
Beta-blockers	79 (65.83)	52 (66.67)	27 (64.29)	0.79
Diuretics	34 (28.33)	20 (25.64)	14 (33.33)	0.37
Statins	97 (80.83)	59 (75.64)	38 (90.48)	**0.048**
Marcoumar	6 (5)	5 (6.41)	1 (2.38)	0.33
Laboratory parameters				
LDL, mg/dL	83 (68-108)	80 (65-101)	91 (79-111)	**0.02**
HDL, mg/dL	49 (41-56)	50 (41-60)	45 (38-51)	**0.03**
Total cholesterol, mg/dL	164 (144-193)	163 (141-191)	166 (148-194)	0.51
Triglycerides, mg/dL	129 (99-197)	123 (95-191)	157 (107-198)	0.27
High-sensitivity CRP, mg/dL	0.3 (0.1-0.6)	0.2 (0.1-0.5)	0.3 (0.2-0.9)	0.03

Table 5.1. Clinical characteristics of the patient cohort. Continuous data are shown as median (interquartile range). Dichotomous data are shown as n (%). The Mann-Whitney test was used for the statistical comparison of continuous variables between symptomatic and asymptomatic plaques and Fisher's exact test for the categorical variables. The Chi-square test was used for categorical variables of more than two classes (ultrasound and histology). †Plaques were characterized as calcified or non-calcified based on the classification by two clinicians. This clinical classification was in 83% agreement with the calcification characterization of a subset of plaques by computed tomography angiography (n=35).

5.3.2 Results of TOPS label-free discovery proteomics

To profile ECM-associated proteins in plaques and expand on the findings of a previous study of our lab (17), we used label-free (TopS) MS proteomics, which maximizes protein identifications (see section 1.2.2). Sequential extractions from two different extracts, GuHCl extract for the core matrisome and NaCl extract for the soluble matrisome, was used to obtain comprehensive coverage of extracellular proteins. Using TopS MS in these two extracts we were able to identify and consistently quantify (quantified in >70% of the samples) 421 ECM proteins in 201 plaque samples (Supplementary Tables 4 and 5, https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis).

Then, we explored the differences in protein abundance between the two different regions of the plaques, the core and the periphery. The proteome of the core and periphery of the plaque was markedly different, highlighting the differences in protein abundance between the early stages (periphery) and advanced stages (core) of disease (**Error! Reference source not found.**).

To overcome the spatial variation in protein abundance within each plaque, we divided the plaque area into core and periphery specimens and investigated the protein differences occurring from calcification and symptomatic status in the core of the plaques. Among the most significantly upregulated proteins in calcified plaques were fetuin-A (FETUA), osteopontin (OSTP), collagens alpha-1 (CO1A1, CO3A1, COLA1, CO8A1, COCA1) and alpha-2 (CO1A2, CO4A2, CO6A2, CO8A2), certain apolipoproteins such as apolipoprotein A-IV (APOA4) and apolipoprotein C (APOC2, APOC3), and gamma-carboxylated proteins such as growth-arrest specific 6 (GAS6), matrix Gla protein (MGP), coagulation factors (FA7, FA9, FA10), prothrombin (THRB) and vitamin K-dependent proteins (PROC, PROZ) (**Error! Reference source not found.**). It is noteworthy that all gamma-carboxylated proteins remained significant after correction for multiple testing.

Error! Reference source not found. depicts the extracellular protein changes between symptomatic and asymptomatic plaques in the core of the plaques. Cathepsins D (CATD), B (CATB), Z (CATZ) and H (CATH), ferritin heavy (FRIH) and light chain (FRIL) and the macrophage / monocyte differentiation antigen CD14 (CD14) were among the

most significantly upregulated proteins in symptomatic plaques, with most of these changes remaining significant after correction for multiple testing (CATD, FRIH, FRIL, CD14). On the other hand, among the most significantly downregulated proteins in symptomatic plaques were gamma-carboxylated proteins (PROC, PROZ, MGP, FA7, FA9, FA10, GAS6), fetuin-A and osteopontin, with gamma-carboxylated proteins remaining significant after correction for multiple testing. Compared to the previous work by Langley et al. (17), this analysis validated 3 out of 4 protein biomarkers for symptomatic carotid plaques (cathepsin D, MMP9 and galectin-3-binding protein). The label-free experiments were also used for the identification of proteotypic peptides to be used for the targeted proteomics pipeline.

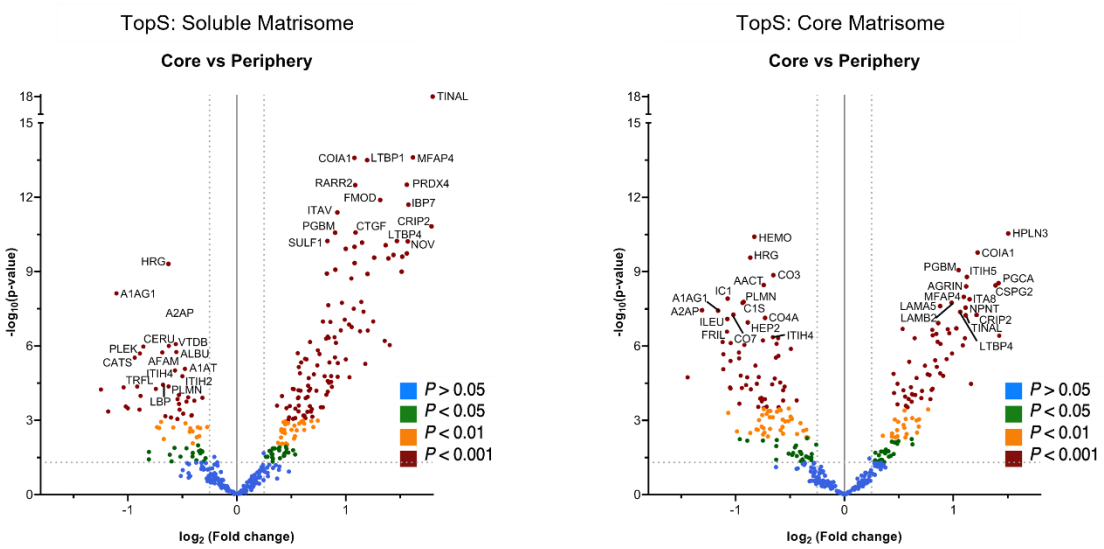


Figure 5.2. Extracellular protein changes between the core and periphery of the plaque in label-free MS proteomics. Volcano plots of significantly dysregulated proteins in the core (n = 100) vs periphery (n = 101) comparison of the plaque in TopS proteomics for the soluble (NaCl) and the core (GuHCl) matrisome respectively. The 15 most significantly dysregulated proteins (P-value<0.001) are labelled. Protein changes with absolute values of fold change <0.25 are labelled in blue and considered not significant.

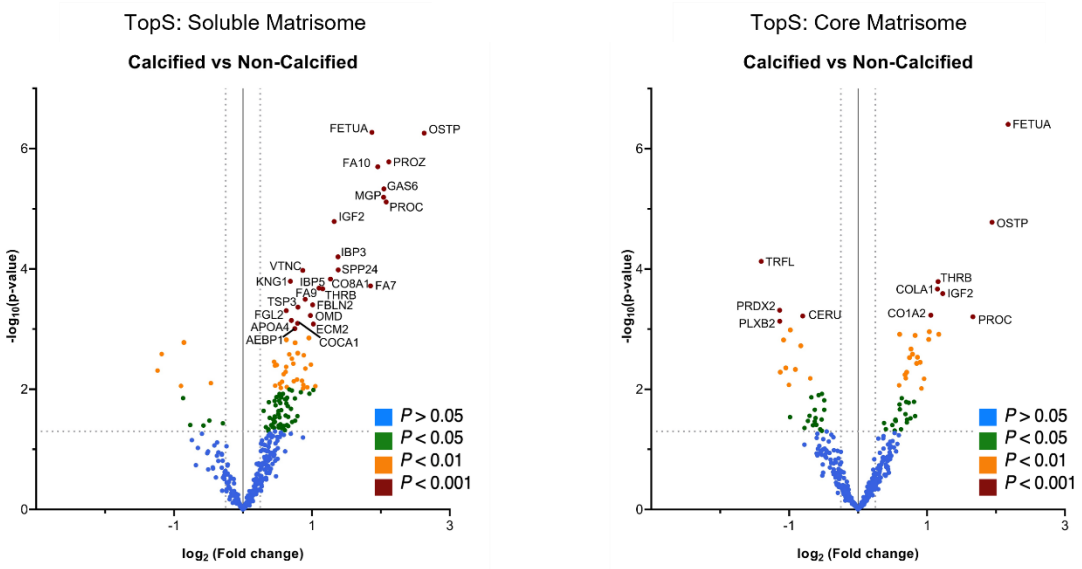


Figure 5.3. Extracellular protein changes in core specimens, between calcified and non-calcified plaques in label-free proteomics. Volcano plots of significantly dysregulated proteins in calcified (n = 57) vs non-calcified (n = 43) plaques in TopS proteomics for the soluble (NaCl) and the core (GuHCl) matrisome respectively. The most significantly dysregulated proteins (P-value<0.001) are labelled. Protein changes with absolute values of fold change <0.25 are labelled in blue and considered not significant.

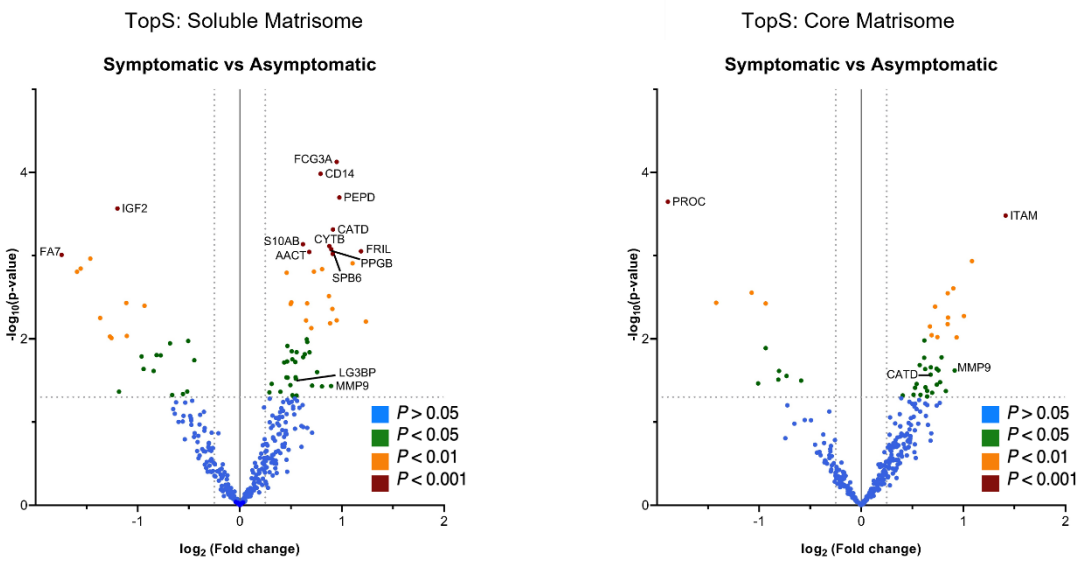


Figure 5.4. Extracellular protein changes in core specimens between symptomatic and asymptomatic plaques in label-free proteomics. Volcano plots of significantly dysregulated proteins in symptomatic (n = 35) vs asymptomatic (n = 65) plaques in TopS proteomics for the soluble (NaCl) and the core (GuHCl) matrisome respectively. The most significantly dysregulated proteins (P-value<0.001), as well as proteins identified as biomarkers for symptomatic plaques in a previous study of our lab(17), are labelled. Protein changes with absolute values of fold change <0.25 are labelled in blue and considered not significant.

5.3.3 Coverage of atherosclerotic plaque proteome and comparison to previous studies

Because of the better quantitative accuracy of labelled MS proteomics compared to label-free MS (258), we used TMT multiplexing. To achieve coverage of cellular and extracellular proteins, sequential extractions were performed for soluble matrisome (NaCl), core matrisome (GuHCl) and cellular (SDS) proteins, as described before (182). The analysis of these three extracts resulted in the consistent (<30% missing values) quantification of 1459 cellular proteins in the SDS extract (Supplemental Table 6, https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis) and 381 extracellular proteins in the NaCl (Supplementary Table 7, https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis) and GuHCl (Supplementary Table 8, https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis) extracts (283 and 286 proteins respectively).

A comparison of the two different proteomics methods used (in the two common extracts) is shown in **Error! Reference source not found.**. As expected, the label-free method was able to provide larger protein coverage than TMT, but most of the identified proteins were common (86.2%, with 12.9% of proteins being identified by the TopS method only). Thus, we continued the rest of our analysis focusing only on TMT datasets.

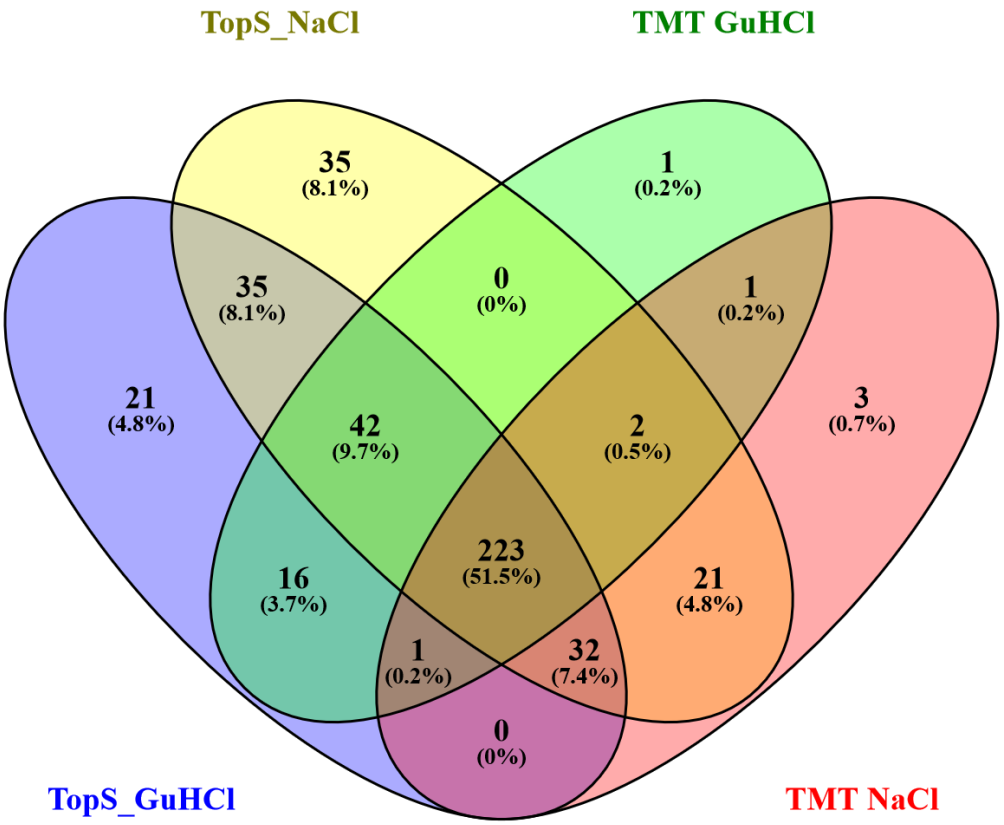


Figure 5.5. Comparison of label-free and TMT MS proteomics methods used. Venn diagram created using the Venny web tool (259).

Error! Reference source not found. depicts the proteomics studies on carotid endarterectomy samples so far, with most of them being small in sample size. The largest study was the one of Lepedda et al. (260)(carotid endarterectomy samples from 48 patients, 19 stable versus 29 unstable plaque samples), which is much smaller than our current study und used MALDI-TOF instead of LC-MS/MS (n=219). Compared to the work of Langley et al. (17), we not only expanded the sample size and confirmed the major findings, but also identified more than 6 times more proteins (**Error! Reference source not found.**).

Reference	Sample Size	Proteomics Method
Lepedda et al. *Atherosclerosis,* 2009 (260)	48 samples	Label-free (MALDI-TOF) MS
Olson et al. *Biochem Biophys Res Commun,* 2010 (261)	20 samples	Label-free (SELDI-TOF) MS
Rocchiccioli et al. *J Transl Med*, 2013 (262)	28 samples	Discovery Label-free MS
Malaud et al. *Atherosclerosis,* 2014 (263)	24 samples	Discovery Label-free MS
Aragones et al. *J Proteome Res,* 2016 (233)	12 samples	Discovery Labelled (ITRAQ-8) MS
Langley et al. *J Clin Invest,* 2017 (17)	12 samples	Discovery Label-free MS
Ucciferri et al. *Talanta,* 2017 (264)	13 samples	Discovery Label-free MS
Hansmeier et al. *J Proteome Res,* 2018 (265)	12 samples	DIA MSy
Ward et al. Biology of Sex Differences, 2018 (266)	20 samples	Discovery Label-free MS

Nehme et al. *Hypertension Research,* 2019 (267)	12 samples	Discovery Label-free MS & Targeted (MRM for 50 proteins) MS

Table 5.2. Proteomics studies on carotid endarterectomy samples.

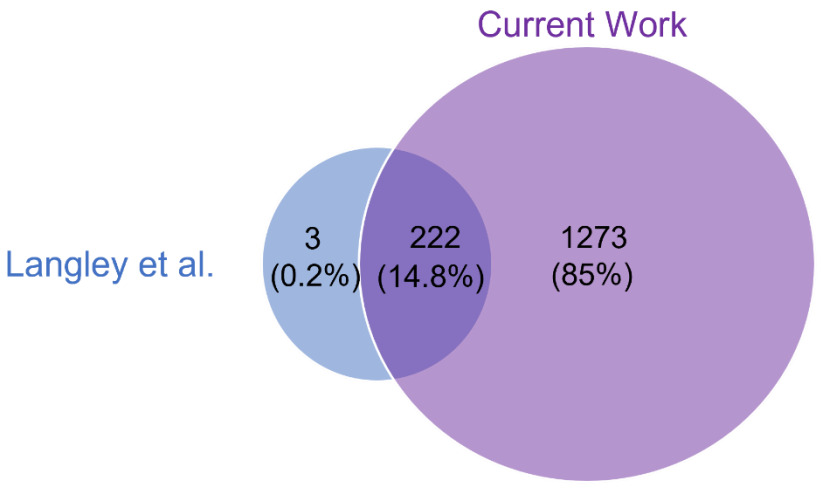


Figure 5.6. Comparison to the previous carotid atherosclerotic plaque study from our lab.

The largest proteomics study on atherosclerosis so far is the one from Herrington et al. (61), who used MS to analyze 200 specimens from coronary arteries and aortas from autopsies of 100 patients. As depicted in **Error! Reference source not found.**, the current study almost doubled the proteomic coverage of human atherosclerotic plaques. Degradation mechanisms of the proteome make studies using autopsies not suitable for the identification of proteomic inflammation and other signatures involved in atherosclerosis.

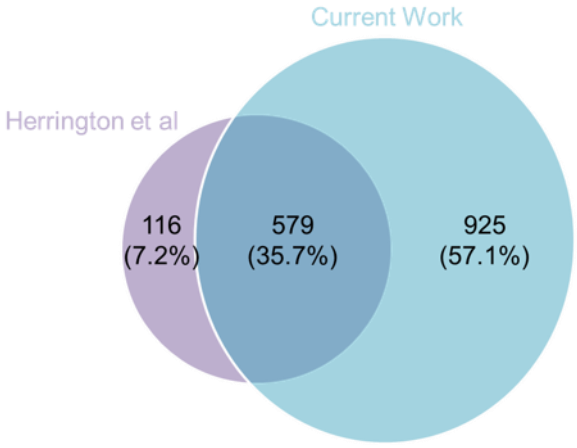


Figure 5.7. Comparison to the largest proteomics study in atherosclerosis so far.

5.3.4 Cellular proteome characteristics of plaque core and periphery

Principal components analysis of cellular proteomics extracts (SDS) revealed a clear separation of the core and the periphery of the plaque (**Error! Reference source not found.**A), with a logistic regression model fitted on the 3 most important principal components being able to classify the core and periphery samples with 93.5% accuracy. Among the most significant changes between the core and periphery (**Error! Reference source not found.**B) were several known markers for SMCs (aortic smooth muscle cell actin - ACTA, transgelin - TAGL, calponin-1 - CNN1, and caldesmon - CALD1) and leukocytes (monocyte differentiation antigen CD14, macrophage scavenger receptor types I and II - MSRE, scavenger receptor cysteine-rich type 1 protein M130 - C163A, macrophage mannose receptor 1 - MRC1, and receptor-type tyrosine-protein phosphatase C - PTPRC or CD45). SMC content was lower in the core of the plaque compared to the periphery, while markers for leukocytes were significantly elevated in the plaque core (**Error! Reference source not found.**C).

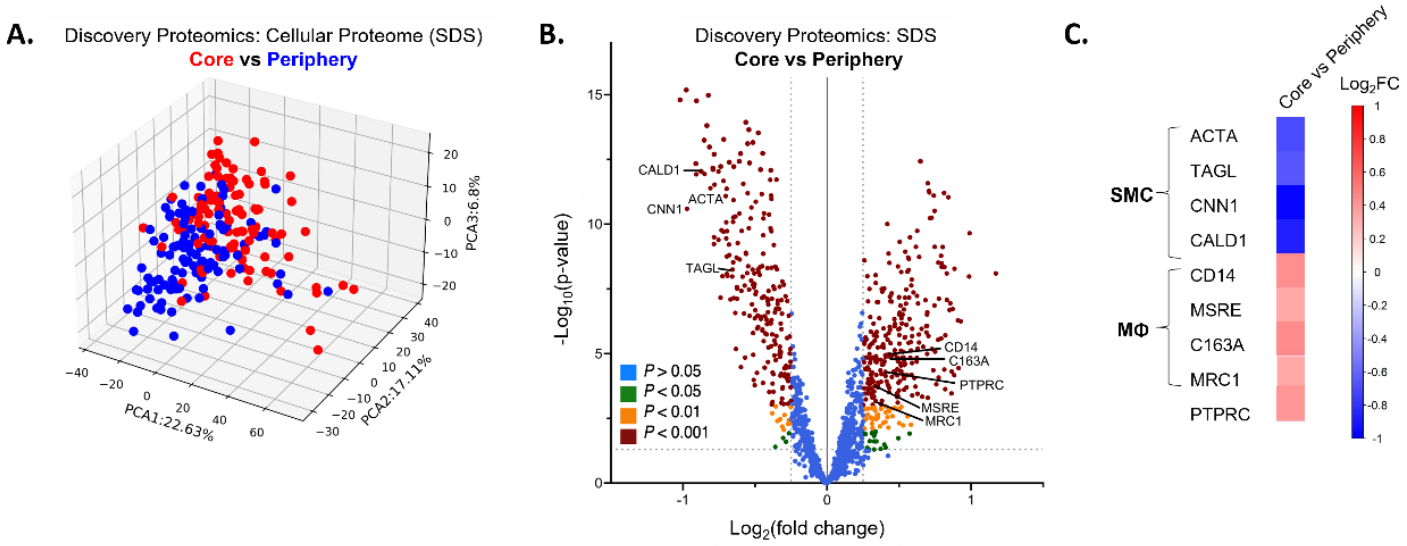


Figure 5.8. Proteomic signatures of the core and periphery of carotid plaques. A. Visualization of the 3 most significant principal components of the carotid plaque samples from the core (red) and periphery (blue) using the proteomics measurements in SDS extract. **B.** Volcano plot depicting the results of the statistical comparison of core (n=101) vs periphery (n=106) of the plaque. Protein changes with absolute values of fold change <0.25 are labelled in blue and considered not significant. Known cellular markers are labelled. **C.** Heatmap depicting the log2 fold changes of the smooth muscle cell and leucocyte (MΦ and hematopoietic cells) markers between the core and periphery of the plaques. PTPRC (CD45) is a marker for hematopoietic cells.

The regional proteomics data were confirmed by spatial RNA-sequencing data (Supplemental Table 9, https://github.com/Cardiovascular-Bioinformatics/MariaHasman_Thesis), which revealed three positional clusters (Figure 5.9. Regional signatures of carotid plaques. A. The three regional clusters (red, blue, green) were revealed from spatial RNA-seq. B. Scatterplot depicting the enrichment of each cluster in protein changes. C1 and C2 corresponded to the protein changes identified in plaque periphery (blue) and core samples (red) (Fisher's Exact test p-value<0.05) and C3 corresponds to an intermediate position. C. Changes in the cellular proteome (SDS extract) of the plaque cores according to cardiovascular risk factors, sex, symptoms and calcification. Scatterplots show the p-values and blue boxes depict the number of proteins with p-value<0.05 in each comparison. D. Scatterplot depicting the pathway enrichment analysis of common protein (SDS: core vs periphery) and RNA (spatial RNAseq: C1 vs C2) changes using Benjamini-Hochberg corrected q-value threshold of 0.01 and Gene Ontology, KEGG and Reactome functional and pathway terms.Figure 5.9.A). Enrichment analysis of the proteomics and transcriptomics data showed that two clusters were associated with changes in the core or periphery of the plaque (Figure 5.9.B). The other cluster was an intermediate region between the core and periphery (Figure 5.9.A, green area). When we compared the proteomics data from the core of the plaques with available clinical characteristics, most changes in cellular proteins were occurring with calcification, symptomatic status, hyperlipidemia and sex respectively (Figure 5.9.C). Enrichment analysis of significantly changing common proteins in the core versus periphery comparison and transcripts in the C2 (core) versus C1 (periphery) clusters, showed that the most pronounced changes were related to ECM and associated proteins, as well as lysosomal degradation (Figure 5.9.D).

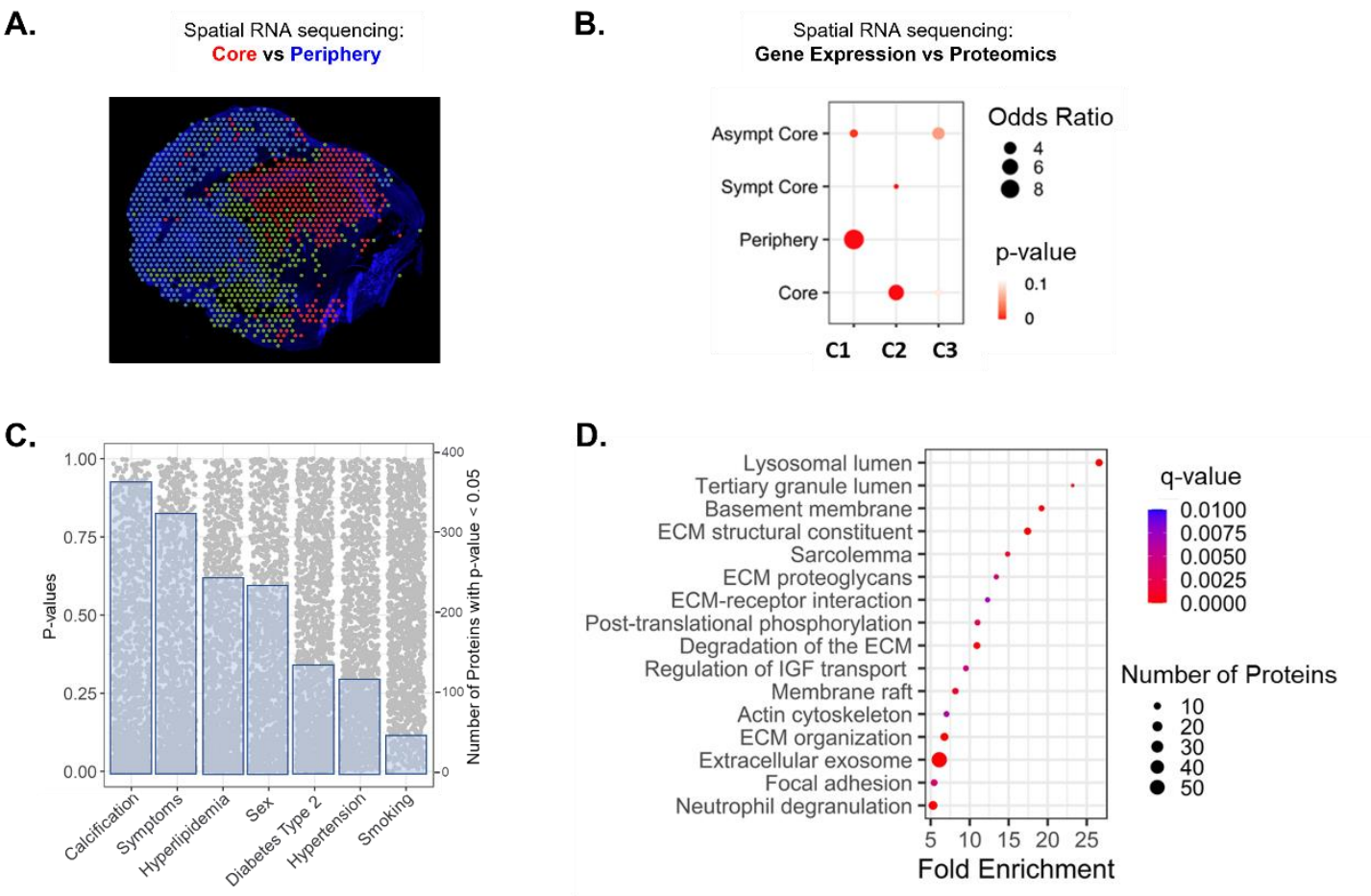


Figure 5.9. Regional signatures of carotid plaques. A. The three regional clusters (red, blue, green) were revealed from spatial RNA-seq. **B.** Scatterplot depicting the enrichment of each cluster in protein changes. C1 and C2 corresponded to the protein changes identified in plaque periphery (blue) and core samples (red) (Fisher's Exact test p-value<0.05) and C3 corresponds to an intermediate position. **C.** Changes in the cellular proteome (SDS extract) of the plaque cores according to cardiovascular risk factors, sex, symptoms and calcification. Scatterplots show the p-values and blue boxes depict the number of proteins with p-value<0.05 in each comparison. **D.** Scatterplot depicting the pathway enrichment analysis of common protein (SDS: core vs periphery) and RNA (spatial RNAseq: C1 vs C2) changes using Benjamini-Hochberg corrected q-value threshold of 0.01 and Gene Ontology, KEGG and Reactome functional and pathway terms.

Because of the enrichment results, we next explored how cell composition could impact the ECM. We assessed the Spearman correlation between known cellular markers from the SDS extract and core ECM proteins from the GuHCl extract, in the core of the plaques (**Error! Reference source not found.**).

We categorized ECM proteins according to their basic function into collagens, proteoglycans, basement membrane, elastin related and proteases/inhibitors of

proteases. We further categorized collagens into fibrillar (CO1A1, CO1A2, CO2A1, CO3A1, CO5A1, CO5A2, CO5A3, COBA1), network-forming (CO4A1, CO4A2, CO8A1, CO8A2), beaded filament (CO6A1, CO6A2, CO6A3), multiplexin (COFA1) and FACITs (COCA1, COEA1) according to their function and used the mean of their abundance to investigate the correlation to the core ECM proteins.

We observed strong positive correlations between SMC markers and structural ECM proteins (collagens, proteoglycans, basement membrane and elastin-related proteins). In contrast, CD14 was inversely related to these proteins but positively correlated to proteases linked to ECM degradation, such as cathepsins B (CATB), D (CATD), G (CATG), Z (CATZ) and macrophage metalloelastase (MMP12), as well as metalloproteinase inhibitors 1 (TIMP1) and 3 (TIMP3). Alternative macrophage markers like C163A, MRC1 and galectin-3 (LEG3) were not associated with a loss of structural ECM proteins. Instead, alternative macrophage markers and PTPRC correlated positively to basement membrane-associated proteins, including laminins (LAMA4, LAMA5, LAMB1, LAMB2, LAMC1), nidogen 2 (NID2), and von Willebrand Factor A domain-containing 1 (VWA1), as well as the beaded filament collagen type VI.

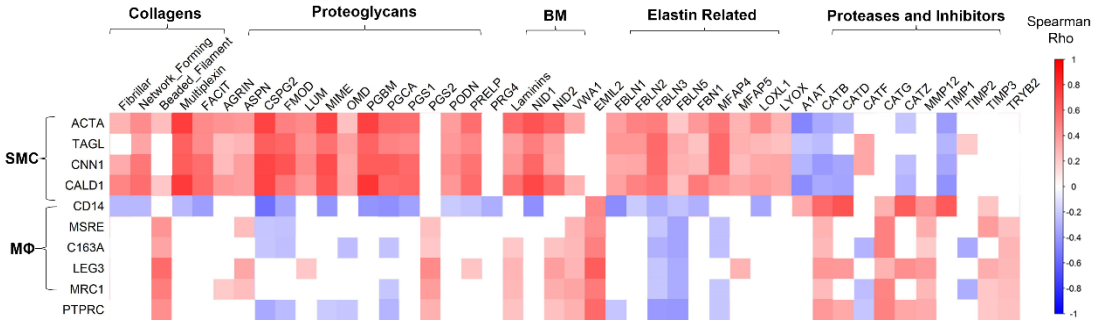


Figure 5.10. Association of cellular markers and core ECM. Heatmap visualises the Spearman correlation between cellular markers quantified in the SDS extract and groups of core ECM proteins quantified in the GuHCl extracts, for the core of the plaques. Only significant correlations (p-value<0.05) are visualized. PTPRC is a marker for hematopoietic cells. BM: basement membrane.

We then filtered the cellular proteome for cell receptor-associated proteins and performed network reconstruction analysis (**Error! Reference source not found.**) and soft clustering to the network. The clustering revealed nine significant clusters, representing CD14+ monocytes/macrophages, neutrophils, leucocytes/macrophages,

SMCs, G-proteins, blood coagulation-associated proteins, and adhesion-related proteins (according to the literature and enrichment analysis performed for the proteins of each cluster, **Error! Reference source not found.**). A filtered network, containing only the proteins that belong to a significant cluster is presented in **Error! Reference source not found.**. Cathepsins were predominantly associated with the CD14 cluster. The neutrophil cluster contained several proteins that we had previously reported as part of the inflammatory signature of symptomatic atherosclerotic plaques (S10A8/A9, MMP9, TIMP1) (17).

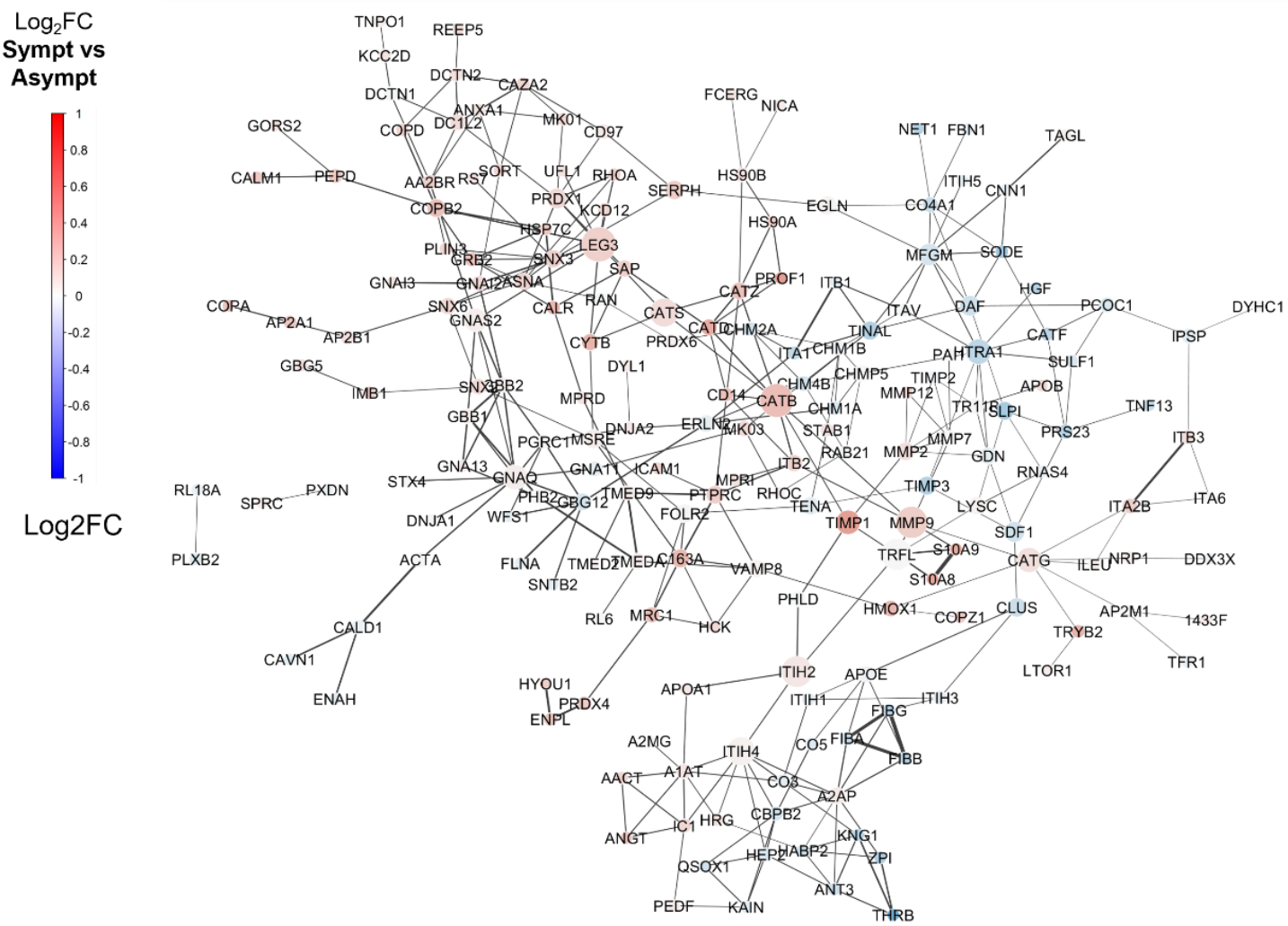


Figure 5.11 Cell receptor-associated protein network. Co-expression network of receptors using the proteomics data from the SDS extract. Node size is relative to the absolute log2 fold change in symptomatic vs asymptomatic plaques comparison. Node color scale is used to depict this log2 fold change. Edge width is relative to the conditional mutual information calculated for the connected proteins.

Cluster	P-value	Proteins
CD14+ cells	0.001	CATB, PROF1, CATD, HS90A, ITB2, CD14, CATZ, CATS, HS90B, PROF1
Neutrophils	0.007	S10A8, S10A9, TIMP1, TRFL, PHLD, MMP9, ITIH2
Leucocytes /Macrophages	0.011	MSRE, VAMP8, HCK, PTPRC, C163A, FOLR2, MRC1, TMEDA, TMED9
SMCs	0.013	CAVN1, CALD1, ENAH, ACTA
G-protein cluster	0.015	DNJA1, GNAQ, GNA13, PHB2, GBB1, GBB2, PGRC1, GNAS2
Blood Coagulation	0.02	KNG1, THRB, HABP2, ANT3, CO3, A2AP, ZPI, A2AP, FIBG, FIBA, FIBB, HRG, ITIH4, A1AT, IC1, ANGT, AACT, CO5, CBPB2, KAIN, HEP2, QSOX1
Adhesion and apoptosis cluster	0.033	COPD, COPB2, MPRD, CALR, GRB2, HSP7C
Unknown cluster 3	0.038	HYOU1, ENPL, PRDX4
Platelet cluster	0.04	ITB3, ITA6, ITA2B

Table 5.3 Cell receptor-associated protein network clustering. Clustering was performed using the soft clustering algorithm ClusterOne.

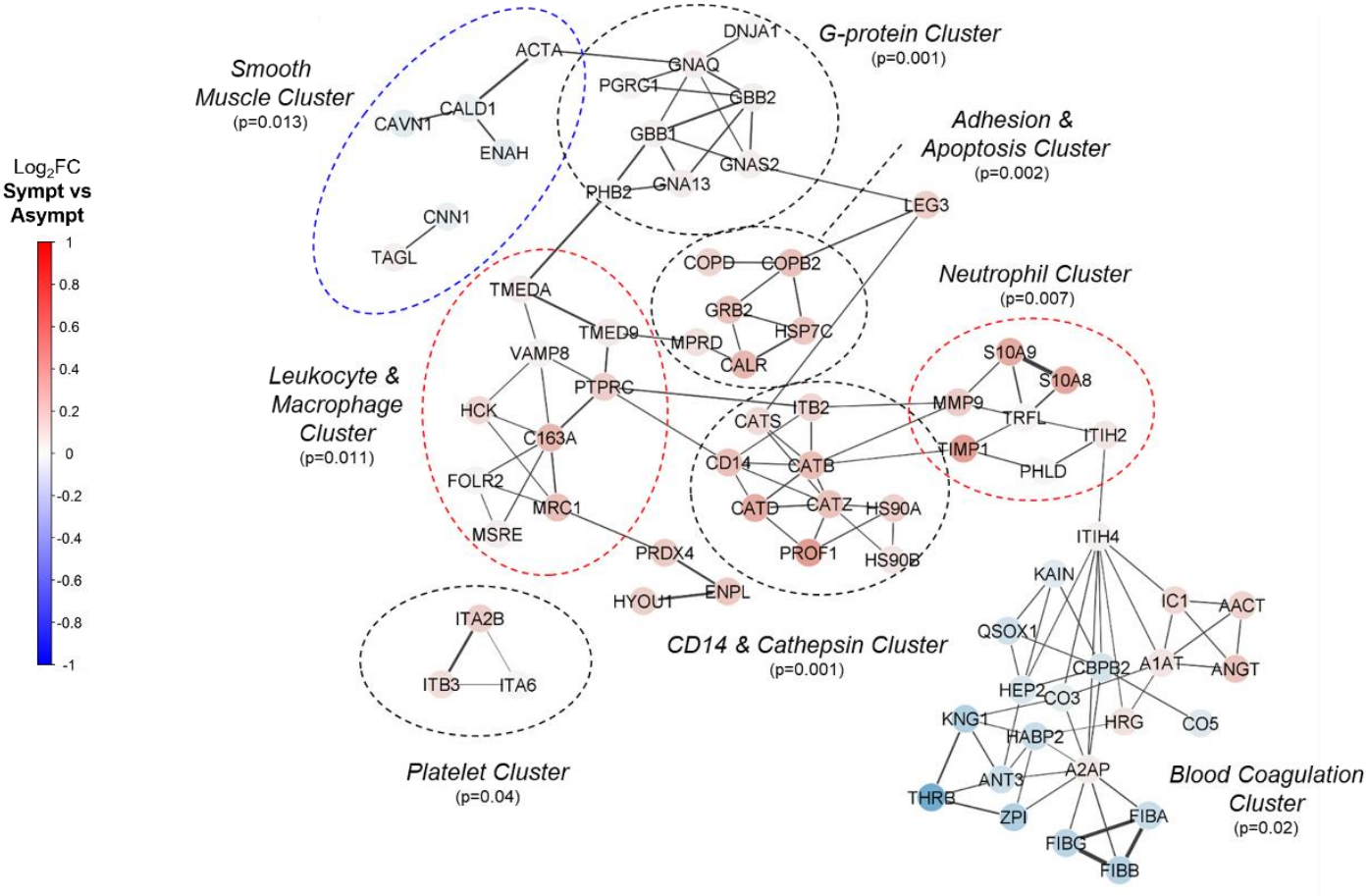


Figure 5.12 Co-expression network of significant clusters of receptor-associated proteins. The color scale is used to depict the log2 fold change in the comparison between plaque cores from symptomatic vs asymptomatic patients based on the SDS extracts. Edge width is relative to the conditional mutual information calculated for the connected proteins. Clustering was conducted using the soft clustering method ClusterOne algorithm (p-value < 0.05). The depicted network is a view of the overall network including the significant clusters and interactions among the proteins composing them.

5.3.5 Extracellular protein changes in calcification

Because of the importance of the ECM in plaque stability and calcification (268–270) we performed statistical comparisons between the core of calcified and non-calcified plaques in the soluble and core matrisome (**Error! Reference source not found.**). Alpha-2-HS-glycoprotein/Fetuin-A was the most pronounced change in both extracts. Along with FETUA, osteopontin, gamma-carboxylated proteins (PROC, PROZ, GAS6, MGP, THRB, FA9, FA10) and some apolipoproteins, such as apolipoproteins C (APOC1, APOC2, APOC3, APOC4) and A-IV (APOA4), were among the significantly upregulated proteins in calcified plaques. On the other hand, plaque calcification was related to

reduced expression of inflammatory proteins such as the calprotectin complex (S10A8/A9), as well as ferritin light and heavy chain.

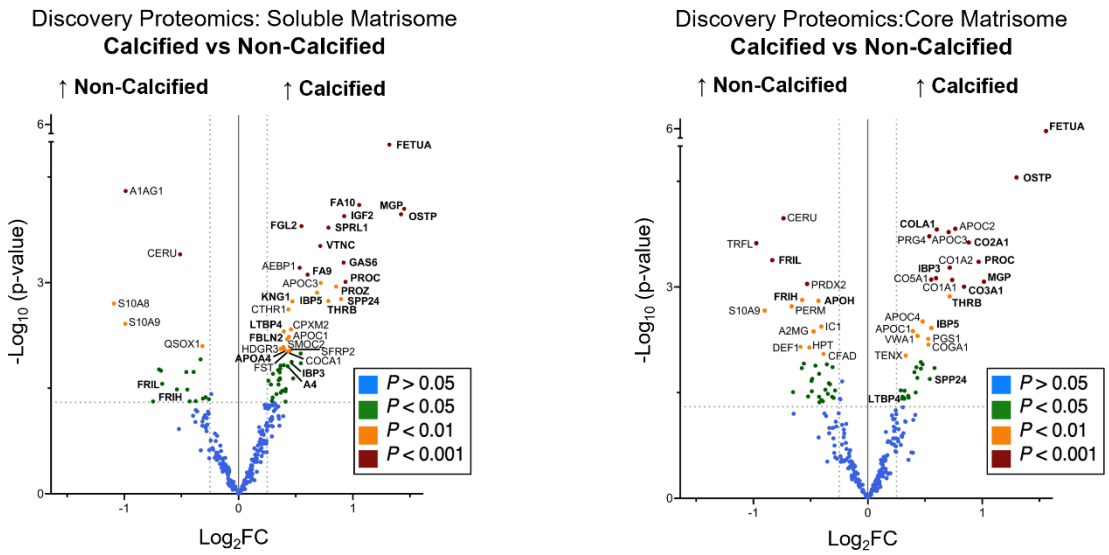


Figure 5.13. Extracellular protein changes in calcified plaques. Volcano plots of significantly dysregulated proteins in the calcified (n = 60) vs non-calcified (n = 46) comparisons of the plaque cores for the soluble matrisome (NaCl) and the core matrisome (GuHCl). Protein changes with absolute values of fold change <0.25 are labelled in blue and considered not significant. Proteins with p-value<0.01 are labelled. Significant proteins in both calcified vs non-calcified and symptomatic vs asymptomatic comparisons are labelled in bold.

We then performed hierarchical clustering of the significantly changing proteins with calcification and revealed three clusters (**Error! Reference source not found.**). Pathways (KEGG), biological processes and molecular functions (Gene Ontology) significantly enriched (q-value<0.05) in the revealed clusters were platelet activation, collagen fibril organization and gamma-carboxylation for the first cluster, complement and coagulation cascades, high-density lipoprotein particle remodelling and platelet degranulation for the second, and inflammatory response for the third cluster, respectively. Therefore, this analysis confirmed that plaque calcification is not just reducing inflammation but also leading to a retention of of gamma-carboxylated proteins and a reduction of other plasma-derived proteins, including proteins of the complement and coagulation cascades and lipoproteins.

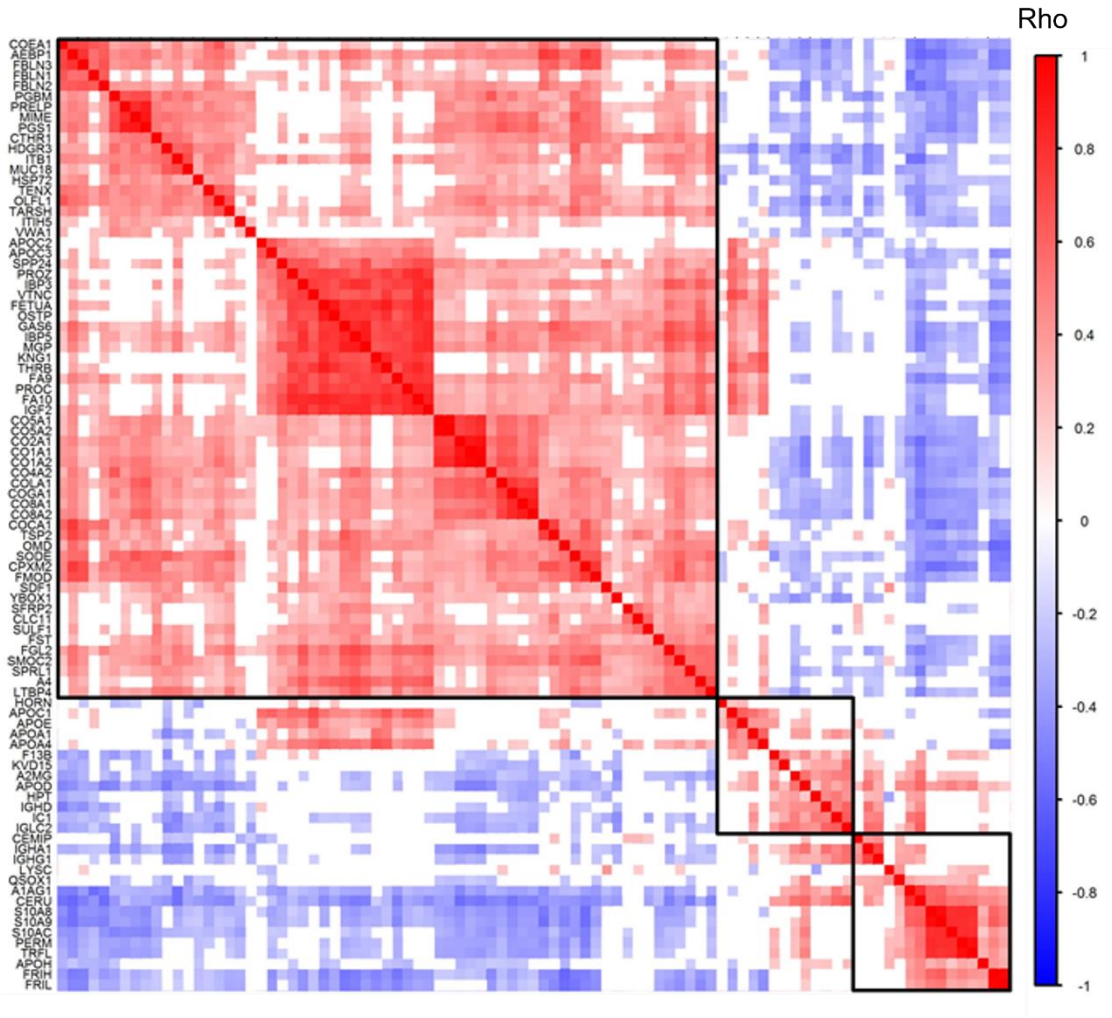


Figure 5.14 Clustering of significant proteins in calcification. Heatmap demonstrating the Spearman correlation and hierarchical clustering of ECM proteins significantly changing in the cores of calcified vs non-calcified plaques.

We finally wanted to see whether the common significant protein changes that appeared with calcification could be validated at the transcript level. We selected the common significant proteins between calcified vs non-calcified and symptomatic vs non-symptomatic comparisons (labelled in bold in **Error! Reference source not found.**). In a publicly available bulk RNA-seq dataset (271) comprising calcified (n=11) and non-calcified (n=9) fibroatheromas from carotid endarterectomy plaque sections, no corresponding significant changes were observed, showing that calcification-related changes are better captured in the protein level (**Error! Reference source not found.**).

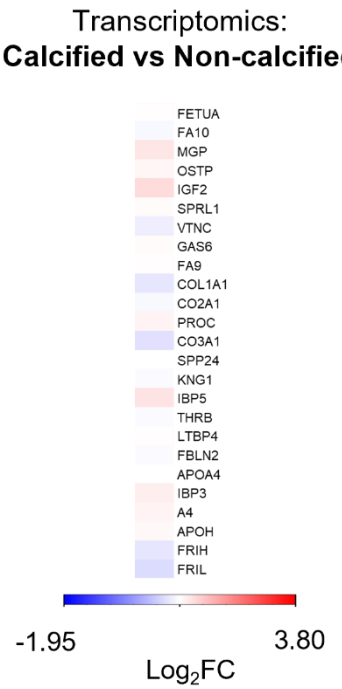


Figure 5.15 The calcification signature in transcriptomics. Heatmap showing the log2 fold changes of the corresponding transcripts for proteins significantly changing in both symptomatic vs asymptomatic and calcified vs non-calcified comparisons, between calcified (n = 11) and non-calcified (n = 9) fibroatheroma plaques from the RNA sequencing dataset GSE104140. None of the significant protein changes was replicated in the calcified vs non-calcified fibroatheroma RNA comparison.

5.3.6 Extracellular protein changes in symptomatic plaques

Significantly changing proteins for the symptomatic vs asymptomatic comparison of the plaque core, are shown in **Error! Reference source not found.**. This analysis validated previously identified (17) significantly upregulated proteins in symptomatic plaques such as cathepsins D and B, and also identified novel changes, including the upregulation of the hyaluronic acid receptor CD44, of adhesion G protein-coupled receptor E5 (CD97 or AGRE5) and ferritin light and heavy chains in symptomatic plaques. The calcification-related proteins observed to be upregulated in calcified plaques were found significantly increased in plaques from asymptomatic patients.

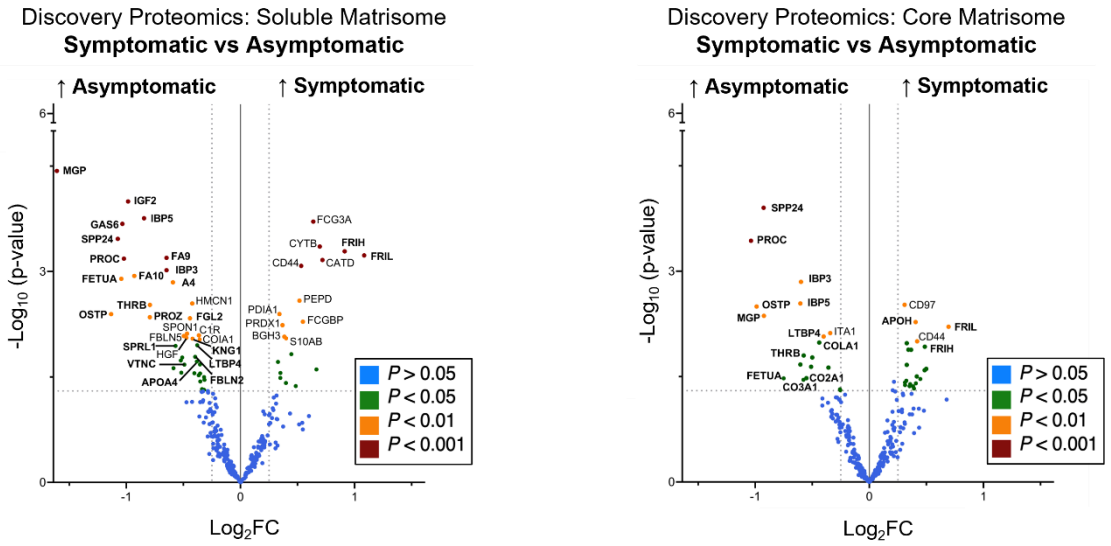


Figure 5.16 Extracellular protein changes in symptomatic plaques. Volcano plots of significantly dysregulated proteins in the symptomatic (n = 36) vs asymptomatic (n = 69) comparisons of the plaque cores for the soluble matrisome (NaCl) and the core matrisome (GuHCl). Protein changes with absolute values of fold change <0.25 are labelled in blue and considered not significant. Proteins with p-value<0.01 are labelled. Significant proteins in both calcified vs non-calcified and symptomatic vs asymptomatic comparisons are labelled in bold.

To assess whether the significant changes in symptomatic plaques were just occurring because of plaque calcification, we excluded calcified plaques from the symptomatic vs asymptomatic comparison (**Error! Reference source not found.**). FRIL and FRIH remained significantly upregulated proteins in symptomatic plaques alongside CATD and MMP9.

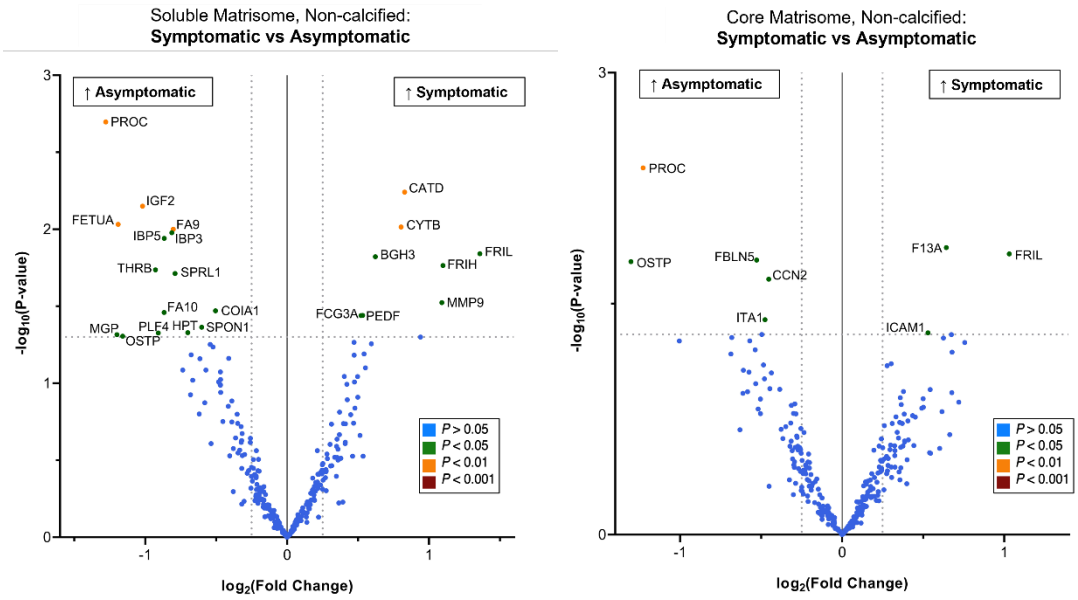


Figure 5.17 Extracellular protein changes in symptomatic plaques of non-calcified plaque samples. Volcano plots depicting results of the statistical comparisons between symptomatic and asymptomatic plaques in the non-calcified plaque cores in the soluble (NaCl, n = 18 symptomatic and n = 27 asymptomatic samples) and core (GuHCl, n = 18 symptomatic and n = 28 asymptomatic samples) matrisome. All significant proteins are labelled.

Hierarchical clustering of the significantly changing proteins revealed three clusters (Figure 5.18). Pathways (KEGG), biological processes and molecular functions (Gene Ontology) significantly enriched (q-value<0.05) in the revealed clusters were collagen-binding for cluster one, which contained plaques from symptomatic patients, gamma-carboxylation, the intrinsic pathway of fibrin clot formation and molecules associated with elastic fibers for cluster two, and complement and coagulation cascades, blood coagulation and intrinsic pathway for cluster three, which contained plaques from asymptomatic patients respectively.

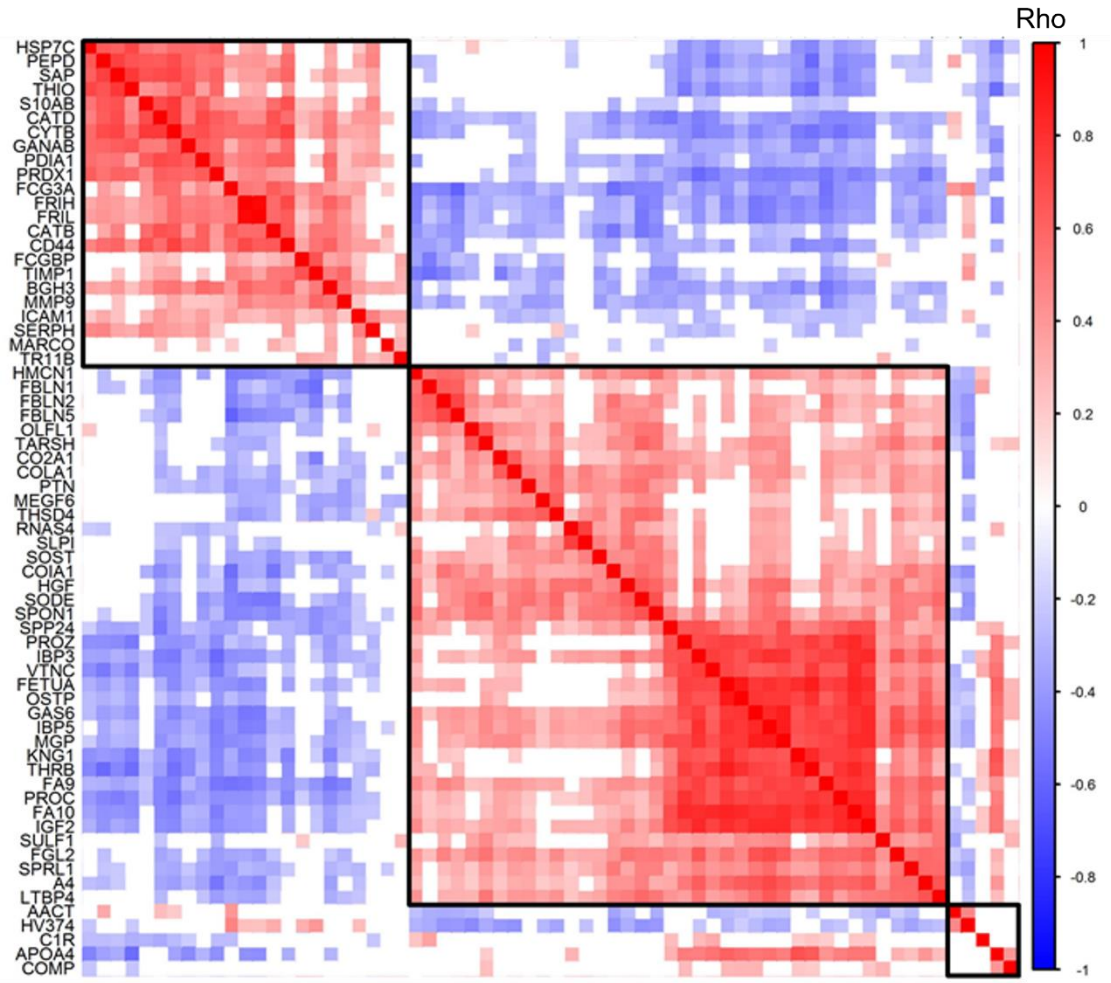


Figure 5.18 Clustering of significant ECM protein in symptomatic plaques. Heatmap demonstrating the Spearman correlation and hierarchical clustering of ECM proteins significantly changing in the core of symptomatic vs asymptomatic plaques.

We compared our proteomics findings to a publicly available bulk RNAseq dataset (272) of stable (n=4) and unstable (n=4) plaques from carotid endarterectomies (**Error! Reference source not found.**). Transcriptomics validated almost half of the changes observed at the protein level, including the upregulation of FRIH and FRIL in symptomatic plaques and the upregulation of GAS6 and MGP in asymptomatic plaques.

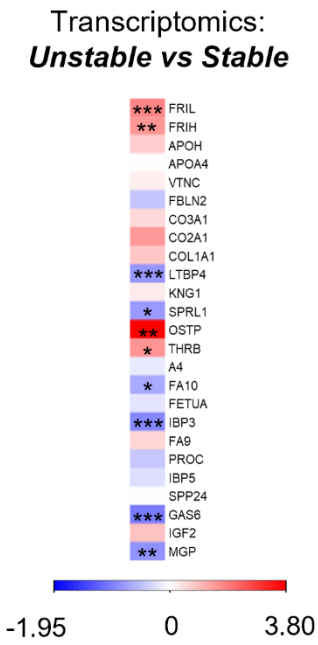


Figure 5.19 The symptomatic signature in transcriptomics. Heatmap showing the log2 fold changes of the corresponding transcripts for proteins significantly changing in both symptomatic vs asymptomatic and calcified vs non-calcified comparisons, between unstable (n = 4) and stable (n = 4) carotid plaques from the RNA sequencing transcriptomics experiment GSE120521.

Finally, we evaluated the correlation between the proteomic signatures of calcification and inflammation (symptomatic status). Consistent with the notion that calcification reduces local vascular inflammation (273), the significant proteins in calcification and inflammation were inversely correlated in both soluble and core matrisome (**Error! Reference source not found.**).

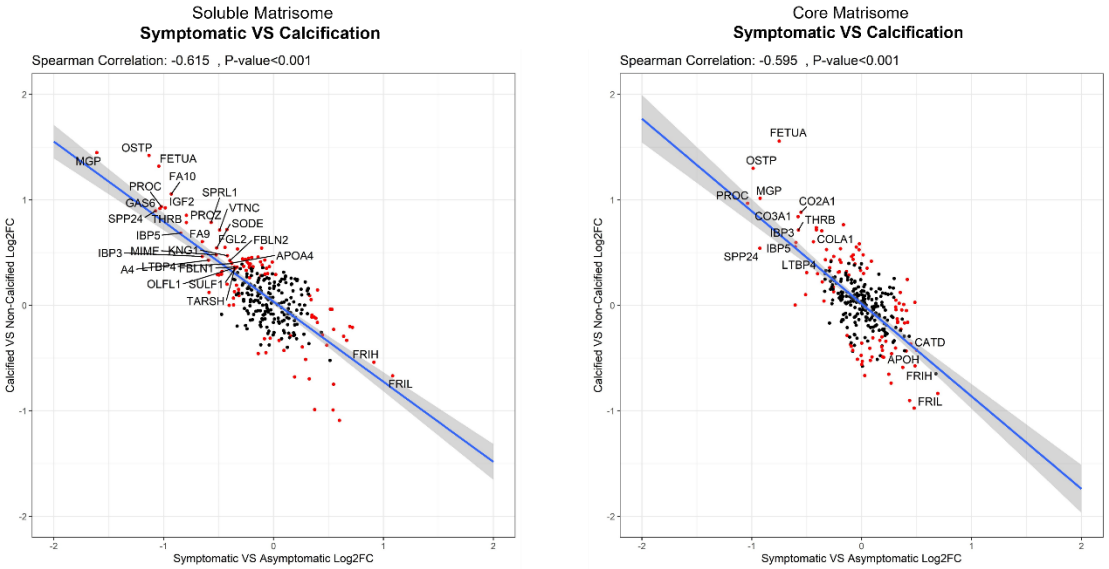


Figure 5.20 Inverse association of calcification with inflammation. Scatter plots depicting the log2 fold changes for symptomatic vs asymptomatic and calcified vs non-calcified comparisons in soluble (NaCl extract) and core (GuHCl extract) matrisome of the plaque cores. Linear regression models were fitted and Spearman correlation analysis was conducted for each extract. Significant changes are highlighted in red. Common significant protein changes in both extracts are labelled. Linear regression plots are depicted with their 95% confidence intervals. Spearman's Rho coefficient and corresponding p-values are shown.

5.3.7 Validation of extracellular changes with targeted proteomics

Targeted proteomics analysis was performed to validate the significant findings of discovery proteomics and relate extracellular protein changes to cell types. We used Parallel Reaction Monitoring (PRM) and quantified 135 ECM and ECM-related proteins of interest (Supplemental Table 1), that were significantly changing in discovery proteomics. For large proteins, we considered peptides coming from different regions of the protein as separate and calculated the statistics respectively (Supplementary Table 1). For example, for the large aggregating proteoglycan versican (CSPG2), we quantified 4 peptides originating from different regions of the protein: two from the G1, one from the G3 and one from the GAG-beta region. As supported by previous work(274), ADAMTS proteases cleave versican into different functional peptides whose abundances are not always correlated because of different degradation mechanisms. Thus in the present work, we considered as three separate read-outs (G1, G3 and GAG-beta respectively).

The agreement between discovery and targeted proteomics was high, with significant correlations exceeding r = 0.6 (**Error! Reference source not found.**).

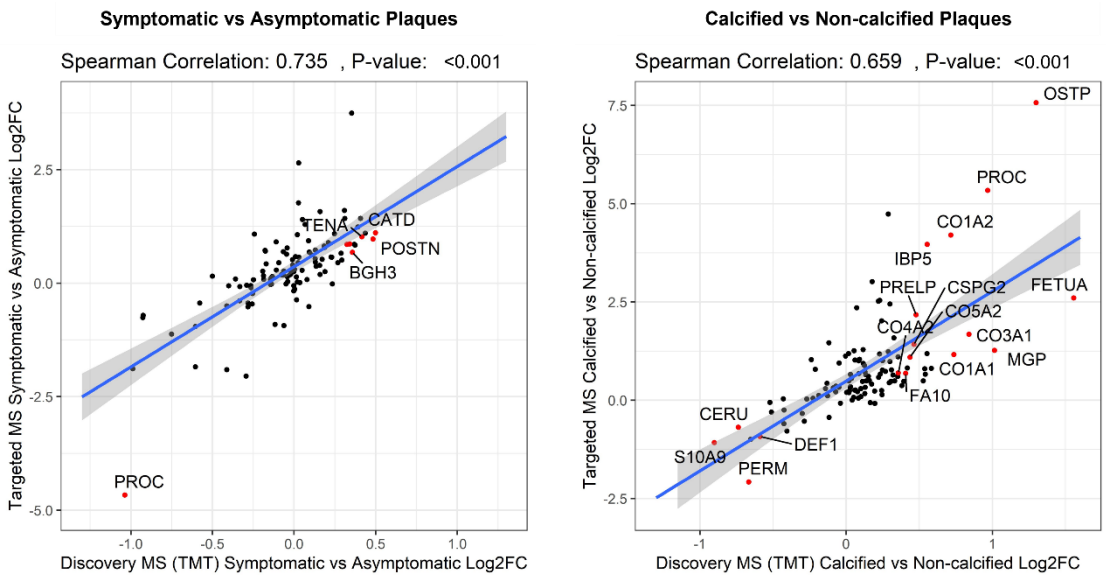


Figure 5.21. Correlation of targeted proteomics to discovery proteomics. Scatterplots depicting the log2 fold changes in the comparisons of symptomatic vs asymptomatic and calcified vs non-calcified plaque cores, between discovery and targeted proteomics. Significant proteins in both comparisons are colored red and labelled while the rest are colored black. Linear regression plots are plotted to depict their 95% confidence intervals.

Error! Reference source not found. shows the calcification signature of the plaques, with osteopontin and vitamin K-dependent protein C (PROC) being among the most pronounced changes. Versican and aggrecan (PGCA), two large aggregating proteoglycans, expanded the calcification signature. Both were higher in calcified plaques by targeted proteomics (p-value 0.0422 and 0.0009 respectively), positively correlating with smooth muscle cell markers and negatively with neutrophils and CD14+ monocytes and cathepsin cluster (**Error! Reference source not found.**.B). Most of the proteins of the calcification signature were also negatively correlated with the neutrophil content of the plaques (**Error! Reference source not found.**.B).

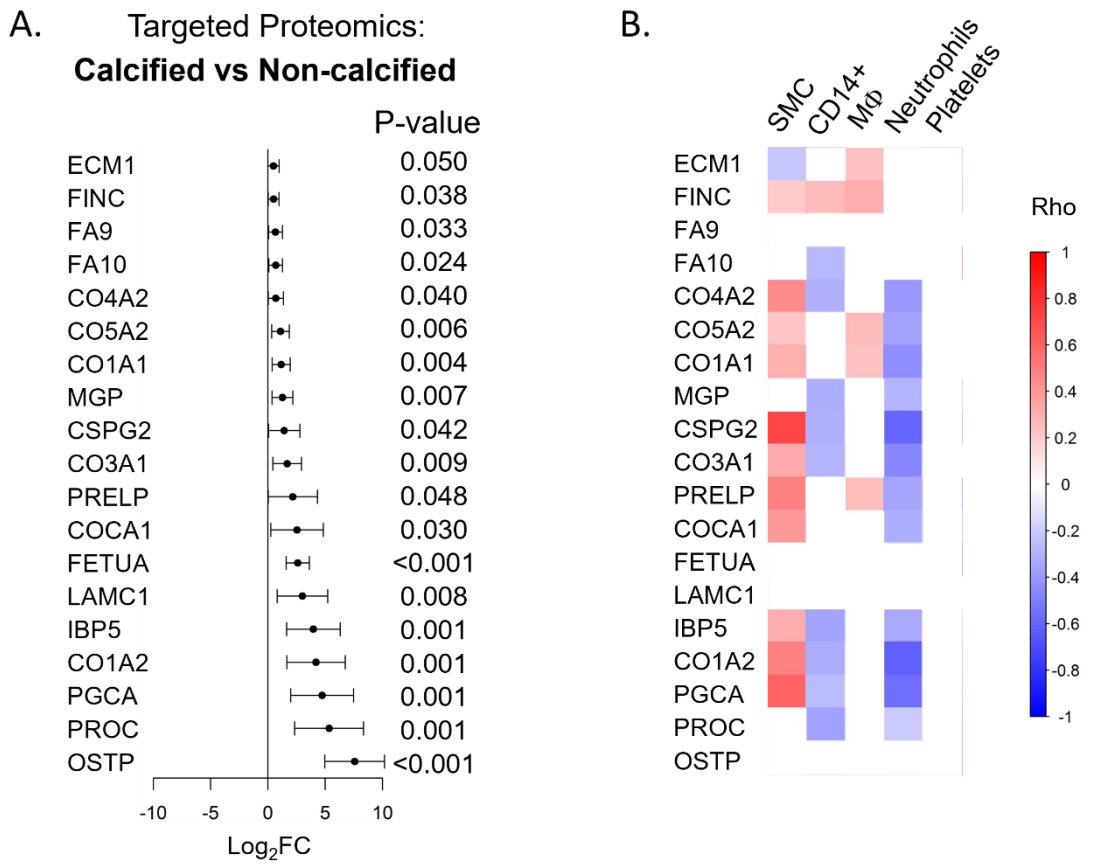


Figure 5.22. Validated calcification signature. **A.** Forest plot depicting log2 fold changes and p-values of the upregulated proteins in the core of calcified (n = 56) vs non-calcified (n = 42) plaques using targeted proteomics. **B.** Spearman correlations of these proteins against the first principal component of cell clusters from the network analysis of intracellular proteins (**Error! Reference source not found.**).

Among the inflammation signature of the core of the plaques, CD14 and myeloperoxidase (PERM) were among the most significant and pronounced changes (**Error! Reference source not found.**.A). Those proteins along with tissue inhibitors of metalloproteinase 1 (TIMP1), calprotectin (S10A8/A9) and neutrophil defensin 1 (DEF1) were mostly correlated with neutrophils (**Error! Reference source not found.**.B). In contrast, cathepsins (CATB, CATD) were highly correlated with CD14+ and macrophage markers (**Error! Reference source not found.**.B).

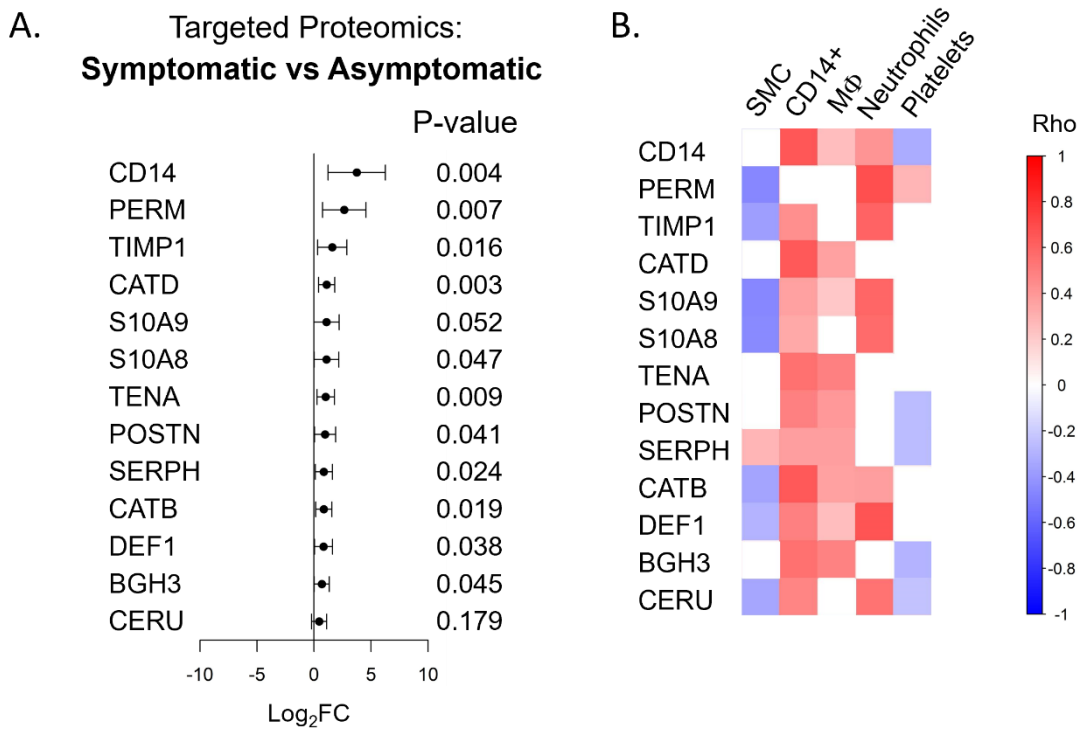


Figure 5.23. Validated inflammation signature. **A.** Forest plot depicting log2 fold changes and p-values of proteins in the core of symptomatic (n = 34) vs asymptomatic (n = 64) plaques using targeted proteomics. **B.** Spearman correlations of these proteins against the first principal component of cell clusters from the network analysis of intracellular proteins (**Error! Reference source not found.**). S10A9 and CERU were included in this list since they were significantly (p-value < 0.05) higher in non-calcified plaques, as determined by targeted proteomics (Supplemental Table 1).

5.3.8 Sex differences in plaques and their association with calcification

One areas of current interest in atherosclerosis research is to identify sex differences. Several studies (18, 133, 275) have reported sex-specific changes in atherosclerosis. Given that sex also emerged as one of the significant covariates in our analysis (**Error! Reference source not found.**.C), we decided to compare plaques from male and female patients. The clinical characteristics of the cohort based on this comparison are shown in **Error! Reference source not found.** below.

	Overall (n=120)	**Male (n=88)**	**Female (n=32)**	**P-value‡**
Demographics				
Age	70 (64-74)	70 (63-73)	70.5 (66-74)	0.44

Sex (male)	88 (73.3)	88 (100.0)	0 (0.0)	-
Characteristics of carotid artery stenosis				
Grade of stenosis	90 (85-90)	90 (85-90)	90 (87.5-95)	0.19
Stenosis grade ≥90%	67 (55.8)	44 (50.0)	23 (71.9)	**0.001**
Contralateral stenosis	34 (28.3)	27 (30.7)	7 (21.9)	0.49
Peak systolic flow velocity, m/s	3.8 (3 – 4.5)	3.8 (3- 4.5)	3.7 (2.8 – 4.5)	0.93
Plaque Morphology				
Echogenic	58 (48.3)	38 (43.2)	20 (62.5)	
Mixed	28 (23.3)	25 (28.4)	3 (9.4)	0.065
Echolucent	34 (28.3)	25 (28.4)	9 (28.1)	
Histological AHA classification				
Type V fibroatheroma	32 (26.7)	21 (23.9)	11 (34.4)	
Type VI complex lesion	52 (43.3)	46 (52.3)	6 (18.8)	**0.002**
Type VII calcified lesion	20 (16.7)	9 (10.2)	11 (34.4)	
Type VIII fibrotic lesion	16 (13.3)	12 (13.6)	4 (12.5)	
Calcified (clinical†)	70 (58.3)	47 (53.4)	23 (71.9)	0.053
Comorbidities and risk factors				
Transient ischemic attack	23 (19.2)	19 (21.6)	4 (12.5)	0.43
Stroke	24 (20.0)	19 (21.6)	5 (15.6)	0.46
History of stroke/TIA	23 (19.2)	20 (22.7)	3 (9.4)	0.12
Acute myocardial infarction	25 (20.8)	22 (25.0)	3 (9.4)	0.08
Coronary artery disease	42 (35.0)	37 (42.0)	5 (15.6)	**0.009**
Peripheral artery disease	49 (40.8)	39 (44.3)	10 (31.2)	0.21
Arterial hypertension	108 (90.0)	79 (89.8)	29 (90.6)	1.00
Diabetes mellitus type 2	36 (30.0)	30 (34.1)	6 (18.7)	0.12
Adipositas (BMI>30)	28 (23.3)	16 (18.2)	12 (37.5)	0.05
Smoking active	29 (24.2)	21 (23.9)	8 (25)	1.00
Past smoker	44 (36.7)	37 (42.0)	7 (21.9)	0.05
Pack-years	20 (0-45)	25 (0-50)	0 (0-30)	**0.02**
COPD	28 (23.3)	21 (23.9)	7 (21.9)	1.00
Medications				
Ace inhibitors	48 (40.0)	36 (40.91)	12 (37.5)	0.74

Angiotensin receptor blockers	41 (34.17)	31 (35.23)	11 (34.38)	0.93
Beta-blockers	79 (65.83)	58 (65.91)	22 (68.75)	0.77
Diuretics	34 (28.33)	22 (25.00)	12 (37.50)	0.18
Statins	97 (80.83)	74 (84.09)	24 (75.00)	0.26
Marcoumar	6 (5.00)	5 (5.68)	1 (3.13)	0.57
Laboratory parameters				
LDL, mg/dL	83 (67-108)	84 (68-109)	80 (56-95)	0.49
HDL, mg/dL	49 (41-56)	46 (39-52)	57 (48-73)	**<0 .001**
Total cholesterol, mg/dL	164 (143-193)	163 (139-195)	170 (149-190)	0.43
Triglycerides, mg/dL	128 (98-197)	141 (96-215)	118.5 (102-162)	0.16
High-sensitivity CRP, mg/dL	0.3 (0.1-0.6)	0.3 (0.1-0.6)	0.4 (0.1-0.7)	0.34

Table 5.4. Clinical characteristics of the discovery cohort in the sex comparison. Continuous data are shown as median (interquartile range). Dichotomous data are shown as n (%). Mann-Whitney test was used for the statistical comparison of continuous variables between plaques from males and females and Fisher's exact test for the categorical variables. The Chi-square test was used for categorical variables of more than two classes (ultrasound and histology). †Plaques were characterized as calcified or non-calcified based on the classification by two clinicians. This clinical classification was in 83% agreement with the calcification characterization of a subset of plaques by computed tomography angiography (n=35). AHA, American Heart Association; BMI, body mass index; COPD, chronic obstructive pulmonary disease; CRP, C-reactive protein; HDL, high-density lipoprotein; LDL, low-density lipoprotein; TIA, transient ischemic attack.

As shown in Table 5.4, most female patients had a high grade of carotid artery stenosis (≥90%) compared to males. There was a statistically significant association between histology classification of the carotid artery and gender, χ^2 (3, N=120) = 15.29, p=.002. Male patients had significantly lower levels of HDL and a higher percentage (42%) had coronary artery disease compared to females (15.6%). When comparing the sex differences to the proteins changing in calcified versus non-calcified plaques, we found a high correlation between the log2 fold changes (**Error! Reference source not found.**).

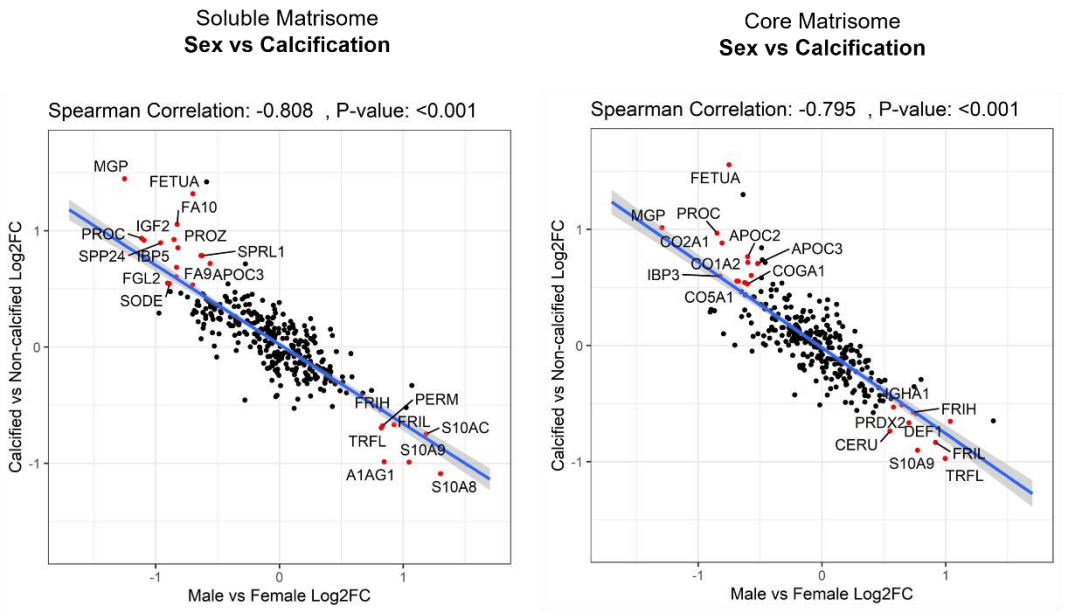


Figure 5.24 Association of calcification with sex differences in the core of the plaques. Scatter plots depicting log2 fold changes for the comparisons of male vs female patients and calcified vs non-calcified plaques in the soluble and core matrisome. Linear regression models were fitted, and Spearman correlation analyses were conducted for both extracts. Significant changes in both comparisons are highlighted in red. Common protein changes in both extracts are labelled. Linear regression plots are depicted with their 95% confidence intervals. Spearman's Rho coefficient and corresponding p-values are shown.

We then wanted to see which changes were sex-specific and not attributed to calcification. We first created a protein correlation network using the core samples of the plaque and all matrisome proteins from both extracts, core and soluble matrisome. The network was then clustered using hierarchical clustering (**Error! Reference source not found.**). **Error! Reference source not found.** presents the protein changes in all clusters of the female versus male comparison (outer circle), along with the calcified versus non-calcified (middle circle) and the symptomatic versus asymptomatic comparisons (the inner circle). Clustering analysis revealed four clusters in the soluble matrisome (C4, C6, C17, C19) and four clusters in the core matrisome (C16, C17, C19, C20) that were significantly altered with sex. However, most of these protein changes were linked to calcification (eg cluster 4 -C4- which contained gamma-carboxylated proteins such as MGP, GAS6, PROC, FA10) or to inflammation (eg cluster 19 -C19- which contained neutrophil-derived proteins such as PERM, S10A8/A9, TIMP1, DEF1, MMP9). Only protein changes constituting clusters 17 and 20 (C17, C20) were independent of calcification and inflammation and can thus

be considered sex-specific. C17 and C20 included the two large-aggregating proteoglycans aggrecan and versican as well as related matrisome proteins, such as link proteins (HPLN1, HPLN3) that bind to hyaluronic acid.

Figure 5.25. Matrisome network of the core of the plaques. Hierarchical clustering on the matrisome protein correlation network of the core of the plaques. Aracne-AP method was used to reconstruct the network. The SIREN algorithm was used to filter negative associations. Core and soluble matrisome (GuHCl and NaCl extracts

respectively) were combined for the network reconstruction. For the matrisome proteins which were quantified in both extracts the one with the highest number of matched spectra was used for this analysis.

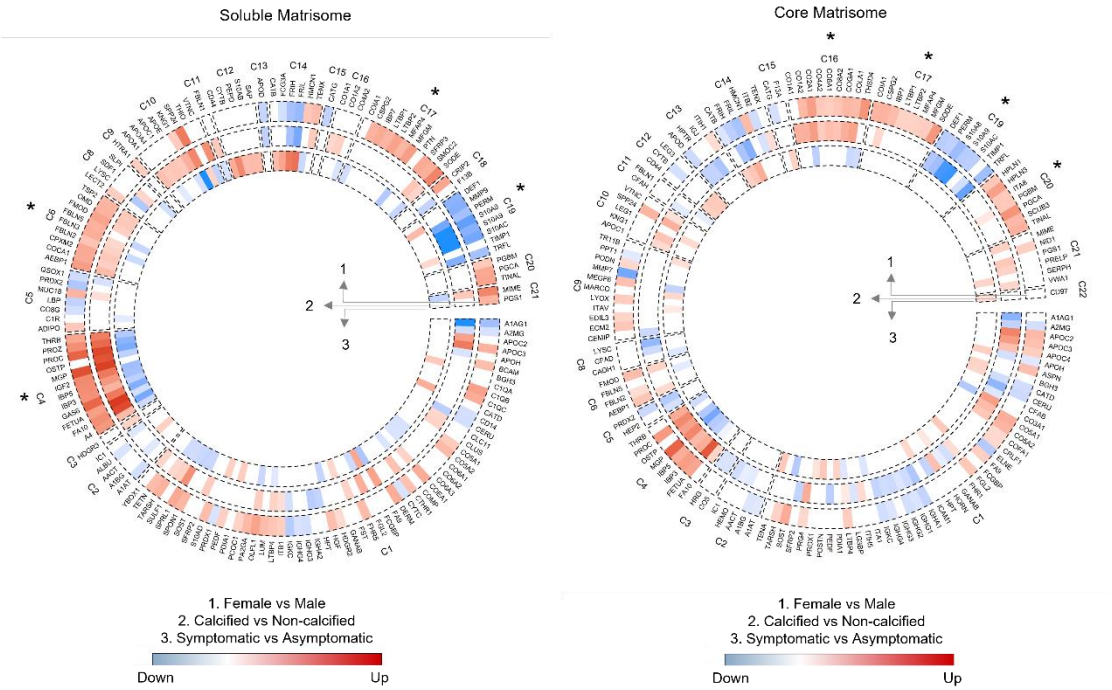


Figure 5.26 Sex-related differences in the core of the plaques and their association with calcification and inflammation. Circular heatmap depicting the differential analysis results for proteins identified in the soluble (NaCl extract) and the core (GuHCl) matrisome of plaque core. Three comparisons are depicted: female (n = 29) versus male (n = 76) – outer circle, calcified (n = 60) versus non-calcified (n = 46) plaques – middle circle, and symptomatic (n = 69) versus asymptomatic (n= 36) plaques – inner circle. Significant proteins in at least one comparison are organized in the circular heatmap according to the reconstructed network and the hierarchical clustering to identify clusters. Clusters significantly enriched for dysregulated proteins in the sex comparison are marked with an asterisk (Fisher's Exact test, p-value <0.05).

We also performed enrichment analysis for significant clusters in at least one of the examined phenotypes to find possible pathways that the proteins belong to (**Error! Reference source not found.**). Among the top significantly enriched pathways or GO functional terms were growth-factor binding for C17 and cell adhesion and hyaluronic acid binding for C20.

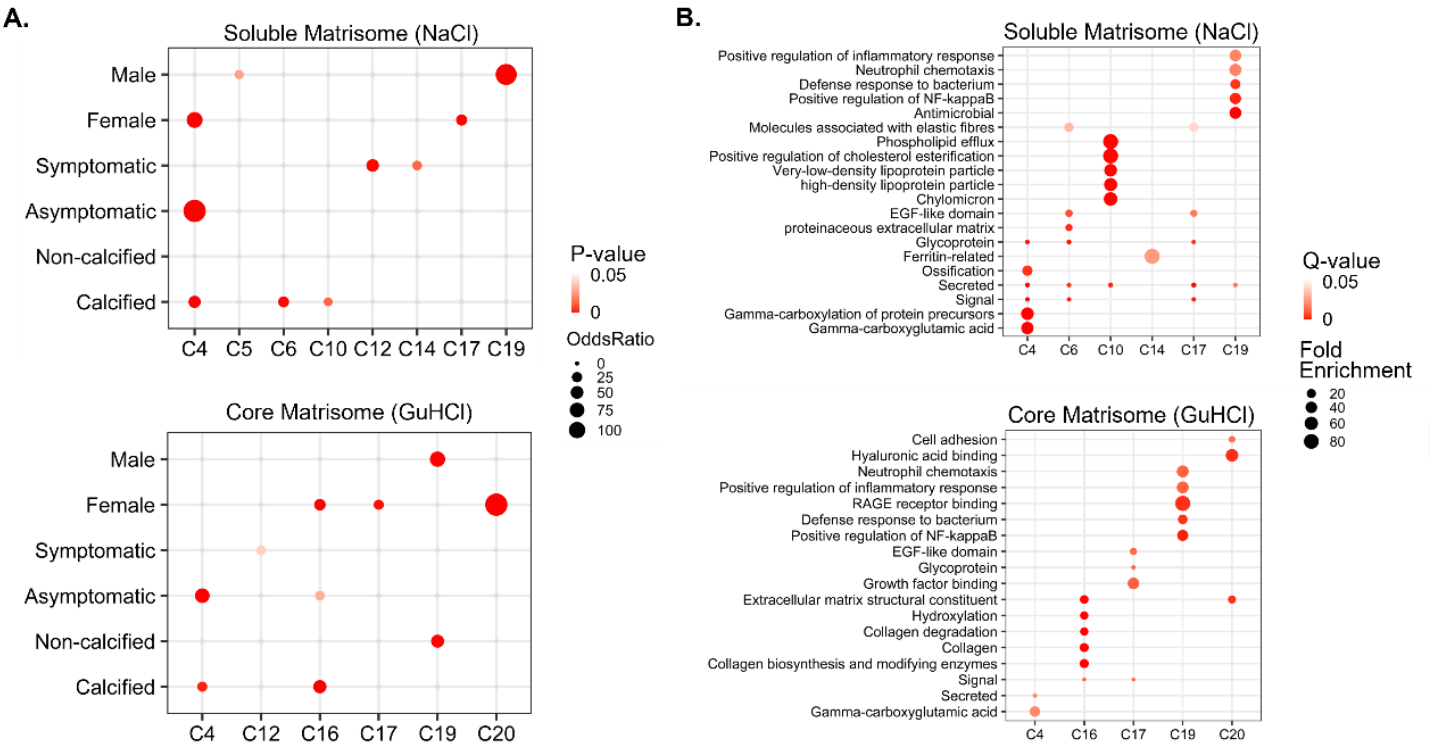


Figure 5.27 Enrichment analysis of significant clusters. A. Results of enrichment analysis for the matrisome clusters which are significantly enriched in one of the examined phenotypes in the soluble and the core matrisome (NaCl and GuHCl extracts, respectively). **B.** Top 5 significantly enriched pathways or GO functional terms for each significant cluster, using a Benjamini Hochberg corrected q-value threshold of 0.05 in the soluble and the core matrisome (NaCl and GuHCl extracts, respectively).

To externally validate the sex-specific protein changes in clusters 17 and 20, as well as the sex changes, attributes to inflammation or calcification (eg clusters 19 or 4), we used label-free proteomics measurements of 200 carotid plaques from the Athero-express cohort, the clinical characteristics of which are shown in **Error! Reference source not found.** below.

	Overall (n=200)	**Male (n=149)**	**Female (n=51)**	**P-value**
Demographics				
Age	70 (62-76)	71 (62-76)	70 (62-75)	0.322
Sex (male)	149 (74.5)	149 (100.0)	0 (0.0)	-
Composite Endpoint and Plaque Vulnerability				
Primary Endpoint in three years	51 (25.5)	44 (29.5)	7 (13.7)	**0.025**

Plaque Vulnerability Index (0-5)	2 (2-3)	3 (2-3)	2 (1-3)	**0.003**
Comorbidities and risk factors				
Stroke	30 (15.0)	25 (16.8)	5 (9.8)	0.229
Coronary artery disease	68 (34.0)	53 (35.6)	15 (29.4)	0.123
Arterial hypertension	139 (69.5)	101 (67.8)	38 (74.5)	0.368
Diabetes mellitus type 2	46 (23.0)	38 (25.5)	8 (15.7)	0.151
Adipositas (BMI>30)	37 (18.5)	26 (17.4)	11 (21.6)	0.428
Smoking active	71 (35.5)	50 (33.6)	21 (41.2)	0.326
Past smoker	94 (47.0)	76 (51)	18 (35.3)	0.052
Medications				
Statins	157 (78.5)	117 (78.5)	40 (78.4)	1.000
Laboratory parameters				
LDL, mg/dL	105 (74-128)	99 (75-129)	102 (65-125)	0.582
HDL, mg/dL	43 (32-52)	40 (32-50)	48 (37-60)	**0.018**
Total cholesterol, mg/dL	174 (142-203)	171 (144-200)	185 (139-216)	0.453
Triglycerides, mg/dL	129 (87-150)	112 (87-150)	124 (88-165)	0.576

Table 5.5 Clinical characteristics of the validation cohort (Athero-express) in the sex comparison. Continuous and ordinal variables are shown as median (interquartile range). Dichotomous variables are shown as n (%). Mann-Whitney test was used for the statistical comparison of continuous and ordinal variables between plaques from male and female patients and Fisher's exact test for the categorical variables. The Chi-square test was used for categorical variables of more than two classes (ultrasound and histology). †Plaques were characterized as calcified or non-calcified based on the classification by two clinicians. This clinical classification was in 83% agreement with the calcification characterization of a subset of plaques by computed tomography angiography (n=35). AHA, American Heart Association; BMI, body mass index; COPD, chronic obstructive pulmonary disease; CRP, C-reactive protein; HDL, high-density lipoprotein; LDL, low-density lipoprotein; TIA, transient ischemic attack.

This analysis confirmed that 11 out of 15 matrisome proteins changing in our sex comparison, including the two large aggregating proteoglycans versican (CSPG2) and aggrecan (PGCA), were also validated in the Athero-express cohort (Figure 5.28.A). Most sex-related proteins were not found significantly changing in the sex comparison of the carotid plaque dataset of chapter 3 (section 3.3.2). Among the significant sex-related proteins, Latent-transforming growth factor beta-binding protein 1 (LTBP1) and Basement membrane-specific heparan sulfate proteoglycan core protein (PGBM)

were also found significantly upregulated in females in the carotid plaques dataset of chapter 3 (section 3.3.2). Lactadherin (MFGM) though was found significantly upregulated in males in the sex comparison of the plaque dataset of chapter 3, showing different directionality of our findings (Figure 5.28.A). This protein was also not validated in the Athero-express cohort. These changes might be explained from the highest sample size of the cohort examined in this chapter, as well as from the fact that these new comparisons are made using the core region of the plaques only. Most proteins of the inflammation signature were also changing between plaques from female and male patients in both cohorts, validating the inverse correlation of sex with inflammation (Figure 5.28.B).

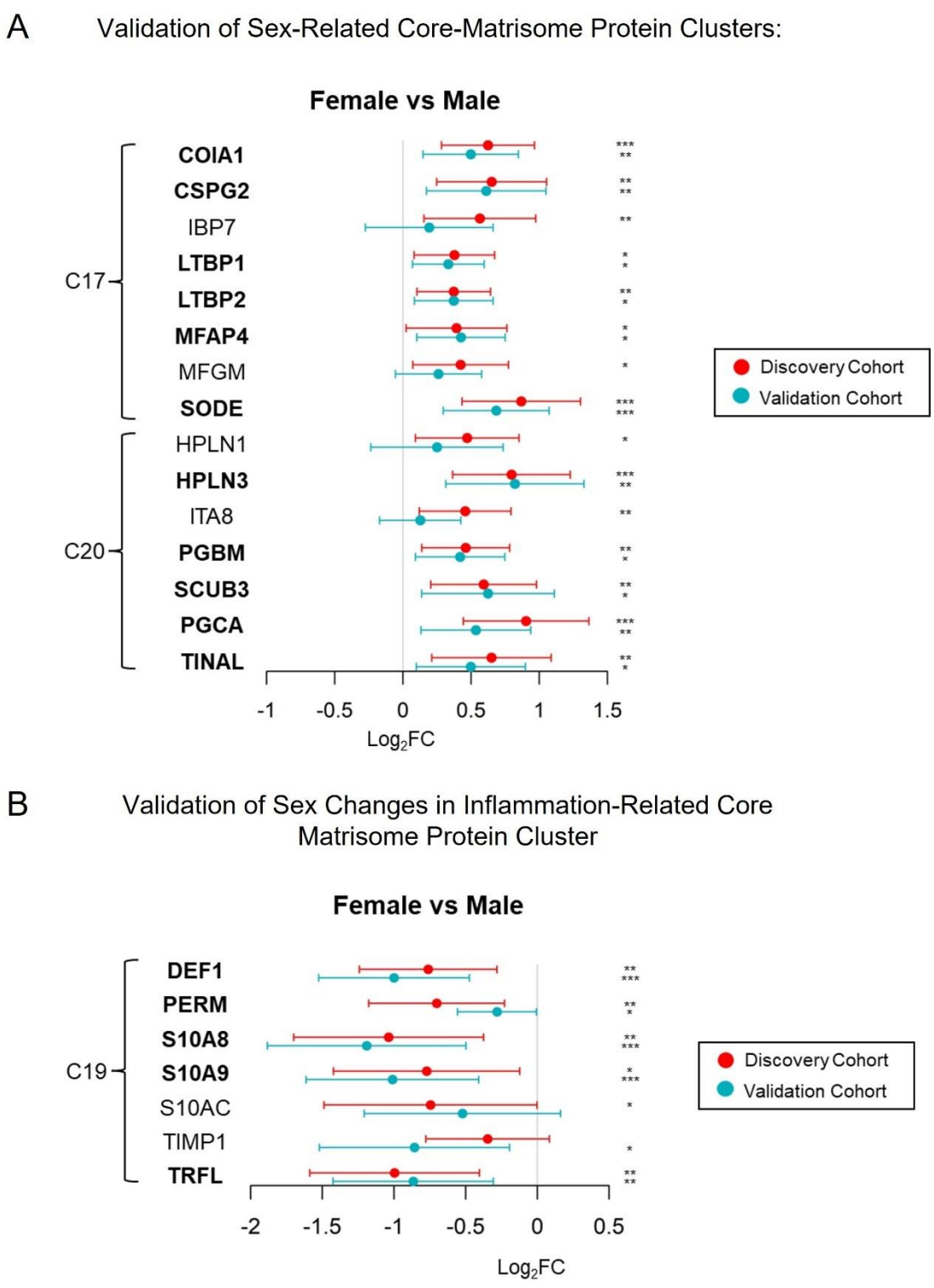


Figure 5.28 Validation of sex-associated changes using an independent proteomics cohort. Forest plot depicting log2 fold changes, 95% confidence intervals and p-values of proteins in the core of plaques between female and male participants in the core matrisome (GuHCl extract) from TMT MS measurements of the discovery cohort (32 females vs 88 males) and label-free measurements in the validation cohort (49 females vs 151 males). Presented proteins include **A.** the sex-associated matrisome clusters C17 and C20 (Figure 5.26) and **B.** the inflammation signature from cluster C19.

Proteins include those from the sex-specific matrisome clusters (**Error! Reference source not found.**) and those validated with targeted MS for inflammation and calcification biosignatures (**Error! Reference source not found.**, **Error! Reference source not found.**). Proteins which are significantly changing between female and male participants in both cohorts appear in bold. Differential protein analysis for both cohorts was conducted using the Ebayes method of the limma package correcting for age and statins. * denotes 0.01 ≤ p-value < 0.05, ** 0.001 ≤ p-value < 0.01, and *** p-value < 0.001.

5.3.9 Analysis of Sex-Specific Networks in Atherosclerosis

We wanted to further explore the sex-specific changes in the plaques, using network analysis. For this reason, we decided to use our network pipeline (described in chapter 4) to further promote our understanding of changes and mechanisms that differentiate between male and female patients in atherosclerotic plaques, using a "system-level" approach which relates proteins to each other and defines hubs and co-expressed proteins that could be functionally related or coordinately regulated

We used the cellular proteome of the plaques (SDS extract), reconstructed sex-specific networks from the core of the plaques and applied clustering. For visualization purposes, we filtered the cellular proteome for cell receptors and associated proteins and included only the significant clusters and interactions between the proteins forming those. The reconstructed networks for males and females are depicted in **Error! Reference source not found.**. Soft clustering revealed eight significant clusters representing smooth muscle cells, leukocytes/macrophages, platelets, G-proteins, adhesion and apoptosis-related proteins, CD14 and cathepsins, neutrophils and proteins involved in the blood coagulation. From the reconstructed networks, the most dysregulated clusters between males and females are those of smooth muscle cells and neutrophils, with both being less connected to other clusters in males than in females.

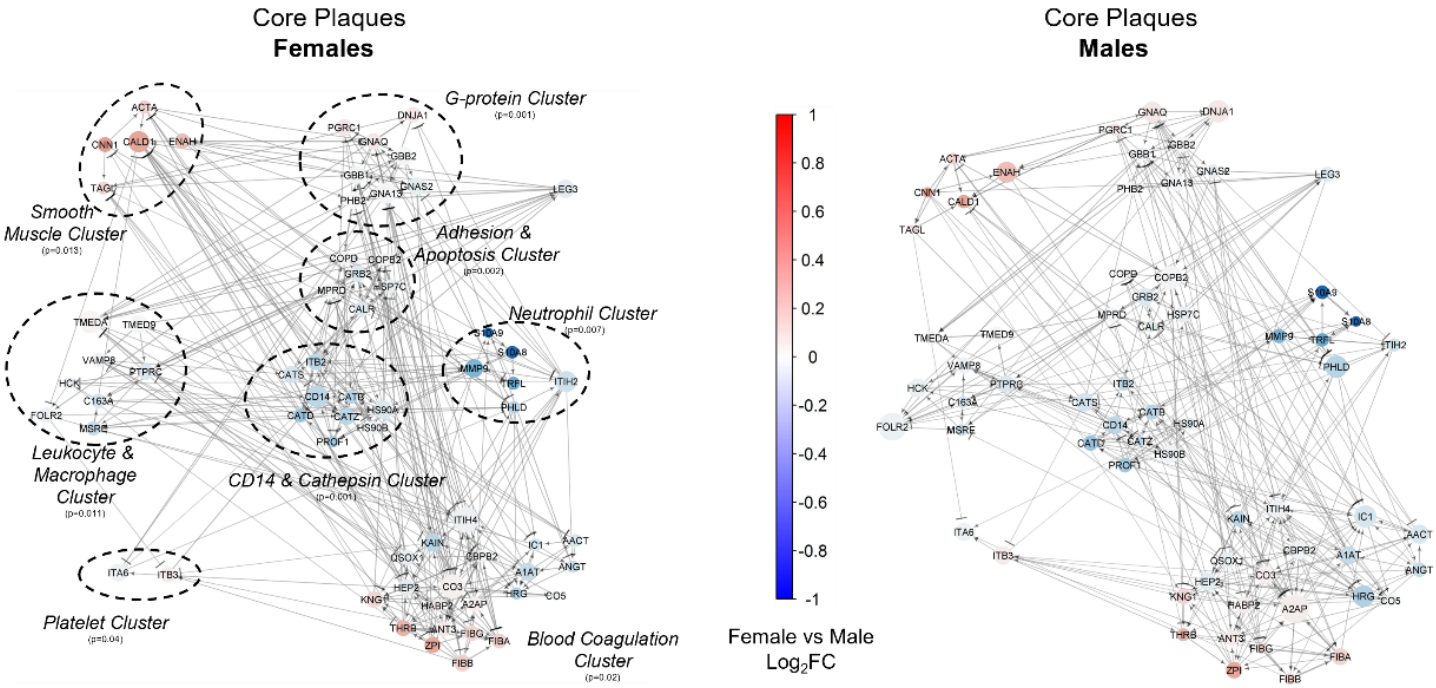


Figure 5.29 Sex-specific co-expression networks of significant clusters of cell receptor-associated proteins in the core of the plaques. The depicted network is a view of the overall network including the significant clusters and interactions among the proteins composing them. Proteins in red represent upregulated proteins in females whereas proteins in blue represent upregulated proteins in males on the SDS extract. The size of nodes is proportional to the betweenness centrality. Edge width is proportional to the mutual information of the connected proteins. The directionality of the edges is depicted by arrows for activation and lines for inhibition. Clustering was conducted with ClusterOne algorithm(79) (p-value<0.05, edge weight=MI).

To identify changes in "superclusters" (communities), we performed community-based clustering using all receptors and receptor-binding proteins (and not only the ones belonging to a significant cluster from **Error! Reference source not found.**) and identified 3 superclusters in the networks (**Error! Reference source not found.**). This analysis confirmed the differentiation of SMC and blood coagulation-related proteins between sexes, as SMC-related proteins clustered together with blood coagulation-related ones in females but not in males, where there also are fewer edges between them, suggesting a reduced interplay between those proteins in male plaques.

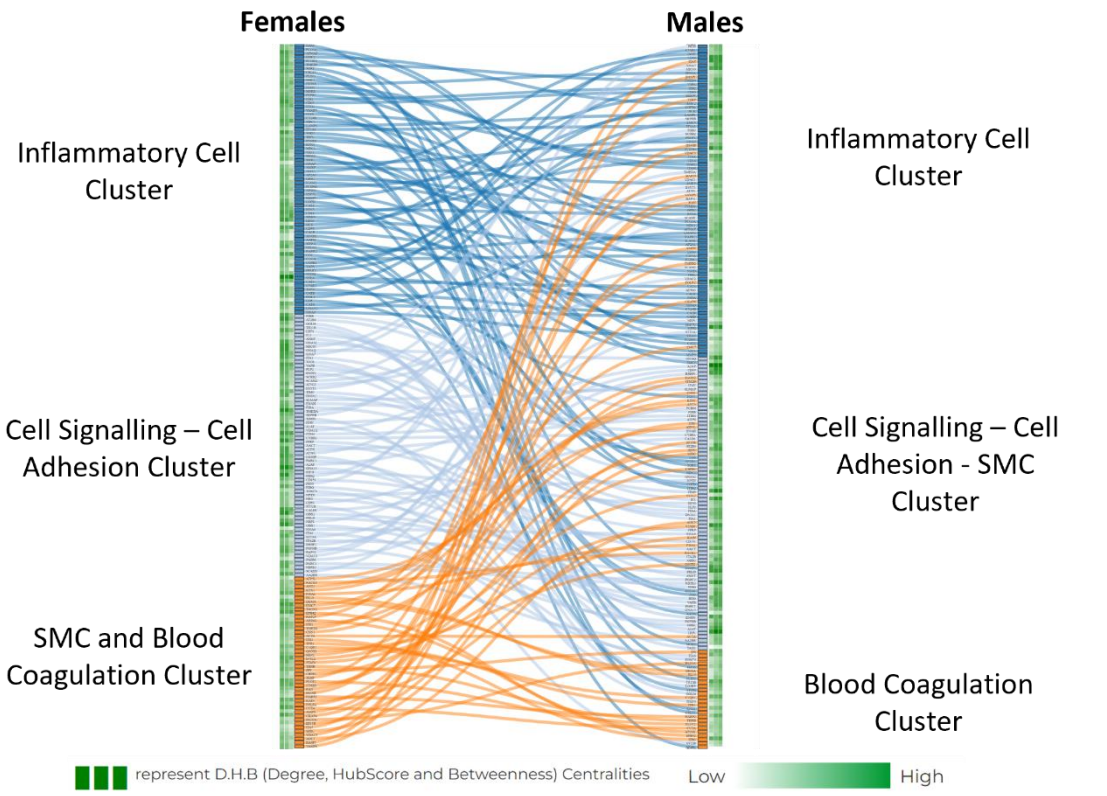


Figure 5.30 Community-based sex-specific network clustering. Sankey's plot depicts the community structure for female and male networks. NetConfer web tool was used to infer network communities (each one represented by a different color) and create the plot. Nodes with the same names across the networks are connected. In both axes communities along with their member nodes are ordered in descending size order. Heatmaps, besides the nodes, represent the three centrality measures rank normalized across the given pair of networks.

Finally, we wanted to identify and compare critical nodes concerning each network. We calculated the betweenness centralities of proteins in each network, using the Shapiro-Wilks test for normality and the 95% confidence interval. We found four proteins with significantly changing betweenness centralities across male and female networks (Table 5.6), among which were MMP9 and galectin-3.

Uniprot ID	Uniprot Accession	Gene Symbol	Protein Name	BC in Female	BC in Male
HS90B_HUMAN	P08238	HSP90AB1	Heat shock protein HSP 90-beta	51.427	163.601

LEG3_HUMAN	P17931	LGALS3	Galectin-3	135.195	22.638
MMP9_HUMAN	P14780	MMP9	Matrix metalloproteinase-9	163.523	51.75
PTPRC_HUMAN	P08575	PTPRC	Receptor-type tyrosine-protein phosphatase C	91.943	235.678

Table 5.6 Description of proteins with significantly changing betweenness centralities across sex-specific networks. P-value<0.01, BC: Betweenness Centrality

5.3.10 Comparison to single-cell RNA sequencing (scRNAseq)

The PlaqView tool (157) was used for a dataset of 38 carotid endarterectomies (247) (**Error! Reference source not found.**) to verify the expression of cell markers used and to relate the inflammation and calcification signatures to cellular markers in the transcript level, analyzed by scRNAseq. This analysis confirmed the expression of the cell markers used (**Error! Reference source not found.**.C), aortic smooth muscle cell actin, transgelin, and caldesmon, in plaque SMCs (cluster 8) while calponin 1 was only expressed in a small subset. Monocyte differentiation antigen CD14 and scavenger receptor cysteine-rich type 1 protein M130 were mostly expressed in macrophages (cluster 3), as expected, whereas galectin-3 was also expressed in other plaque cells, such as SMCs (cluster 8) and endothelial cells (cluster 5). Finally, receptor-type tyrosine-protein phosphatase C (PTPRC / CD45) was expressed in almost all cells of hematopoietic origin.

The majority of proteins of the inflammatory signature (**Error! Reference source not found.**.B) were highly expressed in hematopoietic cells, such as calprotectin (S10A8/A9) that was detected in the monocyte and neutrophil cluster (cluster 4), although cathepsins (CATB, CATD) and metalloproteinase inhibitor 1 (TIMP1) were

found highly expressed in the majority of plaque cells. Other proteins, such as periostin (POSTN) and serpin H1 (SERPH), were also expressed in endothelial cells (cluster 5).

Most proteins of the calcification signature (**Error! Reference source not found.**.D) were related to smooth muscle cells. Exceptions were matrix Gla protein, which was expressed in nearly all cell types, and osteopontin, which was most commonly expressed in macrophages but with a low level of expression. However, OSTP was detected by spatial RNAseq within plaque regions expressing CD14, CD44, and cathepsins D and B (Figure 5.32). Finally, certain gamma-carboxylated proteins, such as coagulation factors 9 and 10 (FA9, FA10), were retained in plaques and hence were better detected at the protein rather than transcript level (Figure 5.32).

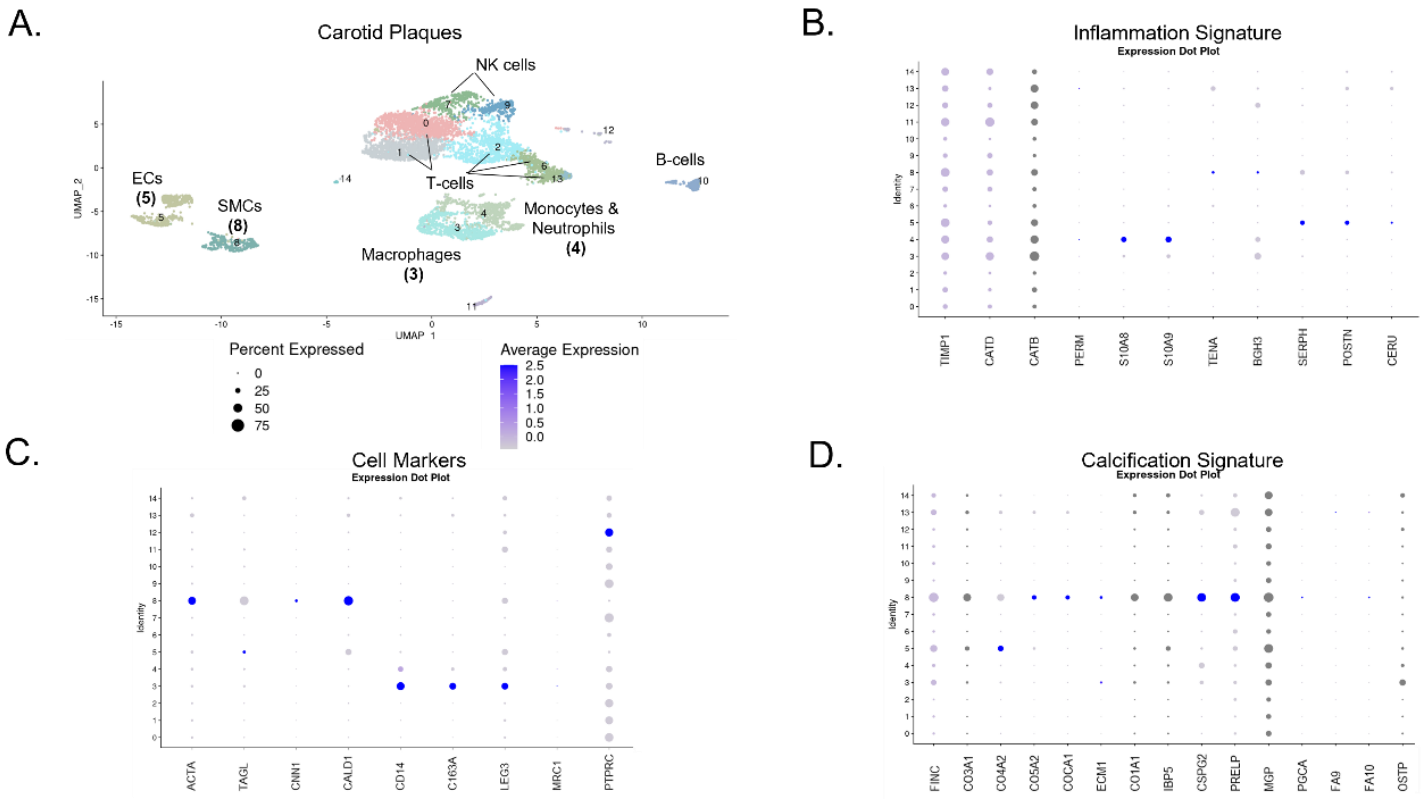


Figure 5.31 Comparison of proteomic signatures to scRNAseq data. A. UMAP of scRNAseq data from carotid endarterectomies (Slenders et al. (247) dataset (n = 38), using PlaqView (157) and the Aran et al. (159) reference-based algorithms to deduce cell identity of individual cells). Clustering was applied to cell types in the following groups: 0: T-cells type 1, 1: T-cells type 2, 2: T-cells type 3, 3: macrophages, 4: monocytes and neutrophils, 5: endothelial cells, 6: T-cells type 4, 7: natural killer cells type 1, 8: smooth muscle cells, 9: natural killer cells type 2, 10: B-cells, 11: common

myeloid progenitor cells, 12: granulocyte/monocyte progenitor cells, 13: T-cells type 5, 14: astrocytes. **B.** Expression dot-plot for the inflammation protein signature. **C.** Expression dot-plot for cellular markers. **D.** Expression dot-plot for the calcification protein signature. Features were clustered using hierarchical clustering based on their average expression values per cell cluster.

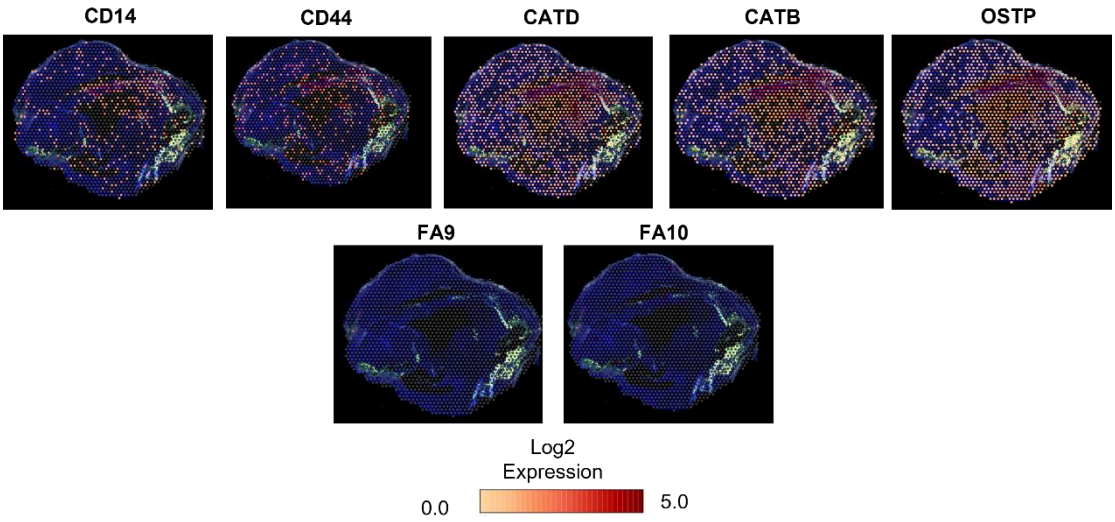


Figure 5.32 Spatial RNAseq. Feature heatmaps of Log2 expression levels of selected genes from cell markers and revealed matrisome biosignatures using Loupe Browser and the regional scRNA-sequencing data. The log2 normalized expression levels of each cell were shown using a yellow to red color scale.

5.3.12 Comparison to imaging classification by ultrasound

Then, the molecular changes that could be captured from the current plaque classification by ultrasound were explored (**Error! Reference source not found.**). Echogenic plaques showed several protein changes, which we validated using targeted proteomics (**Error! Reference source not found.**. A). Echogenic plaques were rich in calcification-related proteins (PROC, OSTP), fibrillar collagens (CO1A1, CO1A2), dermatopontin (DERM) and stromal cell-derived factor 1 (SDF1). The initial comparison of echogenic versus echolucent plaques returned similar protein changes to the ones of the comparison of calcified lesions versus fibroatheroma based on histology. Thus, we wanted to see whether the validated proteins would change in the calcified versus fibroatheroma comparison based on histology data (**Error! Reference source not found.**.B). PROC, OSTP, and SDF1 were also increased in a histology-based comparison of calcified lesions versus fibroatheroma. On the contrary, echolucent plaques had more cathepsin D and nidogen 2, a basal lamina protein. Thus, the

classification by ultrasound predominantly reflects the fibrillar collagen content and the calcification signature of plaques but fails to capture most of the inflammatory signatures in plaques.

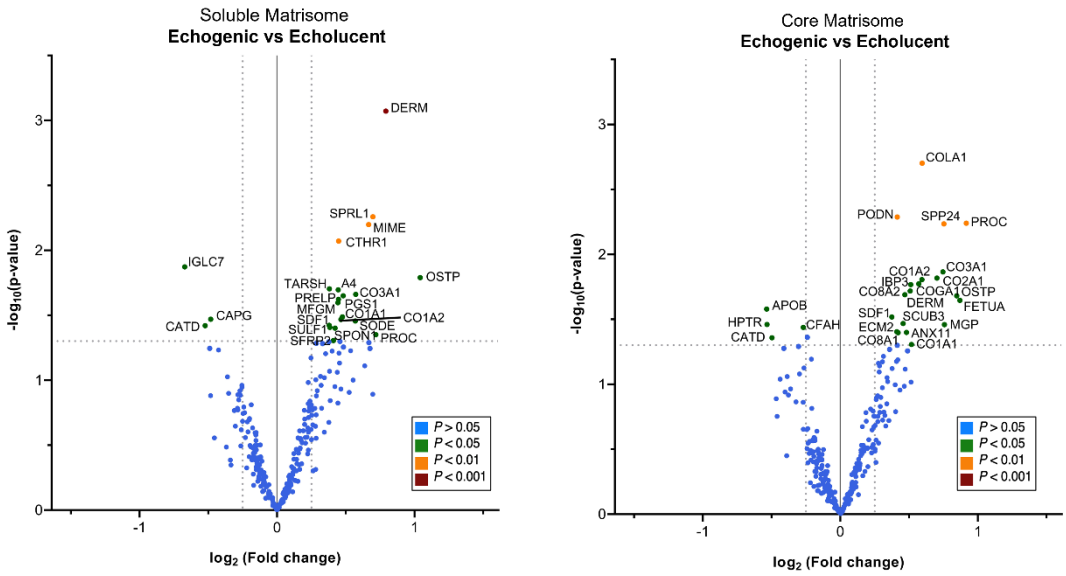


Figure 5.33 Proteomic changes based on ultrasound classification. Volcano plots depicting the differential expression analysis results of echogenic vs echolucent core plaques in soluble (n = 49 echogenic versus n = 26 echolucent plaque samples) and core (n = 51 echogenic versus n = 24 echolucent plaque samples) matrisome.

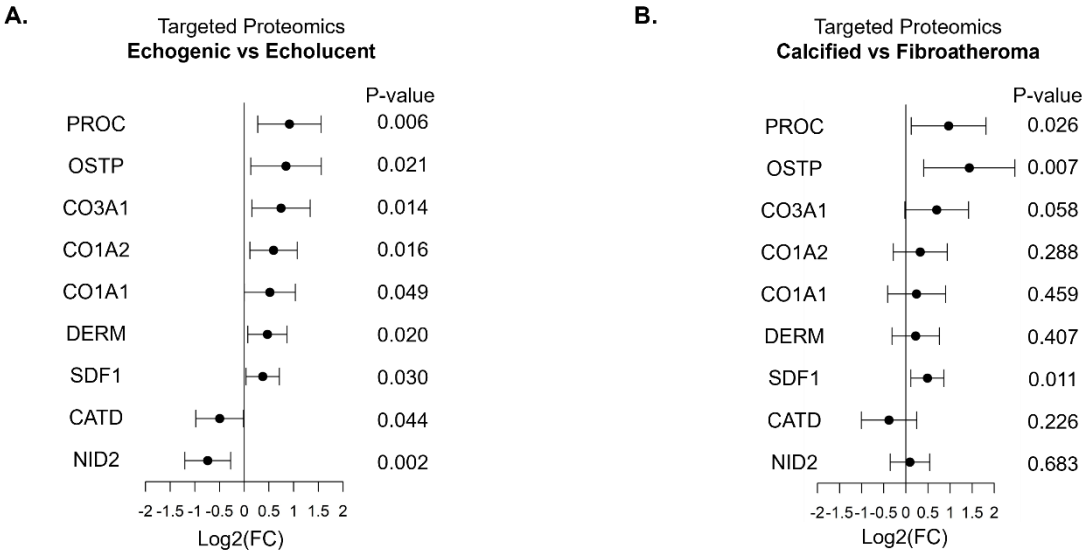


Figure 5.34 Validated proteomic changes based on ultrasound and histology. A. Forest plot depicting the log2 fold changes in the comparison of echogenic (n=43) vs echolucent (n=28) plaque cores that have been validated by targeted proteomics (PRM). **B.** Forest plot depicting the log2 fold changes of the targeted proteomic data

in the comparison of calcified (n=10) vs fibroatheroma (n=19) plaque cores based on the histological characterization of the validated ultrasound signature.

Error! Reference source not found. depicts the reconstructed matrisome network, consisting of the validated significant protein changes in one of the calcified versus non-calcified, symptomatic versus asymptomatic, or echogenic versus echolucent plaque comparisons. Known interactions, such as the one of calprotectin (S10A8 and S10A9), were confirmed. Within the inflammation signature, strong correlations exist among proteases (CATB, CATD) and protease inhibitors (TIMP1). Conversely, the calcification signature is enriched with calcium-binding coagulation factors and serum proteins, i.e., FETUA, experimentally verifying known protein interactions.

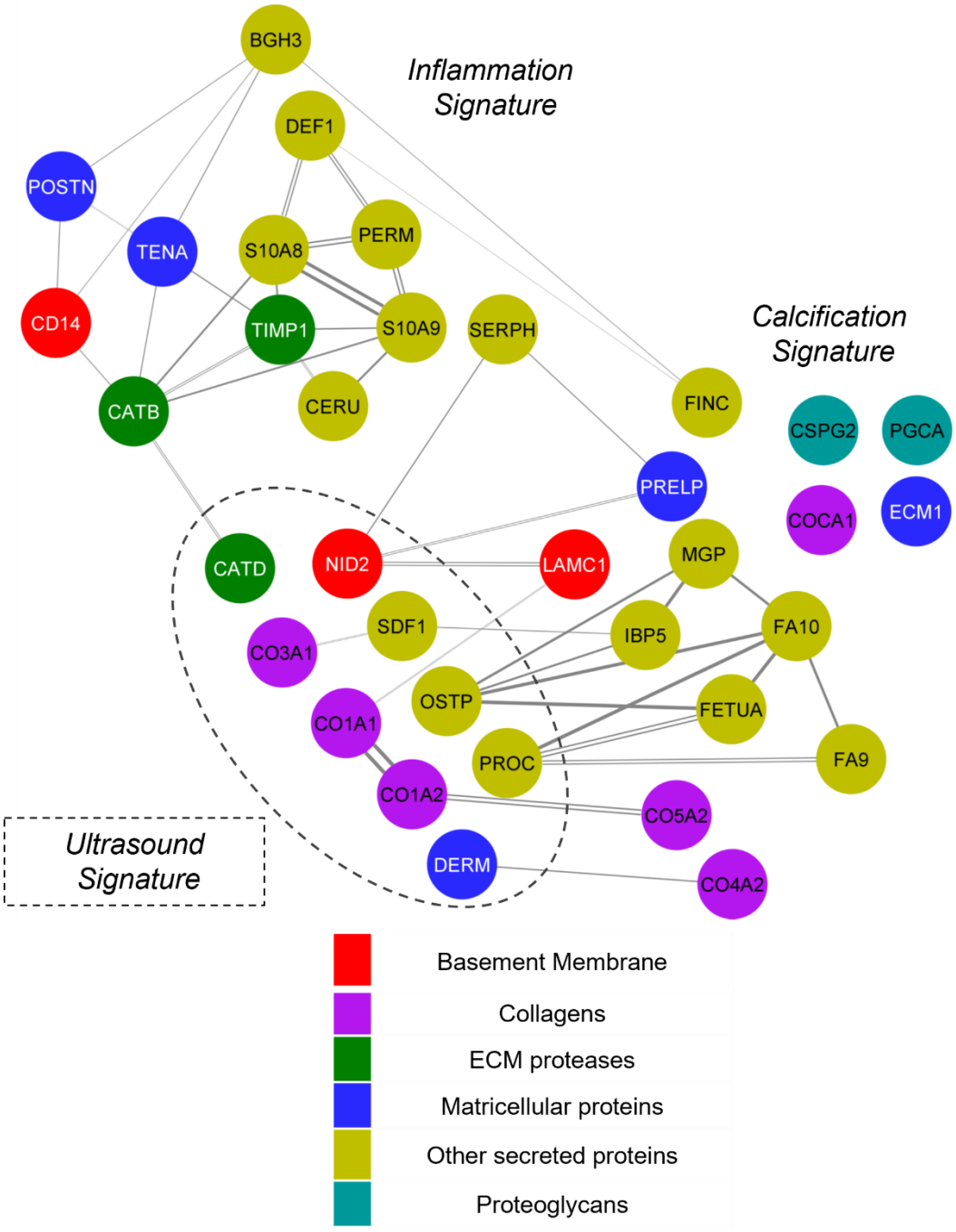


Figure 5.35 Reconstructed matrisome network for validated protein changes. Part of the reconstructed matrisome network included as nodes only extracellular proteins that were validated to significantly change in the symptomatic vs asymptomatic, calcified vs non-calcified, or echogenic vs echolucent comparisons. The ARACNe-AP method was used for the reconstruction. The correlation lines' thickness is proportional to the conditional mutual information between the two proteins. Nodes are colored based on the functional category of each protein. Double lines represent edges with an experimentally verified protein-protein interaction.

5.3.13 Protein changes in the periphery

To study the extracellular changes in the periphery of the plaques where fewer changes would be observed, as the periphery is considered to be the non-diseased area of the plaques, we conducted the same comparisons to the plaque periphery (calcified versus non-calcified and symptomatic versus asymptomatic plaques).

Matrix Gla protein and Alpha-2-HS-glycoprotein/Fetuin-A were among the few upregulated proteins in the periphery of calcified plaques (**Error! Reference source not found.**), and cathepsins B (CATB) and D (CATD) among those in the periphery of symptomatic plaques (**Error! Reference source not found.**). Macrophage metalloelastase (MMP-12) was significantly upregulated only in the periphery (**Error! Reference source not found.**) and not the core (**Error! Reference source not found.**) of symptomatic plaques. In the core plaque area, MMP9 was the most upregulated MMP at the protein level even when calcified plaques were excluded (**Error! Reference source not found.**).

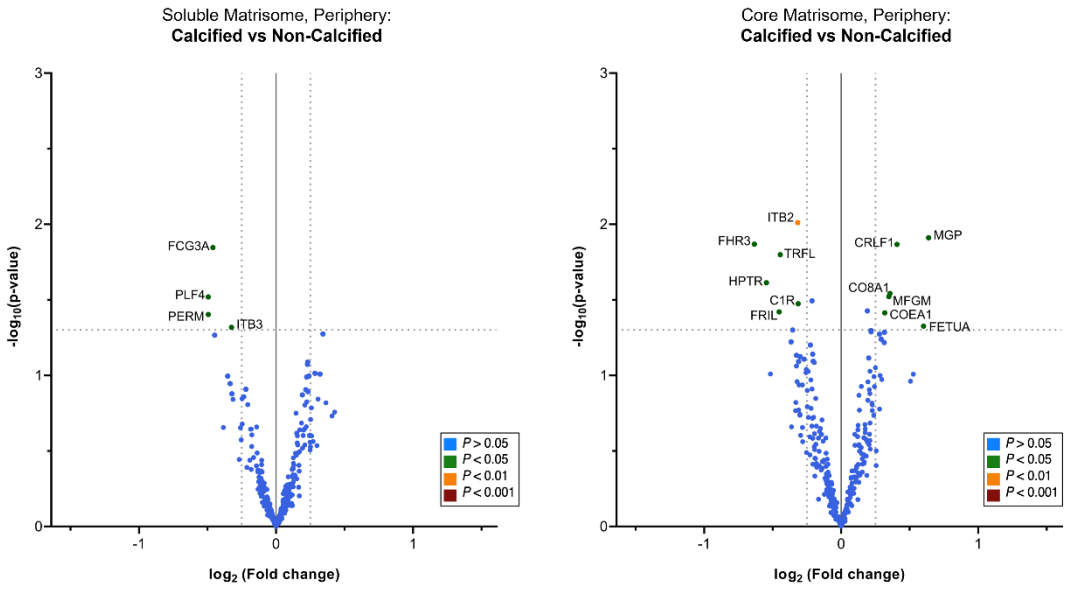


Figure 5.36 Extracellular protein changes in the periphery of calcified plaques. Volcano plots depicting results of the statistical comparison between calcified and non-calcified periphery plaques in the soluble (NaCl extract, n=63 calcified vs n=39 non=calcified) and core (GuHCl extract, n= 66 calcified versus n= 42 non-calcified) matrisome.

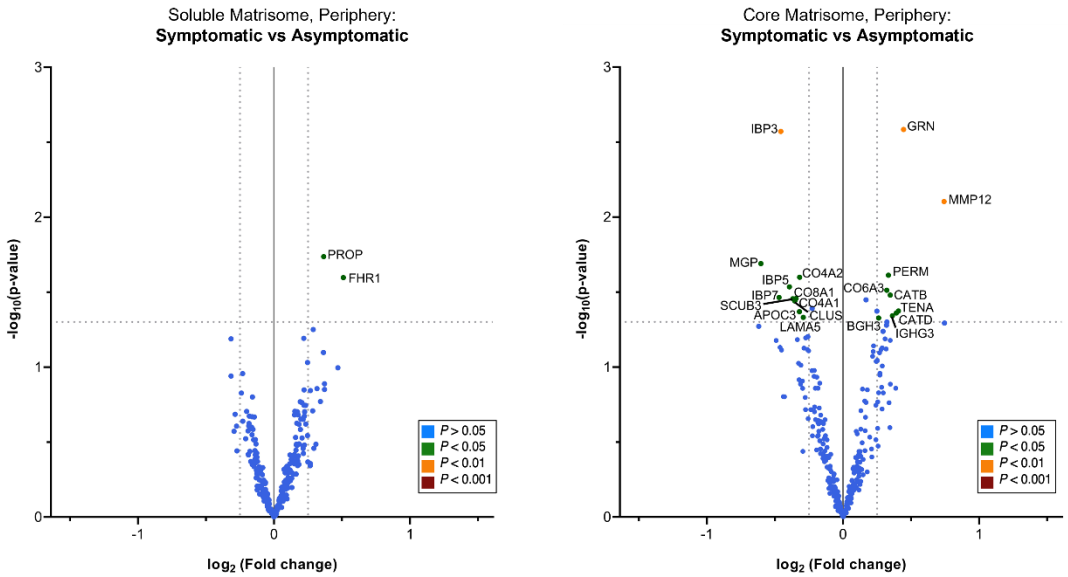


Figure 5.37 Extracellular protein changes in the periphery of symptomatic plaques. Volcano plots depicting results of the statistical comparison between symptomatic vs asymptomatic periphery plaques in the soluble (NaCl extract, n=37 versus n=65 asymptomatic) and core (GuHCl extract, n=40 symptomatic vs n=68 asymptomatic) matrisome.

5.3.14 Results of Olink Proximity Extension Assay Measurements

Olink platform measurements were used to further validate the calcification and inflammation signatures. Certain significant findings of MS proteomics were validated by the platform (Figure 5.38). More specifically, most of the inflammation signature was validated by the Olink platform (eg cathepsin D CATD, neutrophil defensin 1 DEF1, transforming growth factor-beta-induced protein igh3 BGH3). Regarding the calcification signature, only vitamin K-dependent protein C was found upregulated in calcified plaques by the Olink platform.

It is noteworthy that osteopontin was found significantly upregulated in non-calcified plaques by the Olink platform. This contrasts the findings by MS proteomics. Olink measurements are designed for blood samples. Many proteins measured in plaque tissue, such as osteopontin (Figure 5.39) or collagen alpha-1(I) chain (CO1A1), had values exceeding the upper limit of detection (~30%), resulting in a poor correlation between MS and Olink measurements. The reason for this is that samples were not diluted before running the measurements with Olink platform, suggesting that

following a better dilution protocol would substantially increase the number of proteins which would be properly quantified with Olink.

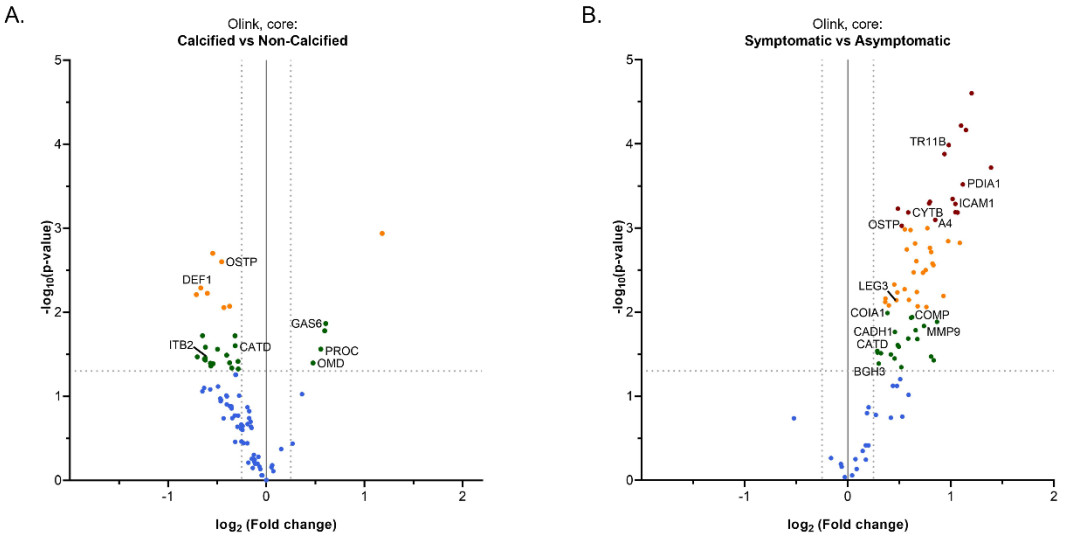


Figure 5.38. Extracellular protein changes in the core of the plaques, validated by the Olink platform. Volcano plots of significantly dysregulated proteins in the calcified (n = 62) vs non-calcified (n = 48) and symptomatic (n = 37) vs asymptomatic (n = 73) comparisons of the plaque cores, using the measurements from the Olink platform. Significant proteins in **A.** calcified vs non-calcified or **B.** symptomatic vs asymptomatic comparisons in TMT MS (in any of the two extracts) are labelled.

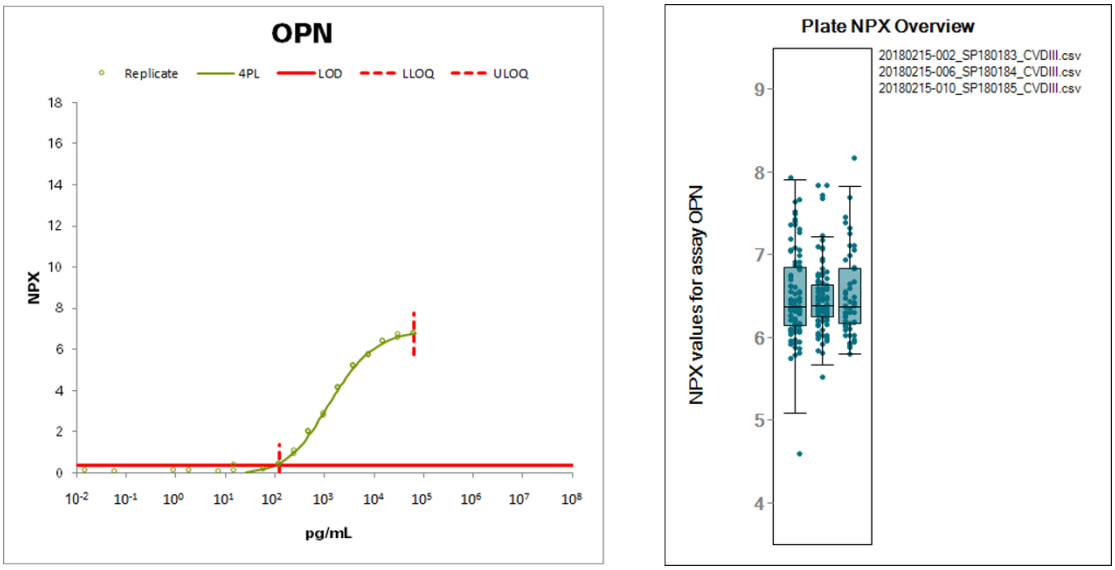


Figure 5.39 Normalised Protein eXpression for osteopontin from Olink platform. Example from relative quantification of osteopontin (OSTP/OPN) from Olink platform. The NPX values are between 6-8 NPX which is at the upper limit of quantification (ULOQ).

Nonetheless, there were significant protein changes in both the calcification (Figure 5.38.A) and symptomatic comparisons (Figure 5.38.B) that were not observed using TMT MS. Thus, we decided to explore the changes in matrisome proteins that were identified only by the Olink platform. In Figure 5.40 we depict the significant protein changes for proteins identified only by Olink in the two basic comparisons. There were in total 15 matrisome proteins identified only by the Olink platform in addition to the ones identified in both MS and Olink. Among these 15 proteins, 6 were significantly changing in the calcified vs non-calcified comparison (Insulin-like growth factor-binding proteins 1 and 2, C-C motif chemokine 14, Coagulation factor VII, Plasminogen activator inhibitor 1 and Serpin B6) (Figure 5.40.A) and 10 in the symptomatic vs asymptomatic comparison (Serpin B6, Plexin-B2, CD166 antigen, Plasminogen activator inhibitor 1, C-C motif chemokine 14, Chitinase-3-like protein 1, Insulin-like growth factor-binding protein 2, Meteorin-like protein, Ficolin-2 and Dystroglycan 1) (Figure 5.40.B). These additional significant changes demonstrated the potential of the Olink platform to expand the coverage of discovery MS and to be used as a complementary method to MS. For example coagulation factor VII was found significantly more abundant in calcified samples (Figure 5.40.A), which could be expected as other coagulation factors such as IX and X were also found significantly more abundant in calcified samples from MS (Figure 5.13). However, these proteins were found significantly changing only by one platform and, therefore, further validation is needed. In contrast, Insulin-like growth factor-binding proteins 1 (IBP1) and 2 (IBP2) had opposing directionality in the calcified vs non-calcified comparison by Olink (Figure 5.40.A).

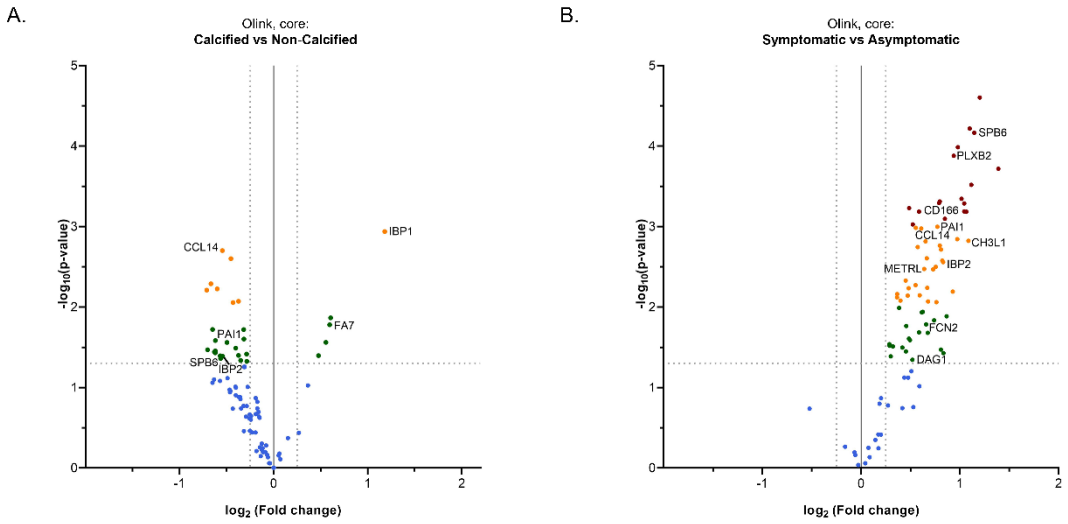


Figure 5.40 Unique extracellular protein changes in the core of the plaques using the Olink platform. Volcano plots of significantly dysregulated proteins in the calcified (n = 62) vs non-calcified (n = 48) and symptomatic (n = 37) vs asymptomatic (n = 73) comparisons of the plaque cores, using the measurements from the Olink platform. Significant proteins in **A.** calcified vs non-calcified or **B.** symptomatic vs asymptomatic comparisons identified only with the Olink platform are labelled.

5.3.15 Non-standard Mass Spectrometry pre-processing searches and analysis

To expand the coverage of the proteome we conducted additional database searches alternative to the standard processing, with variations in digestion enzyme, PTMs, or analysis software.

5.3.15.1 Semi-tryptic Searches

As mentioned above in the methods (5.2.2.6), trypsin was used as a digestion enzyme in our MS analysis. A semi-tryptic search in both GuHCl and NaCl extracts was first performed to expand the coverage of the plaque proteome. Semi-tryptic peptides are the ones cleaved at the C-Terminal side of arginine (R) and lysine (K) by trypsin at one end but not the other. The majority of the new peptides identified were belonging to proteins already quantifiable from the tryptic search and when running statistical analysis for the phenotypes of interest (ie calcification, inflammation), several gamma-carboxylated proteins were found to be significantly changing especially with calcification, as with the normal tryptic searches.

The significant proteomic changes that occurred with the semi-tryptic search are depicted in Figure 5.41 and Figure 5.42. As there were no additional identified proteins

from this search compared to the commonly used tryptic search, we wanted to identify additional significant proteins in our two basic comparisons of calcification and symptoms per extract. Even those proteins were still significant in the other extract of the normal tryptic search (ie A1BG was significant in the calcification comparison in the soluble matrisome in the tryptic search).

We observed that Alpha 1-B glycoprotein (A1BG) was significantly changing in the core matrisome of the semi-tryptic search in both our comparisons, retaining their inverse association too. For this reason, we decided to analyse A1BG at the peptide level and identify whether the average peptide quantity was higher in a certain phenotype of plaques in calcification (Figure 5.43). Most of its peptides were more quantifiable in calcified plaques in both extracts, even though there were minor exceptions.

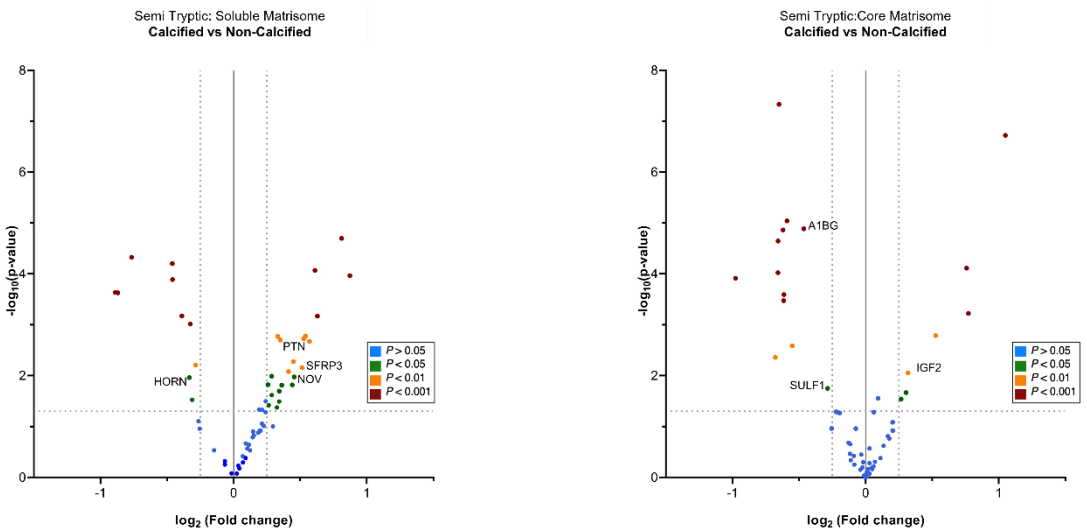


Figure 5.41 Significant changes with calcification in plaque cores, using semi-tryptic search. Volcano plots of significantly dysregulated proteins in the calcified (n = 60) vs non-calcified (n = 46) comparisons of the plaque cores for the soluble matrisome (NaCl) and the core matrisome (GuHCl), using semi-tryptic search. Protein changes with absolute values of fold change <0.25 are colored in blue and considered not significant. Significant proteins (p-value<0.05) that were not found significant in the normal tryptic search in each extract are labelled.

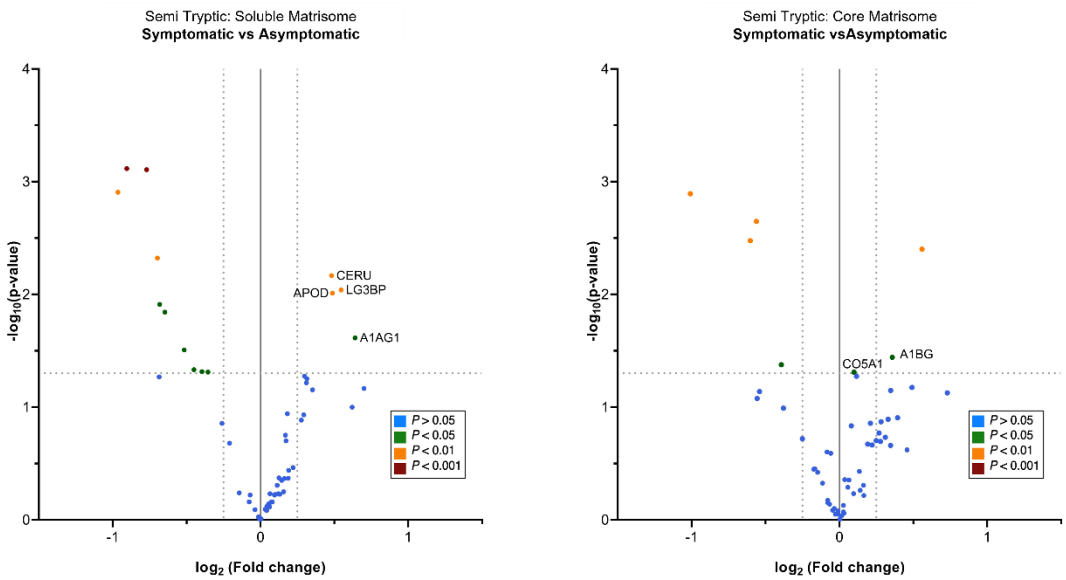


Figure 5.42 Significant changes related to symptoms in plaque cores, using semi-tryptic search. Volcano plots of significantly dysregulated proteins in the symptomatic (n=36) vs asymptomatic (n = 69) comparisons of the plaque cores for the soluble matrisome (NaCl) and the core matrisome (GuHCl), using semi-tryptic search. Protein changes with absolute values of fold change <0.25 are labelled in blue and considered not significant. Significant proteins (p-value<0.05) that were not found significant in the normal tryptic search in each extract are labelled.

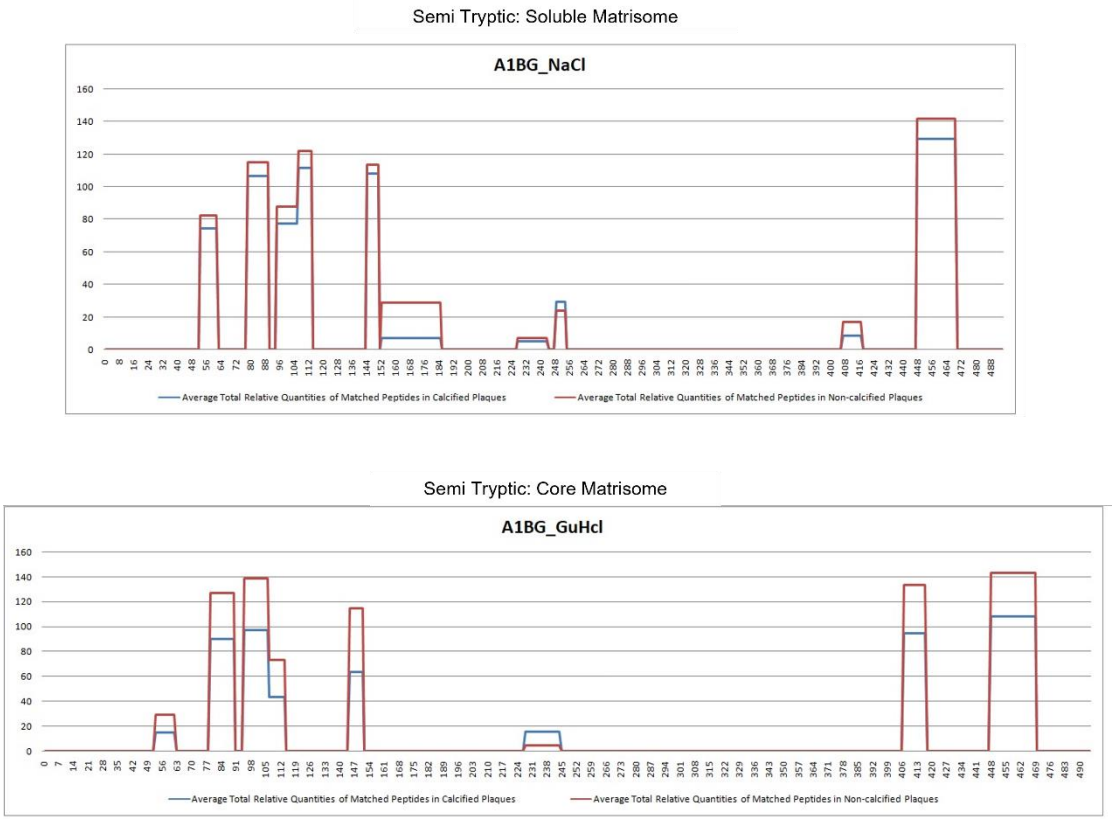


Figure 5.43 Alpha 1-B glycoprotein: peptide quantification in semi-tryptic search. Average relative quantities of A1BG per amino acid between calcified and non-calcified samples in soluble and core matrisome.

5.3.15.2 Gamma-carboxylation Searches

Since gamma-carboxylated proteins were significantly changing in our phenotypes of interest, we proceeded to search for gamma-carboxylation as a post-translational modification in the GuHCl extract. 34 gamma-carboxylated peptides were identified, from 8 different proteins. Of those, only five had less than 30% missing values in the 201 samples used, three of which were belonging to carboxylated proteins of the identified calcification signature (PROC, FA9, FA10). Thus, we explored whether there was a difference in the total quantity of calcified versus non-calcified plaques (Figure 5.44). Most of the calcified samples contained more gamma-carboxylated peptides than non-calcified samples. However, part of protein C carboxylated peptide was more abundant in non-calcified samples (Figure 5.44.C). Since this site of the protein belongs to carboxylated peptides according to the known positions of carboxylation from Uniprot and the literature (aa 48, 49, 56, 58), we would have expected higher abundance in calcified samples. Then further explored the overall amount of

carboxylation, from all consistently identified gamma-carboxylated proteins. This analysis showed that the carboxylated peptides were significantly more abundant in calcified compared to non-calcified samples (p-value=0.0305).

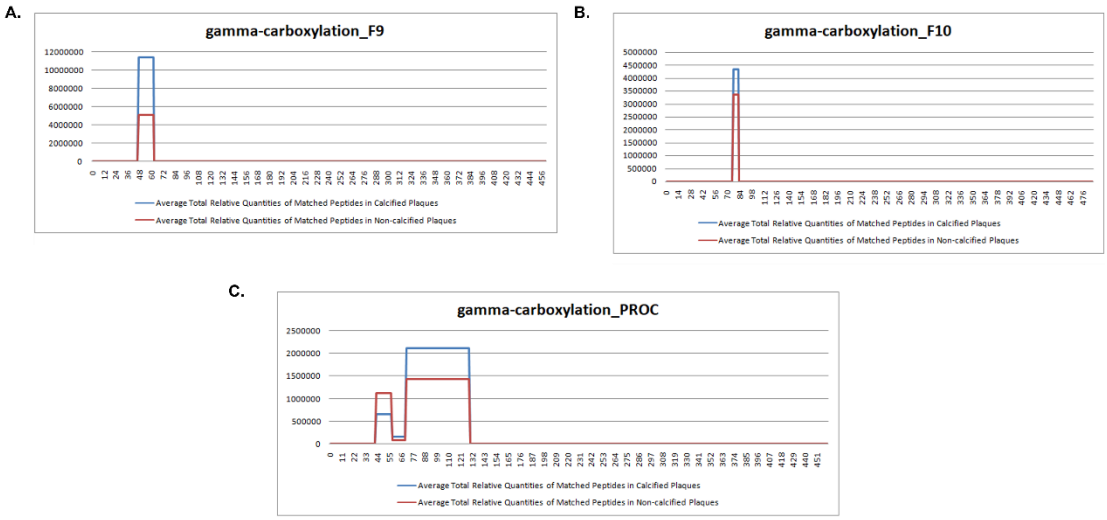


Figure 5.44 Peptide relative quantities in selected gamma-carboxylated proteins. Average relative quantities per amino acid between calcified and non-calcified samples in **A.** coagulation factor 9, **B.** coagulation factor 10 and **C.** vitamin K-dependent protein C.

5.3.15.3 MaxQuant Algorithm comparison results

Apart from the standard processing pipeline for MS with Proteome Discoverer software, there are also alternative search engines which aim to optimize the data quantification process. One of the most widely used is the MaxQuant algorithm (195), a proteomics-based software for the analysis of MS experiments. We proceeded with the use of this algorithm, to compare its performance to the one of Proteome Discoverer (Figure 5.45). As shown in **Error! Reference source not found.**.A, there were many common proteins identified using both workflows. However, both search engines had identified more than 150 unique proteins, with Proteome Discoverer being able to identify 18 unique proteins more than MaxQuant. After data filtering for missing values (30% threshold allowed), almost 50% of identified proteins with Proteome Discoverer were filtered out due to high missingness whereas the corresponding percentage for MaxQuant was 20% of proteins identified. Finally, we wanted to explore the correlation between the commonly identified proteins between the two engines. For this purpose, we used the basic comparison of the core versus

the periphery of the plaques (Figure 5.45.C). There was an excellent correlation between the log fold changes in this comparison between the two engines, probably due to both commonly identified proteins and similar capability and degree of analysis of MS proteomics.

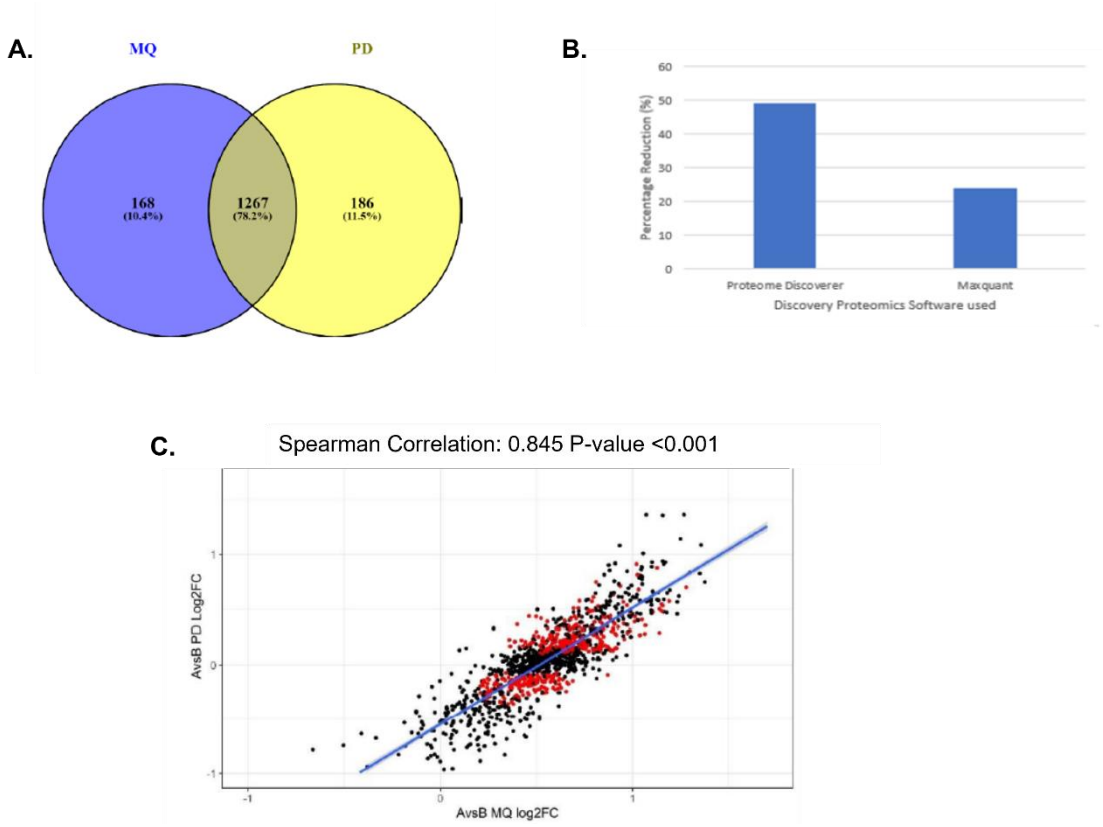


Figure 5.45 Comparison of Proteome Discoverer and MaxQuant algorithm protein identification results. **A.** Venn diagram depicting the identified proteins when using MaxQuant and Proteome Discoverer. **B.** Bar chart showing the percentage reduction between identified and quantified proteins in MaxQuant and Proteome Discoverer after data filtering data for 30% of missing values threshold. **C.** Scatter plot depicting the Spearman correlation between the log2FC of the core versus periphery comparison in MaxQuant and Proteome Discoverer. Significant proteins in both algorithms are colored in red colour.

5.3.16 Molecular plaque phenotypes, biosignatures, and CVD risk prognostic models

Principal component and clustering analysis of the patients using the plaque core samples were performed combining the cellular and matrisome biosignature abundances in the core of the plaques to identify plaque subtypes of potential clinical importance using proteomics. For the clustering we used the KMEANS algorithm,

optimizing the number of clusters by searching between 2 and 20 clusters and selecting the clustering with the highest Calinski-Harabasz metric. We obtained four distinct clusters (Figure 5.46), for which we performed correlation analysis to patient characteristics (Figure 5.47). Cluster 1 was enriched in symptomatic and mixed (imaging) plaques and cluster 2 in echolucent and complex (histology) plaques. Both clusters had more male patients. Cluster 3 was mainly composed of calcified and echogenic plaques in female patients, while cluster 4 showed no correlations with clinical parameters. Further analysis revealed that the first principal component (PC1) is inversely associated with structural ECM proteins, PC2 is negatively associated with the calcification signature, and PC3 is positively associated with SMC markers but negatively with CD14 and OSTP.

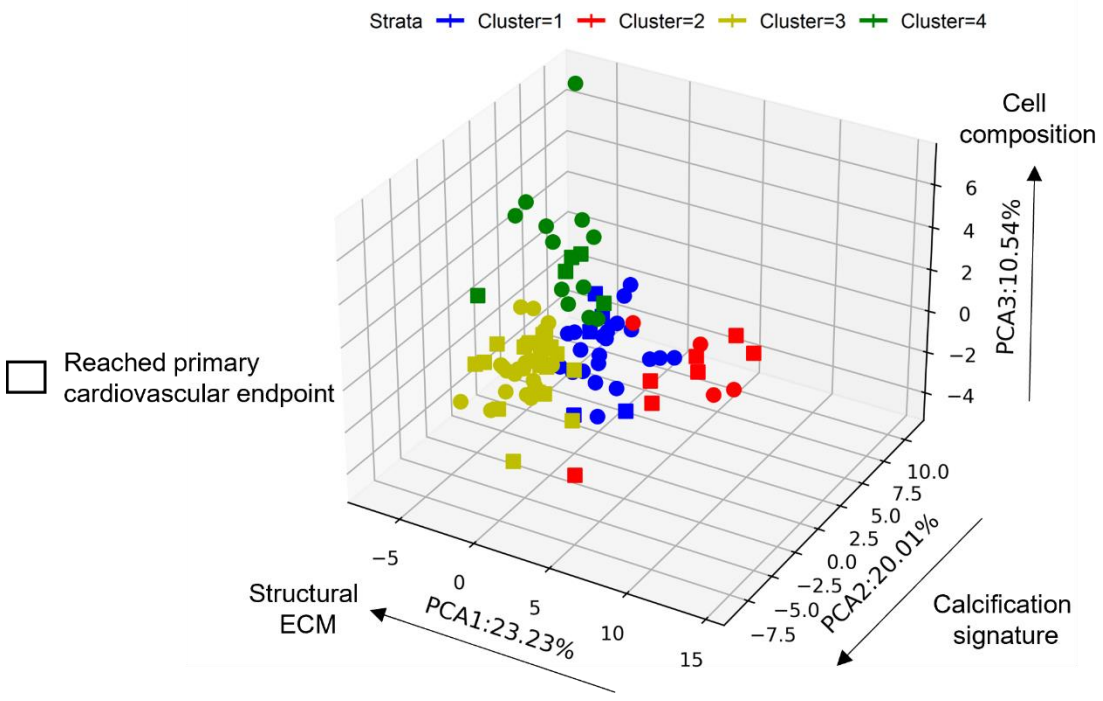


Figure 5.46 Clustering of patients using the proteomic biosignatures. Representation of the plaque cores in a PCA with the 3 most significant PCs including structural ECM proteins (PC1), proteins of the calcification signature (PC2) and cellular marker proteins (PC3). KMEANS algorithm was used for clustering and clusters are visualized with different colors. K was identified as the number of clusters that achieved the highest Calinski-Harabasz distance when examining values from 2-20. Rectangles in all figures represent cases with a primary cardiovascular endpoint over a 9-year follow-up.

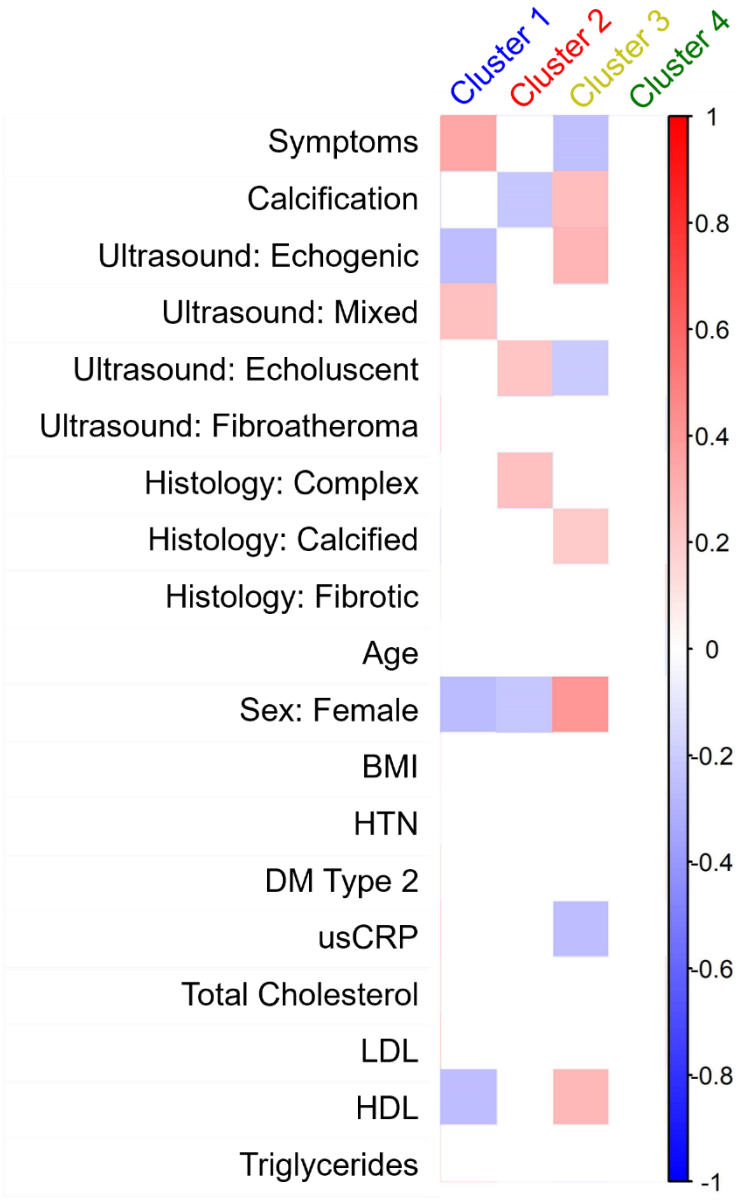


Figure 5.47 Correlation of the proteomics clusters to demographics, clinical, imaging, and histology characteristics of the patients. For binary characteristics, Pearson correlation was used, and significance was measured with Fisher exact test. For continuous characteristics, the point-biserial correlation was used. BMI: body mass index, HTN: hypertension, DM: type 2 diabetes mellitus, usCRP: ultra-sensitive C-reactive protein, LDL: Low-density lipoprotein cholesterol, HDL: High-density lipoprotein cholesterol.

Finally, we explored whether molecular changes identified by proteomics can be associated with patient outcomes. We had a 9-year follow-up for patients, one-third of which reached primary cardiovascular endpoints: composite of cardiovascular

death, myocardial infarction, transient ischemic attacks, or stroke as well as atherosclerosis progression in the coronary or peripheral arteries requiring either interventional (percutaneous coronary intervention or peripheral balloon angioplasty with and without stenting) or surgical revascularization (aortocoronary bypass or peripheral bypass). While the classification by ultrasound and histology did not predict the outcome in our cohort, unsupervised clustering based on proteomics reached an Area under the Curve of 60.4% (**Error! Reference source not found.**). Cluster 2, representing predominantly echolucent plaques with reduced structural ECM proteins in males, was most strongly associated with adverse cardiovascular outcomes (Figure 5.48). In contrast, cluster 1 with symptomatic, mixed plaques, and cluster 4, representing plaques with a high SMC content were associated with good prognosis. Notably, cluster 3 with mainly calcified, echogenic plaques in females showed still a high rate of adverse cardiovascular events during follow-up.

Dataset	Method	AUC†	Precision	Recall
Discovery Cohort	Ultrasound (Echolucent and Mixed vs Echogenic)	46.6%	26.2%	40.7%
	Histology (Complex and Calcified vs Fibroatheroma and Fibrotic)	50.0%	39.0%	65.7%
	Proteomic Clustering (Clusters 1 and 2 vs Clusters 3 and 4)	60.4%	47.2%	71.4%
	Machine Learning Signature by MOEA/SVM: CNN1, PROC, SERPH, CSPG2 *	75.0%	63.4%	74.2%
Validation Cohort – Athero-Express	Plaque Vulnerability Index (Categories 0, 1, and 2 vs Categories 3, 4, and 5)	51.0%	25.0%	50.0%
	Machine Learning Signature by MOEA/SVM: CNN1, PROC, SERPH, CSPG2 *	67.5%	43.1%	86.3%

Table 5.6 Machine learning analysis for the prediction of the 9-year follow-up primary endpoint. CNN1 - calponin, PROC - vitamin-K dependent protein C, SERPH - serpin H1, and CSPG2- versican. MOEA/SVM: Hybrid method combining Multi-Objective Evolutionary optimisation Algorithm and Support Vector Machines. *Performance metrics for the machine learning biosignature in the discovery cohort

were measured using 10-fold cross-validation. †Area Under the Curve (AUC) of the receiver operating characteristic curve.

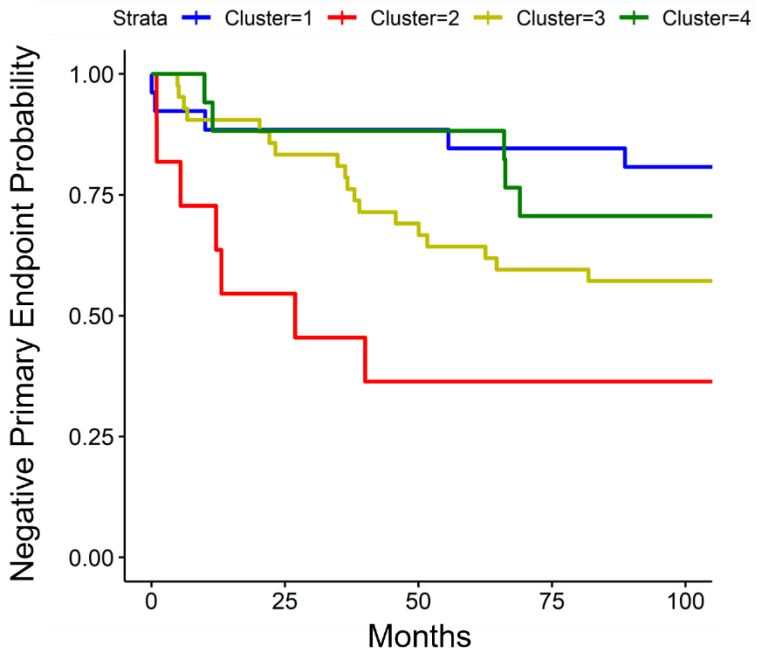


Figure 5.48 Molecular plaque phenotypes associated with cardiovascular outcomes. Kaplan-Meier plot for the survival analysis of the four distinct molecular plaque phenotypes based on the primary composite cardiovascular endpoint over a 9-year follow-up.

We then applied a Multi-Objective Evolutionary Optimization Algorithm to identify the minimum biosignature that performs best when used as input for a non-linear classifier (Radial Basis Function Support Vector Machines). This method revealed a biosignature of four proteins, the SMC marker CNN1, the gamma-carboxylated protein PROC, the collagen-binding protein SERPH, and the large aggregating proteoglycan CSPG2, which in combination provided an AUC of 75.0% using 10-fold cross-validation (**Error! Reference source not found.**).

5.4 Discussion

In this chapter we presented the most comprehensive proteomic human study on carotid atherosclerotic plaques, also profiling their spatial proteome. We analyzed the plaque proteome of 219 plaque samples from 120 patients using different proteomics analysis methods and identified unique molecular signatures for plaque calcification and inflammation. We correlated those signatures to clinical characteristics, such as sex and integrated our proteomics findings with transcriptomics, ultrasound and

histology measurements, and an additional validation cohort. Moreover, we used network and further bioinformatics analyses to further explore phenotype-specific proteomics changes. We also assessed the ability of revealed proteome signatures to predict cardiovascular outcomes and atherosclerosis progression, using machine learning. Our results show that the local plaque protein composition can reflect different atherosclerosis stages and could predict long-term adverse cardiovascular outcomes.

Previous proteomics studies in human carotid atherosclerotic plaques were small in size, with most of them involving less than 30 samples and the largest involving 48 (260). The largest proteomic study in atherosclerosis so far is the one of Herrington et al. (61), who performed a proteomic analysis in 200 samples from coronary artery and aortic specimens from 100 autopsied young adults. Our cohort has a similar size but we used fresh frozen carotid endarterectomy specimens instead of autopsy specimens. Analyzing specimens acquired after death can be challenging due to the extravasation of inflammatory cells which could influence proteomics findings, and to possible effects of death on the stability of proteins. What is more, Herrington et al. (61) used a DIA library of 77 ECM proteins and did not include core ECM proteins such as versican, aggrecan, collagen V chains or protease inhibitors (TIMPs). By using fresh frozen samples in our analysis and performing sequential extraction to enrich for cellular, core and soluble ECM proteins, we doubled the coverage of the human atherosclerotic plaque proteome compared to the study of Herrington et al. (61) Notably we only included proteins quantified in more than 70% of the samples, as opposed to Herrington et al. study, which used and considered proteins quantifiable if they were present in more than 50% of the samples. These two large proteomics studies complement a metabolomic study in carotid plaques from Tomas et al.(276) who used a targeted metabolomics approach to measure 165 metabolites from 159 carotid plaques and found that high-risk plaques show a different metabolic profile which was similar to the metabolic signature of activated leucocytes and cancer cells.

We used three extracts to identify cellular content, soluble and core matrisome proteins from two regions of the plaques (core and periphery) and were able to assess the spatial variation in cellular and extracellular proteins. The plaque core showed a

higher abundance of inflammatory cells like macrophages whereas the plaque periphery showed a higher abundance of smooth muscle cells. The cellular proteome revealed nine clusters, among which cell clusters including SMCs, neutrophils and macrophages. We then used single-cell and spatial RNA-seq to validate the specificity of the cellular markers we chose. Single-cell RNA-seq has the limitation that it needs enzymatic digestion and dissociation of the tissue into single cells to isolate viable cells. This can cause a bias towards cells that can be easily dissociated into single cells and remain viable during extraction. Infiltrating immune cells are more easily dissociated than cells that are embedded into the ECM, making single-cell RNA-seq biased towards inflammatory cells and particularly T-cells. While most T-cells, such as T helper and T regulatory cells, play an important role in atherogenesis, their contribution to plaque proteome is minor (235, 277). Moreover, RNA degradation can occur in viable cells non-uniformly, as sensitivity to RNA degradation is known to vary between tissues and cell types (278). Myeloperoxidase, a transcript most abundantly expressed in neutrophils (279), was barely detected by scRNA-seq. Only calprotectin (S10A8/A9) was detectable in a small number of neutrophils. Spatial RNA-seq is an alternative solution to scRNA-seq which does not require tissue dissociation and cell isolation. Using spatial RNA-seq we were able to identify regional clusters with different gene expressions.

Intimal vascular calcification is thought to occur as an imbalance of calcification inhibitor proteins such as Matrix Gla protein and stimulatory mediators such as osteocalcin (OSTCN) (20) and shares some molecular mechanisms with bone formation (19). The calcification inhibitors, to be functional, contain the gamma-carboxylation post-translational modification, which is dependent on vitamin K and replaces certain glutamate residues (Glu) with gamma-carboxyglutamate ones (Gla). Warfarin, a commonly prescribed oral anticoagulant, is a vitamin K antagonist and is associated with an increased risk of vascular calcification, suggesting that gamma-carboxylation of ECM proteins is a protective response to vascular calcification and inhibition might increase the likelihood of vascular calcification initiation and progression (280). Since none of our patients was under warfarin treatment, the differences we observed in gamma-carboxylated proteins were not due to

medications. We were able to identify a plaque calcification signature consisting of collagens and collagen fibril organization-related proteins such as integrin beta-1 (ITB1), and vitamin K-dependent proteins such as MGP and protein C. MGP is the most well-known calcification inhibitory protein and has been first reported to be highly expressed in atheromatous plaques *in vivo* more than 20 years ago (281). Viegas et al. (280) used immunohistochemistry and quantitative polymerase chain reaction and found that MGP was also upregulated in calcified aortic valve disease. Interestingly, Dhore et al. (282) have shown that inhibitors of calcification such as MGP are expressed in all stages of atherosclerosis whereas activators of calcification such as osteopontin, which we found upregulated in calcified plaques, are present only in advanced calcified lesions. Apart from gamma-carboxylated calcification inhibitors, we also found coagulation factors (such as coagulation factors IX and X) and anticoagulant factors (such as vitamin K-dependent protein C) upregulated in calcified lesions. Kapustin et al. (283) also used our MS data and identified these proteins to be detected in vascular smooth muscle cell-derived exosomes, to be increased in calcifying conditions and to contribute to the inhibition of exosome-mediated calcification. When we searched for the gamma-carboxylation post-translational modification we were able to consistently identify and quantify these coagulation and anticoagulant factor proteins (PROC, FA9, FA10). Peptide level analysis showed that most of the calcified samples contained more gamma-carboxylated peptides than non-calcified ones. The most pronounced protein change in calcified plaques was the increase of Alpha-2-HS-glycoprotein, which is a known inhibitor of vascular and tissue calcification (284). FETUA is produced in the liver, forms a complex with calcium phosphate and has been associated with calcified aortic valve disease (280). Consistent with the notion that calcification dampens vascular inflammation, the signatures linked to plaque calcification and inflammation were inversely correlated. This result indicates a protective role of calcification in atherosclerotic plaques and a correlation of calcification with plaque stability, as calcified lesions are more stable while non-calcified lesions are more vulnerable and prone to plaque rupture. This finding is supported by a recent study, which used computer tomography angiographies to measure high and low-calcified carotid plaques (285). When we assessed our findings changes in the transcript level none of the changes was

significant, thus calcification-related changes are better captured at the protein level. When exploring changes between symptomatic and asymptomatic plaques by using only non-calcified plaques PROC and FETUA were still upregulated in asymptomatic plaques, suggesting a role in the early calcification response or plaque stability.

Inflammatory proteins play an important role in all stages of atherogenesis, from early to advanced atherosclerotic lesions (264). The interplay between SMCs and the proinflammatory milieu impacts the local protease activity and the composition of the plaque ECM (286). Being in line with the results of a previous study from our lab (17), we found cathepsins (CATB, CATD), calprotectin (S10A8/S10A9) and MMP9 to be included in the plaque inflammatory signature. Additional studies have also highlighted the importance of these proteins for symptomatic plaques. Levels of calprotectin have been correlated with plaque instability and recently suggested as a therapeutic target for atherosclerosis in diabetic patients (287). MMP9 has been found in foam cells, and unstable plaques and has been linked to plaque vulnerability and rupture (288). Interestingly, when Sulkava et al. (289) assessed metalloproteinases in different arterial beds MMP9 was found mostly upregulated in femoral plaques, while MMP12 was the most upregulated one in carotid plaque, suggesting a site-specific expression of MMPs. In our results, MMP12 was not found significant in any of the two extracts but was found significantly upregulated in non-calcified, symptomatic plaques using label-free MS (TopS), showing that possibly calcification influenced this result. CATB belongs to cysteine cathepsins. Weiss-Sadan et al. (290) described the role of cysteine cathepsins in atherosclerosis. Cysteine cathepsins are highly expressed in macrophages, endothelial cells and smooth muscle cells within the plaque tissue. They are involved in many activities promoting inflammation, plaque evolution and destabilization, and atherosclerosis progression, such as lipid metabolism, lysosomal dysfunction, ECM remodelling, autophagy and inflammasome activation. CATD was also found upregulated in symptomatic plaque. CATD has been previously found to be significantly changing in atherosclerotic plaque (291) and has been associated with coronary artery disease (292). Cathepsins, matrix metalloproteinases, and their inhibitors were strongly associated with the presence of monocyte/macrophages and neutrophils. MMP-9 was correlated with neutrophils

(293), whereas cathepsins have been associated with macrophage content (294). In the present work, we further expand upon the importance of neutrophils in advanced atherosclerosis by demonstrating that one of the most significantly associated proteins with symptomatic plaques was myeloperoxidase (295). We also found markers for collagen degradation such as Xaa-Pro dipeptidase (PEPD), and leukocyte recruitment such as Adhesion G protein-coupled receptor E5 (CD97), to be upregulated in symptomatic plaques. Finally, we found ferritin light and heavy chains (FRIL, FRIH) increasing in plaques from symptomatic patients. Ferritin can promote plaque destabilization through the oxidation of lipids by iron overload (296, 297). The increase of ferritin in symptomatic plaques might be due to intraplaque haemorrhage or leakage from *vasa vasorum*. When transcriptomics was assessed, half of the significant changes at the protein level were validated. Among these were the upregulation of FRIH and FRIL in symptomatic plaques and the upregulation of GAS6 and MGP in asymptomatic plaques.

Hartman et al. (133) recently identified sex-related changes using gene regulatory network analysis on RNA-seq data and revealed Growth arrest-specific protein 6 and Serpin Family G Member 1 (SERPING1) as potential key SMC driver genes in females compared to males. We found GAS6, which was also a part of the calcification signature, significantly more abundant in the core of the plaques of female patients in our cohort, whereas SERPING1 did not change in the protein level. Calcification was correlated and more common in female patients in our cohort. Moreover, de Bakker et al. (298) used immunohistochemistry to study age and sex-related differences in peripheral atherosclerotic iliofemoral plaques and showed that men had a higher presence of lipid cores and plaque haemorrhage compared to women and women had an age-related increase in plaque calcification and haemorrhage, whereas age did not affect these characteristics in men. When excluding protein changes attributed to calcification (such as GAS6), we only found two significant sex-related clusters (C17, C20) containing the two large aggregating proteoglycans versican and aggrecan. The changes of these two proteins were validated in an independent cohort, together with additional sex-specific matrisome changes, such as the hyaluronan and proteoglycan link protein 3 (HPLN3). Versican, aggrecan and HPLN3 were all upregulated in the

periphery of the plaques compared to the core suggesting that the revealed age differences might be associated with the previous findings of women having smaller lipid cores in their plaques compared to men. Using cellular proteomics and network analysis, we found MMP9 and LEG3 to have a significantly higher betweenness centrality in female patients compared to men, suggesting a potentially more prominent role of neutrophils and macrophages in women. The implication of MMP9 and LEG3 as potential therapeutic targets for atherosclerosis needs further mechanistic experiments and validation.

Carotid duplex ultrasound is the first-line diagnostic procedure in patients to assess carotid artery stenosis (299). Echolucent plaques have been associated with a risk for ischemic cerebrovascular events (300, 301). We assessed the changes between echogenic and echolucent plaques. Among the protein changes corresponding to echogenic plaques were fibrillar collagens and some proteins of the calcification signature such as protein C. Thus, the current imaging classification reflects the collagen content of the plaque, correlates to the calcification signature, but ultrasound imaging cannot discern between lesions with high inflammation and stable fibroatheroma with low-grade inflammation. This could also explain the observed lack of association between ultrasound classification with clinical presentation (symptomatic/asymptomatic) as well as with future cardiovascular events. Proteomic signatures could complement carotid imaging read-outs to improve prognostication.

We assessed different search methods, such as label-free TopS, Olink and MaxQuant algorithms, to compare our findings. In general, different methods agreed for most of the proteins comprising our calcification and inflammation signatures. More specifically, we mostly used TopS to have an initial assessment of our cohort and select proteotypic peptides for our validation targeted method. Most of the inflammation signature proteins were validated by the Olink platform whereas the calcification comparison did not show good agreement. This result was expected, as Olink has designed antibodies that specifically target proteins, and favors less abundant proteins unlike MS and its assays are validated in plasma and serum (33). Thus, since our cohort included tissue samples, many of the measured proteins were out of the analytical range. On the contrary, targeted proteomics by MS validated almost

perfectly the identified significant proteins of discovery proteomics. Olink platform identified 15 matrisome proteins in total not identified by discovery MS, 6 of which were significantly changing in the calcification comparison and 10 of which in the symptomatic versus asymptomatic comparison. Therefore, using Olink in combination with MS rather than validation is more suitable for this cohort. The comparison with the MaxQuant pipeline resulted in the main selected method in our analysis being able to perform slightly better, identifying more unique proteins as expected (302). Both pipelines shared many commonly identified proteins which showed high correlation. The non-standard search performed using semi-tryptic peptides, lead to the identification of very few additional proteins compared with discovery proteomics. Some of those have been previously found to be involved in calcification or atherosclerosis. Secreted frizzled-related protein 3 (SFRP3) and pleiotropin (PTN) were significantly upregulated in calcified plaques in our cohort. SFRP3 was previously found to be predictive of cardiovascular outcome in aortic stenosis and heart failure and its role in atherogenesis need to be further explored (303), whereas pleiotropin has been shown to potentiate chondrogenic differentiation, increase transcription of hypertrophic chondrocyte markers and enhance calcification (304). Ceruloplasmin (CERU) and apolipoprotein D were found significantly upregulated in symptomatic patients. CERU has been previously found to show higher levels of expression in coronary heart disease and has been associated with cardiovascular disease (305) and increased levels of apolipoprotein D were recently linked to poor outcomes in patients with coronary artery disease (306).

Finally, using proteomics we managed to identify four distinct plaque phenotypes. Echolucent plaques with reduced structural ECM proteins had worse long-term cardiovascular outcomes. Ultrasound measurements of the carotid plaque area and intraplaque neovascularization were predictive of future cardiovascular events (307, 308). However, this prediction was based on the separation of early from advanced stages of atherosclerosis. In our cohort, all patients had advanced atherosclerosis. Histology with AHA classification is performed after surgery to assess the morphology of atherosclerotic plaques (309). However, limited data exists on histological plaque properties and patient outcomes in contemporary cohorts (310, 311). A histology-

based plaque vulnerability index predicted outcomes in a previous study (312). Since symptomatic carotid stenosis is defined as the occurrence of symptoms ipsilateral to the stenosis during the preceding 6 months (299), histological findings may not reflect the previous clinical presentation. The morphology of atherosclerotic plaques undergoes continuous changes over time. Most plaques rupture and heal several times at the stage of advanced lesions (313). Previous studies have shown no reduction in stroke risk when the endarterectomy was done more than 12 weeks after an index event in symptomatic patients and that women had no benefit when carotid endarterectomy was performed more than 4 weeks after symptom onset (314, 315). Thus, just 1-3 months after plaque rupture or symptom onset, a symptomatic plaque could have again similar histological morphology as an asymptomatic one. On the contrary, proteomics analyze complex, mutually dependent proteins which cannot be assessed histologically. Here, we were able to identify four distinct plaque clusters based on the protein composition of plaques. Based on this clustering, calcified plaques had the second worst cardiovascular outcome, following echolucent plaques with reduced ECM proteins, as stated above. Calcified plaques were more common in females, further emphasizing the importance of sex-specific differences in atherosclerosis. These plaques are traditionally considered as stable lesions, with a low risk of embolization (316). However, a high level of calcification does not necessarily mean more stable lesions and it has been shown that the incidence of preoperative neurological symptoms was similar in patients with calcified and non-calcified plaques (24). In line with this, we showed that patients with calcified plaques have a high rate of adverse cardiovascular outcomes, similar to the outcomes of patients with vulnerable plaques. Using proteomics, we were able to identify patients with both calcified and non-calcified plaques at risk for future cardiovascular events. In line with the notion that atherosclerosis is a systemic inflammatory disease, local plaque composition is informed on the progression of systemic atherosclerosis. We found a biosignature of four proteins (calponin-1, protein C, serpin H1 and versican) predicting composite cardiovascular endpoint with an AUC of 75% in the discovery cohort and 67.5% in the validation cohort, over 9 and 3-years follow-up period, respectively.

6. General Discussion

6.1 Conclusions

In this thesis, we assessed the existing methods for the reconstruction of biological networks pointing out their strengths and weaknesses, developed novel algorithms for the analysis of MS proteomics, addressed limitations of existing analysis techniques, benchmarked them against widely used open-source and commercial pipelines and applied them to different published and novel datasets. Finally, we analyzed the largest known proteomics dataset of carotid atherosclerotic plaque samples, and created and stored our results in a relational database.

We introduced the DiRec-AP algorithm, a mutual information-based technique for the reconstruction of biological regulatory networks. DiRec-AP overcomes the limitations of existing network reconstruction techniques, setting a different association threshold per protein for inferring protein interactions, creating directional networks and locating negative associations among proteins. We benchmarked the algorithm against the most representative tools of each different network reconstruction technique using different public networks and observed a big improvement in all network metrics. We applied DiRec-AP in the extracellular matrisome of carotid and coronary artery atherosclerotic plaque datasets and reconstructed, to our knowledge, the largest networks with such coverage for the first time. Moreover, the reconstructed cellular protein networks allowed for the discovery of cell-specific clusters and contributed to the interpretation of our MS data. Furthermore, we demonstrated that single-cell RNA-sequencing data can be used to interpret the findings of reconstructed protein networks. Finally, we were able to reconstruct phenotype and sex-specific plaque networks, which can be used to generate new hypotheses on regulatory relationships between matrisome proteins and further explore potential drug targets through the identification of network-specific proteins.

Standard labelled and label free proteomics datasets obtained using DDA Proteomics have high variability, high number of missing values and require further validation targeted and DIA proteomics pipelines have been developed to overcome these limitations. Targeted and DIA proteomics data analysis is a very time-consuming

process which until now involved manual and not optimized processing. To overcome these issues, we created an optimization pipeline, the HOpTar-omics tool, for the analysis of the different types of targeted and DIA MS data, overcoming the limitations of low reproducibility and specificity and high data missingness of such type of data. We applied HOpTar-omics to datasets from three different MS methods and were able to improve substantially the data completeness and quantification accuracy in all examined datasets and facilitate applicability on a large scale. Applying HOpTar-omics in combination with heavy standards in an atherosclerotic plaque apolipoprotein dataset allowed performing the largest absolute quantification study of apolipoproteins in atherosclerotic plaques and to suggest a potential novel therapeutic target (apolipoprotein J or clusterin) against the progression of atherosclerosis, that will be further explored experimentally. Moreover, this targeted proteomics pipeline allowed us to validate the inflammation, sex and calcification signatures revealed from the discovery proteomics.

In the last chapter, we combined discovery and targeted MS proteomics with network reconstruction and other bioinformatics analyses, to identify protein changes in carotid atherosclerotic plaques, and identify and validate phenotype and sex-specific differences. The cellular proteome revealed cells clusters that included SMCs, neutrophils and macrophages. We confirmed that the core of the plaques includes more inflammatory cells like macrophages, whereas the plaque periphery has more SMCs. Possible inflammatory mechanisms involved in atherosclerosis could be, according to our findings, leukocyte recruitment, since we found the CD97 marker to be upregulated in symptomatic patients, and collagen degradation, as PEPD was also upregulated in symptomatic patients. Among the major inflammatory proteins we found being upregulated in symptomatic patients, were neutrophil-derived proteins and more specifically, the also previously detected (17), cathepsins, calprotectin complex and MMP9. We further expanded the inflammation signature with ferritin light and heavy chains. This signature was inversely correlated with calcification and sex. Calcification signatures included collagens, vitamin K-dependent and gamma-carboxylated proteins, as well as coagulation factors, and most of these proteins were significantly downregulated in symptomatic patients. Among sex-associated proteins,

independent of inflammation and calcification, were versican and aggrecan proteoglycans, as well as link proteins that bind to hyaluronic acid such as HPLN3. Finally, based on the plaque proteome, we were able to reveal four distinct clusters related to cardiovascular outcomes and four potential tissue biomarkers to predict composite cardiovascular endpoints: calponin-1, vitamin K-dependent protein C, serpin H1 and versican. Based on our findings, we suggest that a more personalized, sex-specific, proteomics-based approach for risk stratification for systemic atherosclerotic burden should be considered.

6.2 Limitations

The present thesis and the algorithms developed have limitations.

Despite the DiRec-AP pipeline significantly increasing the performance of existing methods in reconstructing protein regulatory networks, as with all in-silico techniques, the reconstructed regulatory networks should always be validated with mechanistic studies and experimental techniques and could only serve towards prioritizing targets and forming hypotheses. The current approach does not consider other important omics data that could significantly increase the accuracy of network findings, such as lipidomics data, genomics data as well as relevant clinical indications, such as body mass index, cholesterol, and triglycerides measurements. The revealed significant phenotype-, sex- and vasculature-specific protein changes should be verified at the RNA level, and their mechanisms of action should also be studied in larger sample size studies, which would allow corrections of the effect of covariates and medications and address more effectively the high intra-patient variability in atherosclerotic plaques.

HOpTar-omics is not a standalone tool but works as an add-on to other open-source (Skyline) or commercial (Spectronaut) tools. Creating an accurate dataset through the optimization process of the HOpTar-omics tool requires other discovery proteomics or clinical measurements. This is not always possible and when it is, not all proteins have corresponding clinical or other measurements. In such cases, the optimization process can be limited or biased towards the proteins that we have measurements for. Nevertheless, recording the best parameters for each MS method (ie PRM, MRM or DIA data with and without heavy standards) and using the best-performing ones to

new datasets without other proteomics or clinical measurements according to the corresponding MS method, would probably facilitate the decision of the best performing parameters. What is more, the normalization process is being done based on the imputed for missing values dataset. This means that the dataset is treated like a discovery MS method, something that in the case of PRM or MRM with no heavy standards is not accurate. Finally, even though the parameters to be optimized remain the same for each dataset, every experiment is different, and HOpTar-omics cannot account for every separate case with the same accuracy in the produced results.

Limitations of the last chapter include the analysis of carotid endarterectomies which represent advanced atherosclerotic plaques which differ from early atherosclerosis (299). Also, future comparisons should include coronary plaques, which are more difficult to attain as fresh frozen samples, and plaques from different locations within the arterial bed, which have been associated with distinct genetic susceptibility loci (317). Despite being the largest proteomics study of human carotid endarterectomies so far, false-positive results are possible with the discovery proteomics. Thus, two separate validation methods, at the RNA and protein level, as well as stringent statistical analysis were employed. RNA expression changes only replicated a few of the changes observed at the protein level. The targeted proteomics PRM method, however, validated the discovery of proteomics data signatures. A strength of this method is that we used multiple proteotypic peptides for larger proteins (such as Versican and Aggrecan), thus information for different regions of the proteins was not missed. Moreover, more experiments are required to clarify the functional role of the identified tissue protein signatures. While we cannot infer a causal relationship between the identified protein signatures with plaque progression, we demonstrated clinical relevance for plaque classification during a long-term follow-up with one-third of patients reaching a cardiovascular endpoint. The validation cohort, nevertheless, included a shorter period for monitoring outcomes (3 instead of 9 years).

6.3 Future Work

Despite the improved results of the developed pipelines in all datasets used compared to state-of-the-art tools, some alterations can be done to make the performance of the algorithms more efficient. Firstly, the parallelization of the codes can be done, with

multiple threads and processes running at the same time. This would make them a lot faster and more computationally efficient. Next, we could expand the DiRec-AP algorithm, to promote further the understanding of the role of inflammatory networks in atherosclerosis. To achieve this, some advanced visualization features (through a link to a visualization tool such as Cytoscape) and methods could be incorporated to include prior knowledge from metabolic pathways databases and protein-protein interaction networks, as well as make DiRec-AP easier to use and more user friendly. Regarding validation of findings, new experiments, such as gene silencing and knockdown or cross-linking MS, could be conducted for the mechanistic interpretation of the directional signalling and regulatory links in the networks that have been reconstructed. New findings from the Olink platform regarding calcification and symptomatic status as well as significant findings from the MaxQuant algorithm search could be further validated with another modality or wet-lab experiments such as ELISA, to ensure that these findings are valid and not artifacts.

6.3.1 PlaqueMS Knowledge Base

To store the processed data of the Vienna Plaques cohort, results and data were obtained from other atherosclerotic plaque-related cohorts, and ease access and usage of our results, we designed and implemented a relational database, the PlaqueMS database.

6.3.1.1 DB Design

We used the Structured Query Language (SQL) to implement the PlaqueMS relational database. The diagram of Entities-Relations is presented in Figure 6.1. In the database are stored more than 6500 proteins identified from more than 1000 samples of different sample types (carotid arteries, coronary arteries, aortas and smooth muscle cell secretome samples from aortic explants), and more than 100 networks from different phenotypes, including male and female patients, calcified or non-plaques, symptomatic or not patients and echogenic or echolucent plaques. The different cohorts and data included in the knowledge base are described in Figure 6.2 and the different comorbidities, characteristics and phenotypes included for each dataset are depicted in Figure 6.3.

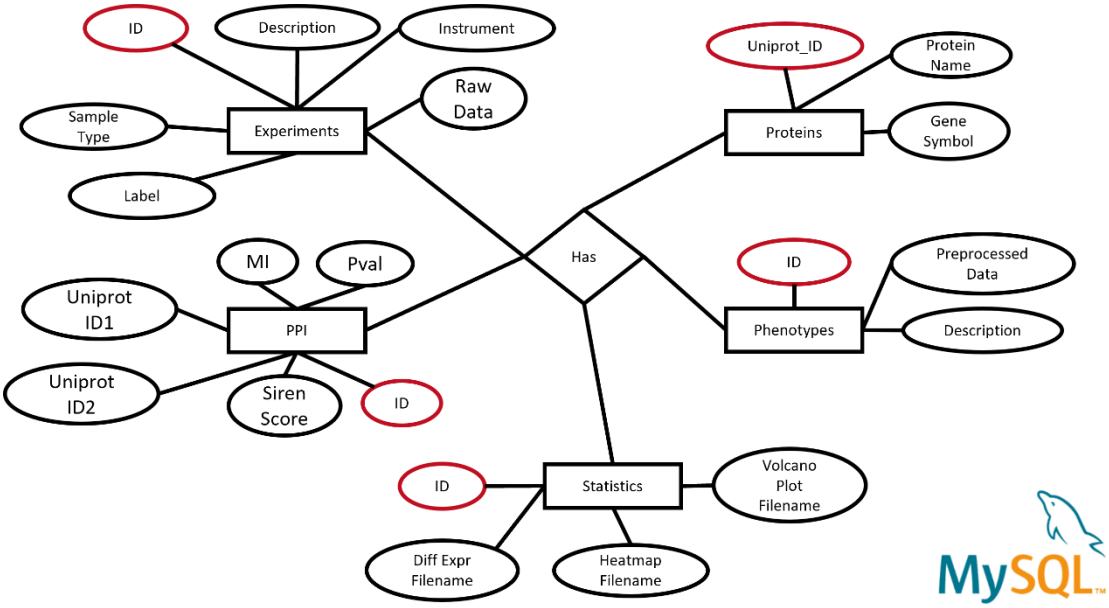


Figure 6.1. ER diagram of PlaqueMS database.

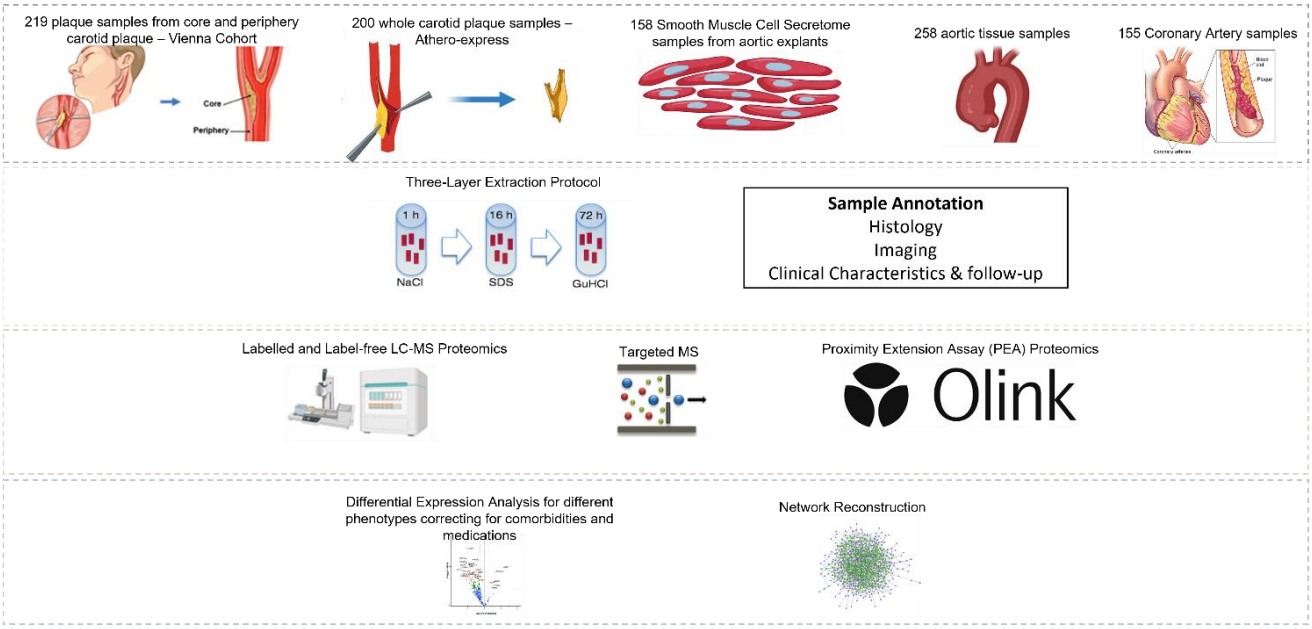


Figure 6.2 PlaqueMS DB cohorts and proteomics data description. In the database are included proteomics datasets from 5 different cohorts and metadata of atherosclerotic tissue and cells.

Tissue Type	Demographics		Positions	Disease	Histology	Ultrasound	Calcification	Outcomes
	Age	Sex						
Vienna Carotid Plaques	✓	✓	Periphery/ Core	Sympt/ Asympt	✓	✓	✓	✓
Carotid Plaques	✓	✓	✗	Sympt/ Asympt	✓	✗	✗	✓
Aortic SMCs	Partial	✓	✗	✗	✗	✗	✗	✗
Thoracic Aortas	✓	✓	Inner/ Outer	✗	✗	✗	✗	✗
Coronary Arteries	✓	✓	LAD/RCA/ Cx	Explants/ Normal, Atherosclerotic / Non-diseased	✗	✗	✗	✗

Figure 6.3 PlaqueMS DB: Phenotypes and characteristics per dataset. The characteristics, comorbidities and phenotypes included in the knowledge base for each dataset.

6.3.1.2 PlaqueMS - Example Query

The database can be searched in different ways and provide the users with the required images. For example, the user can search for “proteins changing between symptomatic and asymptomatic plaques in the core of non-calcified carotid plaques using labelled LC-MS Proteomics and core matrisome” and get as a result the corresponding volcano plot (Figure 6.4), as well as a file with the differential expression results.

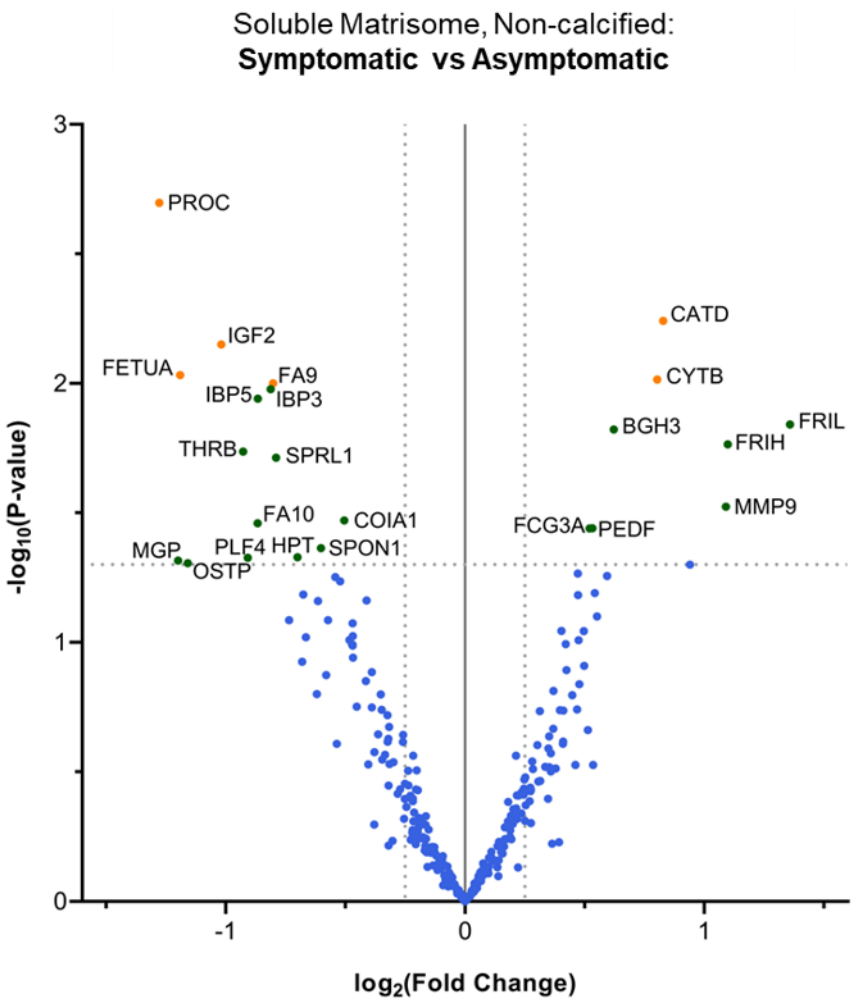


Figure 6.4 Result of PlaqueMS database example query. Volcano plot depicting significant changes between symptomatic and asymptomatic patients in the core of non-calcified plaques, using TMT discovery proteomics and GuHCl extract.

The PlaqueMS knowledge base could be further developed as a web tool, allowing users to mine combined data from atherosclerotic plaque characteristics, histology and proteomics findings, as well as clinical characteristics from different cohorts. Data integration and analysis tools could be included to the PlaqueMS knowledge base allowing for the identification of tissue and disease specific proteomic biosignatures and biosignatures of inflammation, calcification, symptoms and sex which are consistent across different tissues.

7. References

1. Libby, P., Ridker, P. M., & Hansson, G. K. (2011) Progress and challenges in translating the biology of atherosclerosis. *Nature,* 473(7347), 317–325

2. Mayr, M., Mayr, U., Chung, Y. L., Yin, X., Griffiths, J. R., & Xu, Q. (2004) Vascular proteomics: Linking proteomic and metabolomic changes. *Proteomics,* 4(12), 3751–3761

3. Zmysłowski, A., & Szterk, A. (2017) Current knowledge on the mechanism of atherosclerosis and pro-atherosclerotic properties of oxysterols. *Lipids in Health and Disease,* 16,(1) 1–19

4. Gimbrone, M. A., & García-Cardeña, G. (2016) Endothelial Cell Dysfunction and the Pathobiology of Atherosclerosis. *Circulation Research,* 118(4), 620–636

5. Navab, M., Ananthramaiah, G. M., Reddy, S. T., Van Lenten, B. J., Ansell, B. J., Fonarow, G. C., et al. (2004) Thematic review series: The Pathogenesis of Atherosclerosis The oxidation hypothesis of atherogenesis: the role of oxidized phospholipids and HDL. *Journal of Lipid Research,* 45(6), 993–1007

6. Bobryshev, Y. V., Ivanova, E. A., Chistiakov, D. A., Nikiforov, N. G., & Orekhov, A. N. (2016) Macrophages and Their Role in Atherosclerosis: Pathophysiology and Transcriptome Analysis. *BioMed Research International,* 2016

7. Bennett, M. R., Sinha, S., Owens, G. K., Libby, P., Bornfeldt, K. E., & Tall, A. R. (2016) Vascular Smooth Muscle Cells in Atherosclerosis. *Circulation Research,* 118(4), 692–702

8. Libby, P., Buring, J. E., Badimon, L., Hansson, G. K., Deanfield, J., Bittencourt, M. S., et al. (2019) Atherosclerosis. *Nature Reviews Disease Primers,* 5, 56

9. Yurdagul, A., Doran, A. C., Cai, B., Fredman, G., & Tabas, I. A. (2018) Mechanisms and Consequences of Defective Efferocytosis in Atherosclerosis. *Frontiers in Cardiovascular Medicine,* 4, 86

10. Llorente-Cortés, V., Martínez-González, J., & Badimon, L. (2000) LDL Receptor–

Related Protein Mediates Uptake of Aggregated LDL in Human Vascular Smooth Muscle Cells. *Arteriosclerosis, Thrombosis, and Vascular Biology,* 20(6), 1572–1579

11. Libby, P., Lichtman, A. H., & Hansson, G. K. (2013) Immune Effector Mechanisms Implicated in Atherosclerosis: From Mice to Humans. *Immunity,* 38(6), 1092–1104

12. Olie, R. H., van der Meijden, P. E. J., & ten Cate, H. (2018) The coagulation system in atherothrombosis: Implications for new therapeutic strategies. *Research and Practice in Thrombosis and Haemostasis,* 2(2), 188–198

13. Guo, H., Callaway, J. B., & Ting, J. P. Y. (2015) Inflammasomes: Mechanism of action, role in disease, and therapeutics. *Nature Medicine,* 21(7), 677–687

14. Reis Geovanini, G., & Libby, P. (2018) Atherosclerosis and inflammation: overview and updates. *Clinical Science,* 132(12), 1243–1252

15. Bäck, M., & Hansson, G. K. (2015) Anti-inflammatory therapies for atherosclerosis. *Nature Reviews Cardiology,* 12(4), 199–211

16. Bessueille, L., & Magne, D. (2015) Inflammation: A culprit for vascular calcification in atherosclerosis and diabetes. *Cellular and Molecular Life Sciences,* 72(13), 2475–2489

17. Langley, S. R., Willeit, K., Didangelos, A., Matic, L. P., Skroblin, P., Barallobre-Barreiro, J., et al. (2017) Extracellular matrix proteomics identifies molecular signature of symptomatic carotid plaques. *The Journal of Clinical Investigation,* 127(4), 1546–1560

18. Ramanathan, R., Gram, J. B., Sidelmann, J. J., Dey, D., Kusk, M. W., Nørgaard, B. L., & Sand, N. P. R. (2019) Sex difference in fibrin clot lysability: Association with coronary plaque composition. *Thrombosis research,* 174, 129–136

19. Nakahara, T., Dweck, M. R., Narula, N., Pisapia, D., Narula, J., & Strauss, H. W. (2017) Coronary Artery Calcification: From Mechanism to Molecular Imaging.

JACC: Cardiovascular Imaging, 10(5), 582–593

20. Tesfamariam, B. (2019) Involvement of Vitamin K-Dependent Proteins in Vascular Calcification. *Journal of Cardiovascular Pharmacology and Therapeutics,* 24(4), 323–333

21. Cardellini, M., Rovella, V., Scimeca, M., Anemona, L., Bischetti, S., Casella, S., et al. (2019) Chronic kidney disease is linked to carotid nodular calcification, an unstable plaque not correlated to inflammation. *Aging and Disease,* 10(1), 71–81

22. Ruiz, J. L., Weinbaum, S., Aikawa, E., & Hutcheson, J. D. (2016) Zooming in on the genesis of atherosclerotic plaque microcalcifications. *The Journal of Physiology,* 594(11), 2915–2927

23. Shioi, A., & Ikari, Y. (2018) Plaque Calcification During Atherosclerosis Progression and Regression. *Journal of Atherosclerosis and Thrombosis,* 25(4), 294–303

24. Pini, R., Faggioli, G., Fittipaldi, S., Vasuri, F., Longhi, M., Gallitto, E., et al. (2017) Relationship between Calcification and Vulnerability of the Carotid Plaques. *Annals of vascular surgery,* 44, 336–342

25. Raghunathan, R., Sethi, M. K., Klein, J. A., & Zaia, J. (2019) Proteomics, glycomics, and glycoproteomics of matrisome molecules. *Molecular & Cellular Proteomics,* 18(11), 2138–2148

26. Bonnans, C., Chou, J., & Werb, Z. (2014) Remodelling the extracellular matrix in development and disease. *Nature Reviews Molecular Cell Biology,* 15(12), 786–801

27. Chistiakov, D. A., Sobenin, I. A., & Orekhov, A. N. (2013) Vascular extracellular matrix in atherosclerosis. *Cardiology in Review,* 21(6), 270–288

28. Galis, Z. S., Sukhova, G. K., Lark, M. W., & Libby, P. (1994) Increased expression of matrix metalloproteinases and matrix degrading activity in vulnerable

regions of human atherosclerotic plaques. *The Journal of Clinical Investigation,* 94(6), 2493–2503

29. Lusis, A. J. (2000) Atherosclerosis. *Nature,* 407, 233–241

30. Cai, J. M., Hatsukami, T. S., Ferguson, M. S., Small, R., Polissar, N. L., & Yuan, C. (2002) Classification of Human Carotid Atherosclerotic Lesions With In Vivo Multicontrast Magnetic Resonance Imaging. *Circulation,* 106(11), 1368–1373

31. Lechareas, S., Yanni, A. E., Golemati, S., Chatziioannou, A., & Perrea, D. (2016) Ultrasound and Biochemical Diagnostic Tools for the Characterization of Vulnerable Carotid Atherosclerotic Plaque. *Ultrasound in Medicine and Biology,* 42(1), 31–43

32. Erlöv, T., Cinthio, M., Edsfeldt, A., Segstedt, S., Dias, N., Nilsson, J., & Gonçalves, I. (2016) Determining carotid plaque vulnerability using ultrasound center frequency shifts. *Atherosclerosis,* 246, 293–300

33. Joshi, A., Rienks, M., Theofilatos, K., & Mayr, M. (2020) Systems biology in cardiovascular disease: a multiomics approach. *Nature Reviews Cardiology,* 18(5), 313–330

34. Gillette, M. A., & Carr, S. A. (2013) Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. *Nature Methods,* 10(1), 28–34

35. Lubec, G., & Afjehi-Sadat, L. (2007) Limitations and pitfalls in protein identifications by mass spectrometry. *Chemical Reviews,* 107(8), 3568–3584

36. Xie, F., Liu, T., Qian, W. J., Petyuk, V. A., & Smith, R. D. (2011) Liquid chromatography-mass spectrometry-based quantitative proteomics. *Journal of Biological Chemistry,* 286(29), 25443–25449

37. Fahrner, M., Kook, L., Fröhlich, K., Biniossek, M. L., & Schilling, O. (2021) A Systematic Evaluation of Semispecific Peptide Search Parameter Enables Identification of Previously Undescribed N-Terminal Peptides and Conserved

Proteolytic Processing in Cancer Cell Lines. *Proteomes,* 9(2), 26

38. Alves, P., Arnold, R. J., Clemmer, D. E., Li, Y., Reilly, J. P., Sheng, Q., Tang, H., Xun, Z., Zeng, R., and Radivojac, P. (2008) Fast and accurate identification of semi-tryptic peptides in shotgun proteomics. *Bioinformatics,* 24(1), 102–109

39. "Case study: Taxonomic Analysis of a Tryptic Peptide", Unipept, 2020. Available from: https://unipept.ugent.be/clidocs/casestudies/tpa

40. Liu, T., Belov, M. E., Jaitly, N., Qian, W. J., & Smith, R. D. (2007) Accurate mass measurements in proteomics. *Chemical Reviews,* 107(8), 3621–3653

41. Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., et al. (2003) Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry,* 75(8), 1895–1904

42. Shi, T., Song, E., Nie, S., Rodland, K. D., Liu, T., Qian, W. J., & Smith, R. D. (2016) Advances in targeted proteomics and applications to biomedical research. *Proteomics,* 16(15-16), 2160–2182

43. Fernández-Costa, C., Martínez-Bartolomé, S., McClatchy, D. B., Saviola, A. J., Yu, N. K., & Yates, J. R. (2020) Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results. *Journal of Proteome Research,* 19(8), 3153–3161

44. Chan, S. Y., & Loscalzo, J. (2012) The Emerging Paradigm of Network Medicine in the Study of Human Disease. *Circulation Research,* 111(3), 359–374

45. Abdelsayed, M., Kort, E. J., Jovinge, S., & Mercola, M. (2022) Repurposing drugs to treat cardiovascular disease in the era of precision medicine. *Nature Reviews Cardiology,* 19(11), 751–764

46. Sonawane, A. R., Aikawa, E., & Aikawa, M. (2022) Connections for Matters of the Heart: Network Medicine in Cardiovascular Diseases. *Frontiers in Cardiovascular Medicine,* 9, 1174

47. Rader, D. J., & Daugherty, A. (2008) Translating molecular discoveries into new therapies for atherosclerosis. *Nature,* 451(7181), 904–913

48. Gander, J., Sui, X., Hazlett, L. J., Cai, B., Hébert, J. R., & Blair, S. N. (2014) Peer Reviewed: Factors Related to Coronary Heart Disease Risk Among Men: Validation of the Framingham Risk Score. *Preventing Chronic Disease,* 11

49. Ji, E., & Lee, S. (2021) Antibody-Based Therapeutics for Atherosclerosis and Cardiovascular Diseases. *International Journal of Molecular Sciences,* 22(11), 5770

50. "Cardiovascular diseases (CVDs)", WHO, 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

51. Hansson, G. K. (2005) Inflammation, Atherosclerosis, and Coronary Artery Disease. *New England Journal of Medicine,* 352(16), 1685–1695

52. Paci, P., Fiscon, G., Conte, F., Wang, R. S., Farina, L., & Loscalzo, J. (2021) Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. *NPJ Systems Biology and Applications,* 7(1), 1–11

53. Miryala, S. K., Anbarasu, A., & Ramaiah, S. (2018) Discerning molecular interactions: A comprehensive review on biomolecular interaction databases and network analysis tools. *Gene,* 642, 84–94

54. Iacobucci, I., Monaco, V., Cozzolino, F., & Monti, M. (2021) From classical to new generation approaches: An excursus of -omics methods for investigation of protein-protein interaction networks. *Journal of Proteomics,* 230, 103990

55. van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2018) Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics,* 19(4), 575–592

56. Vella, D., Zoppis, I., Mauri, G., Mauri, P., & Di Silvestre, D. (2017) From protein-

protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP Journal on Bioinformatics and Systems Biology,* 2017(1), 1–16

57. Schmidt, H. H., & Menche, J. (2022) The regulatory network architecture of cardiometabolic diseases. *Nature Genetics,* 54(1), 2–3

58. Xi, D., Zhao, J., Lai, W., & Guo, Z. (2016) Systematic analysis of the molecular mechanism underlying atherosclerosis using a text mining approach. *Human genomics,* 10(1), 1–8

59. Koplev, S., Seldin, M., Sukhavasi, K., Ermel, R., Pang, S., Zeng, L., et al. (2022) A mechanistic framework for cardiometabolic and coronary artery diseases. *Nature Cardiovascular Research,* 1(1), 85–100

60. Banik, S. K., Baishya, S., Das Talukdar, A., & Choudhury, M. D. (2022) Network analysis of atherosclerotic genes elucidates druggable targets. *BMC Medical Genomics,* 15(1), 1–12

61. Herrington, D. M., Mao, C., Parker, S. J., Fu, Z., Yu, G., Chen, L., et al. (2018) Proteomic Architecture of Human Coronary and Aortic Atherosclerosis. *Circulation,* 137(25), 2741–2756

62. Bandaru, S., Ala, C., Salimi, R., Akula, M. K., Ekstrand, M., Devarakonda, S., et al. (2019) Targeting Filamin A Reduces Macrophage Activity and Atherosclerosis. *Circulation,* 140(1), 67–79

63. Abe, J. ichi, Ko, K. A., Kotla, S., Wang, Y., Paez-Mayorga, J., Shin, I. J., et al. (2019) MAGI1 as a link between endothelial activation and ER stress drives atherosclerosis. *JCI Insight,* 4(7)

64. Hu, L., Wang, X., Huang, Y. A., Hu, P., & You, Z. H. (2021) A survey on computational models for predicting protein–protein interactions. *Briefings in Bioinformatics,* 22(5), bbab036

65. Liu, X., Hong, Z., Liu, J., Lin, Y., Rodríguez-Patón, A., Zou, Q., & Zeng, X. (2020)

Computational methods for identifying the critical nodes in biological networks. *Briefings in Bioinformatics,* 21(2), 486–497

66. Meng, X., Li, W., Peng, X., Li, Y., & Li, M. (2021) Protein interaction networks: centrality, modularity, dynamics, and applications. *Frontiers of Computer Science,* 15(6), 1–17

67. Cho, D. Y., Kim, Y. A., & Przytycka, T. M. (2012) Chapter 5: Network Biology Approach to Complex Diseases. *PLOS Computational Biology,* 8(12), e1002820

68. Saint-Antoine, M. M., & Singh, A. (2020) Network inference in systems biology: recent developments, challenges, and applications. *Current Opinion in Biotechnology,* 63, 89–98

69. Zhang, P., & Itan, Y. (2019) Biological Network Approaches and Applications in Rare Disease Studies. *Genes,* 10(10), 797

70. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research,* 49(D1), D605–D612

71. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications,* 10(1), 1–10

72. Langfelder, P., & Horvath, S. (2008) WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics,* 9(1), 1–13

73. Villaverde, A. F., Ross, J., Morán, F., & Banga, J. R. (2014) MIDER: Network Inference with Mutual Information Distance and Entropy Reduction. *PLOS ONE,* 9(5), e96732

74. Cui, Y., Leng, C., & Sun, D. (2016) Sparse estimation of high-dimensional correlation matrices. *Computational Statistics & Data Analysis,* 93, 390–403

75. Lachmann, A., Giorgi, F. M., Lopez, G., & Califano, A. (2016) ARACNe-AP: Gene

network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics,* 32(14), 2233–2235

76. Ziebarth, J. D., Bhattacharya, A., & Cui, Y. (2013) Bayesian Network Webserver: a comprehensive tool for biological network modeling. *Bioinformatics,* 29(21), 2801–2803

77. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE,* 5(9), e12776

78. Azad, A., Pavlopoulos, G. A., Ouzounis, C. A., Kyrpides, N. C., & Buluç, A. (2018) HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. *Nucleic Acids Research,* 46(6), e33–e33

79. Nepusz, T., Yu, H., & Paccanaro, A. (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods,* 9(5), 471–472

80. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research,* 42(D1), D358–D363

81. Oughtred, R., Rust, J., Chang, C., Breitkreutz, B. J., Stark, C., Willems, A., et al. (2021) The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science,* 30(1), 187–200

82. Razick, S., Magklaras, G., & Donaldson, I. M. (2008) iRefIndex: A consolidated protein interaction database with provenance. *BMC Bioinformatics,* 9(1), 1–19

83. Dimitrakopoulos, G. N., Klapa, M. I., Moschonas, N. K., & Martelli, P. L. (2021) PICKLE 3.0: enriching the human meta-database with the mouse protein interactome extended via mouse–human orthology. *Bioinformatics,* 37(1), 145–146

84. Kotlyar, M., Pastrello, C., Malik, Z., & Jurisica, I. (2019) IID 2018 update: context-

specific physical protein–protein interactions in human, model organisms and domesticated species. *Nucleic Acids Research,* 47(D1), D581–D589

85. Guirimand, T., Delmotte, S., & Navratil, V. (2015) VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Research,* 43(D1), D583–D587

86. Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., et al. (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database,* 2010

87. Alonso-López, Di., Campos-Laborie, F. J., Gutiérrez, M. A., Lambourne, L., Calderwood, M. A., Vidal, M., & De Las Rivas, J. (2019) APID database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database,* 2019

88. Gioutlakis, A., Klapa, M. I., & Moschonas, N. K. (2017) PICKLE 2.0: A human protein-protein interaction meta-database employing data integration via genetic information ontology. *PLOS ONE,* 12(10), e0186039

89. Sun, T., Zhou, B., Lai, L., & Pei, J. (2017) Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics,* 18(1), 1–8

90. Chen, Y., Wang, W., Liu, J., Feng, J., & Gong, X. (2020) Protein interface complementarity and gene duplication improve link prediction of protein-protein interaction network. *Frontiers in Genetics,* 11, 291

91. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research,* 13(11), 2498–2504

92. Bostanci, N., Selevsek, N., Wolski, W., Grossmann, J., Bao, K., Wahlander, A., et al. (2018) Targeted proteomics guided by label-free quantitative proteome analysis in saliva reveal transition signatures from health to periodontal disease. *Molecular and Cellular Proteomics,* 17(7), 1392–1409

93. Gibson, S. M., Ficklin, S. P., Isaacson, S., Luo, F., Feltus, F. A., & Smith, M. C. (2013) Massive-Scale Gene Co-Expression Network Construction and Robustness Testing Using Random Matrix Theory. *PLOS ONE,* 8(2), e55871

94. González-Domínguez, J., & Martín, M. J. (2017) Fast Parallel Construction of Correlation Similarity Matrices for Gene Co-Expression Networks on Multicore Clusters. *Procedia Computer Science,* 108, 485–494

95. Varghese, R. S., Zuo, Y., Zhao, Y., Zhang, Y. W., Jablonski, S. A., Pierobon, M., et al. (2017) Protein network construction using reverse phase protein array data. *Methods,* 124, 89–99

96. Chai, L. E., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., & Zakaria, Z. (2014) A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine,* 48, 55–65

97. Noor, A., Serpedin, E., Nounou, M., Nounou, H., Mohamed, N., & Chouchane, L. (2013) An Overview of the Statistical Methods Used for Inferring Gene Regulatory Networks and Protein-Protein Interaction Networks. *Advances in Bioinformatics,* 2013

98. Ma, S., Shah, S., Bohnert, H. J., Snyder, M., & Dinesh-Kumar, S. P. (2013) Incorporating Motif Analysis into Gene Co-expression Networks Reveals Novel Modular Expression Pattern and New Signaling Pathways. *PLOS Genetics,* 9(10), e1003840

99. Paroni, A., Graudenzi, A., Caravagna, G., Damiani, C., Mauri, G., & Antoniotti, M. (2016) CABeRNET: A Cytoscape app for augmented Boolean models of gene regulatory NETworks. *BMC Bioinformatics,* 17(1), 1–12

100. Zhang, X., Zhao, J., Hao, J. K., Zhao, X. M., & Chen, L. (2015) Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Research,* 43(5), e31–e31

101. Sella, N., Verny, L., Uguzzoni, G., Affeldt, S., & Isambert, H. (2018) MIIC online: a web server to reconstruct causal or non-causal networks from non-

perturbative data. *Bioinformatics,* 34(13), 2311–2313

102. Yan, T., Jiang, B., Fienberg, S. E., & Leng, C. (2019) Statistical Inference in a Directed Network Model With Covariates. *Journal of the American Statistical Association,* 114(526), 857–868

103. De Landtsheer, S., Lucarelli, P., & Sauter, T. (2018) Using regularization to infer cell line specificity in logical network models of signaling pathways. *Frontiers in Physiology,* 9, 550

104. Deng, Y., Zenil, H., Tegnér, J., & Kiani, N. A. (2017) HiDi: an efficient reverse engineering schema for large-scale dynamic regulatory network reconstruction using adaptive differentiation. *Bioinformatics,* 33(24), 3964–3972

105. Rubiolo, M., Milone, D. H., & Stegmayer, G. (2018) Extreme learning machines for reverse engineering of gene regulatory networks from expression time series. *Bioinformatics,* 34(7), 1253–1260

106. Ben Guebila, M., Weighill, D., Lopes-Ramos, C. M., Burkholz, R., Pop, R. T., Palepu, K., et al. (2022) An online notebook resource for reproducible inference, analysis and publication of gene regulatory networks. *Nature Methods, 2022* 19(5), 511–513

107. Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., et al. (2016) Jupyter Notebooks - a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 2016, 87–90

108. Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research,* 30(7), 1575–1584

109. Moschopoulos, C. N., Pavlopoulos, G. A., Likothanassis, S. D., & Kossida, S. (2008) An enhanced markov clustering method for detecting protein complexes. In *2008 8th IEEE International Conference on BioInformatics and BioEngineering, IEEE 2008*, 1-6

110. Theofilatos, K., Pavlopoulou, N., Papasavvas, C., Likothanassis, S., Dimitrakopoulos, C., Georgopoulos, E., et al. (2015) Predicting protein complexes from weighted protein–protein interaction graphs with a novel unsupervised methodology: Evolutionary enhanced Markov clustering. *Artificial Intelligence in Medicine,* 63(3), 181–189

111. King, A. D., Pržulj, N., & Jurisica, I. (2004) Protein complex prediction via cost-based clustering. *Bioinformatics,* 20(17), 3013–3020

112. Dimitrakopoulos, C., Theofilatos, K., Pegkas, A., Likothanassis, S., & Mavroudi, S. (2016) Predicting overlapping protein complexes from weighted protein interaction graphs by gradually expanding dense neighborhoods. *Artificial Intelligence in Medicine,* 71, 62–69

113. Bader, G. D., & Hogue, C. W. V. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics,* 4(1), 1–27

114. Koutrouli, M., Karatzas, E., Paez-Espino, D., & Pavlopoulos, G. A. (2020) A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology,* 8, 34

115. Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., & Lin, C. Y. (2014) cytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology,* 8(4), 1–7

116. Guimerà, R., & Amaral, L. A. N. (2005) Functional cartography of complex metabolic networks. *Nature,* 433(7028), 895–900

117. Han, J. D. J., Berlin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., et al. (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature,* 430(6995), 88–93

118. Paci, P., Colombo, T., Fiscon, G., Gurtner, A., Pavesi, G., & Farina, L. (2017) SWIM: a computational tool to unveiling crucial nodes in complex biological networks. *Scientific Reports,* 7(1), 1–16

119. Paci, P., & Fiscon, G. (2022) SWIMmeR: an R-based software to unveiling crucial nodes in complex biological networks. *Bioinformatics,* 38(2), 586–588

120. Muetze, T., Goenawan, I. H., Wiencko, H. L., Bernal-Llinares, M., Bryan, K., & Lynn, D. J. (2016) Contextual Hub Analysis Tool (CHAT): A Cytoscape app for identifying contextually relevant hubs in biological networks. *F1000Research,* 5

121. Kuchaiev, O., Stevanović, A., Hayes, W., & Pržulj, N. (2011) GraphCrunch 2: Software tool for network modeling, alignment and clustering. *BMC Bioinformatics,* 12(1), 1–13

122. Kuchaiev, O., Milenković, T., Memišević, V., Hayes, W., & Pržulj, N. (2010) Topological network alignment uncovers biological function and phylogeny. *Journal of The Royal Society Interface,* 7(50), 1341–1354

123. Yu, H., Zhu, X., Greenbaum, D., Karro, J., & Gerstein, M. (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Research,* 32(1), 328–337

124. Brohée, S., Faust, K., Lima-Mendez, G., Sand, O., Janky, R., Vanderstocken, G., et al. (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Research,* 36(suppl_2), W444–W451

125. Nagpal, S., Baksi, K. Das, Kuntal, B. K., & Mande, S. S. (2020) NetConfer: A web application for comparative analysis of multiple biological networks. *BMC Biology,* 18(1), 1–12

126. Jardim, V. C., De Siqueira Santos, S., Fujita, A., & Buckeridge, M. S. (2019) BioNetStat: A tool for biological networks differential analysis. *Frontiers in Genetics*, 10, 594

127. Tesson, B. M., Breitling, R., & Jansen, R. C. (2010) DiffCoEx: A simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics,* 11(1), 1–9

128. Watson, M. (2006) CoXpress: Differential co-expression in gene expression

data. *BMC Bioinformatics,* 7(1), 1–12

129. Gysi, D. M., de Miranda Fragoso, T., Zebardast, F., Bertoli, W., Busskamp, V., Almaas, E., & Nowick, K. (2020) Whole transcriptomic network analysis using Co-expression Differential Network Analysis (CoDiNA). *PLoS ONE,* 15(10), e0240523

130. Amar, D., Safer, H., & Shamir, R. (2013) Dissection of Regulatory Networks that Are Altered in Disease via Differential Co-expression. *PLOS Computational Biology,* 9(3), e1002955

131. Emilsson, V., Ilkov, M., Lamb, J. R., Finkel, N., Gudmundsson, E. F., Pitts, R., et al. (2018) Co-regulatory networks of human serum proteins link genetics to disease. *Science,* 361(6404), 769–773

132. Kamal, A. H. M., Chakrabarty, J. K., Udden, S. M. N., Zaki, M. H., & Chowdhury, S. M. (2018) Inflammatory Proteomic Network Analysis of Statin-treated and Lipopolysaccharide-activated Macrophages. *Scientific Reports,* 8(1), 1–13

133. Hartman, R. J. G., Owsiany, K., Ma, L., Koplev, S., Hao, K., Slenders, L., et al. (2021) Sex-Stratified Gene Regulatory Networks Reveal Female Key Driver Genes of Atherosclerosis Involved in Smooth Muscle Cell Phenotype Switching. *Circulation,* 143(7), 713–726

134. Parker, S. J., Chen, L., Spivia, W., Saylor, G., Mao, C., Venkatraman, V., et al. (2020) Identification of Putative Early Atherosclerosis Biomarkers by Unsupervised Deconvolution of Heterogeneous Vascular Proteomes. *Journal of Proteome Research,* 19(7), 2794–2806

135. Shao, X., Taha, I. N., Clauser, K. R., Gao, Y. (Tom), & Naba, A. (2020) MatrisomeDB: the ECM-protein knowledge database. *Nucleic Acids Research,* 48(D1), D1136–D1144

136. Basiak, M., Kosowski, M., Cyrnek, M., Bułdak, Ł., Maligłówka, M., Machnik, G., & Okopień, B. (2021) Pleiotropic Effects of PCSK-9 Inhibitors. *International Journal of Molecular Sciences,* 22(6), 3144

137. Patel, D. K., & Strong, J. (2019) The Pleiotropic Effects of Sodium-Glucose Cotransporter-2 Inhibitors: Beyond the Glycemic Benefit. *Diabetes therapy,* 10(5), 1771–1792

138. Prattichizzo, F., Giuliani, A., Mensà, E., Sabbatinelli, J., De Nigris, V., Rippo, M. R., et al. (2018) Pleiotropic effects of metformin: Shaping the microbiome to manage type 2 diabetes and postpone ageing. *Ageing Research Reviews,* 48, 87–98

139. Schooling, C. M., Zhao, J. V., Au Yeung, S. L., & Leung, G. M. (2020) Investigating pleiotropic effects of statins on ischemic heart disease in the UK biobank using mendelian randomisation. *eLife,* 9, e58567

140. Rossini, E., Biscetti, F., Rando, M. M., Nardella, E., Cecchini, A. L., Nicolazzi, M. A., et al. (2022) Statins in High Cardiovascular Risk Patients: Do Comorbidities and Characteristics Matter? *International Journal of Molecular Sciences,* 23(16), 9326

141. Sever, P. S., Dahlöf, B., Poulter, N. R., Wedel, H., Beevers, G., Caulfield, M., et al. (2001) Rationale, design, methods and baseline demography of participants of the Anglo-Scandinavian Cardiac Outcomes Trial. *Journal of Hypertension,* 19(6), 1139–1147

142. Willeit, P., Skroblin, P., Moschen, A. R., Yin, X., Kaudewitz, D., Zampetaki, A., et al. (2017) Circulating MicroRNA-122 Is Associated With the Risk of New-Onset Metabolic Syndrome and Type 2 Diabetes. *Diabetes,* 66(2), 347–357

143. Wolska, A., Yang, Z. H., & Remaley, A. T. (2020) Hypertriglyceridemia - new approaches in management and treatment. *Current opinion in lipidology,* 31(6), 331

144. Iwata, H., Goettsch, C., Sharma, A., Ricchiuto, P., Goh, W. W. Bin, Halu, A., et al. (2016) PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. *Nature Communications,* 7(1), 1–19

145. Lempiäinen, H., Brænne, I., Michoel, T., Tragante, V., Vilne, B., Webb, T. R., et

al. (2018) Network analysis of coronary artery disease risk genes elucidates disease mechanisms and druggable targets. *Scientific Reports,* 8(1), 1–14

146. Huan, T., Zhang, B., Wang, Z., Joehanes, R., Zhu, J., Johnson, A. D., et al. (2013) A systems biology framework identifies molecular underpinnings of coronary heart disease. *Arteriosclerosis, Thrombosis, and Vascular Biology,* 33(6), 1427–1434

147. Nakano, T., Katsuki, S., Chen, M., Decano, J. L., Halu, A., Lee, L. H., et al. (2019) Uremic Toxin Indoxyl Sulfate Promotes Proinflammatory Macrophage Activation Via the Interplay of OATP2B1 and Dll4-Notch Signaling. *Circulation,* 139(1), 78–96

148. Xiao, J., Li, F., Yang, Q., Zeng, X. F., & Ke, Z. P. (2020) Co-expression analysis provides important module and pathways of human dilated cardiomyopathy. *Journal of Cellular Physiology,* 235(1), 494–503

149. Schlotter, F., Halu, A., Goto, S., Blaser, M. C., Body, S. C., Lee, L. H., Higashi, H., Delaughter, D. M., Hutcheson, J. D., Vyas, P., Pham, T., Rogers, M. A., Sharma, A., Seidman, C. E., Loscalzo, J., Seidman, J. G., Aikawa, M., Singh, S. A., and Aikawa, E. (2018) Spatiotemporal Multi-Omics Mapping Generates a Molecular Atlas of the Aortic Valve and Reveals Networks Driving Disease. *Circulation,* 138, 377–393

150. Lee, L. Y., Pandey, A. K., Maron, B. A., & Loscalzo, J. (2021) Network medicine in Cardiovascular Research. *Cardiovascular Research* 117(10), 2186–2202

151. Doran, S., Arif, M., Lam, S., Bayraktar, A., Turkez, H., Uhlen, M., et al. (2021) Multi-omics approaches for revealing the complexity of cardiovascular disease. *Briefings in Bioinformatics* 22(5), bbab061

152. Maron, B. A., Wang, R. S., Shevtsov, S., Drakos, S. G., Arons, E., Wever-Pinzon, O., et al. (2021) Individualized interactomes for network-based precision medicine in hypertrophic cardiomyopathy with implications for other clinical pathophenotypes. *Nature Communications,* 12(1), 1–11

153. Barallobre-Barreiro, J., Radovits, T., Fava, M., Mayr, U., Lin, W. Y., Ermolaeva, E., et al. (2021) Extracellular Matrix in Heart Failure: Role of ADAMTS5 in Proteoglycan Remodeling. *Circulation,* 144(25), 2021–2034

154. Khosravi, P., Gazestani, V. H., Pirhaji, L., Law, B., Sadeghi, M., Goliaei, B., & Bader, G. D. (2015) Inferring interaction type in gene regulatory networks using co-expression data. *Algorithms for Molecular Biology,* 10(1), 1–11

155. Hocker, J. D., Poirion, O. B., Zhu, F., Buchanan, J., Zhang, K., Chiou, J., et al. (2021) Cardiac cell type-specific gene regulatory programs and disease risk association. *Science Advances,* 7(20), eabf1444

156. Li, G., Luan, C., Dong, Y., Xie, Y., Zentz, S. C., Zelt, R., et al. (2021) ExpressHeart: Web Portal to Visualize Transcriptome Profiles of Non-Cardiomyocyte Cells. *International Journal of Molecular Sciences,* 22(16), 8943

157. Ma, W. F., Turner, A. W., Gancayco, C., Wong, D., Song, Y., Mosquera, J. V., et al. (2022) PlaqView 2.0: A comprehensive web portal for cardiovascular single-cell genomics. *Frontiers in Cardiovascular Medicine,* 9

158. Pan, H., Xue, C., Auerbach, B. J., Fan, J., Bashore, A. C., Cui, J., et al. (2020) Single-Cell Genomics Reveals a Novel Cell State During Smooth Muscle Cell Phenotypic Switching and Potential Therapeutic Targets for Atherosclerosis in Mouse and Human. *Circulation,* 142(21), 2060–2075

159. Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., et al. (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology,* 20(2), 163–172

160. Yao, C., Chen, G., Song, C., Keefe, J., Mendelson, M., Huan, T., et al. (2018) Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications,* 9(1), 1–11

161. Zhang, J., Dutta, D., Kottgen, A., Tin, A., Schlosser, P., Grams, M. E., et al. (2022) Plasma proteome analyses in individuals of European and African ancestry

identify cis-pQTLs and models for proteome-wide association studies. *Nature Genetics*, 54(5), 593-602

162. Jin, H., Goossens, P., Juhasz, P., Eijgelaar, W., Manca, M., Karel, J. M.,et al. (2021) Integrative multiomics analysis of human atherosclerosis reveals a serum response factor-driven network associated with intraplaque hemorrhage. *Clinical and Translational Medicine,* 11(6), e458

163. Marx, V. (2019) A dream of single-cell proteomics. *Nature Methods,* 16(9), 809–812

164. Jeub, L. G. S., Sporns, O., & Fortunato, S. (2018) Multiresolution Consensus Clustering in Networks. *Scientific Reports,* 8(1), 1–16

165. Gao, C., Huang, Q., Liu, C., Kwong, C. H., Yue, L., Wan, J. B., et al. (2020) Treatment of atherosclerosis by macrophage-biomimetic nanoparticles via targeted pharmacotherapy and sequestration of proinflammatory cytokines. *Nature Communications,* 11(1), 1–14

166. Theofilatos, K., Dimitrakopoulos, C., Alexakos, C., Korfiati, A., Likothanassis, S., & Mavroudi, S. (2016) InSyBio BioNets: an efficient tool for network-based biomarker discovery. *EMBnet journal,* 22, 871

167. Wang, Y. X. R., & Huang, H. (2014) Review on statistical methods for gene network reconstruction using expression data. *Journal of Theoretical Biology,* 362, 53–61

168. Huang, X., Lin, X., Zeng, J., Wang, L., Yin, P., Zhou, L., et al. (2017) A Computational Method of Defining Potential Biomarkers based on Differential Sub-Networks. *Scientific Reports,* 7(1), 1–10

169. del Sol, A., Balling, R., Hood, L., & Galas, D. (2010) Diseases as network perturbations. *Current Opinion in Biotechnology,* 21(4), 566–571

170. Tzfadia, O., Diels, T., De Meyer, S., Vandepoele, K., Aharoni, A., & Van De Peer, Y. (2016) CoExpNetViz: Comparative co-expression networks construction and

visualization tool. *Frontiers in Plant Science,* 6, 1194

171. MacNeil, L. T., & Walhout, A. J. (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research,* 21(5), 645–657

172. Benjamini, Y., & Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological),* 57(1), 289–300

173. Stolovitzky, G., Prill, R. J., & Califano, A. (2009) Lessons from the DREAM2 Challenges. *Annals of the New York Academy of Sciences,* 1158(1), 159–195

174. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols,* 4(1), 44–57

175. Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012) Wisdom of crowds for robust gene network inference. *Nature Methods,* 9(8), 796–804

176. Schaffter, T., Marbach, D., & Floreano, D. (2011) GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics,* 27(16), 2263–2270

177. Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., et al. (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Research,* 39(suppl_1), D1005–D1010

178. Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003) A comparison of normalization methods for high densityoligonucleotide array data based on variance and bias. *Bioinformatics,* 19(2), 185–193

179. Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spinola, M. I., Contreras-Moreira, B., et al. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active

(experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Research,* 36(suppl_1), D120–D124

180. Zhu, C., Byers, K. J. R. P., McCord, R. P., Shi, Z., Berger, M. F., Newburger, D. E., et al. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Research,* 19(4), 556–566

181. MacIsaac, K. D., Wang, T., Gordon, D. B., Gifford, D. K., Stormo, G. D., & Fraenkel, E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics,* 7(1), 1–14

182. Didangelos, A., Yin, X., Mandal, K., Saje, A., Smith, A., Xu, Q., et al. (2011) Extracellular matrix composition and remodeling in human abdominal aortic aneurysms: A proteomics approach. *Molecular and Cellular Proteomics,* 10(8)

183. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research,* 43(7), e47–e47

184. Theofilatos, K. A., Likothanassis, S., & Mavroudi, S. (2015) Quo vadis1 computational analysis of PPI data or why the future isn't here yet. *Frontiers in Genetics,* 6, 289

185. Jaberi, N., Soleimani, A., Pashirzad, M., Abdeahad, H., Mohammadi, F., Khoshakhlagh, M., et al. (2019) Role of thrombin in the pathogenesis of atherosclerosis. *Journal of Cellular Biochemistry,* 120(4), 4757–4765

186. Esmon, C. T. (2013) Targeting factor Xa and thrombin: Impact on coagulation and beyond. *Thrombosis and Haemostasis,* 111(04), 625–633

187. Ngo, A. T. P., Jordan, K. R., Mueller, P. A., Hagen, M. W., Reitsma, S. E., Puy, C., et al. (2021) Pharmacological targeting of coagulation factor XI mitigates the development of experimental atherosclerosis in low-density lipoprotein receptor-deficient mice. *Journal of Thrombosis and Haemostasis,* 19(4), 1001–1017

188. Singhal, A., Chauhan, A., Goyal, P., & Taneja, A. (2021) Transthyretin - A Novel Biomarker for Insulin Resistance and Atherosclerosis Risk in Prediabetics. *The Journal of the Association of Physicians of India,* 69(11), 11–12

189. Ambrosius, W., Michalak, S., Rosinska, J., Lukasik, M., & Kozubski, W. (2018) Transthyretin levels negatively correlate with carotid carotid intima-media thickness and beta-stiffness index. *Neurology,* 90(15 supplement), P5.219

190. Borràs, E., & Sabidó, E. (2017) What is targeted proteomics? A concise revision of targeted acquisition and targeted data analysis in mass spectrometry. *Proteomics,* 17(17-18), 1700180

191. MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., et al. (2010) Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics,* 26(7), 966–968

192. Zauber, H., Kirchner, M., & Selbach, M. (2018) Picky: A simple online PRM and SRM method designer for targeted proteomics. *Nature Methods,* 15(3), 156–157

193. Röst, H. L., Liu, Y., D'Agostino, G., Zanella, M., Navarro, P., Rosenberger, G., et al. (2016) TRIC: An automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nature Methods,* 13(9), 777–783

194. Toghi Eshghi, S., Auger, P., & Mathews, W. R. (2018) Quality assessment and interference detection in targeted mass spectrometry data using machine learning 03 Chemical Sciences 0301 Analytical Chemistry. *Clinical Proteomics,* 15(1), 33

195. Cox, J., & Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology,* 26(12), 1367–1372

196. Martinez-Val, A., Bekker-Jensen, D. B., Hogrebe, A., & Olsen, J. V. (2021) Data Processing and Analysis for DIA-Based Phosphoproteomics Using Spectronaut. *Methods in Molecular Biology*, 2361, 95–107

197. Zhang, F., Ge, W., Ruan, G., Cai, X., & Guo, T. (2020) Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020. *Proteomics,* 20(17-18), 1900276

198. Chiva, C., Olivella, R., Borràs, E., Espadas, G., Pastor, O., Solé, A., & Sabidó, E. (2018) QCloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories. *PLOS ONE,* 13(1), e0189209

199. Vaca Jacome, A. S., Peckner, R., Shulman, N., Krug, K., DeRuff, K. C., Officer, A., et al. (2020) Avant-garde: an automated data-driven DIA data curation tool. *Nature Methods,* 17(12), 1237–1244

200. Kawashima, Y., Watanabe, E., Umeyama, T., Nakajima, D., Hattori, M., Honda, K., & Ohara, O. (2019) Optimization of Data-Independent Acquisition Mass Spectrometry for Deep and Highly Sensitive Proteomic Analysis. *International Journal of Molecular Sciences,* 20(23), 5932

201. Dogu, E., Taheri, S. M., Olivella, R., Marty, F., Lienert, I., Reiter, L., et al. (2018) MSstatsQC 2.0: R/Bioconductor Package for Statistical Quality Control of Mass Spectrometry-Based Proteomics Experiments. *Journal of Proteome Research,* 18(2), 678–686

202. Toussi, C. A., & Haddadnia, J. (2019) Improving protein secondary structure prediction: the evolutionary optimized classification algorithms. *Structural Chemistry,* 30(4), 1257–1266

203. Zhou, C., Hou, C., Wei, X., & Zhang, Q. (2014) Improved hybrid optimization algorithm for 3D protein structure prediction. *Journal of Molecular Modeling,* 20(7), 1–12

204. McAllister, S. R., & Floudas, C. A. (2009) An improved hybrid global optimization method for protein tertiary structure prediction. *Computational Optimization and Applications,* 45(2), 377–413

205. Reynès, C., Sabatier, R., Molinari, N., & Lehmann, S. (2008) A new genetic algorithm in proteomics: Feature selection for SELDI-TOF data. *Computational*

Statistics & Data Analysis, 52(9), 4380–4394

206. Pirhadi, S., Maghooli, K., Moteghaed, N., Garshasbi, M., & Mousavirad, S. (2021) Biomarker discovery by imperialist competitive algorithm in mass spectrometry data for ovarian cancer prediction. *Journal of Medical Signals and Sensors,* 11(2), 108–119

207. Carreno, J. F., & Qiu, P. (2020) Feature selection algorithms for predicting preeclampsia: A comparative approach. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2626–2631

208. Corthésy, J., Theofilatos, K., Mavroudi, S., Macron, C., Cominetti, O., Remlawi, M., et al. (2018) An Adaptive Pipeline To Maximize Isobaric Tagging Data in Large-Scale MS-Based Proteomics. *Journal of Proteome Research,* 17(6), 2165–2173

209. Ji, J., Xiao, H., & Yang, C. (2020) HFADE-FMD: a hybrid approach of fireworks algorithm and differential evolution strategies for functional module detection in protein-protein interaction networks. *Applied Intelligence,* 51(2), 1118–1132

210. Jaffe, J. D., Feeney, C. M., Patel, J., Lu, X., & Mani, D. R. (2016) Transitioning from Targeted to Comprehensive Mass Spectrometry Using Genetic Algorithms. *Journal of the American Society for Mass Spectrometry,* 27(11), 1745–1751

211. Martin Bland, J., & Altman, D. G. (1986) Statistical Methods for Assessing Agreement between two Methods of Clinical Measurement. *The Lancet,* 327(8476), 307–310

212. Broudy, D., Killeen, T., Choi, M., Shulman, N., Mani, D. R., Abbatiello, S. E., et al. (2014) A framework for installable external tools in Skyline. *Bioinformatics,* 30(17), 2521–2523

213. Sharma, V., Eckels, J., Taylor, G. K., Shulman, N. J., Stergachis, A. B., Joyner, S. A., et al. (2014) Panorama: A Targeted Proteomics Knowledge Base. *Journal of Proteome Research,* 13(9), 4205–4210

214. Sharma, V., Eckels, J., Schilling, B., Ludwig, C., Jaffe, J. D., MacCoss, M. J., & MacLean, B. (2018) Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline. *Molecular & Cellular Proteomics,* 17(6), 1239–1244

215. Pino, L. K., Searle, B. C., Bollinger, J. G., Nunn, B., MacLean, B., & MacCoss, M. J. (2020) The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrometry Reviews,* 39(3), 229–244

216. Krokhin, O. V., Craig, R., Spicer, V., Ens, W., Standing, K. G., Beavis, R. C., & Wilkins, J. A. (2004) An Improved Model for Prediction of Retention Times of Tryptic Peptides in Ion Pair Reversed-phase HPLC: Its Application to Protein Peptide Mapping by Off-Line HPLC-MALDI MS. *Molecular & Cellular Proteomics,* 3(9), 908–919

217. Escher, C., Reiter, L., Maclean, B., Ossola, R., Herzog, F., Chilton, J., et al. (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics,* 12(8), 1111–1121

218. Searle, B. C., Egertson, J. D., Bollinger, J. G., Stergachis, A. B., & MacCoss, M. J. (2015) Using Data Independent Acquisition (DIA) to Model High-responding Peptides for Targeted Proteomics Experiments. *Molecular & Cellular Proteomics,* 14(9), 2331–2340

219. Bereman, M. S., Johnson, R., Bollinger, J., Boss, Y., Shulman, N., MacLean, B., et al. (2014) Implementation of Statistical Process Control for Proteomic Experiments Via LC MS/MS. *Journal of The American Society for Mass Spectrometry,* 25(4), 581–587

220. Reiter, L., Rinner, O., Picotti, P., Hüttenhain, R., Beck, M., Brusniak, M.Y., et al. (2011) mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature Methods,* 8(5), 430–435

221. Gandhi, T., Bernhardt, O., Müller, S., Leinert, I., Verbeke, L., Bober, M., et al. (2019) Boosting PRM based Targeted Proteomics using SpectroDive. *Target (Light),* 2, 08

222. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., & Ralser, M. (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods,* 17(1), 41–44

223. Dupuis, N., Muller, S., Treiber, T., & Escher, C. (2019) Evaluation of PQ500, a 500-plasma protein blood panel in NSCLC subjects using high-throughput MRM mass spectrometry. *Journal of Clinical Oncology,* 37(8_suppl), 110–110

224. Gutmann, C., Takov, K., Burnap, S. A., Singh, B., Ali, H., Theofilatos, K., et al. (2021) SARS-CoV-2 RNAemia and proteomic trajectories inform prognostication in COVID-19 patients admitted to intensive care. *Nature Communications,* 12(1), 1–17

225. Zhang, B., Whiteaker, J. R., Hoofnagle, A. N., Baird, G. S., Rodland, K. D., & Paulovich, A. G. (2018) Clinical potential of mass spectrometry-based proteogenomics. *Nature Reviews Clinical Oncology,* 16(4), 256–268

226. Joshi, A., & Mayr, M. (2018) In Aptamers They Trust: Caveats of the SOMAscan Biomarker Discovery Platform From SomaLogic. *Circulation,* 138(22), 2482–2485

227. Calderón-Celis, F., Encinar, J. R., & Sanz-Medel, A. (2018) Standardization approaches in absolute quantitative proteomics with mass spectrometry. *Mass Spectrometry Reviews,* 37(6), 715–737

228. Cáceres Sepúlveda, G., Ochoa, S., & Thibault, J. (2022) Pareto domain: An invaluable source of process information. *Chemical Product and Process Modeling,* 17(1), 29–53

229. Kamstrup, P. R., Tybjærg-Hansen, A., & Nordestgaard, B. G. (2013) Extreme Lipoprotein(a) Levels and Improved Cardiovascular Risk Prediction. *Journal of the American College of Cardiology,* 61(11), 1146–1156

230. Pechlaner, R., Tsimikas, S., Yin, X., Willeit, P., Baig, F., Santer, P., et al. (2017) Very-Low-Density Lipoprotein–Associated Apolipoproteins Predict Cardiovascular Events and Are Lowered by Inhibition of APOC-III. *Journal of the*

American College of Cardiology, 69(7), 789–800

231. Wei, W. Q., Li, X., Feng, Q., Kubo, M., Kullo, I. J., Peissig, P. L., et al. (2018) LPA variants are associated with residual cardiovascular risk in patients receiving statins. *Circulation,* 138(17), 1839–1849

232. Bäck, M., Yurdagul, A., Tabas, I., Öörni, K., & Kovanen, P. T. (2019) Inflammation and its resolution in atherosclerosis: mediators and therapeutic opportunities. *Nature Reviews Cardiology,* 16(7), 389–406

233. Aragonès, G., Auguet, T., Guiu-Jurado, E., Berlanga, A., Curriu, M., Martinez, S., et al. (2016) Proteomic Profile of Unstable Atheroma Plaque: Increased Neutrophil Defensin 1, Clusterin, and Apolipoprotein e Levels in Carotid Secretome. *Journal of Proteome Research,* 15(3), 933–944

234. Koller, L., Richter, B., Winter, M. P., Sulzgruber, P., Potolidis, C., Liebhart, F., et al. (2017) Clusterin/apolipoprotein J is independently associated with survival in patients with chronic heart failure. *Journal of Clinical Lipidology,* 11(1), 178–184

235. Wang, D., Wang, Z., Zhang, L., & Wang, Y. (2017) Roles of Cells from the Arterial Vessel Wall in Atherosclerosis. *Mediators of Inflammation,* 2017

236. Holm Nielsen, S., Jonasson, L., Kalogeropoulos, K., Karsdal, M. A., Reese-Petersen, A. L., auf dem Keller, U., et al. (2020) Exploring the role of extracellular matrix proteins to develop biomarkers of plaque vulnerability and outcome. *Journal of Internal Medicine*, 287(5), 493-513

237. Olink Proteomics. Available from: https://www.olink.com/

238. Hellings, W. E., Moll, F. L., Kleijn, D. P. V. de, & Pasterkamp, G. (2012) 10-years experience with the Athero-Express study. *Cardiovascular Diagnosis and Therapy,* 2(1), 63

239. Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and

database search. *Analytical Chemistry,* 74(20), 5383–5392

240. Shevchenko, A., Wilm, M., Vorm, O., & Mann, M. (1996) Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Analytical Chemistry,* 68(5), 850–858

241. Nesvizhskii, A. I., Keller, A., Kolker, E., & Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry,* 75(17), 4646–4658

242. Kusebauch, U., Campbell, D. S., Deutsch, E. W., Chu, C. S., Spicer, D. A., Brusniak, M. Y., et al. (2016) Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell,* 166(3), 766–778

243. Lundberg, M., Eriksson, A., Tran, B., Assarsson, E., & Fredriksson, S. (2011) Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids Research,* 39(15), e102–e102

244. Orsburn, B. C. (2021) Proteome Discoverer—A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes,* 9(1), 15

245. Bern, M., Kil, Y. J., & Becker, C. (2012) Byonic: Advanced peptide and protein identification software. *Current Protocols in Bioinformatics*, 40(1), 13-20

246. Bern, M., Cai, Y., & Goldberg, D. (2007) Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Analytical chemistry,* 79(4), 1393–1400

247. Slenders, L., Landsmeer, L. P. L., Cui, K., Depuydt, M. A. C., Verwer, M., Mekke, J., et al. (2022) Intersecting single-cell transcriptomics and genome-wide association studies identifies crucial cell populations and candidate genes for atherosclerosis. *European Heart Journal Open,* 2(1), oeab043

248. "What is Space Ranger?", Space Ranger, Support - Spatial Gene Expression - Software, 10x Genomics. Available from:

https://support.10xgenomics.com/spatial-gene-expression/software/pipelines/latest/what-is-space-ranger

249. "What is Loupe Browser?", Loupe Browser, Support - Single Cell Gene Expression - Software, 10x Genomics. Available from: https://support.10xgenomics.com/single-cell-gene-expression/software/visualization/latest/what-is-loupe-cell-browser

250. Prism, GraphPad Software, San Diego, California USA. Available from: https://www.graphpad.com/

251. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods,* 17(3), 261–272

252. Haw, R., & Stein, L. (2012) Using the reactome database. *Current Protocols in Bioinformatics,* 38(1), 8-7

253. Kanehisa, M., & Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research,* 28(1), 27–30

254. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000) Gene Ontology: tool for the unification of biology. *Nature genetics,* 25(1), 25-29

255. Pedregosa F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011) Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. *Journal of Machine Learning Research,* 12, 2825–2830

256. Smola, A. J., & Schölkopf, B. (2004) A tutorial on support vector regression. *Statistics and Computing,* 14(3), 199–222

257. Owens, G. K., & Pasterkamp, G. (2019) PlaqOmics Leducq Fondation Trans-Atlantic Network: Defining the Roles of Smooth Muscle Cells and Other

Extracellular Matrix-Producing Cells in Late-Stage Atherosclerotic Plaque Pathogenesis. *Circulation Research,* 124(9), 1297–1299

258. Ankney, J. A., Muneer, A., & Chen, X. (2018) Relative and Absolute Quantitation in Mass Spectrometry–Based Proteomics. *Annual review of analytical chemistry,* 11, 49–77

259. Oliveros, J.C. (2007-2015) Venny. An interactive tool for comparing lists with Venn diagrams. Available from: *http://bioinfogp.cnb.csic.es/tools/venny/index.html*

260. Lepedda, A. J., Cigliano, A., Cherchi, G. M., Spirito, R., Maggioni, M., Carta, F., et al. (2009) A proteomic approach to differentiate histologically classified stable and unstable plaques from human carotid arteries. *Atherosclerosis,* 203(1), 112–118

261. Olson, F. J., Sihlbom, C., Davidsson, P., Hulthe, J., Fagerberg, B., & Bergström, G. (2010) Consistent differences in protein distribution along the longitudinal axis in symptomatic carotid atherosclerotic plaques. *Biochemical and Biophysical Research Communications,* 401(4), 574–580

262. Rocchiccioli, S., Pelosi, G., Rosini, S., Marconi, M., Viglione, F., Citti, L., et al. (2013) Secreted proteins from carotid endarterectomy: An untargeted approach to disclose molecular clues of plaque progression. *Journal of Translational Medicine,* 11(1), 1–15

263. Malaud, E., Merle, D., Piquer, D., Molina, L., Salvetat, N., Rubrecht, L., et al. (2014) Local carotid atherosclerotic plaque proteins for the identification of circulating biomarkers in coronary patients. *Atherosclerosis,* 233(2), 551–558

264. Ucciferri, N., Rocchiccioli, S., Comelli, L., Marconi, M., Ferrari, M., Pelosi, G., & Cecchettini, A. (2017) Extracellular matrix characterization in plaques from carotid endarterectomy by a proteomics approach. *Talanta,* 174, 341–346

265. Hansmeier, N., Buttigieg, J., Kumar, P., Pelle, S., Choi, K. Y., Kopriva, D., & Chao, T. C. (2018) Identification of Mature Atherosclerotic Plaque Proteome

Signatures Using Data-Independent Acquisition Mass Spectrometry. *Journal of Proteome Research,* 17(1), 164–176

266. Ward, L. J., Olausson, P., Li, W., & Yuan, X. M. (2018) Proteomics and multivariate modelling reveal sex-specific alterations in distinct regions of human carotid atheroma. *Biology of Sex Differences,* 9(1), 1–12

267. Nehme, A., Kobeissy, F., Zhao, J., Zhu, R., Feugier, P., Mechref, Y., & Zibara, K. (2019) Functional pathways associated with human carotid atheroma: a proteomics analysis. *Hypertension Research,* 42(3), 362–373

268. Lynch, M., Barallobre-Barreiro, J., Jahangiri, M., & Mayr, M. (2016) Vascular proteomics in metabolic and cardiovascular diseases. *Journal of Internal Medicine,* 280(4), 325–338

269. Gialeli, C., Shami, A., & Gonçalves, I. (2021) Extracellular matrix: paving the way to the newest trends in atherosclerosis. *Current Opinion in Lipidology,* 32(5), 277

270. Ngai, D., Lino, M., & Bendeck, M. P. (2018) Cell-Matrix Interactions and Matricrine Signaling in the Pathogenesis of Vascular Calcification. *Frontiers in Cardiovascular Medicine,* 5, 174

271. "Series GSE104140", NCBI - GEO - Accession Display. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104140

272. Mahmoud, A. D., Ballantyne, M. D., Miscianinov, V., Pinel, K., Hung, J., Scanlon, J. P., et al. (2019) The Human-Specific and Smooth Muscle Cell-Enriched LncRNA SMILR Promotes Proliferation by Regulating Mitotic CENPF mRNA and Drives Cell-Cycle Progression Which Can Be Targeted to Limit Vascular Remodeling. *Circulation Research,* 125(5), 535–551

273. Dweck, M. R., Aikawa, E., Newby, D. E., Tarkin, J. M., Rudd, J. H. F., Narula, J., & Fayad, Z. A. (2016) Noninvasive Molecular Imaging of Disease Activity in Atherosclerosis. *Circulation Research,* 119(2), 330–340

274. Rodríguez-Manzaneque, J. C., Fernández-Rodríguez, R., Rodríguez-Baena, F. J., and Iruela-Arispe, M. L. (2015) ADAMTS proteases in vascular biology. *Matrix Biology,* 44, 38–45

275. Liang, W., Ward, L. J., Karlsson, H., Ljunggren, S. A., Li, W., Lindahl, M., & Yuan, X. M. (2016) Distinctive proteomic profiles among different regions of human carotid plaques in men and women. *Scientific Reports,* 6(1), 1–10

276. Tomas, L., Edsfeldt, A., Mollet, I. G., Matic, L. P., Prehn, C., Adamski, J., et al. (2018) Altered metabolism distinguishes high-risk from stable carotid atherosclerotic plaques. *European heart journal,* 39(24), 2301–2310

277. Tabas, I., & Lichtman, A. H. (2017) Monocyte-Macrophages and T Cells in Atherosclerosis. *Immunity,* 47(4), 621–634

278. Gallego Romero, I., Pai, A. A., Tung, J., & Gilad, Y. (2014) RNA-seq: Impact of RNA degradation on transcript quantification. *BMC Biology,* 12(1), 1–13

279. Aratani, Y. (2018) Myeloperoxidase: Its role for host defense, inflammation, and neutrophil function. *Archives of Biochemistry and Biophysics,* 640, 47–52

280. Viegas, C. S. B., Rafael, M. S., Enriquez, J. L., Teixeira, A., Vitorino, R., Luís, I. M., et al. (2015) Gla-rich protein acts as a calcification inhibitor in the human cardiovascular system. *Arteriosclerosis, Thrombosis, and Vascular Biology,* 35(2), 399–408

281. Shanahan, C. M., Cary, N. R. B., Metcalfe, J. C., & Weissberg, P. L. (1994) High expression of genes for calcification-regulating proteins in human atherosclerotic plaques. *Journal of Clinical Investigation,* 93(6), 2393–2402

282. Dhore, C. R., Cleutjens, J. P., Lutgens, E., Cleutjens, K. B., Geusens, P. P., Kitslaar, P. J., et al. (2001) Differential expression of bone matrix regulatory proteins in human atherosclerotic plaques. *Arteriosclerosis, Thrombosis, and Vascular Biology,* 21(12), 1998–2003

283. Kapustin, A. N., Schoppet, M., Schurgers, L. J., Reynolds, J. L., McNair, R., Heiss, A., et al. (2017) Prothrombin Loading of Vascular Smooth Muscle Cell-Derived Exosomes Regulates Coagulation and Calcification. *Arteriosclerosis, Thrombosis, and Vascular Biology,* 37(3), e22–e32

284. Jahnen-Dechent, W., Heiss, A., Schäfer, C., & Ketteler, M. (2011) Fetuin-A regulation of calcified matrix metabolism. *Circulation Research,* 108(12), 1494–1509

285. Karlöf, E., Seime, T., Dias, N., Lengquist, M., Witasp, A., Almqvist, H., et al. (2019) Correlation of computed tomography with carotid plaque transcriptomes associates calcification with lesion-stabilization. *Atherosclerosis,* 288, 175–185

286. Basatemur, G. L., Jørgensen, H. F., Clarke, M. C., Bennett, M. R., & Mallat, Z. (2019) Vascular smooth muscle cells in atherosclerosis. *Nature Reviews Cardiology,* 16(12), 727–744

287. Flynn, M. C., Pernes, G., Lee, M. K. S., Nagareddy, P. R., & Murphy, A. J. (2019) Monocytes, Macrophages, and Metabolic Disease in Atherosclerosis. *Frontiers in Pharmacology,* 10, 666

288. Packard, R. R. S., Lichtman, A. H., & Libby, P. (2009) Innate and adaptive immunity in atherosclerosis. *Seminars in Immunopathology,* 31(1), 5–22

289. Sulkava, M., Raitoharju, E., Levula, M., Seppälä, I., Lyytikaïnen, L. P., Mennander, A., et al. (2017) Differentially expressed genes and canonical pathway expression in human atherosclerotic plaques-Tampere Vascular Study. *Scientific Reports,* 7(1), 1-10

290. Weiss-Sadan, T., Gotsman, I., & Blum, G. (2017) Cysteine proteases in atherosclerosis. *FEBS Journal,* 284(10), 1455–1472

291. Fasehee, H., Fakhraee, M., Davoudi, S., Vali, H., & Faghihi, S. (2019) Cancer biomarkers in atherosclerotic plaque: Evidenced from structural and proteomic analyses. *Biochemical and Biophysical Research Communications,* 509(3), 687–693

292. Mohammadpour, A. H., Salehinejad, Z., Elyasi, S., Mouhebati, M., Mirhafez, S. R., Samadi, S., et al. (2018) Evaluation of serum cathepsin D concentrations in coronary artery disease. *Indian Heart Journal,* 70(4), 471–475

293. Ionita, M. G., Van Den Borne, P., Catanzariti, L. M., Moll, F. L., De Vries, J. P. P., Pasterkamp, G., et al. (2010) High neutrophil numbers in human carotid atherosclerotic plaques are associated with characteristics of rupture-prone lesions. *Arteriosclerosis, Thrombosis, and Vascular Biology,* 30(9), 1842–1848

294. Abd-Elrahman, I., Meir, K., Kosuge, H., Ben-Nun, Y., Sadan, T. W., Rubinstein, C., et al. (2016) Characterizing cathepsin activity and macrophage subtypes in excised human carotid plaques. *Stroke,* 47(4), 1101–1108

295. Zhang, R., Brennan, M. L., Fu, X., Aviles, R. J., Pearce, G. L., Penn, M. S., et al. (2001) Association Between Myeloperoxidase Levels and Risk of Coronary Artery Disease. *JAMA,* 286(17), 2136–2142

296. Kiechl, S., Willeit, J., Egger, G., Poewe, W., & Oberhollenzer, F. (1997) Body Iron Stores and the Risk of Carotid Atherosclerosis. *Circulation,* 96(10), 3300–3307

297. Tsimikas, S., Willeit, J., Knoflach, M., Mayr, M., Egger, G., Notdurfter, M., et al. (2009) Lipoprotein-associated phospholipase A2 activity, ferritin levels, metabolic syndrome, and 10-year cardiovascular and non-cardiovascular mortality: results from the Bruneck study. *European Heart Journal,* 30(1), 107–115

298. de Bakker, M., Timmerman, N., van Koeverden, I. D., de Kleijn, D. P. V., de Borst, G. J., Pasterkamp, G., et al. (2020) The age- and sex-specific composition of atherosclerotic plaques in vascular surgery patients. *Atherosclerosis,* 310, 1–10

299. Aboyans, V., Ricco, J. B., Bartelink, M. E. L., Björck, M., Brodmann, M., Cohnert, T., et al. (2018) 2017 ESC Guidelines on the Diagnosis and Treatment of Peripheral Arterial Diseases, in collaboration with the European Society for Vascular Surgery (ESVS): Document covering atherosclerotic disease of extracranial carotid and vertebral, mesenteric, renal, upper and lower

extremity arteries Endorsed by: the European Stroke Organization (ESO)The Task Force for the Diagnosis and Treatment of Peripheral Arterial Diseases of the European Society of Cardiology (ESC) and of the European Society for Vascular Surgery (ESVS). *European Heart Journal,* 39(9), 763-816

300. Arbab-Zadeh, A., & Fuster, V. (2015) The myth of the "vulnerable plaque": Transitioning from a focus on individual lesions to atherosclerotic disease burden for coronary artery disease risk assessment. *Journal of the American College of Cardiology,* 65(8), 846–855

301. Wardlaw, J. M., Chappell, F. M., Stevenson, M., De Nigris, E., Thomas, S., Gillard, J., et al. (2006) Accurate, practical and cost-effective assessment of carotid stenosis in the UK. *Health Technology Assessment (Winchester, England),* 10(30), iii–iv

302. Palomba, A., Abbondio, M., Fiorito, G., Uzzau, S., Pagnozzi, D., & Tanca, A. (2021) Comparative Evaluation of MaxQuant and Proteome Discoverer MS1-Based Protein Quantification Tools. *Journal of Proteome Research,* 20(7), 3497–3507

303. Ueland, T., Caidahl, K., Askevold, E. T., Karlsson, T., Hartford, M., & Aukrust, P. (2015) Secreted Frizzled-Related Protein 3 (sFRP3) in acute coronary syndromes. *International Journal of Cardiology,* 190, 217–219

304. Bouderlique, T., Henault, E., Lebouvier, A., Frescaline, G., Bierling, P., Rouard, H., et al. (2014) Pleiotrophin Commits Human Bone Marrow Mesenchymal Stromal Cells towards Hypertrophy during Chondrogenesis. *PLOS ONE,* 9(2), e88287

305. Arenas De Larriva, A. P., Limia-Pérez, L., Alcalá-Díaz, J. F., Alonso, A., López-Miranda, J., & Delgado-Lista, J. (2020) Ceruloplasmin and Coronary Heart Disease—A Systematic Review. *Nutrients,* 12(10), 3219

306. Annema, W., Gawinecka, J., Muendlein, A., Saely, C. H., Drexel, H., & von Eckardstein, A. (2022) Elevated levels of apolipoprotein D predict poor outcome

in patients with suspected or established coronary artery disease. *Atherosclerosis,* 341, 27–33

307. Huang, R., DeMarco, J. K., Ota, H., Macedo, T. A., Abdelmoneim, S. S., Huston, J., et al. (2021) Prognostic Value of Intraplaque Neovascularization Detected by Carotid Contrast-Enhanced Ultrasound in Patients Undergoing Stress Echocardiography. *Journal of the American Society of Echocardiography,* 34,(6) 614–624

308. Mitchell, C., Korcarz, C. E., Gepner, A. D., Kaufman, J. D., Post, W., Tracy, R., et al. (2018) Ultrasound carotid plaque features, cardiovascular disease risk factors and events: The Multi-Ethnic Study of Atherosclerosis. *Atherosclerosis,* 276, 195–202

309. Stary, H. C. (2000) Natural History and Histological Classification of Atherosclerotic Lesions. *Arteriosclerosis, Thrombosis, and Vascular Biology,* 20(5), 1177–1178

310. Mura, M., Schiava, N. Della, Long, A., Chirico, E. N., Pialoux, V., & Millon, A. (2020) Carotid intraplaque haemorrhage: pathogenesis, histological classification, imaging methods and clinical value. *Annals of Translational Medicine,* 8(19), 1273–1273

311. Wendorff, C., Wendorff, H., Pelisek, J., Tsantilas, P., Zimmermann, A., Zernecke, A., et al. (2015) Carotid Plaque Morphology Is Significantly Associated with Sex, Age, and History of Neurological Symptoms. *Stroke,* 46(11), 3213–3219

312. Goncalves, I., Sun, J., Tengryd, C., Nitulescu, M., Persson, A. F., Nilsson, J., & Edsfeldt, A. (2021) Plaque vulnerability index predicts cardiovascular events: A histological study of an endarterectomy cohort. *Journal of the American Heart Association,* 10(15), e021038

313. Kubo, T., Maehara, A., Mintz, G. S., Doi, H., Tsujita, K., Choi, S. Y., et al. (2010) The Dynamic Nature of Coronary Artery Lesion Morphology Assessed by Serial Virtual Histology Intravascular Ultrasound Tissue Characterization. *Journal of*

the American College of Cardiology, 55(15), 1590–1597

314. Rothwell, P. M., Eliasziw, M., Gutnikov, S. A., Fox, A. J., Taylor, D. W., Mayberg, M. R., et al. (2003) Analysis of pooled data from the randomised controlled trials of endarterectomy for symptomatic carotid stenosis. *The Lancet,* 361(9352), 107–116

315. Rothwell, P. M., Eliasziw, M., Gutnikov, S. A., Warlow, C. P., & Barnett, H. J. M. (2004) Sex Difference in the Effect of Time From Symptoms to Surgery on Benefit From Carotid Endarterectomy for Transient Ischemic Attack and Nondisabling Stroke. *Stroke,* 35(12), 2855–2861

316. Huang, H., Virmani, R., Younis, H., Burke, A. P., Kamm, R. D., & Lee, R. T. (2001) The Impact of Calcification on the Biomechanical Stability of Atherosclerotic Plaques. *Circulation,* 103(8), 1051–1056

317. Hartiala, J. A., Han, Y., Jia, Q., Hilser, J. R., Huang, P., Gukasyan, J., et al. (2021) Genome-wide analysis identifies novel susceptibility loci for myocardial infarction. *European Heart Journal,* 42(9), 919–933

Appendix

CURRICULUM VITAE

2018-22 King's College London, UK

PhD in Cardiovascular Science.

2016-18 Joint Postgraduate Programme, University of Patras, Greece

Master of Science in Informatics for Life Sciences (sector: Bioinformatics)

Thesis: "Cardiovascular Disease Risk Assessment with a Multi-omics Approach and Computational Techniques for the Construction and Analysis of Correlation Networks"

2010-2015 Computer Engineering and Informatics department, University of Patras, Greece

Bachelor's Degree in Computer Engineering and Informatics

Thesis: "Storyteller Application for Windows Phone"

Presentation of PhD work

30/04/2021 – Internal: Lunchtime Seminar, Cardiovascular sciences (virtual)
08/06/2021 - Leducq PlaqOmics Meeting (virtual)
12/04/2022 – King's BHF Centre Postgraduate Symposium

Publications

1. Konstantinos Theofilatos, Stefan Stojkovic, **Maria Hasman**, Sander W. van der Laan, Ferheen Baig, Javier Barallobre-Barreiro, Marion Groeger, Lukas Schmidt, Siqi Yin, Xiaoke Yin, Sean Burnap, Bhawana Singh, Svitlana Demyanets, Christoph Neumayer, Stephanie Kampf, Maja Carina Nackenhorst, Martin Bilban, Christian Hengstenberg, Kurt Huber, Gerard Pasterkamp, Johann Wojta, Manuel Mayr. *A Proteomic Atlas of Atherosclerosis: Signatures of Plaque Inflammation, Calcification and Sex Differences and their Association with Outcomes.* Under Review in Nature Cardiovascular Research (2022) (Based on Chapter 5)
2. **Maria Hasman**, Manuel Mayr & Konstantinos Theofilatos. *Uncovering protein networks in clinical proteomics.* Under Review in MCP (2022) (Based on Chapter 2)
3. **Maria Hasman**, Xiaoke Yin, Sophia Tsoka, Seferina Mavroudi, Manuel Mayr & Konstantinos Theofilatos. *Resolving atherosclerotic networks with directional regulatory network reconstruction with adaptive partitioning pipeline*. Under Review in Bioinformatics (Oxford) (2022) (Based on Chapter 3)

4. Elisa Duregotti, Christina M. Reumiller, Ursula Mayr, **Maria Hasman**, Lukas E. Schmidt, Sean A. Burnap, Konstantinos Theofilatos, Javier Barallobre-Barreiro, Arne Beran, Maria Grandoch, Alessandro Viviano, Marjan Jahangiri & Manuel Mayr. *Reduced secretion of neuronal growth regulator 1 contributes to impaired adipose-neuronal crosstalk in obesity.* Nat Commun 13(1), 1-16 (2022). https://doi.org/10.1038/s41467-022-34846-w (Use of the same proteomics data preprocessing and statistical analysis pipeline developed for Chapter 5)
5. Javier Barallobre-Barreiro, Tamás Radovits, Marika Fava, Ursula Mayr, Wen-Yu Lin, Elizaveta Ermolaeva, Diego Martínez-López, Eric L. Lindberg, Elisa Duregotti, László Daróczi, **Maria Hasman**, Lukas E. Schmidt, Bhawana Singh, Ruifang Lu, Ferheen Baig, Aleksandra Malgorzata Siedlar, Friederike Cuello, Norman Catibog, Konstantinos Theofilatos, Ajay M. Shah, Maria G. Crespo-Leiro, Nieves Doménech, Norbert Hübner, Béla Merkely & Manuel Mayr. *Extracellular Matrix in Heart Failure: Role of ADAMTS5 in Proteoglycan Remodeling.* Circulation 144, 2021-2034 (2021). https://doi.org/10.1161/CIRCULATIONAHA.121.055732 (application of earlier version of Chapter 3 network reconstruction method)
6. Clemens Gutmann, Kaloyan Takov, Sean A. Burnap, Bhawana Singh, Hashim Ali, Konstantinos Theofilatos, Ella Reed, **Maria Hasman**, Adam Nabeebaccus, Matthew Fish, Mark JW. McPhail, Kevin O'Gallagher, Lukas E. Schmidt, Christian Cassel, Marieke Rienks, Xiaoke Yin, Georg Auzinger, Salvatore Napoli, Salma F. Mujib, Francesca Trovato, Barnaby Sanderson, Blair Merrick, Umar Niazi, Mansoor Saqi, Konstantina Dimitrakopoulou, Rafael Fernández-Leiro, Silke Braun, Romy Kronstein-Wiedemann, Katie J. Doores, Jonathan D. Edgeworth, Ajay M. Shah, Stefan R. Bornstein, Torsten Tonn, Adrian C. Hayday, Mauro Giacca, Manu Shankar-Hari & Manuel Mayr. *SARS-CoV-2 RNAemia and proteomic trajectories inform prognostication in COVID-19 patients admitted to intensive care.* Nat Commun 12, 3406 (2021). https://doi.org/10.1038/s41467-021-23494-1. (Application of part of Chapter 4 DIA preprocessing method)