BIOMARKER DISCOVERY IN PARKINSON'S DISEASE AND CENTENARIANS

Proteomic studies of neurodegeneration and healthy ageing by mass spectrometry and machine learning

JENNY CECILIA HÄLLQVIST

UCL Great Ormond Street Institute of Child Health

Thesis submitted for the degree of Doctor of Philosophy (PhD) awarded by University College London

May 2022

I, Jenny Cecilia Hällqvist, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed.....

Date.....

This thesis is dedicated to my parents,
Lars and Cecilia Hällqvist

Abstract

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterised by motor and cognitive symptoms. The pathology hallmarks include alpha-synuclein aggregation and predominant dopaminergic cell loss in the midbrain. Advancing age is the main risk-factor and the reasons behind development of non-hereditary Parkinson's disease remain largely unknown. Although much effort has gone into finding biomarkers, there are currently no specific biomarkers allowing for screening of PD.

Aiming to discover new biomarkers and affected pathways for PD, and to probe the divergence between healthy and non-healthy ageing, discovery proteomics was performed and followed by a targeted validation. The protein expression associated with Parkinson's disease and healthy ageing was explored using a label-free, bottom-up mass spectrometry-based discovery methodology applied to serum, plasma and urine from Parkinson's disease patients, and plasma from cognitively healthy centenarians, all groups matched with controls. The discovery phase identified several proteins putatively related to Parkinson's disease and to longevity in the centenarians. Pathway analysis suggested an altered inflammatory response in both groups. The biomarker targets which emerged from the discovery phase were developed into a mass spectrometric, multiple reaction monitoring-based assay, augmented with inflammatory proteins from the literature, and applied to new and larger sample cohorts. Several proteins from the pathways were successfully confirmed in the targeted validation phase, and the results indicated activation of the unfolded protein response, reduced Wnt signalling and increased complement-mediated inflammation in the Parkinson's patients. In the centenarians, a longevity-promoting protein expression consisting of downregulated C3 and upregulated A2M and ADIPOQ was identified. Supervised machine learning models were trained to classify individuals as PD or healthy controls, and when predicting new samples, Parkinson's disease patients and controls could be discriminated perfectly in plasma and with 85.1% accuracy in urine.

Impact statement

Worldwide, approximately 1% of the population over the age of 65 develop Parkinson's disease, and roughly 4% after the age of 85. Therefore, there is a need for biomarkers of PD, not only for diagnosis and to follow the disease progression, but to also monitor new therapies as they are developed. Although significant efforts have gone into finding disease-specific biomarkers, none have yet proven successful and consequently, the diagnosis is set clinically, often at a relatively late stage of the disease when a large proportion of the dopaminergic neurons have been lost. The discovery of new biomarkers, especially for prodromal and early stages of PD, would have a significant impact on the understanding of the mechanisms behind the disease and also the possibilities of finding new drug targets. Most fluid biomarker studies of PD have focused on cerebrospinal fluid because of its proximity to the central nervous system. However, because of the highly invasive nature of the lumbar punctures used to sample cerebrospinal fluid, it may not be the most ideal choice of sample for routine screening of Parkinson's disease.

Due to their less invasive sampling methodology and ease of access, the biofluids blood and urine were evaluated for feasibility in neurodegenerative biomarker discovery in this work. Several putative biomarkers were identified by untargeted mass spectrometry and developed into a mass spectrometric targeted test. Using machine learning modelling, a panel of biomarkers measured in the targeted test could discriminate perfectly between newly diagnosed Parkinson's disease patients and healthy controls in plasma. In urine, the discriminating accuracy was 85.1%. Although further refining and testing for robustness and reproducibility is needed, these results show great promise and speak in favour of using urine and plasma in biomarker discovery studies of neurodegenerative disease.

A mass spectrometric blood- or urine-based test for risk of Parkinson's disease would be highly beneficial, given the relatively non-invasive sampling compared to lumbar punctures, and because of the speed, accuracy, and cost-efficiency of targeted mass spectrometric assays. This test could be used for screening and confirmation of Parkinson's disease and could aid in capturing individuals in early stages of the disease, before severe dopaminergic neuron loss has occurred. Thus, treatment could be commenced earlier.

Finally, the protein expression observed in the targeted analysis pointed towards a reduction in Wnt/ß-catenin signalling in the PD patients. Wnt signalling is critical for brain homeostasis and, importantly, maintenance of dopaminergic neurons. Wnt signalling in the early stages of Parkinson's disease should be explored further and may better the understanding of the disease.

Table of contents

List of figures

List of tables

List of equations

Glossary

ACN Acetonitrile

AD Alzheimer's disease

ANOVA Analysis of variance

AUC Area under the curve

BBB Blood-brain-barrier

BEH Ethylene bridged hybrid column chemistry

BLAST Basic local alignment search tool

CSF Cerebrospinal fluid

CSH Charged surface hybrid column chemistry

CV Coefficient of variation

DAVID Database for annotation, visualization and integrated discovery (a pathway and gene ontology tool)

DMSO Dimethylsulfoxide

DNP de novo Parkinson's disease (patients very recently diagnosed with Parkinson's disease)

DTE Dithioerythritol

ER Endoplasmic reticulum

ERAD Endoplasmic-reticulum-associated protein degradation

ESI Electrospray ionisation

FA Formic acid

FDR False discovery rate

GO Gene ontology

GUI Guided user interface

GWAS Gene wide association study

HEPT Height equivalent to one theoretical plate

HILIC Hydrophilic interaction liquid chromatography

HPLC High performance liquid chromatography

IAA Iodoacetamide

IEF Isoelectric focusing

IMS Ion mobility

IPA Ingenuity pathway analysis

KEGG Kyoto encyclopedia of genes and genomes (a pathway database)

LC Liquid chromatography

LCMS Liquid chromatography coupled with mass spectrometry

LDA Linear discriminant analysis

LOWESS Locally weighted scatterplot smoothing

MAD Median absolute deviation

MRM Multiple reaction monitoring (a mass spectrometric acquisition technique)

MS Mass spectrometry

MSA Multiple system atrophy

MS^E A mass spectrometric acquisition technique, altering between low and high fragmentation

MSMS Tandem mass spectrometry

MWCO Molecular weight cut-off

OND Other neurological disorders

OPLS Orthogonal partial least squares

OPLS-DA Orthogonal partial least squares discriminant analysis

PCA Principal component analysis

PD Parkinson's disease

PLS Projection to latent structures/partial least squares

PQN Probabilistic quotient normalising

PSP Progressive supranuclear palsy

QC Quality control

QTOF Quadrupole time of flight

RBD Rapid eye movement sleep behaviour disorder

REM Rapid eye movement

RFE Recursive feature elimination

ROC Receiver operating characteristic

ROS Reactive oxygen species

SD Standard deviation

SDS Sodium dodecyl sulphate

SIMCA A tool for multivariate analysis

SPE Solid phase extraction

STRING Database of known and predicted protein-protein interactions

SVD Singular value decomposition

SVM Support vector machine

TCA Trichloroacetic acid

TFA Trifluoroacetic acid

TOF Time of flight

UPLC Ultra performance liquid chromatography

UPR Unfolded protein response

Acknowledgements

The completion of this project would not have been possible without the contribution from many brilliant people I have been fortunate enough to work with.

Firstly, a huge thank you to my primary supervisor, Prof Kevin Mills. I could not have asked for a better mentor, and I have so enjoyed working together. I am amazed by your encyclopaedia-like knowledge of all things biochemistry and mass spectrometry, your wisdom, endless enthusiasm, and kindness. Thank you for taking me on as a student, for always finding the time to discuss problems and for being supportive and interested, regardless of the subject being Viking invasions, politics or science.

Thank you so much to my three secondary supervisors, Professors Selina Wray, Henrik Zetterberg and Philippa Mills. As the project took a different turn than originally intended, we ended up not performing many of the experiments initially planned, still I knew that you were available if I ever wanted to discuss anything. I am also enormously grateful for all the thoughtful input during the writeup of this thesis.

This PhD project was done part-time, and I am greatly indebted to Prof Tom Warner and Dr Wendy Heywood who both, to no gain of their own, granted me the time to complete it while working on their projects. Apart from thanking them for this generosity, I also want to express how much I enjoyed, and still do, working together.

My heartfelt thanks to all the collaborators in the Adage and Propage-ageing consortia. I am especially grateful to Professors Kailash Bhatia, Brit Mollenhauer and Claudio Franceschi and their brilliant research groups, who provided the sample cohorts analysed in this project. I can only hope that other collaborations will be as great as these have been.

Thank you so much to all the wonderful collaborators and colleagues I have been lucky to meet and work with during my time at UCL. Thank you, Amanda Heslegrave, Ross Paterson, Rohan de Silva, Kin Mok, Natalia Barahona Torres, Sebastian Schreglmann, Eoin Mulroy, Peter Clayton, Robert Clayton, Matt Wilson, Anna Baud, Nadia Ashrafi, Youssef Khalil, Sree Vootukuri, Sonam Gurung, Valeria Nikolaenko, Harriet Gunn, Fatimah Almousawi, Eva Sedlak, Elena Alvarez Mediavilla, Katharina Iwan, Claire Leckey, Heloise Vinette, Tom Baldwin, Carolin Huynh and Simon Pope. I especially want to thank my lab partners and dear friends Ivan Doykov and Justyna Spiewak for all the times we shared both inside and outside of the lab. We have worked on so many projects together and you have made every single one of them better and more enjoyable.

Thank you to my friends and former colleagues at the Swedish Metabolomics Centre, Jonas Gullberg, Thomas Moritz, Inga-Britt Carlsson, Annika Johansson, Hans Stenlund, Siv Sääf, Maria Ahnlund, Anders Nordström and Krister Lundgren. You are all a large part of the reason I ended up working with the remarkable technique of mass spectrometry and you taught me so many of the things I know.

Lastly, I want to thank my wonderful and loving family for all the support and encouragement. To my parents Cecilia and Lars – the greatest acknowledgement of all is for you. I can't thank you enough for constantly encouraging me to follow my dreams and for, without failure, being 100% supportive, even when that meant me moving almost 2000 km away from you. Know that this would not have been possible without you. To my fantastic brother Oscar and my equally fantastic sister-in-law Annelie, thank you for our enjoyable scientific discussions, some better and some worse..., and for all the great times. Bea and Leo, thank you for the joy and happiness you bring. Finally, to my beloved Rui - thank you for the unwavering support and encouragement in all things, for your immense patience and love of sharing knowledge. I don't know how to thank you enough for all the input and suggestions you've given me during this project and for your interest in everything we discuss. I am so grateful for all the things I've learnt from you and in awe of your ability to always look on the bright side.

Thank you, diolch yn fawr, tack, obrigada, dziękuję ci, blagodaria, grazie, danke, gracias, merci, shukran, köszönöm, spasiba!!!

Proteomic biomarker discovery by mass spectrometry: introduction, background and theory

1

1.1 WHAT IS BIOMARKER DISCOVERY?

In 1998, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as:

> *"a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention"* [1]

An ideal biomarker should be specific to the disease in question and preferably reflect the pre-symptomatic stage and disease progression. It must be stable over time, and it is desirable that it changes in response to disease-modifying treatment. Biomarkers can be metabolites, proteins, or genes [2], but also other types of chemical and biological entities. Biomarkers are routinely used in clinical practice and have several meaningful applications, ranging from disease screening to diagnosis of different conditions and indication of disease severity. Examples of established biomarkers commonly utilised in clinical medicine include blood glucose elevation used to diagnose diabetes [3], creatinine levels in urine and blood to indicate renal failure [4], and levels of C-reactive protein as a marker of inflammation and response to antibiotic treatment [5, 6]. One of the most successful applications of biomarker monitoring can be found in inborn errors of metabolism. These inherited conditions are often caused by a defect in a protein-coding gene and lead to a lack of functioning protein, thus causing a blockage in a certain metabolic pathway and accumulation or decrease of specific metabolites. Many countries therefore perform population screening, monitoring relevant metabolites using targeted mass spectrometry [7, 8].

Biomarkers are undoubtedly valuable and therefore much sought after. Tremendous efforts have been applied into finding biomarkers for diagnosis in common neurodegenerative conditions, such as Alzheimer's and Parkinson's diseases. In these complex and heterogenous phenotypes, where the cause of disease is largely unknown, this is a combined endeavour of finding biomarkers not only for diagnosis, but also to identify biochemical mechanisms and affected pathways thereby allowing for better understanding of the disease and development of treatments. For some diseases, such as the ones observed in inborn errors of metabolism, single key biomarkers capable of diagnosing disease have been identified. However, especially for complex conditions where there has not been one pivotal discovery pinpointing the cause of disease, the efforts are increasingly going into finding panels of compounds, a "fingerprint", which can differentiate disease from control [9].

Biomarker discovery can be performed utilising any of the "omics" techniques, where genomics studies DNA molecules, transcriptomics studies RNA, proteomics studies proteins and metabolomics study metabolites [10]. Figure 1-1 gives a simplified overview of the omics technologies and examples of the instrumentation used.

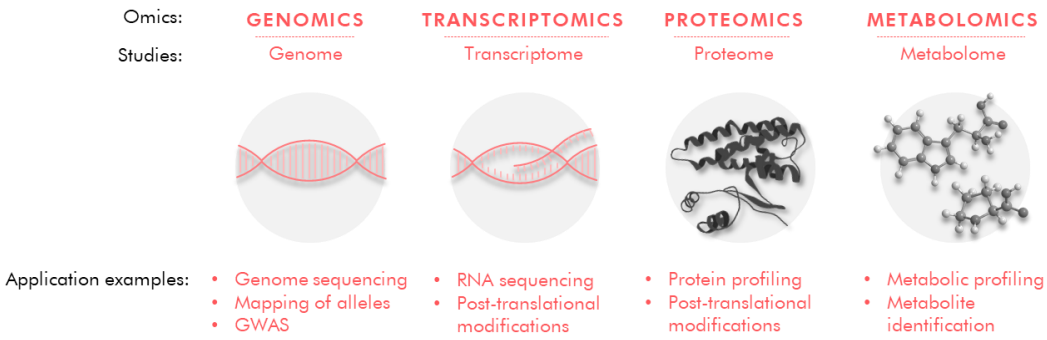


Figure 1-1. Major omics technologies and examples of their applications. *Genomics studies the genome and applications include genomic sequencing and GWAS studies. Transcriptomics studies the transcriptome and is often applied to sequence RNA and to study post-translational modifications. Proteomics studies the proteome, applications include protein profiling and post-translational modification studies. Metabolomics studies the metabolome and is applied for metabolic profiling, and also the identification of metabolites.*

The experiments in this thesis investigate proteomic biomarkers of Parkinson's disease and healthy ageing using mass spectrometry-based techniques, therefore the focus of this introductory chapter will be on proteomics, the instrumentation and analytical pipelines used, and the current status of biomarkers in the two fields.

1.1.1 *Steps involved in the biomarker discovery workflow*

The pipeline for any biomarker discovery study can be divided into three distinct umbrella phases: (i) discovery, (ii) validation/verification and (iii) clinical evaluation [11]. In the initial discovery phase, a small number of well-characterised samples are analysed in an unbiased manner, attempting to measure a maximum of endogenous compounds present in a sample. In the validation/verification phase, putative biomarkers from the discovery phase are developed into a targeted test utilising an analytical platform different from the one used in the discovery study. This is then applied to a new and larger set of samples, typically consisting of more than one hundred subjects. The putative biomarkers which are successfully validated can thereafter be evaluated in much larger sample sets to assess their robustness and viability in a clinical test. Figure 1-2 shows a graphical illustration of the numbers of samples and analytes typically involved in the three steps of process.

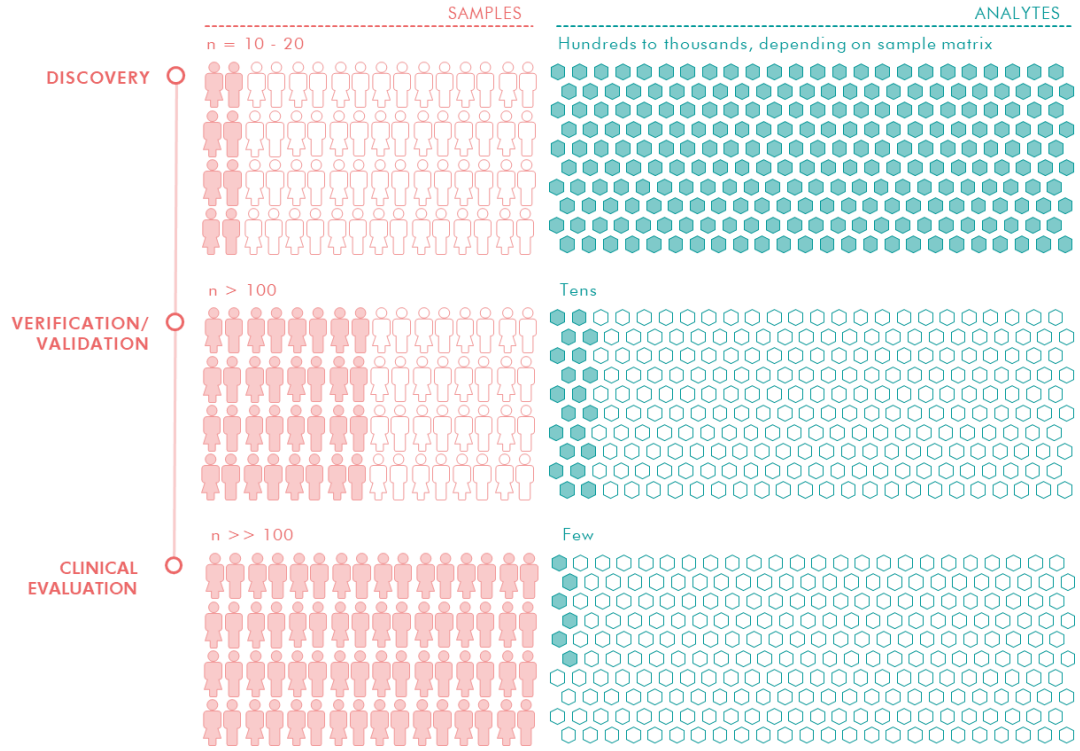


Figure 1-2. Biomarker discovery workflow describing the numbers of samples and analytes in the different phases. *In the discovery phase, a small group of well-defined samples are analysed, generating a large number of analytes, represented by filled hexagons. In the validation/verification phase, a larger number of samples, typically more than 100, are analysed, measuring only the compounds of interest selected from the discovery phase. In the clinical evaluation phase, the biomarkers which were verified in the validation phase are measured in very large numbers of samples to ensure that the markers are robust*

1.2 LIQUID CHROMATOGRAPHY COUPLED WITH MASS SPECTROMETRY

Proteomic analyses, both targeted and untargeted, are frequently performed using liquid chromatography (LC) paired with mass spectrometry (MS). The LC system separates compounds based on their chemical properties and the mass spectrometer separates and detects the compounds based on their molecular weights and charges. The retention time a molecule elutes from a column is extremely consistent using modern pumps and chromatographic columns, allowing another dimension of identification. The coupling of the two techniques provides a powerful tool for the separation and detection of proteins present in a sample.

1.2.1 *Liquid Chromatography*

Liquid chromatography is a separation technique which can be used to separate and purify analytes of interest. It is commonly used to reduce the complexity of the sample reaching a detector, and/or to separate compounds to facilitate identification. A rudimentary LC system consists of a column packed with a stationary solid phase and a

liquid mobile phase pumped through the column. The mobile phase customarily consists of two separate solutions, delivered to the column in different proportions throughout the run to manipulate the analyte separation. The chemistry of the stationary and mobile phases can be tailored to suit the intended application depending on its requirements. There is a wide range of column chemistries and the utility of them can broadly be divided into two umbrella categories – normal phase applications and reversed phase applications. The normal phase applications are generally tailored towards polar compounds, making them useful for biochemical analyses of sample matrices containing a large proportion of polar analytes, such as urinary metabolites [12]. Reversed phase LC has a wide range of applications and is often employed in studies of more hydrophobic molecules, such as metabolite and peptide screening experiments, due to its all-around ability to separate compounds based on their polarities [13, 14].

Reversed phase is arguably the most common procedure for liquid chromatography separations. Reversed phase applications encompass a non-polar stationary phase with spherical silica-based particles bonded with hydrocarbon chains, generally alkyls with four to 18 carbons, and a mobile phase with high aqueous content. The separation of analytes is mainly achieved by hydrophobic interaction and solute transfer between the aqueous-rich mobile phase and non-polar stationary phase. The order of analyte elution goes from more polar to less polar [15].

Normal phase applications generally consist of a silica-based polar stationary phase and an aqueous or organic mobile phase, depending on the application. An important and much used sub-category of normal phase chromatography is hydrophilic interaction chromatography (HILIC), an efficient technique for separating polar compounds. In HILIC, the stationary phase can be neutral, charged or zwitterionic, and the mobile phase is composed of a high amount of organic solvent. The separation of analytes is chiefly determined by the analytes' hydrophilic partitioning between the highly organic mobile phase and an aqueous layer formed on the stationary phase [16]. The order of analyte elution generally goes from less polar to more polar.

Additionally, there are hybrid column chemistries utilising a combination of the reversed and normal phase properties. There are also applications combining different column chemistries in series, thus allowing for highly specific separations. Figure 1-3 shows an illustration of different column chemistries and mobile phase systems commonly utilised to separate analytes in biochemical applications.

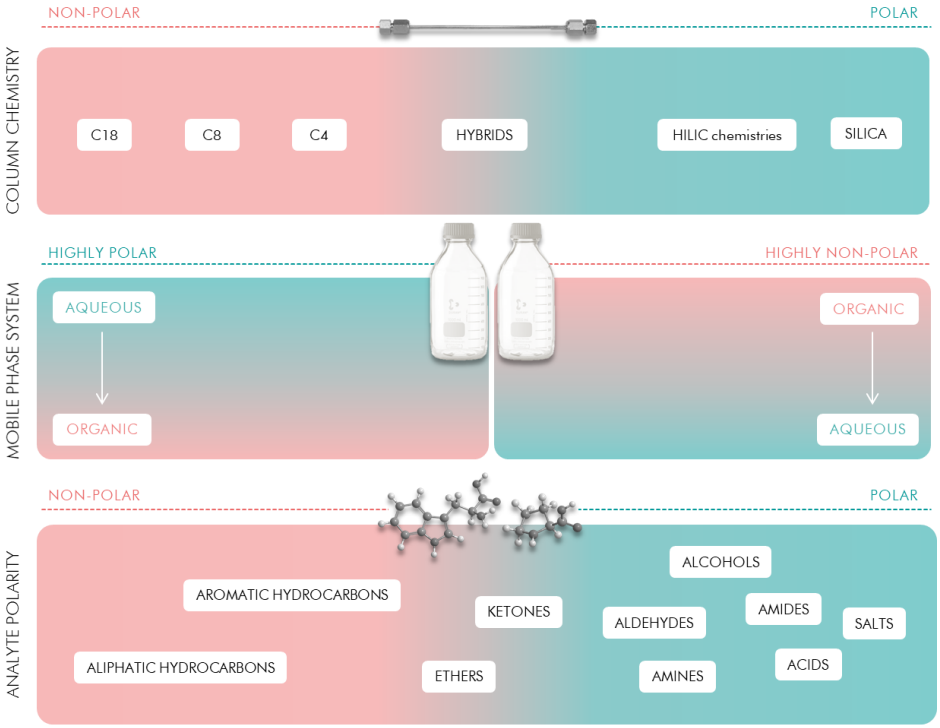


Figure 1-3. Illustration of liquid chromatography column chemistries, mobile phase systems and analytes. *Column chemistry shows stationary phase polarity, from non-polar C18 commonly used in reversed phase applications to polar HILIC and silica, used in normal phase applications. Mobile phase system shows the highly polar system used for reverse phase applications, where the mobile phase gradient starts as highly aqueous and moves towards organic, and the highly non-polar mobile phase system utilised in HILIC applications, initiated with high organic and moving towards aqueous. The analyte polarity shows a scale of increasing polarity of functional groups.*

1.2.1.1 *Theory of liquid chromatography separations*

Liquid chromatography separations are generally performed with the aim of obtaining the best possible resolution in the shortest amount of time. The success of this aim is governed by chemical and physical properties of the column and mobile phase, and by the applied temperature and flow rate [17].

The time an analyte remains on a column is determined by its affinity for the mobile and stationary phases. By utilising a suitable mobile phase and a well-designed elution gradient, the elution can be manipulated to render selectivity and maximum separation between analytes. The time a non-retained analyte takes to reach the detector is known as the dead time (t_0) of the system while the time it takes a retained analyte to reach the detector is known as retention time (t_R). The base peak width is denoted by w and the width of a peak at half of its heigh is denoted by $w_{1/2}$. Figure 1-4 shows an example chromatogram of two analyte peaks including the terms used to describe the properties of the elution and separation.

The flow rate with which the mobile phase is pumped through the column, the system's dead volume, column temperature, the diameter of the column and particle size affect separation, sensitivity and resolution, and must be therefore be carefully optimised [17]. The chromatographic separation of analytes prior to detection is crucial in most applications, but especially so in untargeted biochemical analyses as the samples are highly complex and the overarching goal is to measure the largest number possible of compounds present in a sample. Although high resolution mass spectrometers are capable of detecting multiple coeluting compounds simultaneously, chromatographic separation reduces the impact of issues such as matrix suppression and detector saturation. Moreover, the analysis time should be kept as short as possible. Therefore, it is necessary to employ a liquid chromatography method which efficiently separates and distributes the analytes across the chromatogram as evenly as possible, thereby allowing the detector sufficient time to detect the greatest number of compounds possible in the shortest amount of time.

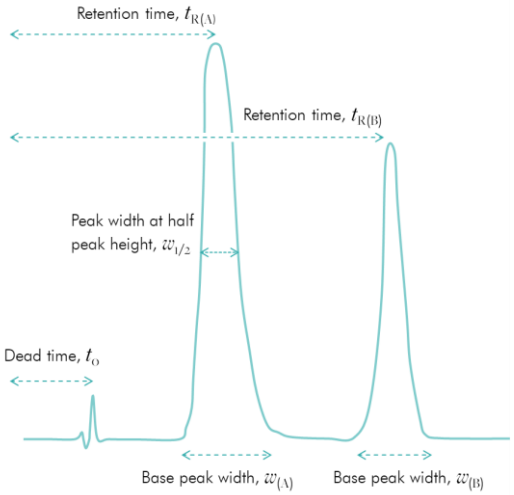


Figure 1-4. Artificial chromatogram describing the nomenclature of dead time, retention time and peak width.

1.2.1.2 *Ultra performance liquid chromatography*

Modern LC systems used for biochemical analyses often employ a technology known as ultra performance liquid chromatography (UPLC). This is a technique which builds on the advances of its predecessor, high performance liquid chromatography (HPLC). The major differences between the two techniques are system pressure tolerance, stationary phase particle size and column diameter. In HPLC, the maximum system pressure is around 500 bar whereas in UPLC, it is around 1000 bar. In HPLC, the stationary phase particle size is 5 – 10 µm, while in UPLC, it is below 2 µm. The column diameter in HPLC is typically 3.0 – 4.6 mm and in UPLC it is typically 1.0 – 2.1 mm [18]. Smaller particle size generates more efficient chromatography and since system pressure is directly proportional to solvent viscosity, column length and flow rate, higher pressure tolerance limits allow for more rapid applications. Overall, UPLC grants increased sensitivity, more efficient chromatographic separation in a shorter amount of time, with the added advantage of reduced solvent consumption due to smaller column diameters [19].

1.2.1.3 *Two-dimensional nano liquid chromatography*

As described in section 1.2.1.1, chromatographic separation is essential as it reduces the complexity of the sample reaching the mass spectrometer and thereby allows more proficient detection of analytes present in a sample. In untargeted proteomics, which is used for discovery studies, the overall aim is to achieve the greatest possible coverage of the proteins present in a sample at a certain time, therefore maximum sensitivity is crucial. To achieve this, two-dimensional nano-LC applications (2D-LC) are commonly applied. In nano-LC, the inner diameter of the column is generally 50 – 75 µm and the flow rate is between 200 and 400 nL/min. The increased sensitivity of nano applications compared to UPLC is mainly achieved by reduced dilution of the chromatographic bands, thus analytes elute from the column in more concentrated bands [20]. Additional advantages of the miniaturisation are reduced solvent consumption and minimal required sample volumes.

To further enhance sensitivity in proteomics applications, sample fractionation is often performed. The aim of this procedure is to reduce the complexity of the sample by dividing it into smaller parts which are analysed separately. Sample fractionation can be performed "offline" as part of the sample preparation, or "in-line" during the chromatographic separation. Both strategies have advantages and disadvantages. For the experiments described in this thesis, in-line fractionation was performed. In this methodology, two separate LC systems are operated in series and the sample is separated in two dimensions, two-dimensional LC (2D-LC). The analytes are separated by two different columns with high orthogonality and different separation mechanisms. A common setup, and the one used in our experiments, utilises a reversed phase column and a basic mobile phase in the first dimension, followed by a reversed phase column and an acidic mobile phase in the second dimension [20, 21]. The difference in separation selectivity between the two LC dimensions is due to altered charge distribution in the peptide chains with alteration of the mobile phase pH [22]. In practical terms, the separation is performed by loading the sample onto the first-dimension column and eluting the first fraction of analytes with a high pH mobile phase onto the second-dimension column, wherefrom the analytes are eluted with a low pH mobile phase and sent to the mass spectrometer for detection. The second fraction is thereafter eluted from the high pH system in a series of "slugs" to the low pH system, and this procedure is repeated until all fractions have been analysed.

In conclusion, 2D-LC is an efficient strategy of increasing the coverage of proteins in a sample as it greatly reduces the sample's complexity for downstream mass spectrometric detection.

1.2.2 *Mass Spectrometry*

The first mass spectrometer was developed by J.J. Thompson in 1912 and was at that time called a "parabola spectrograph" [23]. Since its invention, mass spectrometry has become an invaluable tool in a wide range of fields - from studies of the solar system to biology. Mass spectrometry is often described as a somewhat enigmatic technology, and a textbook on the subject humorously concludes:

> *"despite occasional mysteries, mass spectrometry is still highly useful"* [24]

The fundamental principle of any mass spectrometer can be divided into three stages [25]:

1) Generate ions in the gas phase
2) Separate the ions with respect to their mass-to-charge ratio (m/z)
3) Count the number of formed ions using a detector

Figure 1-5 shows a highly simplified schematic diagram of the process for two compounds, A and B. These are first introduced to the MS through the sample inlet into the ion source. From there, they are pulled from the ion source towards the mass analyser by the system's vacuum. The ionised compound A has a mass of 100 but since it is doubly charged (z = 2), its resulting mass-to-charge ratio becomes m/z = 50, while for the singly charged compound B with a mass of 80, the mass-to-charge ratio is m/z = 80.

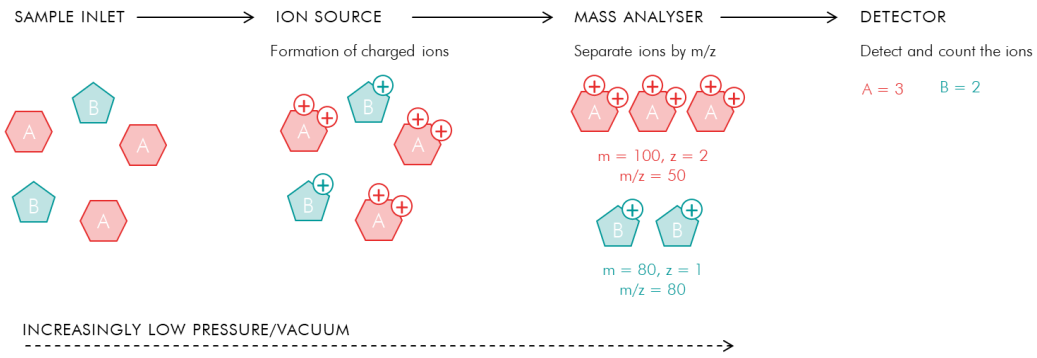


Figure 1-5. Basic principle of mass spectrometry. *Compounds A and B are introduced to the MS by the sample inlet. They are ionised in the ion source where A has a positive double-charge and B has a positive single-charge, from there A and B are pulled by the increasingly low pressure towards the mass analyser where they are separated based on their m/z ratios (m/z = 50 for A and m/z = 80 for B), and finally, the ions are detected and counted by a detector, three for A and two for B.*

There are several different techniques utilised for ionisation and mass analyser separation; some examples include time-of-flight (TOF), orbitrap, quadrupole, matrix-assisted laser desorption/ionization (MALDI), and electrospray ionisation (ESI) techniques [26]. Hybrid-instruments are common, where two or more of the mass analyser techniques are combined, thus benefitting from their respective best traits.

1.2.2.1 *Ionisation*

Ionisation is a key step in any MS application, as only charged molecules can be detected. Depending on the application, different ionisation techniques can be utilised. Apart from the already mentioned ESI and MALDI, other techniques, tailored towards the ionisation of compound classes with certain characteristics, include atmospheric pressure chemical ionisation, atmospheric pressure photo ionisation and desorption electrospray ionisation [27]. Throughout the experiments performed in this thesis, ESI was utilised, therefore the focus will be on this technique.

In proteomic applications, soft ionisation techniques such as ESI dominate as they readily produce multiply charged ions, thereby making it possible to analyse large molecules. ESI has been said to provide "*the wings of a molecular elephant*" [28] as molecules with high masses can be detected while intact. The soft ionisation is partly achieved because ion transfer is initiated at atmospheric pressure and allows ions to move gradually towards the vacuum of the mass analyser [25]. Ions are formed from the LC eluate which is converted into an aerosol in the mass spectrometer's ion source and sprayed into an electrostatic field with large potential before entering the vacuum system of the mass analyser [29].

The formation of ions in ESI happens in three steps: creation of an electrically charged spray, reduction of the droplets' size, and lastly desolvation and liberation of the ions. The sample is introduced to the ion source under atmospheric pressure, sprayed through a capillary. A high potential is applied (typically 3-4 kV) to the capillary, which leads to the formation of charged droplets. The solvent of the droplets is evaporated by applying a hot, inert gas, thus continuously producing smaller and more concentrated droplets. The tightly packed droplets containing ions of the same polarity will eventually break up in a process termed Coulomb explosion. This occurs when the force of repulsion between the ions exceeds that of the droplet's surface tension and the droplets break into smaller drops [30]. The mechanism behind ion formation in ESI is not entirely understood. There are two theories – the charged residue model and the ion evaporation model. There are claims that the first model is valid for larger molecules while the latter applies to small ones [25, 31]. In the ion evaporation model, ions eventually desorb into the gas phase after the

droplets are subjected to Coulomb explosion [30], while in the charge residue model the solvent is completely evaporated post Coulomb explosion [32]. Figure 1-6 illustrates a simplified scheme of the ion formation models.

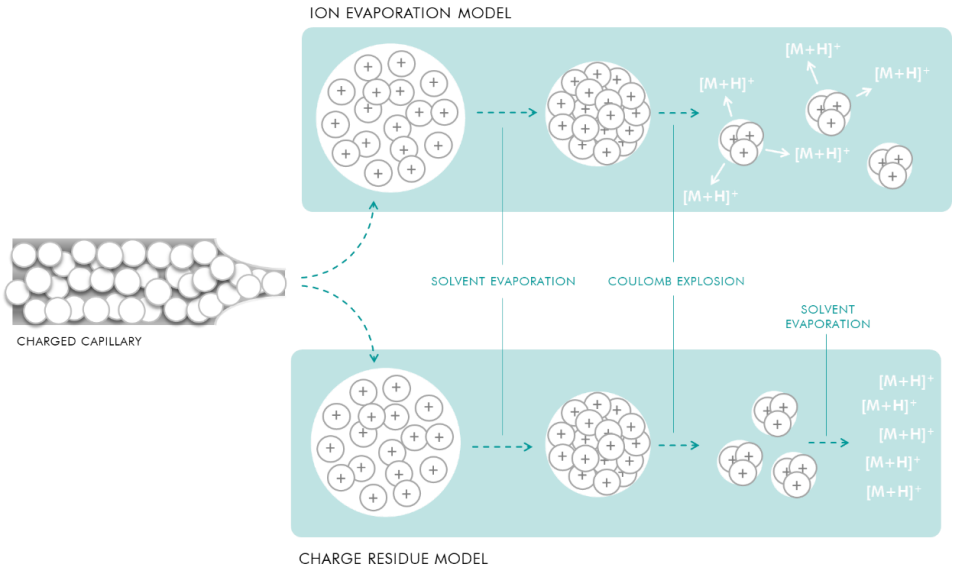


Figure 1-6. Simplified illustration of the ionization in ESI+. *Ions of the same polarity form a Taylor cone on the surface of the eluate solution, charged droplets are ejected in a jet and break up in droplets moving towards the mass detector.*

1.2.2.2 *Mass analyser separation*

The mass analyser is a central component of a mass spectrometer and where the mass-to-charge separation of ions takes place [26]. In the experiments performed in this thesis, quadrupole and time-of-flight mass analysers were used and thus the focus will be on these techniques.

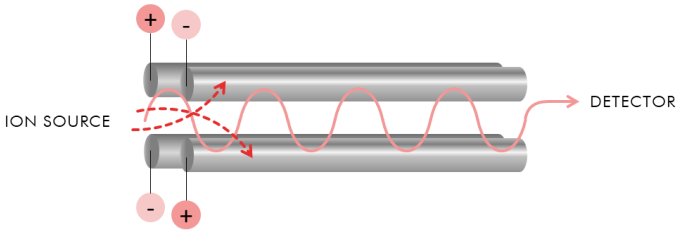


Figure 1-7. Quadrupole mass analyser. *The quadrupole rods have different potentials applied, thus generating an electric field in which ions develop an oscillating motion with frequencies and amplitudes related to the ions' m/z. Only ions with stable trajectories will pass through the quadrupole to the detector*

The quadrupole mass analyser was developed by the Nobel prize winner Wolfgang Paul and was first described in 1956 [33]. A quadrupole consists of four cylindrical rods in parallel with a cavity in between them. The pairs of opposing rods have a radio frequency voltage with a direct current offset applied to them, thereby creating an electric field. As ions enter the quadrupole from the ion source, they are subjected to the electric field which causes them to oscillate at unique frequencies and amplitudes related to their m/z ratios [34]. Only ions with stable trajectories will pass through the quadrupole [35]. By altering the radio frequency and the direct current applied to the quadrupole, ions of different m/z will develop a stable oscillation thus making it possible to create a mass-to-

charge filter for either an m/z range or a set of specific m/z values. Apart from being used as mass filters, quadrupoles can also be utilised for fragmentation of ions. The fragmentation is achieved by only applying radio frequency to the quadrupole, either as an energy ramp or as a fixed value. Quadrupoles are utilised in both targeted and untargeted applications. Figure 1-7 shows an illustration of a mass filter quadrupole.

The idea of time-of-flight as a mass analyser technique had been around for some time, but was first applied in practice by A.E. Cameron and D.F. Eggers in 1948, at the time it was called "Ion Velocitron" [36]. In its early days, the time-of-flight analyser was known for low resolution and poor sensitivity but thanks to electronical and computational advances, it is today recognised as a high sensitivity instrument with nearly unlimited mass range [37]. In time-of-flight mass analysers, ions are accelerated by an electric filed, thereby providing ions of the same charge with equivalent kinetic energy. The ions will disperse in respect to their m/z when flying down a field-free flight tube. The time-of-flight for the ions is recorded as they hit a detector [37], and is directly proportional to their mass-to-charge ratios [38], as demonstrated by equation (1-1):

$$t = d\sqrt{\frac{m}{2eEz}} \tag{1-1}$$

where t is time-of-flight, d is the distance from the ion source to the detector, m is the ion mass, z is the ion charge, e is a universal constant and E is an electric field. Ions arrive at the ion detector in order of increasing weight, thus by recording their time of flight, the exact molecular weights can be determined. Time-of-flight analysers are useful for measuring compounds distributing over a wide m/z range and are therefore often employed in screening studies. Figure 1-8 gives a simplified schematic illustration of the process.

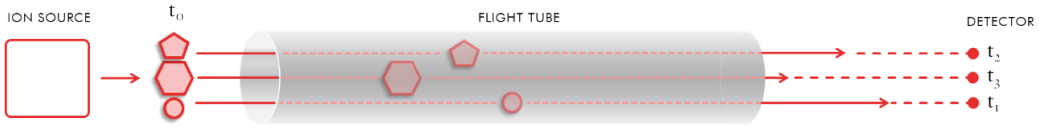


Figure 1-8. Time-of-flight mass analyser. *Ions of different mass enter the flight tube at time t_0. As they travel through the flight tube, they are dispersed in space based on their mass-to-charge ratios, molecules with lower m/z ratios reaching the detector first as visualised by the time points t_1, t_2 and t_3.*

The resolving power of a time-of-flight mass analyser is constrained by the slight variation in kinetic energy for ions of the same m/z. This variation is caused by a distribution in initial kinetic energy of the ions and as a result, ions of the same m/z will have minor differences in velocity and flight time, thus reaching the detector at slightly different times [39]. Several strategies were attempted to solve this problem. Arguably, the most successful was the reflectron, introduced by B.A. Mamyrin in 1973 [40]. A reflectron consists of a series of electronic lenses containing an electrostatic field held at increasingly higher potential that

can reflect the ions to the detector. Ions with higher kinetic energy will penetrate deeper into the reflectron and have a longer flight path before reaching the detector, while less energetic ions will be reflected earlier, allowing all ions of the same m/z, but slightly different kinetic energies, to be refocused so they reach the detector simultaneously regardless of small differences in their initial kinetic energy [17]. The efficiency of a time-of-flight analyser depends on its capability to accurately measure short time intervals in which ions of different m/z reach the detector [29] and the use of a reflectron significantly increases the resolution. Figure 1-9 shows a simplified illustration of a TOF coupled with a reflectron.

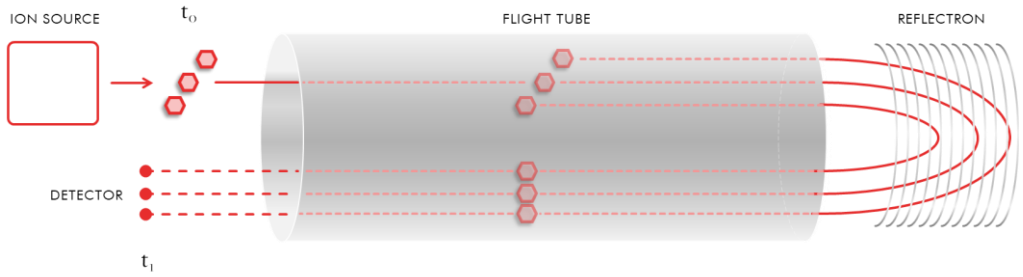


Figure 1-9. Time-of-flight mass analyser utilising a reflectron. *Ions with identical m/z but slight differences in initial kinetic energy are accelerated at t_0 and will travel through the flight tube at different velocities. As they reach the reflectron, the ions are refocused and the kinetic energy equalised so that they reach the detector at the same time, t_1.*

A technique often coupled with mass analysers in untargeted applications is "ion mobility" (IMS). Ion mobility is a way to further enhance the resolution and separation of ions by their structural conformation. The technique was pioneered by E.W. McDaniel and developed in the 1950s and 1960s [41]. In ion mobility, ions are subjected to a flow of inert gas in a low electric field. The velocity of the ions is, in simple terms, constrained by collisions with the inert gas [42]. Bulkier ions will have a greater number of collisions and thus a longer drift time, making it possible to separate ions with the same m/z by their structural configuration.

Detection is the final stage of the ions' journey through a mass spectrometer. As ions exit the mass analyser, they hit a detector where the number of ions of a certain m/z at a specific time are counted and the relative abundance of the ion species can be determined. An ideal detector should have a fast response time, good collection efficiency, wide dynamic range, low noise, high amplification and produce the same response regardless of mass [43]. In the untargeted experiments presented in this thesis, an ultra-fast electron multiplier detector was used, and in the targeted applications the detector was a photomultiplier. Apart from these two detectors, there are numerous others [44].

1.2.3 *Instrumentation used in untargeted mass spectrometry for discovery proteomics: 2D-LC coupled with Q-TOF-IMS-MS*

In untargeted proteomics discovery studies, the aim is to achieve the greatest possible coverage of the proteins present in a sample and since the samples are highly complex, sensitivity and resolution are critical. The instrumentation is therefore highly specialised to achieve this goal. In the discovery experiments performed in this thesis, the LC-MS arrangement consisted of a two-dimensional nano-LC system (Waters nanoAcquity, Manchester, UK) coupled to a mass spectrometer (Waters Synapt-G2-S*i*) equipped with an ESI ion source, an ion mobility module to separate the ions based on their structural conformation, a collision cell for alternating between high and low energy, thereby passing both intact and fragmented ions, and a time-of-flight mass analyser utilising a reflectron to accurately record the ions' mass-to-charge ratios. Figure 1-10 gives a schematic diagram of the modules and their functions.

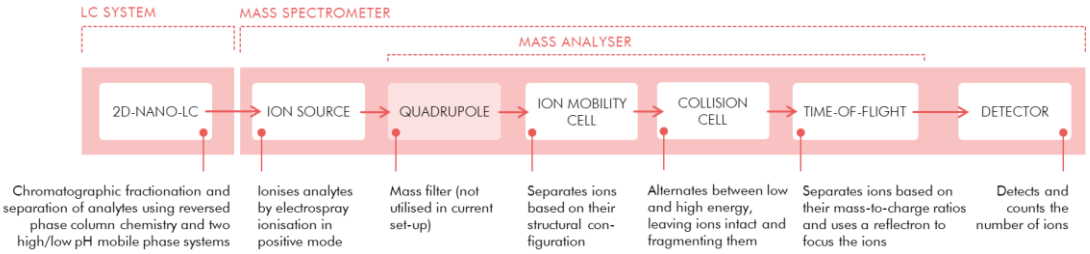


Figure 1-10. Diagram of the instrumentation utilised in the untargeted proteomics experiments of this thesis. *The sample is fractionated on a reversed phase column with high pH mobile phase and thereafter separated on a reversed phase column with a low pH mobile phase before entering the mass spectrometer. The analytes are ionised in positive mode in the ion source and thereafter pulled towards the mass analyser by lower pressure. The ions first enter the ion mobility module, where the ions are separated by their structural conformations. The ions thereafter reach the collision cell, where high and low energy alternate, thus sending both intact and fragmented ions to the flight tube, which is equipped with a reflectron. The time of flight of the intact and fragmented ions are recorded by the detector, and m/z values and relative intensities for each species are recorded.*

The resulting output data from the LC-MS system is a three-dimensional representation of intensity, retention time and mass-to-charge ratio for each chromatographic peak in each sample. The output also contains drift time given by the ion mobility separation. Moreover, each peak is represented by a precursor ion (the non-fragmented ion) and the fragments produced from the precursor ion. Figure 1-11 shows a demonstration of the information which is extracted. How the data is used for identifying proteins is described in detail in section 1.3 of this chapter.

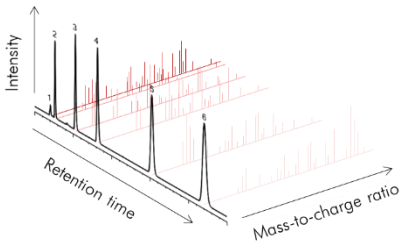


Figure 1-11. Representation of output data from untargeted LC-MS. *The illustration shows peaks eluting at certain retention times, the m/z of the fragments resulting from the ions, and the intensities.*

1.2.4 *Instrumentation for targeted proteomics mass spectrometry: UPLC coupled with triple quadrupole MS*

In targeted applications where selectivity, dynamic range and speed of analysis are important, it is common to use a single chromatographic HPLC or UPLC system coupled to a triple quadrupole (QqQ) mass spectrometer. The invention of the triple quadrupole mass spectrometer is attributed to C.G Enke and R.A. Yost and they published their first report on the instrument in 1978 [45].

A triple quadrupole MS is a tandem mass spectrometer and consists of two quadrupole mass analysers with a non-resolving, radio frequency only, collision-cell, present between them. Arguably, the most common operating mode of triple quadrupole instruments is a method known as "multiple reaction monitoring" (MRM). In this methodology, the first (Q1) and last (Q3) quadrupoles operate as mass analysers and the middle (q2) quadrupole as a collision cell that can be flooded with inert gas, typically argon. In practical terms, it works by utilising the first quadrupole as a mass filter, allowing only precursor ions of a certain pre-defined mass-to-charge ratio to pass. The ions are thereafter collided with a flow of inert gas in the collision cell, producing fragmentation of the ions. Only the selected product ions are allowed to pass through the third quadrupole, also acting as a mass filter, and on to the detector [35]. By varying the energy applied to the second quadrupole, the number of collisions can be regulated, and the degree of fragmentation controlled. Lower collision energy typically generates little or no fragmentation of the precursor ions whereas higher collision energies can result in ion cleavage and molecular rearrangements [46]. Figure 1-12 shows a simplified diagram for the precursor selection, fragmentation and product ion selection employed in triple quadrupole multiple reaction monitoring.

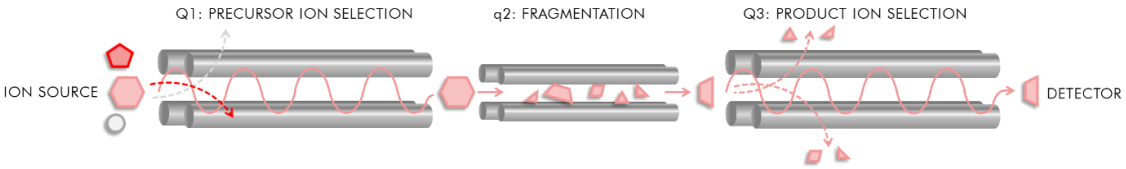


Figure 1-12. Triple quadrupole operating in MRM mode. *Ions exit the ion source and enter the first quadrupole, Q1, where only ions of a specified m/z develop stable trajectories and thus move on to the second, non-resolving, quadrupole q2. Here, the ions are fragmented by application of energy and collisions with an inert gas. From q2, the ions move to the third quadrupole, Q3, where a mass filter is applied, thereby only allowing ion fragments of a specified m/z to develop stable trajectories and finally reach the detector.*

MRM allows for superior sensitivity and specificity, which can be tailored to fit a variety of different purposes. It is also easily translatable and can be used in clinical environments. A triple quadrupole's mass resolution is lower than that of the high resolution instruments previously described, typically around or less than 0.1 Dalton [47], but as the technique is

highly selective due to the chromatographic separation and the MRM mode, high resolution is generally not required.

In the targeted experiments performed in this thesis, a UPLC system (Waters UPLC Acquity) was coupled to a triple quadrupole mass spectrometer (Waters Xevo TQ-S) operating in positive MRM mode.

1.3 PROTEOMICS BY MASS SPECTROMETRY

Mass spectrometry has been utilised for protein and peptide sequencing since the 1960s [48]. The two predominant approaches in mass spectrometric protein analyses are "top-down" and "bottom-up" proteomics. In the top-down methodology, intact proteins are introduced to the mass spectrometer's gas phase and fragmented, whereas in bottom-up proteomics, proteins are digested into peptides before entering the mass spectrometer [49]. In the work presented in this thesis, we have used a predominantly bottom-up proteomics approach, therefore this is the methodology which will be focused on.

1.3.1 *Bottom-up discovery proteomics*

Discovery proteomics can be defined as the global characterising of the entire protein content in a tissue, bio-fluid, cell line or organism [50]. The use of highly specific proteases to cleave a protein into smaller fragments, peptides, allows for the peptides to be separated chromatographically before entering the MS, thereby reducing the complexity of the eluting species. Cleavage of proteins can be performed with multiple types of reagents and enzymes, but trypsin is arguably the most common. Trypsin cleaves amino acid sequences after lysine and arginine unless any of them are followed by a proline, an exception known as the Keil rule [51].

In the discovery study experiments performed in this thesis, an acquisition mode known as "MS^E" was utilised. This method rapidly alternates between low and high energy applied to the collision cell and results in spectra from both the non-fragmented precursor peptide and its fragments being acquired simultaneously, thereby making it especially suitable for peptide sequencing [52, 53]. Post-acquisition, the low and high energy spectra are aligned by retention and IMS drift times, thus matching the precursor ions with their fragments [53]. In practical terms, this means that information about a peptide's retention time, molecular mass and charge, and amino acid sequence is obtained.

Figure 1-13 gives an example of the non-fragmented and fragmented spectra obtained from a peak eluting at a specific retention time. The spectra contain structural information

about the peptides and make it possible to deduce the amino acid sequence. As the data is processed by application-specific software, the fragments belonging to a certain peptide are collated and the protein from which the peptides are originated is determined using sequencing databases.

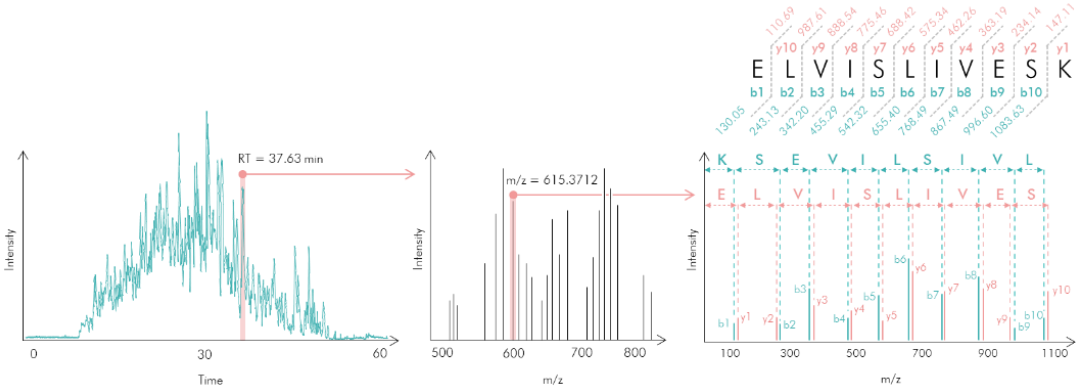


Figure 1-13. MS (middle) and MS/MS (right) spectra of the peptide ELVISLIVESK. *The MS spectrum shows the doubly charged precursor ion with m/z 615.3712 and the MSMS spectrum shows all the possible fragmented y and b product ions and their m/z.*

1.3.2 *Targeted proteomics*

Targeted proteomics has been around as an analytical technique since the mid-1970s [48] and was awarded the journal Nature's "Method of the year" in 2012 [54]. In targeted proteomics, the procedure differs significantly from the discovery approach. Here, specific peptides are chosen to represent the protein in question. Synthetic peptide standards are generally used to determine optimal analytical conditions such as the ideal ion fragments to monitor and the collision energy to apply to the non-resolving quadrupole. The standards are moreover used to develop a suitable gradient for the LC system and to determine retention times. Advantages of targeted proteomics include shorter analysis times, translatability between labs, the option of absolute quantitation, and the possibility of highly specific and tailored analyses [54].

The most common choice of analysis method for targeted studies is multiple reaction monitoring as it allows for highly specific monitoring of the transitions between precursor and product ions. It is further common to develop "multiplexed" methods, where several peptides are measured in the same run. The sensitivity of multiplexed methods can be improved by timing the different MRM functions so that they only record data within the time window of the elution of the peptides.

1.3.3 *Proteomic biomarkers - from discovery to a clinical test*

The process of finding and translating new biomarkers for novel clinical tests is a costly, complicated and time consuming process [11]. Biomarker studies are often interdisciplinary as expertise in several fields is required [55], therefore it frequently involves a collaboration between clinicians, analytical chemists and bioinformaticians.

The first step of the biomarker study involves identifying a suitable discovery cohort for the disease in question and matching it with appropriate controls. Any confounding factors, such as differences in age, sex or collection site between the disease and control groups, should be avoided. Secondly, the sample preparation and instrumental method for the discovery study need to be chosen with care to enable the detection of the largest possible number of analytes. After the data collection, appropriate statistical and modelling methods must be selected for the data analysis. Once putative biomarkers have been identified in the discovery phase, the next step is to validate the findings in a targeted test. This involves finding suitable peptides to represent the proteins, ordering peptide standards, finding the best ion fragments and ideal MRM settings (called "tuning" of the peptides), developing an LC gradient which efficiently separates the analytes, and finally merging the mass spectrometric and chromatographic parameters into a targeted, multiplexed LCMS method. Also in the validation study, great care needs to be taken when selecting the patient cohort. After instrumental analysis and quantitation of the analytes, the validity of the putative biomarkers can be evaluated in statistical tests. They can also be modelled by machine learning algorithms to assess their predictive ability. It is common to utilise a panel of proteins which together can differentiate between disease and control. After the validation study, the proteins demonstrating differential expression between the groups, or predictive ability, are kept in the assay while analytes showing no discriminating capacity are excluded. The refined assay can thereafter be evaluated clinically, to ascertain if it is feasible to utilise as a clinical test. Figure 1-14 shows a detailed illustration of the workflow and the steps involved.

Figure 1-14. Detailed proteomics biomarker discovery workflow. *The workflow includes proteomic discovery studies to find putative biomarkers in a small, deeply phenotyped sample set, followed by validation of the putative biomarkers in a larger set of samples and finally clinical evaluation of the biomarkers in a large set of samples. The important steps within and between the three phases are described.*

1.3.4 *Challenges involved in proteomic biomarker discovery studies*

There are several challenges involved in the biomarker discovery workflow, and especially in the discovery phase. In discovery proteomics, the aim is to achieve the greatest coverage possible of all proteins present in a sample, thereby placing high demands on both the sample preparation and the analytical method.

1.3.4.1 *Dynamic range of blood-based samples*

Different sample matrices pose different challenges in regard to sample preparation, with plasma and serum being among the more complex. One of the major challenges in blood-based proteomics biomarker discovery is the dynamic range of protein concentrations. The difference between the most abundant proteins - albumin, immunoglobulins and transferrin, and low abundant proteins such as interleukins and cytokines are several

orders of magnitude [56]. In addition, albumin is a large protein, weighing 69 kDa. When comparing this to a small and low abundant protein such as a chemokine, weighing 8-10 kDa the number of moles of tryptic peptides produced during digestion will differ significantly (illustrated in Figure 1-15). As a modern mass spectrometer's normal operating range can detect around four orders of magnitude of difference in abundance, detection of low-abundant species present in a sample becomes a challenging task. Furthermore, the diversity of the high-abundant proteins is minimal although they represent around 95% of the total mass of plasma proteins, meaning that a limited number of proteins make up the majority of the total protein mass [57]. One way of getting around this problem is to deplete the sample of abundant proteins using antibodies or immuno-depletion, selectively removing the more abundant proteins thereby reducing the dynamic range, de-complexing the sample and improving the possibility to detect low-abundant species. The depletion techniques are based on antibodies selectively pulling out certain highly abundant housekeeping proteins, leaving the low abundant proteins more readily available for detection. The depleted proteins are commonly, all or a subset of, albumin, immunoglobulins, fibrinogen, transferrin, alpha-1-antitrypsin, alpha-1-acid glycoprotein, alpha-2-macroglobulin, apolipoprotein A1 and A2, and haptoglobin. The enrichment strategies work by utilising beads with a number of binding sites for each protein, effectively meaning that high abundant proteins will be diluted, and low abundant proteins will be concentrated, thus making the proteins present in a sample approach an equimolar concentration. There is a wide range of commercially available products for depletion and enrichment.

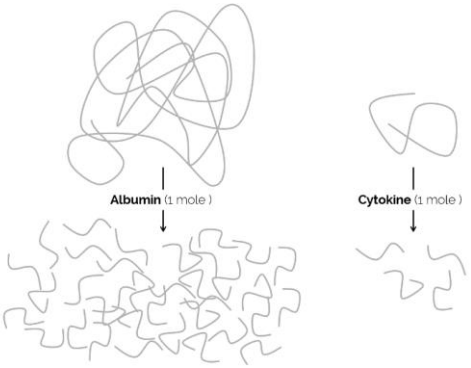


Figure 1-15. Equimolar digestion of albumin and a small cytokine, illustrating the difference in number of peptides available.

1.3.4.2 *Instrumental drift*

The analysis of a sample in the discovery phase is generally a relatively lengthy process due to the extensive fractionation and long LC gradient. To achieve the greatest possible protein coverage, long run times are necessary, but they do pose technical challenges. As a mass spectrometer is operated for long periods of time, parts become contaminated and sensitivity decreases, meaning that a low-abundant peptide possible to detect in the beginning of a run may not be possible to detect towards the end of a run. The instrumental drift is mainly caused by samples coming into direct contact with components of the mass spectrometer. Experiments have shown that the changes in

response are generally non-linear over time and not comparable between different peaks [58]. Strategies to overcome these problems include quality control sample-based correction algorithms to remove variation in the data related to instrumental drift post analysis and breakdown of sample sets into smaller batches [59]. Although the correction algorithm-based strategies are useful, they do not solve the problem altogether since some types of variation, such as severe drops in sensitivity, cannot be corrected for afterwards. Furthermore, addition of fractionated quality control samples to a run would prolong the run time, thus introducing even more drift. Since instrumental conditions by necessity need to be kept the same during an experiment in order to allow for a non-biased comparison of the results in the end, there is only so much that can be done. One crucial part of the experimental set-up is therefore the run order, which must be done in a controlled randomised fashion thereby reducing the risk of instrumental drift being interpreted as biological difference between the samples [60]. In the experiments performed in this work, the samples were randomised using a "constrained" randomising strategy, where paired samples were kept as a unit and randomised within and between the pairs. In the experiments where the samples were not paired, the run order was randomised within and between the sample groups.

1.4 MACHINE LEARNING

Machine learning is a discipline in the intersection between statistics and computer science. It has a wide range of applications, from speech recognition to modelling of biological systems. In the machine learning methodology, computer algorithms find patterns in data, predict continuous or class outcomes and select relevant variables to explain a phenomenon. Artificial-intelligence and machine learning are powerful tools increasingly used to model medical data, utilising prior knowledge to predict future outcomes [61, 62].

Machine learning algorithms can broadly be divided into unsupervised and supervised methodologies [62]. In the unsupervised methodology, models are constructed based on input data ($\mathbf{X}$) only, without any definition of a potential output ($\mathbf{Y}$). The aim is often to find clusters and structures in the data, and/or dimensional reduction into a smaller number of latent variables. There are numerous unsupervised machine learning algorithms, such as PCA, which is described in section 1.4.1. The general methodology for supervised strategies involves the training of a model using a matrix of independent variables ($\mathbf{X}$) and one or several dependent variables ($\mathbf{Y}$), allowing the model to learn how to best model the data, while generalising well to be able to predict new data. There

are many supervised algorithms, such as (O)PLS and (O)PLS-DA, Ridge regression, linear discriminant analysis and support vector machine, all described in sections 1.4.2 - 1.4.6. Machine learning is a relatively young field under rapid development. Here, the models used in this thesis are introduced.

1.4.1 *Principal component analysis*

Principal component analysis (PCA) can be considered the "bread and butter" of many data analysis toolboxes. It is an unsupervised machine learning model, and a useful instrument for discerning major trends, groups, and outliers in the data. It models co-variation between the variables (e.g., protein expression) and expresses this as a score, for each object (or sample) in function of sets of weighted variables (the loadings).

PCA was developed in 1901 by the English statistician and founder of the Department of Applied Statistics at University College London, Karl Pearson [63, 64]. PCA is a projection method that summarises the systematic variation in a matrix **X** into a low dimensional model hyperplane built up by latent variables or principal components (PCs). The largest variation in the data is described by the first PC, while the second largest variation is described by the second PC, and so on. A PC is a bi-linear decomposition consisting of two separate vectors. The scores (T) approximate the observations (or samples) in **X** and a score scatter plot will reveal trends, groups, and outliers in the samples. Observations in close proximity to each other will have similar properties and analogously, observations far from each other will have dissimilar properties. The loadings (P) describe the relation between the variables in **X**. A loadings scatter plot will show the weight of individual variables in each component and can be related to the score scatter plot to reveal which variables are responsible for the variation in the observations. A sample with a high score value in one component means that the sample has high levels on the variables with strong weight on that component, and vice-versa [65, 66].

Figure 1-16 shows a simple example of PCA applied to a dataset containing quantities of foods consumed in different European countries. Shown in the figure are the scores - describing the samples (countries), and the loadings – describing the variables (food quantities). It can be seen in the scores that Portugal, Italy, Austria and Spain are separated horizontally, along the first principal component, from England, Sweden, Holland and Denmark. Interpreting why the countries separate, we investigate the loadings of the first principal component. There, it can be seen that "Garlic" and "Olive oil" are on the left-hand side and therefore, these foods are consumed in larger quantities in Portugal, Italy, Austria

and Spain. Analogously, "Tinned fruit", "Sweetener", "Tinned soup" and "Tea", which are found on the right-hand side of the plot, are consumed in higher quantities in England, Sweden, Holland and Denmark. Vertically - along the second principal component - France, England, Ireland and Switzerland separate from Sweden, Finland, Denmark and Norway. To understand why, the loadings from the second principal component are investigated. In the second principal component's loadings, it can be noted that "Instant coffee", "Powder soup" and "Yoghurt" are consumed in higher quantities in France, England, Ireland and Switzerland, while higher quantities of "Crisp bread", "Frozen fish", "Frozen vegetables" and "Ground coffee" are consumed in Sweden, Finland, Denmark and Norway.

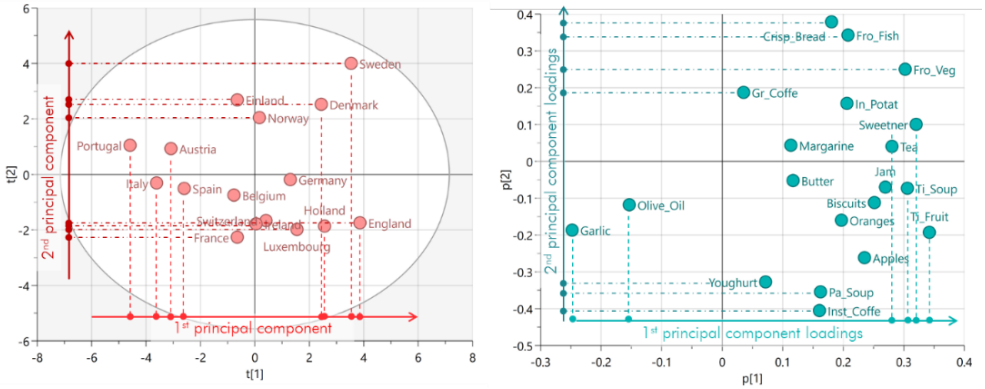


Figure 1-16. PCA scores and loadings from an example comparing the consumption of different foods between European countries. *The scores (left) show the first and second principal components (t[1] versus t[2]), showing which countries are more similar and dissimilar in their food consumption. The loadings plot (right) show the first and second principal components' loadings (p[1] versus p[2]), where it can be seen which foods are more related to which countries.*

1.4.2 *(Orthogonal) partial least squares projections to latent structures*

Partial least squares, or projections to latent structures (PLS), is the regression extension of PCA and useful when the data matrix of independent variables, **X**, can be related to a response matrix or vector of dependent variable(s), **Y**, containing quantitative information. PLS is a supervised machine learning algorithm developed by the Swedish-Norwegian statistician Herman O.A. Wold and his son, Svante Wold [67]. PLS is a multivariate regression model relating **X** to **Y** by partial least squares. The scores in PLS are representations of the observations in the **X** matrix correlated to the **Y** matrix. To visualise the variables' contribution to the PLS model, the weights for **X** and **Y** can be viewed together. An extension of PLS is orthogonal PLS (OPLS), developed by the Swedish chemometrician Johan Trygg in 2002 [68]. In OPLS, the variation in **X** is divided into two parts; one that is related to **Y**, called predictive variation and one that is orthogonal and not related to **Y**, called orthogonal variation [65, 69]. OPLS and PLS models have the same predictive ability, but OPLS has advantages for interpretability.

PLS and OPLS are very useful modelling strategies when relating independent variables to a response, or when determining if there is a relationship between for instance age and protein expression. Figure 1-17 shows an example of a proteomics dataset modelled with OPLS to determine the relationship between age and the protein expression. In the scores, the horizontal predictive principal component shows how the samples relate to age. Observing the predictive component's loadings, the levels of proteins closer to the y-variable, "Age" increase with age, while the opposite is observed for variables in the opposite side to the "Age" variable. Vertically, the orthogonal principal component models variation not correlated with age. A significant benefit of the orthogonality between the principal components is that model interpretation is simplified since it is possible to distinguish between structured variation correlated to a response as well as uncorrelated structure variation. For this reason, the predictive component's loadings are often represented in a bar plot where only the variation related to the y-variable is shown.

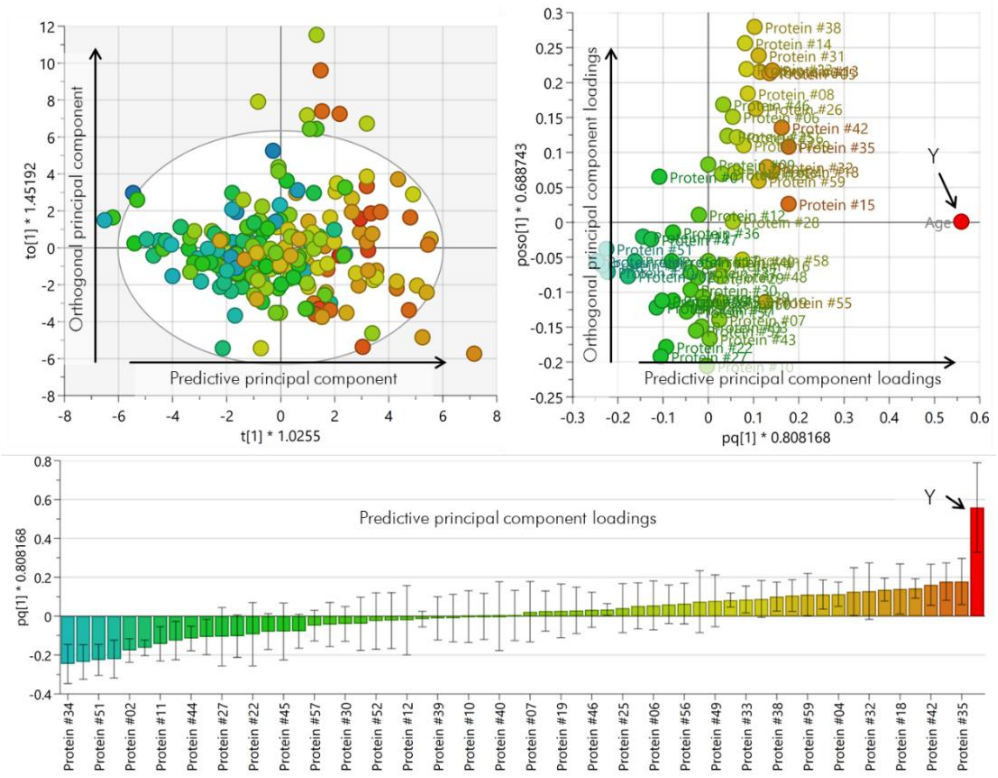


Figure 1-17. OPLS scores and loadings from an example relating protein expression to age. *The scores (top left) show the predictive principal component (t[1]) on the x-axis, describing the relation to age, and the orthogonal principal component (to[1]) on the y-axis, describing variation in the samples not related to age. The loadings plot (top right) shows the predictive component's loadings (pq[1]) on the x-axis, representing the correlation between the proteins and age, and the orthogonal component's loadings (poso[1]), which represents variation in the protein expression not related to age. The predictive component's loadings are moreover presented as a bar plot (bottom), where the proteins on the righthand side, closer to the bar representing the y-variable, are positively correlated with age and proteins on the lefthand side are negatively correlated with age. The error bars are the standard error, obtained by calculating multiple models using different subsets of the data. The proteins with the largest loadings values have the largest impact on the model, while proteins with a standard error crossing zero are not significant to modelling age. The scores are coloured continuously according to age, and the loadings according to correlation with the y-variable.*

1.4.3 *Discriminant (O)PLS*

Discriminant analysis PLS models (PLS-DA) are used when the observations can be divided into discrete classes. In PLS-DA, the response matrix **Y** is of qualitative nature (binary vectors 0/1 for each class). Moreover, the discriminant models can be used to predict the outcomes of new samples. As in the regular (O)PLS, the variation in **X** can be divided into components, or latent variables, related and unrelated to **Y** (OPLS-DA) [65, 69].

Figure 1-18 shows an example of the European food consumption dataset from section 1.4.1, modelled by OPLS-DA with the countries Portugal, Spain, Italy and France set as class ("South") and the countries Sweden, Norway, Denmark and Finland set as class ("North"). All other countries were excluded from the dataset. The predictive latent variable scores show how the samples separate according to the classification "North" and "South", and the orthogonal principal component's scores show how the samples distribute within the classes according to variation not specific for the discrimination of "North"/"South". In the loadings, it is shown that the predictive component separates the samples according to higher consumption of "Garlic", "Olive oil" and "Instant coffee" for the countries belonging to the south-class, and higher consumption of "Sweetener", "Tea", "Ground coffee" and "Crisp bread", among others, in the north-class. Just as in OPLS, it is common to represent the predictive component's loadings in a bar plot where only the variation related to the class separation is shown.

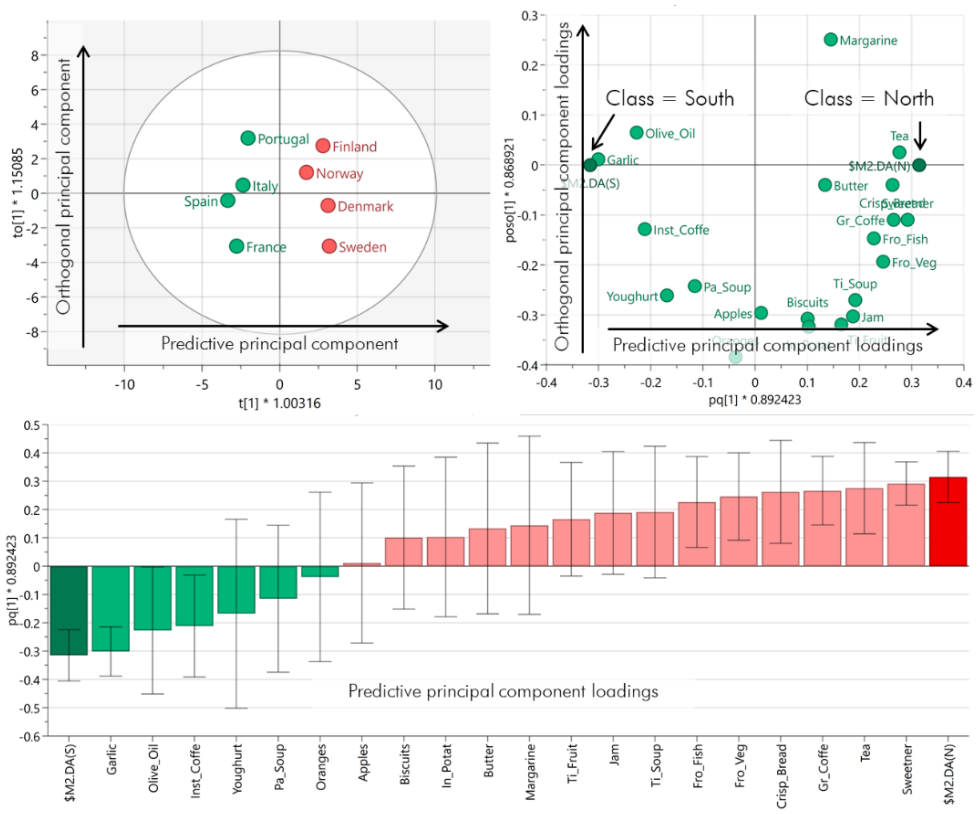


Figure 1-18. OPLS-DA scores and loadings from an example of northern and southern European countries modelled as classes based on the consumption of different foods. *The scores (top left) show the predictive principal component (t[1]) on the x-axis, describing the variation related to the classes "North" and "South", and the orthogonal principal component (to[1]) on the y-axis, describing variation in the samples not related to the classes. The loadings plot (top right) shows the predictive component's loadings (pq[1]) on the x-axis, representing the correlation between the different food quantities and the classes "North" and "South". The orthogonal component's loadings (poso[1]) represents variation in food consumption not related to the classes. The predictive component's loadings are moreover presented as a bar plot (bottom), where the different foods are in closer proximity to the class where their values are highest, "South" to the left in green, and "North" to the right in red. The error bars are the standard error obtained through cross-validation. The variables with the largest loadings values have the largest impact on the class separation, while variables with a standard error crossing zero do not contribute to the separation of the classes.*

1.4.4 *Ridge regression*

Ridge regression was introduced by Art Hoerl and Robert Kennard in 1970 [70]. It is a methodology based on ordinary least squares regression, thus the values of the regression coefficients (ß) for each variable are found by minimising the residual sum of squares. Additionally, it addresses the problems of collinearity between variables and of overfitting by introducing a regularisation parameter. The regularisation is introduced by what is known as the "L2 penalty". It shrinks the beta coefficients of the predictors with little influence in the model towards, but not to, zero [71]. The regularisation is complex and the feature importance ranking for a model is not straight forward. For this reason, it is required to optimise the model, ideally using a cross validation approach to select the settings which minimise the sum of squares of the residuals. There are extensions of Ridge regression, such as Lasso [72] and Elastic Net [73], where the regularisation is different. In

Lasso, beta coefficients of variables with little importance are shrunk to zero (which also makes Lasso an efficient feature selection method), while Elastic Net utilises the regularisation of both Ridge regression and Lasso.

1.4.5 *Linear discriminant analysis*

Linear discriminant analysis (LDA) is a supervised machine learning method, originally developed by Ronald Fisher in 1936 [74]. It uses linear combinations of variables to explain the data given a certain output and thus falls under the dimensional reduction strategies umbrella, like PCA. LDA finds the dimension in which the projected discriminant classes have the maximum between-class separation relative to the minimum within-class variance [75]. Firstly, LDA computes the between-class variance (S_B), and the within-class variances (S_W). From the between- and within-class variances, transformation matrices are calculated, and the eigenvectors and eigenvalues from the transformation matrices are utilised for projecting the data, aiming to reduce the variation within the classes while maximising the variation between them [76]. New observations can be classified as belonging to one of the classes depending on where in the model's space they are projected. Figure 1-19 shows a simplified illustration of data from two classes which cannot be separated by the variables x1 and x2. After LDA projection, the within-class variances have been minimised, and the classes can be separated in the LDA space.

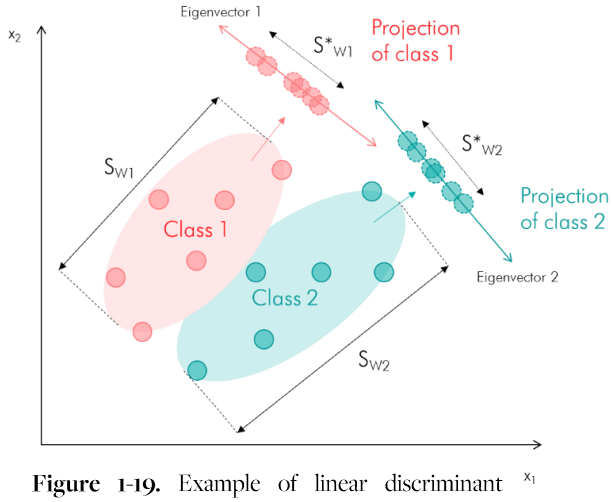


Figure 1-19. Example of linear discriminant analysis, showing the effect of projection for data from two classes. *The figure shows classes 1 (■) and 2 (■) for the variables x1 and x2 and their within-class variances S_{W1} and S_{W2} for the non-projected data, and the classes projected on eigenvectors 1 and 2, the within-class variances S^*_{W1} and S^*_{W2} for the transformed data. As can be noted, the variances of the projected classes have been minimised, and they can be separated in the LDA space.*

1.4.6 *Support vector machine*

Support vector machines (SVM) are supervised machine learning models, learning the best algorithm to classify data. The algorithm was originally developed by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963 [77]. SVM uses the training data to build an algorithm which maps points in space and utilises a hyperplane to maximise the separation between classes [78]. The mathematical foundations of SVMs are complex and increase in intricacy with the number of variables. In simplified terms, the algorithm bases

its decision function on a hyperplane which maximises the distance between the support vectors of the discriminant classes. The margins of the hyperplane are determined by how strictly the limits for misclassification are set, where a larger margin (called a "soft margin") will render the model more generalisability but also more erroneous classifications [79]. This is generally the preferred scenario, as a too strictly set margin may lead to overfitting of the model, thereby making it difficult to predict new data. The misclassifications generated by the soft margins can be restrained by a user-specific parameter, controlling the number of permitted misclassifications during the model's training. After the SVM model has been built, trained, and optimised, new observations can be predicted as belonging to one class or the other depending on which side of the hyperplane they are mapped to. Figure 1-20 shows an illustration of the decision boundary in SVM, where the classes 1 and 2 cannot be separated by either of the variables x1 and x2 but can be separated by the SVM's hyperplane.

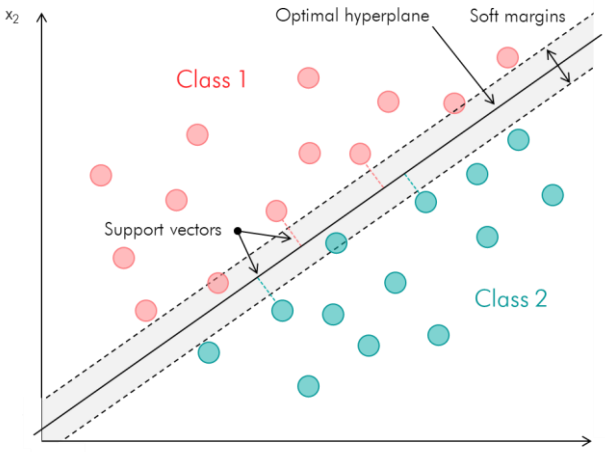


Figure 1-20. Illustration of the decision function in support vector machines. *The figure shows classes 1 and 2, not separable in either of the vectors x1 or x2, but separated by the SVM's optimal hyperplane. The soft margins of the hyperplane are also shown, allowing a degree of misclassifications in favour of generalisability of the model.*

1.4.7 *Additional tools*

1.4.7.1 *Feature selection*

In machine learning modelling, feature selection are methods that reduce the number of independent variables in a matrix X included in a model by selecting variables related to the dependent variable(s) Y, while leaving non-descriptive variables out. The rationale behind the process is to remove irrelevant and/or redundant variables, while simultaneously reducing the model's complexity, minimising its noise and curtailing the risk of multicollinearity and overfitting [80]. In the feature selection process, a subset of features from the original dataset are kept in the model, while redundant features are excluded [81]. There are numerous strategies to perform feature selection, including recursive feature elimination, Lasso, Elastic Net, and receiver operating characteristic (ROC) curve.

1.4.7.2 *Cross-validation*

Cross-validation (CV) is a model validation methodology used to assess how well a model will perform when inserting new observations. In practice, cross-validation is performed by evaluating the results of several models built with different fractions of the data. The held-out fractions of data are predicted in the respective models, allowing for evaluation of the model's performance [82]. There are several ways of partitioning data for cross-validation, examples including:

- Leave-one-out cross-validation: Here, models are built using the dataset while leaving one observation out and then evaluating its predicted result versus the actual value [83]. Consequently, the number of models to build will be equal to the number of observations, making leave-one-out CV useful for smaller datasets.
- Leave-p-out cross-validation: At each cross-validation round, p observations are left out and a model is built using the remaining data. The p observations are then predicted, and the process is repeated until all possible combinations of p from the original data have been predicted [84]. Although a thorough CV methodology, the leave-p-out CV quickly becomes computationally expensive and may therefore not be feasible for larger datasets.
- k-fold cross-validation: The dataset is partitioned into k equally sized subparts, building models with all data but one part which is predicted, and repeating the procedure until all k parts have been predicted [85]. k-fold CV can also be stratified, where roughly equal proportions of the classes in the model are used in each k-fold [86]. Stratified k-fold CV is thus useful for discriminant class models, as it ensures that the model is trained and tested using data from the different classes. k-fold CV is computationally inexpensive and allows for a systematic cross validation of all samples, making it an attractive choice for larger datasets.

1.5 PARKINSON'S DISEASE

Parkinson's disease (PD) is a progressive neurological disorder characterised by motor dysfunction resulting in bradykinesia, rigidity, tremor, and postural instability. Non-motor symptoms are common, including apathy, depression, mood disturbances, anxiety, sleep problems, constipation, loss of smell and taste and dementia [87, 88]. Even though it is possible to reduce symptoms by treatment, PD remains a progressive disease [89, 90]. PD is the second most common neurodegenerative disorder after Alzheimer's disease (AD). After the age of 65, the prevalence of PD is approximately 1% and after the age of 80 it rises to more than 3%. Sporadic PD is more common than heritable forms and makes up approximately 90% of the cases. In most populations, PD is roughly twice as common in men as in women. Risk factors of idiopathic PD may include external elements such as

head trauma and exposure to pesticides, whereas certain habits such as smoking and coffee drinking appear to be protective [91, 92]. One of the earliest manifestations of PD is rapid eye movement sleep behaviour disorder (RBD) which can occur long before any motor symptoms manifest [93].

PD was first described around two hundred years ago by the English surgeon and apothecary James Parkinson in the work "An Essay on the Shaking Palsy", 1817 [94]. Parkinson's observations led him to suggest that PD is caused by lesions in the spinal cord. Today we know that the neuropathology in most cases is characterised by selective loss of dopaminergic neurons in the substantia nigra pars compacta and intracellular accumulation of alpha-synuclein-containing Lewy bodies in the surviving neurons. The loss of dopamine causes a deregulation in the basal ganglia circuits and leads to the cardinal motor symptoms of PD [95]. There are also atypical forms of parkinsonism such as progressive supranuclear palsy (PSP) and multiple system atrophy (MSA). PSP has no alpha-synuclein pathology, but rather tau pathology with neurofibrillary tangles as in AD instead of Lewy bodies. MSA does have alpha-synuclein pathology, but in oligodendroglia instead of neurons [96]. Less common conditions such as Perry syndrome and heritable Parkinsonism caused by different gene mutations also occur with various manifestations.

1.5.1 *Disease mechanisms and gene deficits known to be causes of and risk factors for PD*

As new discoveries are made, it is becoming increasingly clear that Parkinson's disease is a multifactorial and heterogeneous condition with a wide spectrum of different pathologies. The mechanisms behind classical PD have been suggested to include alpha-synuclein proteostasis and degradation, lysosomal autophagy system, mitochondrial dysfunction, activated microglia, inflammation, and oxidative stress.

1.5.1.1 *Alpha-synuclein*

The protein alpha-synuclein is coded for by the gene *SNCA*. It consists of 140 amino acids and is highly abundant in the brain, but also found in other tissues such as red blood cells. Alpha-synuclein has a wide range of functions, playing a role in synaptic activity by vesicle transport and neurotransmitter release, being involved in the regulation of dopamine neurotransmission [97, 98], among others. Alpha-synuclein gains neurotoxic properties when soluble monomers oligomerise and subsequently aggregate to form large, insoluble fibrils known as Lewy bodies. Degradation of alpha-synuclein is governed by the ubiquitin-proteasome system and the lysosomal autophagy system. As alpha-synuclein aggregates, clearance is vital. Dysfunctional lysosomal clearance is believed to play an

important role in PD pathology. A vicious cycle of alpha-synuclein aggregation combined with deficiencies in the ubiquitin-proteasome system and lysosomal autophagy system has been suggested to contribute to intracellular alpha-synuclein accumulation [90]. Further supporting this theory, certain mutations in the genes coding for LRRK2 and GBA are associated with reduced lysosomal autophagy system function.

1.5.1.2 *Oxidative stress*

Reactive oxygen species (ROS) are produced continuously in all tissues. Oxidative stress damage can occur when the activity of ROS clearing antioxidants and produced reactive oxygen species become imbalanced. This state has been proposed as an underlying mechanism of mitochondrial dysfunction and is believed to play a role in the dopaminergic neurotoxicity [99]. One example is mitochondrial complex 1, the activity of which has been found reduced in several tissues from PD patients. It has been suggested that accumulation of alpha-synuclein inside the mitochondria leads to a mitochondrial complex 1 defect and oxidative stress. Nigral dopaminergic neurons may be especially susceptible to oxidative stress as they have long unmyelinated axons with a large number of synapses requiring large amounts of energy.

1.5.1.3 *Neuroinflammation*

Neuroinflammation is suggested to contribute to neuronal degradation, production of pro-inflammatory cytokines and oxidative stress via activation of microglia, the central nervous system's first line of defence. It is proposed that microglia are activated when trying to clear alpha-synuclein aggregates, leading to the production of pro-inflammatory species. This ultimately ends up becoming neurotoxic due to the production of ROS [100]. An increase in pro-inflammatory cytokines such as IL-1β, IL-2, IL-6, TGF-α and TGF-β has been shown in PD brains and there is evidence that the immune response is not restricted to microglia, but can be seen in other tissues as well [101].

1.5.1.4 *Genetics of PD*

A number of gene variants have been found associated with increased risk of PD. Some of the most commonly identified risk factors are the proteins alpha-synuclein, leucine rich repeat kinase 2 (LRRK2) and lysosomal acid glucosylceramidase (GBA). Alpha-synuclein (SNCA) is the major component of Lewy bodies, providing a link between familial and sporadic PD. To date, five different missense mutations in the gene *SNCA* have been discovered to be causal of PD. The neuropathology consists of neuronal degradation in the substantia nigra and extensive Lewy body formation in the cerebral cortex and

brainstem. Mutations in the *LRRK2* gene occur in approximately 4% of the familial PD cases. A number of mutations have been suggested to be strongly associated with PD. The phenotype of LRRK2-linked Parkinsonism is commonly the same between sporadic and familial cases, with Lewy bodies in the brainstem and neuronal loss in the substantia nigra. In some cases, neurofibrillary tangles and neuronal loss are observed without the formation of Lewy bodies. The gene *GBA* codes the enzyme GBA. Homozygous mutations in the gene can result in a deficiency of GBA which causes Gaucher disease [102]. However, it is also the most common genetic risk factor for developing PD, typically with earlier disease onset and atypical clinical manifestations [103].

1.5.2 *Diagnosis and treatment of PD*

In clinical terms, PD is defined by bradykinesia and at least one other motor symptom - tremor or rigidity [104]. The criteria for diagnosing PD were devised by several different organisations, including the United Kingdom Parkinson's Disease Society Brain Bank, Gelb and the International Parkinson and Movement Disorder Society (MDS-PD). It is estimated that roughly 90% of the cases are correctly clinically diagnosed using the MDS-PD criteria [105]. However, total certainty of PD diagnosis during life is currently not possible [104], partly due to the lack of reliable biomarkers for PD, meaning that diagnosis will be affected by a certain bias depending on the criteria used and the symptoms manifested. Hence, there is a need for better and more informative tests for PD.

It is important to note that diagnosis and treatment are limited by the late onset of cardinal symptoms, often occurring after a substantial portion of the dopaminergic neurons have already been lost. Figure 1-21 gives an overview of the timescale and degree of disability for motor and non-motor symptoms [87, 90].

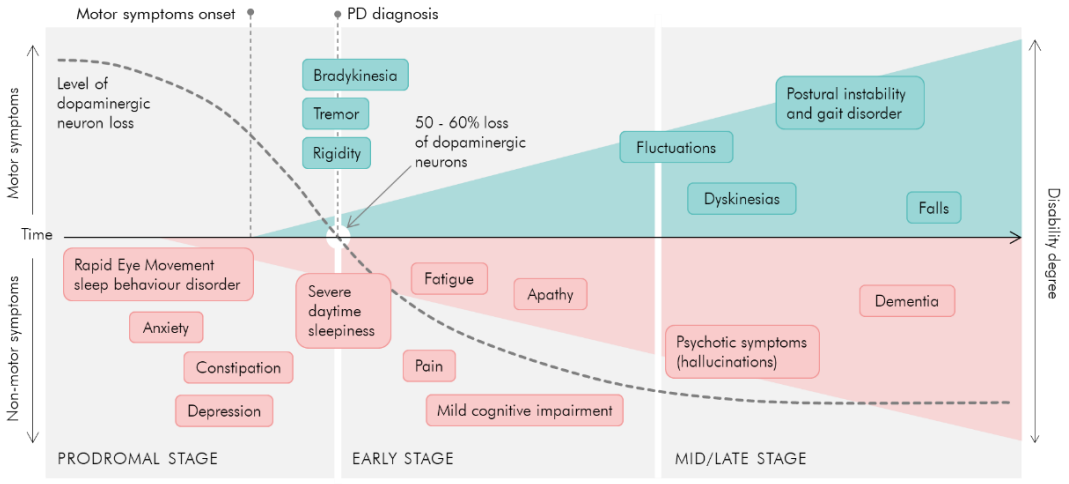


Figure 1-21. The different stages, progression, and clinical manifestations in Parkinson's disease. *Motor- and non-motor symptoms, the level of disability and the level of dopaminergic neuron loss are shown.*

Levodopa is the most common drug used to treat motor symptoms in PD and has been used since the 1960s. It is able to cross the blood brain barrier (BBB) where it is metabolised to dopamine. Levodopa is often combined with other drugs to prolong its clearance half-life which is only one to three hours, and to enhance its efficacy [106]. There are however drawbacks associated with levodopa, such as dyskinesia, motor fluctuations and "off" time (times when the drug has no effect, and the patient may experience motor symptoms such as spasms and tremor). An alternative to levodopa is apomorphine, a dopamine receptor agonist which also reduces motor symptoms. It has been reported that patients treated with apomorphine infusions suffered from less frequent dyskinesia and a reduction in "off" time compared to placebo [107]. Another alternative is deep-brain stimulation, a surgical technique which involves the implantation of electrodes at specific sites of the brain. Deep-brain stimulation has shown an overall anti-Parkinsonian effect and a reduction of dyskinesia and motor fluctuations [108].

1.5.3 *Current status of biomarkers in Parkinson's disease*

There is undoubtedly a need for new and better biomarkers of Parkinson's disease to aid in diagnosis, especially of the early stages before significant loss of dopaminergic neurons has occurred. A look at the current status of biomarkers for Parkinson's disease demonstrate their scarcity. Although great efforts have gone into finding candidates that can diagnose PD, only a few putative biomarkers have been tested in a clinical setting [109].

The status as reported by Cova and Priori in 2018 [110], demonstrates that the majority of the diagnostic markers are clinical. Moreover, most are late indicators, meaning that they do not capture the disease early enough, or also common in the general population (such as constipation and depression), therefore lacking specificity.

1.6 HEALTHY AGEING AND CENTENARIANISM

One of the major risk factors for neurodegenerative diseases, including Parkinson's disease, is ageing. Thus, a study of ageing and in particular longevity in centenarians was included in this work. A better understanding of the divergence between healthy and unhealthy ageing would be beneficial for providing insights into the mechanisms responsible for the development of neurodegenerative disease.

The human population is reaching an increasingly older average age, mainly due to improved life-conditions (access to food, water, less hard labour, antibiotics, disease-modifying medication, and vaccines, to mention a few). According to the World Health

Organization (WHO), there are more than 125 million people aged 80 years or older. WHO further predicts that the world's population over 60 years will more than double between 2015 and 2050 [111]. While this is undoubtedly positive, age-related diseases are the leading cause of mortality and represent a substantial healthcare cost. Diseases where age is a risk factor include, but are not limited to, dementia and diabetes, and cardiovascular, respiratory, and renal diseases [112, 113].

1.6.1 *Successful/healthy ageing*

Rowe and Khan are the first authors to use the term "successful ageing" in their publication with the same name from 1987. They make a distinction between *usual* and *successful* non-pathological ageing, where usual ageing is defined as non-pathological ageing, but with high risk of morbidity and mortality, and successful/healthy ageing is defined as low morbidity/mortality risk and high function. The authors state that successful ageing lies in the intersection between avoiding disease and disability, high cognitive and physical function, and engagement with life [114, 115]. Centenarians, a term used for people who reach the age of 100 years or more, are often used as a model for successful ageing. The prevalence of centenarianism is approximately 0.01 to 0.02% in economically developed countries [116]. In most populations, centenarianism is four to seven times more common in women compared to men [117]. In a study from 2008, the highest concentrations of centenarians were found in Japan, Bulgaria and Italy [118].

Centenarians are described as having a marked delay of life-threatening diseases such as cancer or Alzheimer's disease [119]. Albeit not all centenarians fulfil the definition of successful/healthy ageing towards the very end of their lives, they do in general reach an older age with intact cognitive abilities and better physical health than the general population [120]. Moreover, they have avoided or survived the most severe pathologies causing mortality in the older population. For these reasons, centenarians are a valuable source of information in the study of longevity and successful ageing. It further poses the question - do centenarians reach an advanced age due to a specific genotype or due to lifestyle and environment, or a combination of both?

A vast number of theories describing the process of ageing have emerged throughout the history of gerontology studies. The theories mainly fall under two, often described competing with each other, umbrella-categories; *programmed* or *error/damage*. The programmed theories are essentially evolutionary and based on the idea that ageing occurs according to a biological timetable aimed at promoting successful reproduction and that changes in gene expression affect and regulate maintenance, defence, and repair

processes. The error/damage theories have an environmental footing and highlight external exposure to an organism as capable of inducing cumulative damage, thereby causing ageing [121].

A review by Weinert and Timiras from 2003 describes theories of ageing, classifying them as evolutionary, molecular, cellular or systemic [122]. It is possible, even likely, that ageing cannot be explained by evolutionary or damage/error classified theories exclusively, but rather a combination of both nature and nurture. Ageing research has been undertaken for more than a century, yet a comprehensive theory remains absent. It could be hypothesised that an organism is subject to both types of ageing to cope with external changes - at times prioritising reproduction and at times prioritising the individual. This hypothesis would encourage a more non-biased approach to ageing research, thus opening up for a better chance of elucidating the underlying mechanisms of ageing and allowing for the development of treatments for age-related diseases [123].

1.6.2 *The ageing processes*

As we grow older, a number of processes occur. The decline of muscle mass and increase of inflammation and oxidative stress are often quoted. Skeletal muscle mass starts to decline with a rate of approximately 5% per decade after the age of 40 years. Severe loss of muscle mass and strength, called sarcopenia, occurs in approximately 50% of individuals over the age of 80 years. Skeletal muscle is not only vital for strength; it also functions as a reserve for energy and is important for fatty acid oxidation and carbohydrate metabolism. Therefore, age-related loss of muscle mass can provoke frailty and metabolic side effects [124].

1.6.2.1 *Inflammageing and oxidative stress*

Age-related, low-grade, chronic inflammation (termed "inflammageing") has been reported in several studies, with subjects presenting upregulated levels of the inflammatory blood markers IL-6 and/or CRP [125, 126]. Inflammageing has been identified as a global indicator of poor health status and specifically as a risk for frailty, cardiovascular disease, kidney disease, diabetes, cancer, and sarcopenia. The causes of inflammageing are not fully elucidated, however, research suggests that it may be triggered by a number of factors, including visceral obesity, genetic predisposition, changes in the gut microbiome and chronic infections [127]. Inflammation and free radical damage often go hand in hand, with high levels of oxidative stress correlating with reduced longevity [128]. It has been suggested that oxidative stress in the mitochondria can bring upon a vicious cycle where damaged mitochondria produce even more reactive oxygen species,

thereby augmenting the damage. Oxidative stress can moreover refer to processes occurring outside the mitochondria, such as oxidation of lipids, proteins, and DNA [129]. Oxidative stress is suggested to be involved in the development of obesity and has further been implicated in the neuronal damage observed in neurodegenerative disorders. There is a strong relationship between oxidative stress and inflammation as reactive oxygen species can activate pro-inflammatory species such as NF-κB [130].

1.6.2.2 *Telomere length and mitochondria*

Chromosomes become increasingly damaged with age. Telomeres, which are nucleoproteins sitting at the tips of chromosomes like caps, can alleviate this damage. However, every time DNA replicates, the telomeres get shorter. The enzyme *telomerase* can add telomere units to the existing telomeres, thereby cancelling out some of this effect albeit not completely, the consequence being that telomeres shorten with age. When telomere-length becomes too short, it will trigger DNA damage signals, the protein p53 among others, leading to apoptosis or senescence [131]. A mouse study linked telomere damage and increased levels of p53 to mitochondria dysfunction [132].

In conclusion, ageing is undoubtedly a multi-factorial and highly complex process. The study of centenarians and how their protein expression compares to that of patients suffering from neurodegenerative disease may help us to better understand the divergence between healthy and unhealthy ageing. Centenarians have been shown to develop neurodegenerative diseases at a lower rate than the normal population, and when they do, it is with a markedly delayed onset [133]. It has been proposed that they may have protective mechanisms providing resilience in place [134], and thus the comparison between centenarians and Parkinson's patients may provide additional information and insights relating to both groups.

Materials and methods for sample preparation, instrumental analysis and data analysis

2

2.1 SAMPLES

The sample cohorts included in the studies presented in this thesis were provided through the European Union-funded Horizon 2020 projects PROPAG-AGEING and HUMAN.

2.1.1 *Samples included in the discovery proteomics studies*

In the discovery proteomics study of centenarians, plasma from centenarians and controls were collected by IRCCS Istituto delle Scienze Neurologiche di Bologna, Italy. In the Parkinson's disease discovery studies, two groups of blood-based samples were included (i) plasma from healthy controls and newly diagnosed, treatment-naïve Parkinson's disease patients from the DeNoPa (de novo Parkinson's disease) cohort [135], and (ii) serum from The Swedish Twin Registry [136] from homozygous twins sampled prior to the development of Parkinson's disease in one of the twins from each pair. In the urine discovery study, samples from idiopathic PD patients, controls, symptomatic and asymptomatic LRRK2 variant carriers collected by the National Hospital of Neurosurgery and Neurology, were included. The discovery proteomics samples are summarised in Table 2-1 and described in detail in the Methods and Materials sections of Chapters 4, 5 and 6.

Table 2-1. Summary of the samples included in the discovery proteomic studies of centenarians and Parkinson's disease. *Sample matrix, and the total number of samples included are listed.*

Sample group	Matrix	Number of samples
Centenarians and controls	Plasma	20
Newly diagnosed PD patients and healthy controls	Plasma	20
Homozygous twin pairs discordant for developing PD post sampling	Serum	18
Idiopathic PD patients, symptomatic and asymptomatic LRRK2 mutation carriers, and controls	Urine	31

2.1.2 *Samples included in the targeted proteomics studies*

In the targeted proteomic centenarian study, plasma samples from centenarians, centenarian offspring and controls from IRCCS Istituto delle Scienze Neurologiche di Bologna were included. In the targeted proteomic Parkinson's disease studies, paired plasma and urine from newly diagnosed Parkinson's disease patients, patients with idiopathic rapid eye movement sleep disturbance disorder, patients with other (non-PD) neurological disorders, and healthy controls from the DeNoPa cohort were included. The targeted proteomics samples are summarised in Table 2-2 and described in detail in the Methods and materials sections of Chapters 4, 5 and 6.

Table 2-2. Summary of the samples included in the targeted proteomic studies of centenarians and Parkinson's disease. *Sample matrix, and the total number of samples included are listed.*

Sample group	Matrix	Number of samples
Centenarians, offspring, and controls	Plasma	186
Newly diagnosed PD patients, iRBD patients, non-PD neurological disorders patients, and healthy controls	Plasma	211
Newly diagnosed PD patients, iRBD patients, non-PD neurological disorders patients, and healthy controls	Urine	211

2.2 MODULES IN THE SAMPLE PREPARATION FOR UNTARGETED AND TARGETED PROTEOMICS

Unless stated otherwise, all chemicals and reagents were purchased from Sigma Aldrich.

2.2.1 *Freeze drying*

100 µL MilliQ water was added to precipitated pellet samples while samples already in aqueous solutions were processed directly. The samples were placed on dry ice for a minimum of 20 minutes to ensure that they were frozen completely solid. The samples were freeze dried using a Modulyo freeze dryer from Edwards (Burgess Hill, UK), equipped with a C/RVpro6 vacuum pump from Welch (Fürstenfeldbruck, Germany), set to operate at -40 °C under vacuum.

2.2.2 *Digestion*

The digestion protocol was performed sequentially, in the order of the sections 2.2.2.1 to 2.2.2.4.

2.2.2.1 *Solubilising proteins*

The buffer used to bring the freeze-dried samples' proteins back into solution has two functions - to lyse the cells, making membrane and cytoplasm proteins available, and to solubilise the proteins in the samples. The buffer consisted of 6 M urea, 2 M thiourea and 2% amidosulfobetain-14 in 100 mM Tris, pH 7.8. 20 µL of this digest buffer was added to the freeze-dried samples which were shaken for 60 minutes on an orbital shaker, 1500 rpm.

2.2.2.2 *Reduction of protein disulphide bridges*

In a protein, disulphate bridges are formed between cysteine moieties. To break these bridges and open the protein's conformation up, reducing agents are utilised. Reduction of the disulphate bonds was performed by adding 45 µg dithioerythriol (DTE) to the samples, which were thereafter shaken for 60 minutes on an orbital shaker, 1500 rpm. The

DTE solution was made up fresh, moments prior to usage, as 30 µg/µL in 100 mM Tris, pH 7.8.

2.2.2.3 *Alkylation of reduced cysteine*

After cysteine disulphate bridges have been reduced, thus breaking the disulphate bonds in the proteins, they must be prevented from reforming. This is done by introducing an alkylating reagent that binds covalently to the cysteine's sulphur moiety. Carbamidomethylation was performed by adding 108 µg iodoacetamide (IAA) to the samples, which were thereafter shaken for 45 minutes on an orbital shaker, 1500 rpm, covered from light. The IAA solution was made up fresh, moments prior to usage, as 36 µg/µL in 100 mM Tris, pH 7.8.

2.2.2.4 *Dilution of urea concentration and addition of the digestion enzyme trypsin*

Urea concentrations higher than 1 M hamper the activity of trypsin. To lower the urea concentration, 160 µL Milli-Q water was added to the sample, thereby bringing the concentration below 1 M. To digest the proteins into peptides, 1 µg trypsin gold from Promega (Mannheim, Germany) was added and the samples were incubated for 16 hours in a +37 °C water bath. The trypsin solution was made up fresh, moments prior to usage, as 0.1 µg/µL in 50 mM acetic acid.

2.2.3 *Solid phase extraction*

Solid phase extraction (SPE) was employed to purify the peptides, cleaning the solution of digested peptides from salts and polar low-molecular weight compounds. Depending on the number of samples processed, individual cartridges or 96-well plates were utilised. Prior to SPE, the samples were adjusted to a trifluoroacetic acid (TFA) v/v concentration of 0.1%.

2.2.3.1 *Solid phase extraction using individual cartridges*

100 mg Isolute C_{18} cartridges from Biotage (Uppsala, Sweden) were utilised. The stationary phase of the cartridges was solvated by two 1 mL aliquots of acetonitrile. Each aliquot was eluted off the cartridge by applying gentle positive pressure. Equilibration of the stationary phase was performed by adding two 1 mL aliquots of 0.1% TFA, again each aliquot was eluted by applying gentle positive pressure. The sample was loaded on the cartridge and the peptides bound to the stationary phase. Salts and low-molecular weight compounds were washed off the cartridges by two 1 mL aliquots of 0.1% TFA, each eluted

by applying gentle positive pressure. Finally, the peptides were eluted off the stationary phase by two 250 µL aliquots of 60% acetonitrile, 0.1% TFA. The samples were eluted into 1.5 mL centrifugal tubes from Eppendorf.

2.2.3.2 *Solid phase extraction using 96-well plates*

100 mg Bond Elute C_{18} 96-well plates from Agilent (Santa Clara, CA, USA) were utilised. The stationary phase was solvated by two 1 mL aliquots of acetonitrile. Each aliquot was eluted off the plate by centrifugation on a Sorvall Legend RT centrifuge (Thermo Fisher Scientific, Waltham, Massachusetts, USA) at 44 x *g* for four minutes. Equilibration of the stationary phase was performed by adding two 1 mL aliquots of 0.1% TFA, again each aliquot was eluted by centrifugation at 44 x *g* for four minutes. The samples were loaded to the wells and the peptides bound to the stationary phase during gentle centrifugation at 36 x *g* for five minutes. Salts and low-molecular weight compounds were washed off by adding two 1 mL aliquots of 0.1% TFA, each eluted by centrifugation at 36 x *g* for five minutes. Finally, the peptides were eluted off the stationary phase by two 250 µL aliquots of 60% acetonitrile, 0.1% TFA, the first aliquot eluted by centrifugation at 36 x *g* for three minutes and the second by 64 x *g* for three minutes. The samples were eluted into a 700 µL 96-well sample plate from Agilent.

2.2.3.3 *Evaporation of solvents*

Solvents were evaporated from the SPE-cleaned samples using a vacuum rotator, a Concentrator Plus from Eppendorf (Hamburg, Germany). The samples were sealed and stored at -80 °C until further use.

2.2.4 *Colorimetric peptide assay*

In the instances where the total peptide concentration of a sample was required prior to instrumental analysis, a colorimetric peptide assay was utilised, Pierce Quantitative Colorimetric Peptide Assay from Thermo Fisher Scientific. The samples were first reconstituted in a solution matching the initial conditions of the liquid chromatography gradient. From the reconstituted samples, 5 µL was aliquoted into a new centrifugal tube and 45 µL Milli-Q water was added. 20 µL of the diluted sample was added to a flat-bottom 96-well plate from Costar (Corning, Glendale, Arizona, USA) in duplicates. An eight-point standard curve was added in triplicate. The standard curve consisted of serially diluted digested peptide standard, provided in the kit, ranging from 0 to 1000 µg/mL. A working solution was constructed by mixing the provided reagents A, B and C in the following ratio: 50 volumes A / 48 volumes B / 2 volumes C. 180 µL of the working

solution was added to the samples and the standard curve. The plate was covered with a lid and incubated in room temperature for 30 minutes. The absorbance in each well was read on an Infinite F200 plate reader from Tecan (Männedorf, Switzerland) equipped with a 475 nm absorbance filter. The absorbance measurements of the calibrants were related to their concentrations to construct a calibration curve utilised to determine the total peptide concentrations in the samples.

2.3 SAMPLE PREPARATION FOR DISCOVERY PLASMA/SERUM PROTEOMICS

The sample preparation for discovery plasma/serum proteomics consisted of Top12 depletion followed by freeze-drying, digestion and SPE clean-up. The steps in the process are illustrated in Figure 2-1. The method development performed to optimise the sample preparation workflow is described in Chapter 3, section 3.2.1.

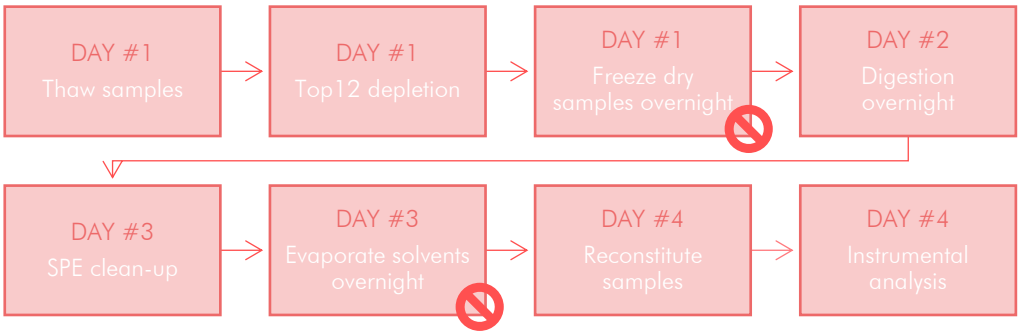


Figure 2-1. Illustration describing the process of preparing plasma samples for discovery proteomics. *The steps include: on day one - thaw the samples and deplete by Top12 followed by freeze drying overnight, on day two - digest the samples and incubate overnight, on day three - SPE clean the samples and evaporate solvents overnight. The stop signs indicate steps after which it is possible to pause the process.*

2.3.1 *Procedure*

The plasma samples were thawed at room temperature. The Top12 depletion cartridges from Thermo Scientific were equilibrated to room temperature for a minimum of one hour. The plasma was vortexed for five seconds and 10 µL was added to the Top12 cartridges which were vortexed briefly to ensure that the resin had not settled at the bottom of the cartridge. The samples were incubated for one hour, allowing the high-abundant species to bind with the antibodies in the resin, under gentle end-to-end rotation. The depleted sample was separated from the resin by centrifugation in room temperature for two minutes at 1000 x *g* using a Micro-centrifuge 5424 R from Eppendorf. The samples were freeze dried overnight as per section 2.2.1, followed by digestion with overnight incubation as per section 2.2.2 and SPE clean-up using cartridges as per section 2.2.3.1 and overnight evaporation of solvents as per section 2.2.3.3. The

depleted, digested and SPE-cleaned samples were reconstituted in 50 µL 3% acetonitrile, 0.1% TFA before instrumental analysis by 2D-LC-MS.

2.4 SAMPLE PREPARATION FOR DISCOVERY URINE PROTEOMICS

The sample preparation for discovery urine proteomics consisted of centrifugation to remove sediments, filtering and concentration of urinary proteins followed by acetone precipitation and freeze-drying, digestion and SPE clean-up. The steps in the process are illustrated in Figure 2-2. The method development performed to optimise the sample preparation workflow is described in Chapter 3, section 3.2.2.

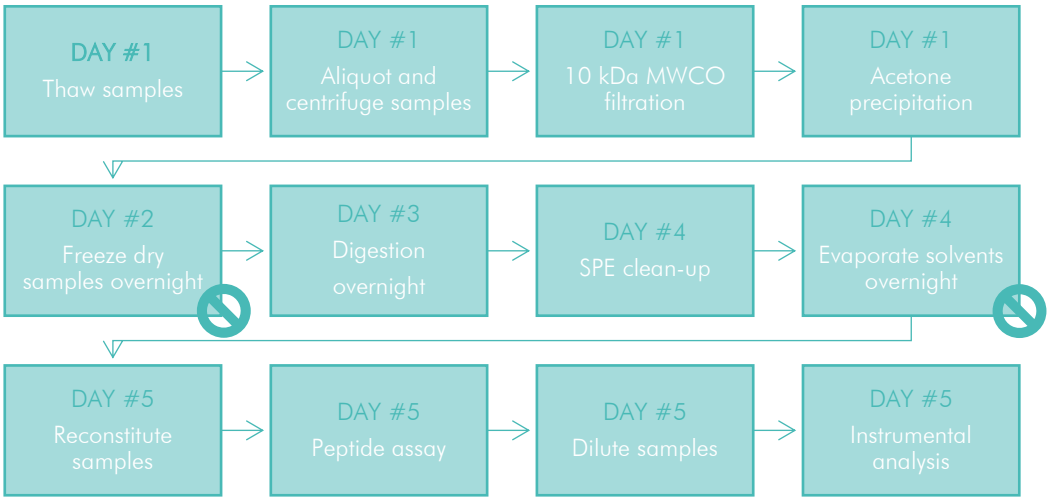


Figure 2-2. Illustration describing the process of preparing urine samples for discovery proteomics. *The steps include: on day one - thaw the samples, aliquot and centrifuge to remove sediments, filter the urine in 10 kDa MWCO filters followed by acetone precipitation overnight and on day two freeze drying overnight, on day three - digest the samples and incubate overnight, on day four - SPE clean the samples and evaporate solvents overnight. On the day of instrumental analysis, the peptide concentration in each sample is determined and the samples diluted to equal peptide concentration. The stop signs indicate steps after which it is possible to pause the process.*

2.4.1 *Procedure*

The urine samples were thawed at room temperature. The samples were vortexed for five seconds and 4 mL of urine was aliquoted into 5 mL centrifuge tubes from Eppendorf. The 4 mL aliquots were centrifuged at room temperature at 3761 x*g* for 30 minutes to separate the urinary sediment from solution using a Sorvall Legend RT centrifuge. 2 mL of the supernatant was transferred to Amicon Ultra-4 10 kDa molecular weight cut-off filters from Merck Millipore (Burlington, Massachusetts, USA) and 2 mL Milli-Q water was added to give a final volume of 4 mL. To concentrate the urinary proteins, the samples were centrifuged at room temperature for one hour at 4444 x*g* (Sorvall Legend RT). The concentrate was transferred to a 1.5 mL centrifuge tube (Eppendorf). To ensure maximum recovery, the filters were washed with 100 µL 50 mM ammonium bicarbonate

which was pooled with the concentrate. 800 µL ice-cold acetone was added to the pooled concentrate and the samples were vortexed for five seconds before overnight incubation in -20 °C. In order to separate the supernatant from the protein pellet, the samples were centrifuged for ten minutes at +4 °C and 16900 x *g* using a Micro-centrifuge 5424 R (Eppendorf). The supernatant was carefully pipetted off and discarded. The pellet was air dried in a fume hood for 20 minutes to evaporate residual acetone. 100 µL Milli-Q water was added to the samples and the protein pellet broken up by vigorous vortexing. The samples were thereafter freeze dried overnight as per section 2.2.1, followed by digestion with overnight incubation as per section 2.2.2, SPE clean-up using cartridges as per section 2.2.3.1, and overnight evaporation of solvents as per section 2.2.3.3. The digested and SPE-cleaned samples were reconstituted in 50 µL 3% acetonitrile, 0.1% TFA and a peptide assay was performed as per section 2.2.4. The peptide concentration in the samples was normalised to 1000 ng/µL before instrumental analysis by 2D-LC-MS.

2.5 INSTRUMENTAL ANALYSIS OF DISCOVERY PROTEOMICS SAMPLES BY 2D NANO-LC IMS MS^E

The instrumental analysis of the discovery cohorts was performed utilising a nano-2D-LC system coupled to a time-of-flight mass spectrometer equipped with ion mobility (IMS) separation. Figure 2-3 shows a diagram of the 2D-LC set-up. The method development performed to optimise the parameters for instrumental analysis is described in Chapter 3, section 3.3.

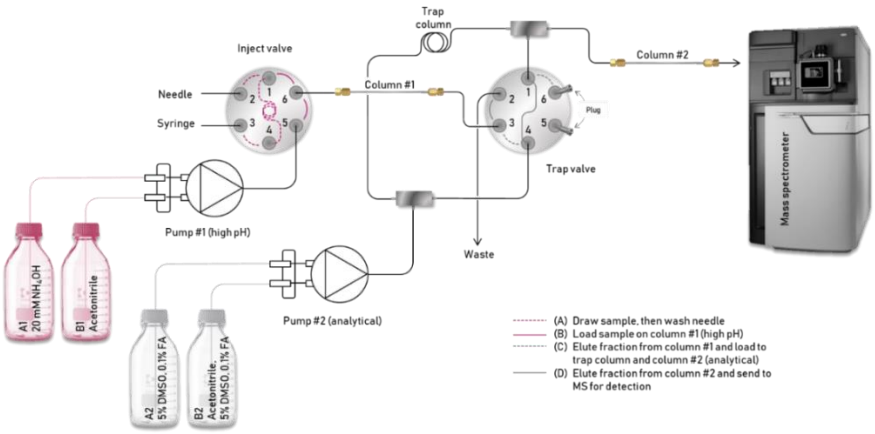


Figure 2-3. Configuration of the 2D-LC fractionation utilised in discovery proteomics. *The diagram shows the set-up of the high pH and analytical pumps and their valve configurations. In configuration (A) of the inject valve, the sample is drawn by the auto sampler after which the sample needle is washed, drawing wash solution through the syringe. In inject valve position (B), the sample is loaded onto column #1 and fraction-wise elution by pump #1, providing high pH mobile phase, commenced. The eluted fraction enters the trap valve and is loaded onto the trap column and subsequently column #2 in configuration (C). Pump #1 is thereafter bypassed, and the fraction eluted from column #2 sent to the mass spectrometer for detection in configuration (D). The events associated with valve configurations (B) - (D) are repeated until all fractions have been eluted from columns #1 and #2 and the analytes have been detected by the mass spectrometer, before the subsequent sample is injected, and the process starts over again.*

2.5.1 *Two-dimensional liquid chromatography separation*

The peptides were separated using a 2D-NanoAquity liquid chromatography system (Waters, Manchester, UK). All samples were fractionated online into ten fractions over a 12-hour period. The fractionation was performed by reversed phase chromatography, utilising the orthogonality in peptide separation between high and low pH mobile phases. The mobile phase in the first, fractionating, chromatographic system consisted of A1: 10 mM ammonium hydroxide titrated to pH 9 with hydrochloric acid (12 M) and B1: acetonitrile. The second, analytically separating, chromatographic system's mobile phase was A2: 5% dimetylsulfoxid (DMSO), 0.1% formic acid, and B2: acetonitrile with 5% DMSO, 0.1% formic acid. 2D-LC fractionation was performed by loading the sample onto a 300 µm x 50 mm, 5 µm Peptide BEH C18 column (Waters). The peptides were thereafter eluted fraction-wise from the first to the second column at a flow rate of 2 µL/min. The initial condition of the gradient elution was 3% B, held over 0.5 minute before linearly increasing the proportion of organic solvent B, fraction per fraction, according to Table 2-3 over 0.5 minute. The conditions thereafter remained static for 4 minutes before returning to the initial conditions over 0.5 minute and equilibration prior to the next elution for 10 minutes. The eluted peptides from the first dimensional column were loaded into a 180 µm x 20 mm, 5 µm Symmetry C_{18} trap column (Waters) before entering the analytical column, a 75 µm x 150 mm, 1.7 µm Peptide BEH C_{18} (Waters). The column temperature was +45 °C. The gradient elution applied to the analytical column started at 3% B and was linearly increased to 40% B over 40 minutes after which it was increased to 85% B over 2 minutes and washed for 2 minutes before returning to initial conditions over 2 minutes followed by equilibration before the subsequent injection for 15 minutes. The eluted peptides were detected using a Synapt-G2-S*i* (Waters) equipped with a nano-electrospray ion source.

Table 2-3. Percentage of eluent B for the first dimension of the 2D-LC fractionation before loading onto the second, analytical system. *The B eluent on the high pH column was acetonitrile and the B eluent on the analytical column was acetonitrile with 0.1% formic acid.*

Fraction number	High-pH column	Analytical column	
	Percentage B at endpoint	Percentage B at midpoint	Percentage B at endpoint
# 1	7.4	5.7	40
# 2	10.8	9.9	40
# 3	12.6	11.9	40
# 4	14.0	13.4	40
# 5	15.3	14.6	40
# 6	16.7	15.9	40
# 7	18.3	17.3	40
# 8	20.4	18.9	40
# 9	23.5	22.0	40
# 10	60.0	42.0	60

2.5.2 *Detection by time-of-flight IMS MS^E mass spectrometry*

The mass spectrometer was a Synapt-G2-Si (Waters). Prior to analysis, a detector set-up in positive mode was performed utilising leu-enkephalin at a concentration of 200 pg/µL, infused at a flow rate of 0.3 µL per minute. The mass spectrometer was thereafter calibrated in positive mode for resolution in the mass range m/z 50 - 2000 utilising [glu1]-fibrinopeptide B with a concentration of 500 fmol/µL, infused at a flow rate of 0.3 µL per minute. At least 13 of the 15 theoretical fragments were required to match for the calibration to be accepted.

Data were acquired in positive MS^E mode from 0 to 60 minutes within the m/z range 50 - 2000. The capillary voltage was set to 3 kV and the source temperature to + 100 °C. The cone gas consisted of nitrogen with a flow of 50 L/h, the desolvation temperature was set to + 200 °C. The purge and desolvation gas consisted of nitrogen, operated at a flow rate of 600 mL/h and 600 L/h respectively. The gas in the IMS cell was helium with a flow rate of 90 mL/h. The low energy acquisition was performed applying a constant collision voltage of 4 V with a 1 second scan time. High energy acquisition was performed by applying a collision energy ramp, from 15 to 40 V, the scan time was 1 second. The lock mass consisted of 500 fmol/µL [glu1]-fibrinopeptide B, continuously infused at a flow rate 0.3 µL/min and acquired every 30 seconds. The doubly charged precursor ion, m/z 785.8426, was utilised for mass correction.

2.6 DATA PROCESSING FOR DISCOVERY PROTEOMICS

2.6.1 *Identification and relative quantitation of proteins*

After acquisition, data were imported to Progenesis QI for proteomics (Waters) and the individual fractions 1 - 10 were individually processed before all results were merged into one experiment. The Ion Accounting workflow was utilised, with UniProt Canonical Human Proteome (exported 2017) as data base. The digestion enzyme was set as trypsin. Carbamidomethyl on cysteines was set as a fixed modification; deamidation of glutamine and asparagine, oxidation of tryptophan and pyrrolidone carboxylic acid on the N-terminus were set as variable modifications. The identification tolerance was restricted to at least two fragments per peptide, three fragments per protein and one peptide per protein. A false discovery rate of 4% or less was accepted. The individual fractions 1 - 10 for each sample were combined in Progenesis, using the multi-fraction experiment workflow. Generally, at least two unique peptides per protein and an identification confidence score

(a Progenesis specific value denoting the identification confidence, where a higher value implies a more certain protein identification) larger than 15 were set as thresholds for classifying a protein as a confident identification. The resulting identifications and relative quantitation were exported to Microsoft Excel.

2.6.2 *Quality control*

The data sets were visualised using the multivariate tool principal component analysis using the software SIMCA, version 15 (Umetrics Sartorius Stedim, Umeå, Sweden) and inspected for issues such as instrumental drift and outliers. In the event that severe instrumental drift was observed, a drift correction filter utilising locally estimated scatterplot smoothing (LOWESS) was applied (strategy described in Chapter 3, section 3.5.2). The variables were tested for normal distribution by D'Agostino's K^2 test. In the cases of non-normal distribution, the variables were transformed to normality using an in-house Box-Cox script written in Python [137]. The data were once more inspected in SIMCA to ensure that eventual corrections had been successful.

2.7 SAMPLE PREPARATION FOR TARGETED PLASMA PROTEOMICS

The sample preparation for targeted plasma proteomics consisted of Top2 depletion followed by freeze-drying, digestion and SPE clean-up. The steps in the process are illustrated in Figure 2-4. The method development performed to optimise the sample preparation workflow is described in Chapter 3, section 3.4.2.

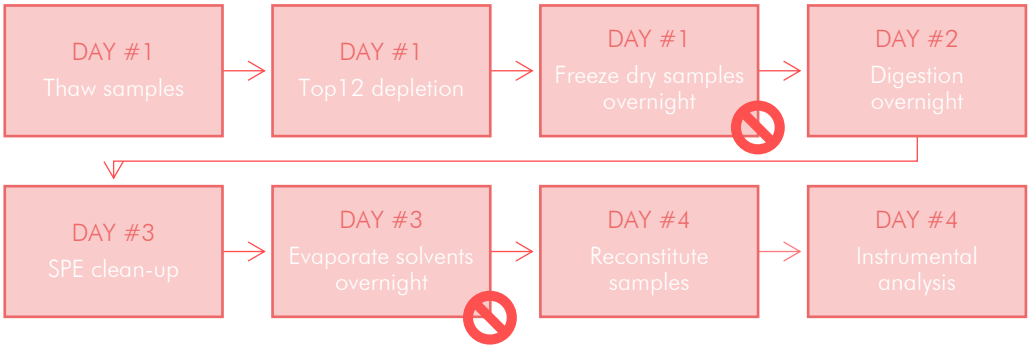


Figure 2-4. Illustration describing the process of preparing plasma samples for targeted proteomics. *The steps include: on day one - thaw the samples and deplete by Top2 followed by freeze drying overnight, on day two - digest the samples and incubate overnight, on day three - SPE clean the samples and evaporate solvents overnight. The stop signs indicate steps after which it is possible to pause the process.*

2.7.1 *Procedure*

The plasma samples were thawed at room temperature. The Top2 depletion cartridges from Thermo Scientific were equilibrated to room temperature for a minimum of one

hour. The plasma was vortexed for five seconds and 10 µL was added to the Top2 cartridges. 150 ng whole protein ENO1 (yeast) was added as an internal standard accounting for sample preparation. The samples were vortexed briefly to ensure that the resin had not settled at the bottom of the cartridge. The samples were incubated for one hour, allowing the high-abundant species to bind with the antibodies in the resin, under gentle end-to-end rotation. The depleted sample was separated from the resin by centrifugation at room temperature for two minutes at 1000 x *g* using a Micro-centrifuge 5424 R from Eppendorf. The samples were freeze dried overnight as per section 2.2.1, followed by digestion with overnight incubation as per section 2.2.2 and SPE clean-up using 96-well plates as per section 2.2.3.2 and overnight evaporation of solvents as per section 2.2.3.3. The depleted, digested and SPE-cleaned samples were reconstituted in 30 µL 3% acetonitrile, 0.1% TFA, containing 0.1 µM of heavy isotope labelled peptides from the proteins ALDOA, C3, GSTO1, RSU1 and TSP1 (full amino acid sequences given in Table 2-4) before instrumental analysis by UPLC-MS/MS.

2.8 SAMPLE PREPARATION FOR TARGETED URINE PROTEOMICS

The sample preparation for targeted urine proteomics consisted of centrifugation to remove sediments, filtering and concentration of urinary proteins followed by acetone precipitation and freeze-drying, digestion and SPE clean-up. The steps in the process are illustrated in Figure 2-5. The rationale behind the sample preparation workflow is described in Chapter 3, section 3.4.3.

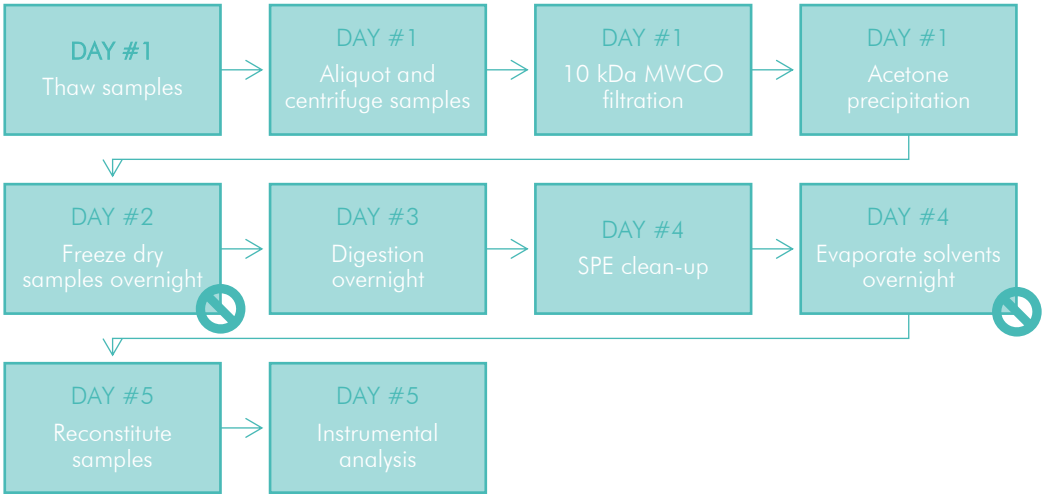


Figure 2-5. Illustration describing the process of preparing urine samples for targeted proteomics. *The steps include: on day one - thaw the samples, aliquot and centrifuge to remove sediments, filter the urine in 10 kDa MWCO filters followed by acetone precipitation overnight and on day two freeze drying overnight, on day three - digest the samples and incubate overnight, on day four - SPE clean the samples and evaporate solvents overnight. The stop signs indicate steps after which it is possible to pause the process.*

2.8.1 *Procedure*

The urine samples were thawed at room temperature. The samples were vortexed for five seconds and 4 mL of urine was aliquoted into 5 mL centrifuge tubes from Eppendorf. The 4 mL aliquots were centrifuged at room temperature at 3761 x *g* for 30 minutes to separate the urinary sediment from solution using a Sorvall Legend RT centrifuge. 3 mL of the supernatant was transferred to Amicon Ultra-4 10 kDa molecular weight cut-off filters from Merck Millipore and 1 mL Milli-Q water was added to give a final volume of 4 mL. 150 ng whole protein ENO1 (yeast) was added as an internal standard accounting for sample preparation. To concentrate the urinary proteins, the samples were centrifuged in room temperature for one hour at 4444 x *g* (Sorvall Legend RT). The concentrate was transferred to a 1.5 mL centrifuge tube (Eppendorf). To ensure maximum recovery, the filters were washed with 100 µL 50 mM ammonium bicarbonate which was pooled with the concentrate. 800 µL ice-cold acetone was added to the pooled concentrate and the samples were vortexed for five seconds before overnight incubation at -20 °C. In order to separate the supernatant from the protein pellet, the samples were centrifuged for ten minutes at +4 °C and 16900 x *g* using a Micro-centrifuge 5424 R (Eppendorf). The supernatant was carefully pipetted off and discarded. The pellet was air dried in a fume hood for 20 minutes to evaporate residual acetone. 100 µL Milli-Q water was added to the samples and the protein pellet broken up by vigorous vortexing. The samples were thereafter freeze dried overnight as per section 2.2.1, followed by digestion with overnight incubation as per section 2.2.2, SPE clean-up using 96-well plates as per section 2.2.3.2 and overnight evaporation of solvents as per section 2.2.3.3. The concentrated, digested and SPE-cleaned urine samples were reconstituted in 50 µL 3% acetonitrile, 0.1% TFA, containing 0.1 µM of heavy isotope labelled peptides from the proteins ALDOA, C3, GSTO1, RSU1 and TSP1 (full amino acid sequences given in Table 2-4) prior to instrumental analysis by UPLC-MS/MS.

2.9 TARGETED INSTRUMENTAL ANALYSIS OF THE VALIDATION SAMPLES BY UPLC-MS/MS

The instrumental analysis of the targeted cohorts was performed utilising a UPLC-LC system coupled to a triple quadrupole mass spectrometer. Figure 2-6 shows a diagram of the UPLC set-up. The development of the targeted multiple reaction monitoring method is described in Chapter 3, section 3.4.1.

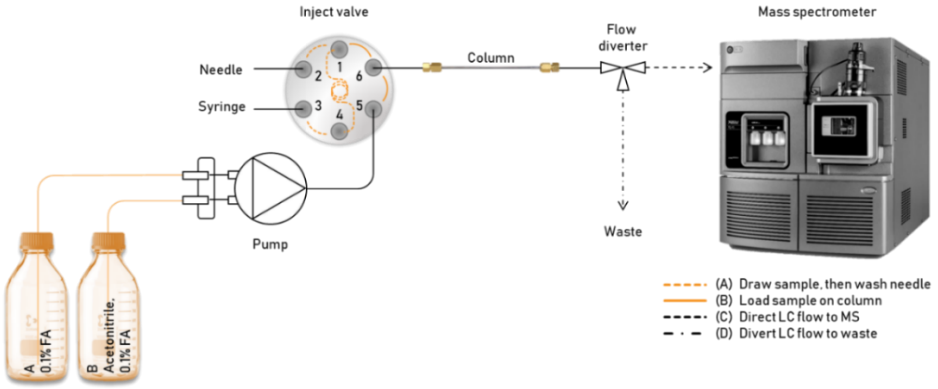


Figure 2-6. Configuration of the LC system utilised in targeted proteomics. *In configuration (A) of the inject valve, the sample is drawn by the auto sampler after which the sample needle is washed, drawing wash solution through the syringe. In inject valve position (B), the sample is loaded onto the column and eluted by the mobile phase provided by the pump. A flow diverter valve post-column can direct the eluent either to the mass spectrometer in configuration (C) or divert it to waste in configuration (D).*

2.9.1 *Liquid chromatography separation*

Chromatographic separation of the peptides was performed on a binary UPLC Acquity system (Waters) utilising a 1 x 100 mm, 1.7 µm ACQUITY UPLC® Peptide CSH C_{18} column (Waters). The mobile phase consisted of A: 0.1% formic acid and B: 0.1% formic acid in acetonitrile pumped at a flow rate of 0.2 mL/min. The column temperature was set to +55 °C. The starting conditions of 3% B were kept static for 0.8 minutes, before initialising the linear gradient utilised to elute and separate the peptides over 7.6 minutes to 25% B. B was thereafter linearly increased to 80% over 0.5 minutes and held for 1.9 minutes to wash the column and elute the most apolar peptides, before returning to the initial conditions over 0.1 minutes followed by equilibration for 6 minutes prior to the subsequent injection.

2.9.2 *Detection by MS/MS*

The LC system was coupled to a Waters Xevo-TQ-S triple quadrupole mass spectrometer for multiple reaction monitoring (MRM) detection in positive electrospray ionisation mode. The capillary voltage was set to 2.8 kV, the source temperature to 150 °C, the desolvation temperature to 600 °C, the cone gas and desolvation gas flows to 150 and 1000 L/hour, respectively. The collision gas consisted of nitrogen and was set to 0.15 mL/ min. The nebuliser operated at 7 bars. The cone voltage was set to 35 V and the collision voltages varied depending on the optimal settings for each peptide. The LC flow was diverted to waste outside of the acquisition windows. The MRM method consisted of 189 unique peptides and was split over two injections to ensure adequate acquisition of the transitions. Figure 2-7 illustrates the retention time-based MRM segments for injections one and two.

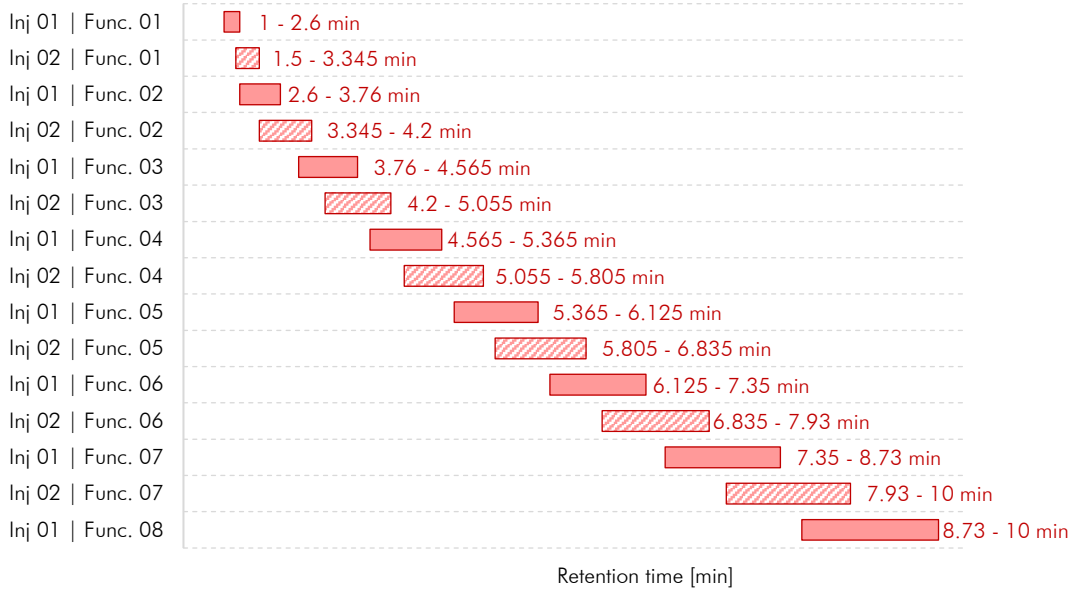


Figure 2-7. MRM segments of the targeted proteomics method for the two injections. *The segments were distributed to allow for the largest possible acquisition windows, thereby minimising the risk of failed measurements due to retention time drift. Each segment was populated by a maximum of 25 peptides. The segments belonging to injection #1 are filled, while the segments belonging to injection #2 are striped. Each function is labelled by its time window.*

The monitored amino acid sequences, the transitions of precursor to product ions and the cone and collision voltages for all peptides included in the assay are presented in Table 2-4.

Table 2-4. Analytical details of the peptides included in the targeted assay. *The precursor proteins of the peptides are represented as gene names. Amino acid sequences, the transitions of precursor to product ions and the cone and collision voltages are presented. A: Alanine, R: Arginine, N: Asparagine, D: Aspartic acid, C: Cysteine, Q: Glutamine, E: Glutamic acid, G: Glycine, H: Histidine, I: Isoleucine, L: Leucine, K: Lysine, M: Methionine, F: Phenylalanine, P: Proline, S: Serine, T: Threonine, W: Tryptophan, Y: Tyrosine, and V: Valine*

Gene	Amino acid sequence	Precursor ion [m/z]	Product ion #1 [m/z]	Product ion #2 [m/z]	Collision voltage, ion #1/#2 [V]
A2M	AFQPFFVELTMPYSVIR	1023.30	1079.59	1208.63	37 / 37
ADIPOQ	IFYNQQNHYDGSTGK	591.27	666.32	1106.49	20 / 20
ANXA2	TNQELQEINR	622.82	659.35	772.43	22 / 22
ANXA2	TPAQYDASELK	611.80	825.40	1024.49	22 / 22
APOE	LGPLVEQGR	484.78	588.31	701.39	17 / 17
APP	GLTTRPGSGLTNIK	472.27	572.33	622.85	16 / 16
APP	YLETPGDENEHAHFQK	638.96	767.39	1010.48	22 / 22
ATIC	HVSPAGAAVGIPLSEDEAK	616.65	847.44	888.43	21 / 21
ATIC	DVSELTGFPEMLGGR	804.39	759.38	963.47	29 / 29
BCAP29	SSTSRPDAYEHTQMK	435.20	468.72	504.23	14 / 14
BCHE	YGNPNETQNNSTSWPVFK	1041.98	490.30	874.92	37 / 37
BCHE	YLTLNTESTR	599.31	707.33	921.46	21 / 21
C15orf62	EQFPSEPSF	534.24	350.17	718.30	19 / 19
C15orf62	LPLWGDEQPR	605.81	701.32	887.40	21 / 21
C3	TVMVNIENPEGIPVK	820.44	853.48	982.52	29 / 29
CAPN2	SDTFINLR	483.26	515.33	763.45	17 / 17

Gene	Amino acid sequence	Precursor ion [m/z]	Product ion #1 [m/z]	Product ion #2 [m/z]	Collision voltage, ion #1/#2 [V]
CAPN2	NFFLTNR	456.24	650.36	797.43	16 / 16
CCL13	SYVITTSR	463.75	577.33	676.40	16 / 16
CCL17	DAIVFVTVQGR	602.84	806.45	905.52	21 / 21
CCL2	WVQDSMDHLDK	458.54	845.38	960.41	15 / 15
CCL22	HFYWTSDSCPRPGVVLLTFR	610.31	627.89	649.40	21 / 21
CCL24	GQQFCGDPK	518.73	576.22	723.29	16 / 16
CCL26	SYEFTSNSCSQR	733.30	838.35	939.39	26 / 26
CCL4	NFVVDYYETSSLCSQPAVVFQTK	894.76	889.47	1104.56	27 / 27
CD200R1	QITQNYSK	491.09	370.66	740.45	14 / 18
CD22	EVQFFWEK	556.77	609.30	756.37	20 / 20
CHI3L1	VTIDSSYDIAK	606.31	696.36	783.39	21 / 21
COL4A2	GLDGYQGPDGPR	616.29	541.27	726.35	22 / 22
COL4A2	SVSIGYLLVK	539.83	692.43	892.55	19 / 19
COL6A3	QLGTVQQVISER	679.38	859.46	1116.60	24 / 24
CPS1	FVHDNYVIR	388.20	458.74	779.40	13 / 13
CRP	APLTKPLK	289.86	586.39	699.48	9 / 9
CRP	ESDTSYVSLK	564.77	696.39	797.44	20 / 20
CS	ALGVLAQLIWSR	663.92	448.06	561.10	16 / 16
CSF1R	NVLLTNGHVAK	583.34	625.34	726.39	21 / 21
CSF1R	VVEATAFGLGK	546.31	764.43	893.47	19 / 19
CST3	ALDFAVGEYNK	613.81	780.39	927.46	22 / 22
CTHRC1	GDASTGWNSVSR	618.78	805.40	906.44	22 / 22
CUL5	YVEQLLTLFNR	698.39	763.45	876.53	25 / 25
CYCS	TGPNLHGLFGR	390.21	505.78	534.29	13 / 13
CYCS	GIIWGEDTLMEYLENPK	670.00	763.35	892.39	18 / 18
DCXR	TQADLDSLVR	559.39	474.52	888.86	26 / 16
DCXR	AVIQVSQIVAR	592.36	772.47	900.53	21 / 21
DKK3	EVPDEYEVGSFMEEVR	957.92	843.87	1053.50	34 / 34
DKK3	SAVEEMEAEEAAAK	732.83	818.39	1078.47	26 / 26
EFNA5	VFDVNDK	418.71	590.28	737.35	14 / 14
EFNA5	VENSLEPADDTVHESAEPSR	728.00	755.84	820.36	25 / 25
ENDOU	YGSEQEFVDDLK	715.33	865.43	1266.58	25 / 25
EPO	TITADTFR	462.74	609.30	710.35	16 / 16
EPO	EVWQGLALLSEAVLR	842.78	674.61	1141.94	22 / 28
FABP5	ELGVGIALR	464.28	529.35	685.44	16 / 16
FABP5	TTQFSTLGEK	556.29	781.41	909.47	20 / 20
FGA	AQLVDMK	402.72	492.25	605.33	14 / 14
FGA	GLIDEVNQDFTNR	760.87	894.41	1237.54	27 / 27
FGF2	LESNNYNTYR	637.29	716.34	830.38	23 / 23
FGF21	EDGTVGGAADQSPESLLQLK	672.34	464.28	701.46	23 / 23
FGF21	YLYTDDAQQTEAHLEIR	689.33	813.89	895.42	23 / 23
GRN	ENATTDLLTK	553.29	690.40	791.45	19 / 19
GRN	AVALSSSVMCPDAR	732.35	1022.44	1109.47	26 / 26
HBE1	MNVEEAGGEALGR	666.81	730.38	859.43	24 / 24
HBE1	EFTPEVQAAWQK	717.36	830.45	1056.55	25 / 25

Gene	Amino acid sequence	Precursor ion [m/z]	Product ion #1 [m/z]	Product ion #2 [m/z]	Collision voltage, ion #1/#2 [V]
HELLS	LISQIQPEVDR	649.36	743.37	984.51	23 / 23
HPX	YYCFQGNQFLR	748.34	862.45	1009.52	27 / 27
HSP90AB1	SIYYITGESK	580.80	634.34	797.40	20 / 20
HSPA1L	VEIIANDQGNR	614.82	703.31	774.35	22 / 22
HSPA5	ITITNDQNR	537.78	747.34	860.42	19 / 19
HSPA8	DAGTIAGLNVLR	600.34	742.46	855.54	21 / 21
HSPD1	VTDALNATR	480.76	645.37	760.39	17 / 17
ICAM1	LLGIETPLPK	540.84	797.48	854.50	19 / 19
ICAM1	ASVSVTAEDEGTQR	725.34	834.36	1006.44	26 / 26
IFNG	AIHELIQVMAELSPAAK	911.15	386.13	473.19	25 / 25
IFNG	IMQSQIVSFYFK	745.89	903.50	1031.56	27 / 27
IL1A	ESMVVVATNGK	567.79	589.33	688.40	20 / 20
IL1B	SLVMSGPYELK	612.32	793.41	924.45	22 / 22
IL2	DLISNINVIVLELK	791.97	813.54	927.59	28 / 28
IL4	EANQSTLENFLER	775.88	807.40	920.48	28 / 28
IL5	TLLIANETLR	572.34	703.37	816.46	20 / 20
IL6	YILDGISALR	560.82	616.38	844.49	20 / 20
IL7	LNDLCFLK	511.77	680.38	795.41	18 / 18
IL10	AMSEFDIFINYIEAYMTMK	1159.71	379.34	510.43	26 / 24
IL10	DQLDNLLLK	536.31	715.43	828.52	19 / 19
IL12A	TSTVEACLPLELTK	781.41	700.40	1173.60	26 / 26
IL12B	EDGIWSTDILK	638.82	676.39	862.47	23 / 23
IL13	ELIEELVNITQNQK	835.95	944.52	1057.60	30 / 30
IL15	TEANWVNVISDLK	744.89	788.45	887.52	26 / 26
IL16	DPGVSESPPPGR	597.79	739.37	925.47	21 / 21
IRAK4	SANILLDEAFTAK	696.87	781.37	894.46	25 / 25
ITIH2	TEVNVLPGAK	514.29	698.42	797.49	18 / 18
ITIH2	VQFELHYQEVK	473.91	666.35	803.40	16 / 16
LMO7	KPQDQLVIER	409.23	417.25	710.38	13 / 13
MAPK12	VTGTPPAEFVQR	651.35	472.25	846.45	23 / 23
MASP2	AGYVLHR	408.23	425.26	524.33	14 / 14
MASP2	WPEPVFGR	494.26	575.33	801.43	17 / 17
MMP3	GNQFWAIR	496.26	545.32	692.39	17 / 17
MMP3	TYFFVEDK	524.75	637.32	784.39	18 / 18
MSN	EDAVLEYLK	540.28	552.30	665.39	19 / 19
MUC5B	ATNSTATPSSTLGTTR	783.39	919.48	1020.53	28 / 28
NCAM1	LEGQMGEDGNSIK	689.32	819.38	950.42	24 / 24
NCAM1	DGQLLPSSNYSNIK	768.39	1009.49	1122.58	27 / 27
NDRG1	ISGWTQALPDMVVSHLFGK	696.03	615.32	787.45	24 / 24
NDRG1	MADCGGLPQISQPAK	786.88	868.49	1038.59	28 / 28
NEFL	YEEEVLSR	512.75	603.35	732.39	18 / 18
NEFL	VLEAELLVLR	577.86	613.44	813.52	20 / 20
NEFM	SIELESVR	466.76	490.26	732.39	16 / 16
NEFM	FVEEIIEETK	618.82	732.41	990.50	22 / 22
NEFH	LEQEHLLEDIAHVR	426.23	595.37	710.39	14 / 14
NEFH	TSVSSVSASPSR	582.80	604.30	877.44	20 / 20

Gene	Amino acid sequence	Precursor ion [m/z]	Product ion #1 [m/z]	Product ion #2 [m/z]	Collision voltage, ion #1/#2 [V]
NFATC2	YQQQNPAAVLYQR	789.90	820.47	917.52	28 / 28
NFKBIZ	ASGQAVDDFK	519.25	694.34	879.42	18 / 18
NLRP3	YLEDLEDVDLK	676.34	831.45	946.47	24 / 24
OLR1	QQAEEASQESENELK	573.93	719.36	848.40	19 / 19
PGAM1	AMEAVAAQGK	488.25	644.37	773.42	17 / 17
PGK1	VLPGVDALSNI	549.50	466.54	885.84	8 / 10
PKM	ITLDNAYMEK	599.29	755.34	791.39	21 / 21
PLAU	SDALQLGLGK	501.28	615.38	728.47	17 / 17
PLD3	LLISCWGHSEPSMR	558.27	667.27	723.82	19 / 19
PLD3	ALLNVVDNAR	542.81	574.29	787.41	19 / 19
PPP3CB	GLTPTGMLPSGVLAGGR	792.43	656.86	813.46	28 / 28
PRDX3	GLFIIDPNGVIK	643.38	742.41	855.49	23 / 23
PRG4	GFGGLTGQIVAALSTAK	795.95	1058.62	1159.67	28 / 28
PTGDS	AQGFTEDTIVFLPQTDK	955.51	363.41	588.56	42 / 26
PTGDS	TMLLQPAGSLGSYSYR	872.44	989.47	1157.56	31 / 31
PTGES2	QWADDWLVHLISPNVYR	704.54	735.46	848.62	20 / 24
RANGAP1	AFNSSSFNSNTFLTR	564.94	637.34	751.38	16 / 16
RANGAP1	VINLNDNTFTEK	704.36	968.43	1195.56	25 / 25
SAA1	EANYIGSDK	498.74	682.34	796.38	17 / 17
SAA1	SFFSFLGEAFDGAR	776.22	303.33	822.63	34 / 22
SELE	YTHLVAIQNK	396.22	573.34	672.40	13 / 13
SERPINA1	LSSWVLLMK	538.81	603.39	876.50	19 / 19
SERPINA3	LYGSEAFATDFQDSAAAK	946.44	952.44	1053.48	34 / 34
SERPINF2	LGNQEPGGQTALK	656.85	674.38	771.44	23 / 23
SERPINF2	HQMDLVATLSQLGLQELFQAPDLR	908.65	500.13	1112.82	18 / 18
SERPING1	LVLLNAIYLSAK	659.41	765.45	992.58	23 / 23
SMC4	SNNIINETTTR	631.82	721.35	834.43	22 / 22
SNAP25	AWGNNQDGVVASQPAR	557.27	558.30	629.34	19 / 19
SNAP25	HMALDMGNEIDTQNR	582.26	633.30	746.38	20 / 20
SOD1	LACGVIGIAQ	501.18	201.01	218.10	22 / 18
SOD2	LTAASVGVQGSGWGWLGFNK	1018.20	1064.53	1208.58	37 / 37
SOD3	AGLAASLAGPHSIVGR	492.95	582.83	618.35	16 / 16
SOD3	VTGVVLFR	445.78	534.34	690.43	15 / 15
SPP2	DALSASVVK	445.25	503.32	590.35	15 / 15
SPP2	VNSQSLSPYLFR	705.87	695.39	782.42	25 / 25
TEK	IVDLPDHIEVNSGK	512.61	548.28	662.33	17 / 17
THY1	HVLFGTVGVPEHTYR	428.73	401.70	479.74	14 / 14
TLR6	DMPSLEILDVSWNSLESGR	1074.49	948.79	1035.89	32 / 30
TNF	DNQLVVPSEGLYLIYSQVLFK	1213.38	570.24	669.35	27 / 27
TNF	ANALLANGVELR	620.85	758.42	871.50	22 / 22
TNFB	MHLAHSTLKPAAHLIGDPSK	531.79	331.20	619.35	18 / 18
TNNT3	DLMELQALIDSHFEAR	629.98	764.89	830.41	21 / 21
TOLLIP	LNITVVQAK	493.31	645.39	758.48	17 / 17
TOLLIP	GPVYIGELPQDFLR	802.43	775.41	1074.56	29 / 29
TRAP1	ELGSSVALYSR	591.31	609.34	882.47	21 / 21

Gene	Amino acid sequence	Precursor ion [m/z]	Product ion #1 [m/z]	Product ion #2 [m/z]	Collision voltage, ion #1/#2 [V]
TUBA4A	DVNAAIAAIK	493.29	586.39	771.47	17 / 17
TUBA4A	AVFVDLEPTVIDEIR	858.46	942.53	1299.68	31 / 31
TXN	LEATINELV	501.22	474.11	528.24	9 / 9
UBC	TITLEVEPSDTIENVK	894.47	905.46	1002.51	32 / 32
VCAM1	LHIDEMDSVPTVR	756.59	251.23	472.42	28 / 24
VCAM1	NTVISVNPSTK	580.32	732.39	944.54	20 / 20
VEGFA	SWSVYVGAR	512.76	565.31	751.41	18 / 18
VEGFC	FAAAHYNTEILK	459.91	580.31	615.83	15 / 15
VEGFD	FAATFYDIETLK	709.86	718.40	881.46	25 / 25
Internal standards					
ENO1 (yeast)	GNPTVEVELTTEK	709.06	623.49	948.68	18 / 20
RSU1	ALYLSDNDFEILPPDIGK	1014.30	633.37	746.37	36 / 36
C3	SSLSVPYVIVPLK	704.43	357.25	934.60	25 / 25
GSTO1	GSAPPGPVPEGSIR	664.36	556.81	1015.56	23 / 23
TSP1	TIVTTLQDSIR	627.36	940.52	1039.59	22 / 22

2.10 TARGETED PROTEOMICS DATA PROCESSING

After acquisition, data were peak picked using an in-house Python-based guided user interface (GUI) peak picking application (development described in Chapter 3, section 3.5.1) or the MassLynx (version 4.1) module TargetLynx (Waters). In the GUI application workflow, the .raw files were converted to .txt files and imported to the application. Peaks were aligned if necessary, and thereafter integrated. When TargetLynx was used, data were imported to TargetLynx and quantitative methods were created and applied to the data.

The integrated peak areas were exported to Microsoft Excel where firstly the ratio between quantifier and qualifier peak areas were calculated and evaluated to ensure that the correct peaks had been integrated. The digestion efficiency was evaluated by monitoring the presence of the yeast enolase peptide from ENO1 (GNPTVEVELTTEK). After the initial quality assessment, the quantifier area was divided by the area of a heavy isotope labelled peptide internal standard to yield a ratio used for the determination of relative concentrations. Any compound that also showed an intensity signal in the blank samples had the blank signal subtracted from the analyte peak intensity. Pooled quality control samples were additionally evaluated to assess the robustness of the run.

The final data were evaluated through correlation analysis and by comparing the groups through multivariate and univariate analyses. The results were assessed and visualised

using the software packages GraphPad Prism (version 6.0.1, GraphPad Software, San Diego, California USA, www.graphpad.com), Simca and various Python scripts.

2.11 GENERAL PROTOCOLS

2.11.1 *Analysis of creatinine for normalising urinary biomarkers*

The metabolite creatinine is commonly utilised for normalising urinary biomarkers, expressing them as a ratio to the urinary creatinine concentration. Creatinine concentrations were determined in all urine samples according to the here described procedure. All creatinine measurements were performed by Mrs Justyna Spiewak (The Biological Mass Spectrometry Group, Institute of Child Health, UCL, London).

A working solution consisting of 0.25 mM D_3-creatinine (CDN Isotopes, Quebec, Canada) in water was prepared. 100 µL of urine was aliquoted to a 1.5 mL centrifuge tube from Eppendorf. The samples were centrifuged at room temperature for ten minutes at 16900 x *g* on a Biofuge Pico (Heraeus, Hanau, Germany). 10 µL of the urine supernatant was added to 300 µL Chromacol micro insert vials from Thermo Fisher Scientific. 200 µL working solution was added and the samples shaken briefly. The creatinine concentrations were determined on a Waters Micromass Quattro Micro triple quadrupole mass spectrometer. The mobile phase was A: 4mM Ammonium Acetate + 0.052% heptafluorobutyric acid and B: methanol. The column was a Discovery HS F5-5 (5 cm x 2.1 mm, 5 µm) equipped with a guard column: Discovery HS F5 Supelguard Cartridge (2 cm × 2.1 mm, 5 µm) both from Sigma-Aldrich. The column temperature was +20 °C. The gradient elution was initialised with 5% B for 1.9 minutes and a flow rate of 0.2 mL/min. B was thereafter linearly increased to 100% over 0.1 minute, at a flow rate of 0.5 mL/min. B was held at 100% for 1.9 minutes at a flow rate of 0.6 mL/min, before returning to the initial composition of 5% B over 0.01 minute. The column was equilibrated before the subsequent injection for 1.99 minutes at a flow rate of 0.6 mL/min. Detection was performed in positive ESI mode. The capillary voltage was 3.5 kV, the cone voltage was 22 V and the extractor and RF lens were set to 2 V and 1.7 V, respectively. The source temperature was 120 °C and the desolvation temperature 350 °C. The desolvation gas flow was 950 L/hour. The transition m/z 114.03 to m/z 43.7 was monitored for endogenous creatinine and the transition m/z 117.199 to m/z 46.72 for the internal standard d_3-creatinine. The cone voltage and collision energy were 28 V and 14 V, respectively, for both compounds. The creatinine concentration in the samples was determined by relating the

ratio of endogenous creatinine to d_3-creatinine to a calibration curve ranging from 0 to 40 mM.

2.11.2 *Top-down protein fractionation by molecular weight*

Top-down fractionation of intact proteins was performed as part of the method development for discovery proteomics. A GELFrEE 8100 fractionation system from Expedeon (Cambridgeshire, UK) was used. The utilised method is here described.

The samples were prepared for fractionation by combining sample (neat or treated plasma), 30 µL sample buffer (provided by the supplier), 8 µL 1 M DTE and a volume of Milli-Q water rendering a final sample volume of 150 µL. The sample was heated for ten minutes at +50 °C, then allowed to cool to room temperature. The fractionation cartridge consisted of 8% Tris Acetate and was prepared by firstly removing the storage buffer and secondly rinsing out residual storage buffer with running buffer (provided by the supplier). 150 µL running buffer was added to the collection chambers and the cathode buffer reservoirs were filled with 6 mL running buffer. The sample was loaded to the loading chamber and the cartridge attached to the instrument. The program demonstrated in Table 2-5 was utilised to fractionate the proteins based on their molecular weights. The first migrated load was discarded. After this, fractions were collected as per Table 2-5. After each fraction had been collected, the collection chamber was washed carefully before the next step. The fractions were freeze dried before digestion and mass spectrometric analysis.

Table 2-5. Settings for GELFrEE system in the fractionation of plasma proteins

Fraction #	Load	1	2	3	4	5	6	7	8	9	10
Voltage (V)	50	50	50	100	100	100	100	100	100	100	100
Time (min)	16	41.5	7	2	2	3	5	7	10	15	20
Total time (min)	16	57.5	64.5	66.5	68.5	71.5	76.5	83.5	93.5	108.5	128.5

2.11.3 *One-dimensional gels to visualise protein preparation results*

Many of the method development steps for discovery proteomics were evaluated visually on gels. The preparation and running of the gels are described in this section.

To prepare the samples, 4x Laemmli buffer from Bio-Rad (Hercules, CA, USA) was diluted four times with 62.5 mM DTE. Equal volumes of sample and Laemmli buffer were mixed, and the sample heated at +95 °C for five minutes. The sample was cooled to room temperature before 10 µL was loaded onto a Mini-PROTEAN TGX precast gel (Bio-

Rad). The gel was run using a Mini-PROTEAN Tetra Cell (Bio-Rad). The running buffer was 0.1% sodium dodecyl sulphate (SDS) and the voltage was set to 200 V. A pre-stained protein ladder was also run. After completion, the gel was Comassie blue stained for 15 minutes with InstantBlue Protein Stain from Expedeon and photographed.

Optimising sample preparation, instrumental parameters and data processing for analysis of low-abundant proteins in urine and blood

3

Abstract. *The overall aim of the experiments carried out throughout this thesis was to identify proteomic biomarkers and affected pathways in Parkinson's disease and healthy ageing. The greatest possible coverage of low- and medium-abundant proteins in these biofluids is crucial to find novel protein targets.*

In this chapter, several preparatory, instrumental, and processing parameters were evaluated and optimised to increase the number of detectable proteins in the mass spectrometric discovery analyses of blood and urine. It was found that by depleting the twelve most abundant plasma proteins, followed by chromatographic online separation of digested peptides into ten fractions, the greatest number of proteins could be detected in blood. In urine, it was concluded that the use of a low molecular weight cut-off filter, followed by online chromatographic fractionation into ten fractions of digested peptides, allowed for the largest number of proteins to be identified. A targeted, MRM-based multiplexed method was developed from the proteins identified in the discovery phase, where blood and urine from PD patients, and blood from centenarians were analysed. A total of 127 proteins were included in the validation assay. A meticulous process was implemented to select unique peptides for each protein, also ensuring that they were compatible with LC-MS/MS analysis. The assay was further augmented with several known pro- and ani-inflammatory proteins from the literature. The preparation and analysis of the samples for the validation phase were optimised and it was decided that depletion of albumin and IgG was the most ideal method for the preparation of plasma samples. For targeted urine preparation, the same protocol as in the discovery phase was utilised. The developed targeted assay was later applied to larger validation cohorts to confirm the targets from the discovery phase. To allow for consistent and high throughput analysis of the targeted validation data, several in-house scripts were developed and greatly expedited the data analysis.

3.1 INTRODUCTION AND AIMS

The samples analysed in this thesis were serum, plasma, and urine. None of these biofluids are optimal samples for studying neurological disease as proteins of "interest" of disease mechanisms are of low abundance relative to the house-keeping proteins. Therefore, to study these proteins, every step of these analyses was optimised to maximise the efficiency of extracting low-abundant proteins for analysis.

Sample preparation and analytical methods have a large impact on the quality of a dataset. From the point of samples arriving in a laboratory to the moment of instrumental analysis, a vast number of decisions must be taken. These decisions include the amount of sample to use, which reagents to utilize and at what concentrations, timings and temperatures of different processing and incubation steps, purification strategies and fractionation methods. The instrumental analysis further requires careful attention to the set-up of the chromatographic separation - a suitable column needs to be selected and paired with an adequate mobile phase and eventual buffering additives. Moreover, the gradient elution must be optimised to achieve sufficient separation between compounds, maximise the number of theoretical plates, minimise band broadening and avoid carry-over between injections. It must also be designed in such a way that retention times remain stable, and the system does not easily reach overpressure. After the chromatic separation, compounds are detected by mass spectrometry and also here a number of important parameters must be considered. In targeted assays, the multiple reaction monitoring method must be constructed in such a way that it allows for a sufficient number of data points to be measured for each peak. The time each data point is acquired must also be large enough. Moreover, suitable fragments must be selected for detection and the collision energy for each of these fragments must be optimised. In untargeted analyses, the mass spectrometer's detector needs to be set up at a sufficiently high voltage to provide a suitable response for the analytes, the instrument needs to be well-calibrated and capable of adequately detecting and measuring compounds in the defined mass range. After acquisition, the untargeted data is peak picked and identified through a largely automatic process. The targeted data is peak picked manually meaning great care must be taken in integrating the correct peaks. Once the data table of feature intensities has been exported, the data must pass quality control checks for sufficient sensitivity, instrumental drift, and robustness of the run before any significance testing and biological interpretation can be undertaken.

This chapter is dedicated to method development and describes the experiments performed to circle in on the best way to prepare and analyse urine and blood samples by targeted and untargeted LC-MS proteomics. The aims of the method development performed in this chapter were to:

- Develop a sample preparation method for the untargeted discovery analysis of plasma and serum which reduces the presence of house-keeping proteins
- Develop a sample preparation method for the untargeted discovery analysis of urine
- Evaluate and improve the instrumental parameters of untargeted LC-MS to allow for the identification of the maximum number of proteins
- Develop a targeted MRM-based method that can measure the putative proteomic biomarkers identified in the discovery phase
- Develop the most ideal sample preparation methods for plasma and urine for targeted proteomics
- Develop a pipeline for data analysis

3.2 METHOD DEVELOPMENT FOR SAMPLE PREPARATION OF URINE AND PLASMA/SERUM TO BE ANALYSED BY UNTARGETED PROTEOMICS

To maximise the number of detectable proteins in the untargeted discovery phase, several steps included in sample preparation were evaluated and optimised. The method development included strategies to decomplex and concentrate the samples, evaluation of the ideal digestion enzyme and incubation time, and fractionation steps. Figure 3-1 shows a graphical summary of the evaluated and optimised steps.

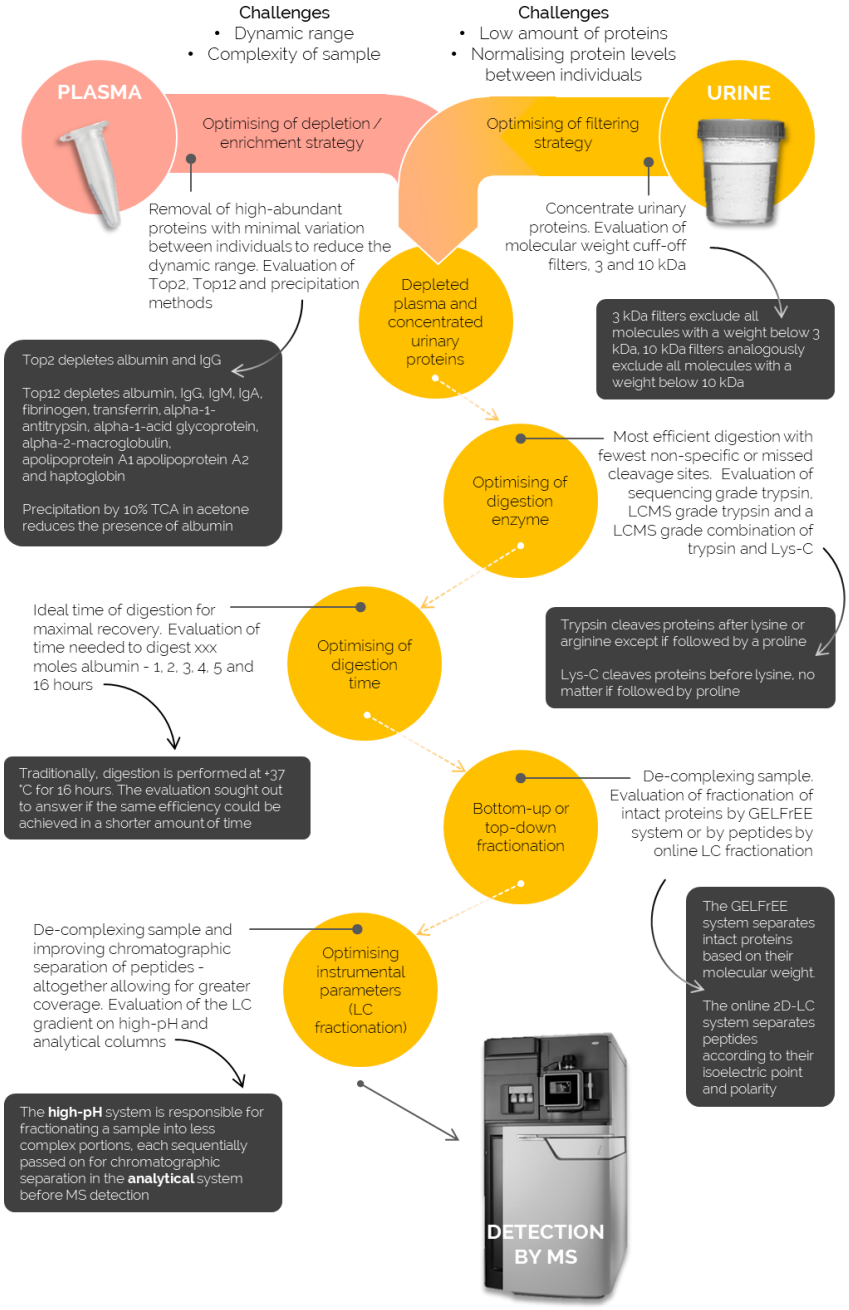


Figure 3-1. Graphical summary of the optimised steps in the preparation for untargeted urine and plasma/serum proteomics.

3.2.1 *Optimising the sample preparation of plasma/serum for untargeted discovery proteomics*

Plasma and serum are complicated matrices for performing proteomic biomarker discovery, primarily due to the high levels of house-keeping proteins such as albumin, immunoglobulins, and other compounds in high concentration. These proteins are often considerable in size and prone to generate an abundance of tryptic peptides per mole of protein, thereby creating an even more complex environment to detect small and low-

abundant proteins in [138]. Without removing the high-abundant proteins prior to tryptic digestion, detection of low-abundant species is extremely difficult, therefore a depletion/enrichment strategy was considered [139]. The initial part of the plasma method development was to assess the use of different depletion techniques and several strategies to de-complex the sample matrix, altogether aiming to develop a protocol allowing for improved protein detection and coverage of low-abundant species. The following preparation steps were evaluated and optimised:

- Depletion strategy
- Enzyme for digestion
- Digestion time
- Top-up or bottom-down fractionation

3.2.1.1 *Assessment of depletion strategy*

Depletion of high-abundant proteins using Pierce™ Top2 and Top12 columns from Thermo Fisher Scientific (Waltham, Massachusetts, US) was evaluated. These columns use immunoaffinity to selectively extract high-abundant proteins from biofluids, leaving behind the lesser abundant proteins. A strategy including enrichment by proteominer beads from BioRad was considered but discarded as it is not compatible with heparin treated plasma. "Top2 columns" deplete albumin and IgG while "Top12 columns" deplete albumin, immunoglobulins (IgG, IgM and IgA), fibrinogen, transferrin, alpha-1-antitrypsin, alpha-1-acid glycoprotein, alpha-2-macroglobulin, apolipoprotein A1 and A2, and haptoglobin. Selective precipitation of albumin with 10% trichloroacetic acid (TCA) in acetone was also performed.

Top2 and *Top12* column depletions were performed according to the manufacturer's instructions. In short, 10 µL pooled plasma were added to the depletion columns which were end-over-end mixed for 1 hour at room temperature before elution by centrifugation, 2 minutes, 2000 x *g*. Acetone/TCA precipitation was performed by adding 40 µL acetone, 10% TCA to 10 µL plasma. The sample was placed in -20 °C for 16 hours before centrifugation at +4 °C, 16900 x *g* and the supernatant was discarded. The samples were freeze-dried prior to digestion. To solubilise the proteins, 20 µL of digest buffer were added (100 mM Tris (pH 7.8), 6 M urea, 2 M thiourea and 2% ASB14). To reduce sulphide bonds, dithioerythriol (45 µg) was added and the samples were shaken for 60 minutes. To prevent the sulphide bonds from reforming, 108 µg iodoacetic acid were added and the samples were shaken for 60 minutes. MilliQ water was added to dilute the concentration of urea and 1 µg trypsin was added before the samples were placed in a +37 °C water bath

for 16 hours. Solid phase extraction (SPE) was performed using 100 mg C_{18} cartridges from Biotage (Uppsala, Sweden). Prior to SPE clean up, the samples were adjusted to a concentration of 0.1% trifluoroacetic acid (TFA). In brief, the cartridges were washed with two 1 mL aliquots of 70% ACN, 0.1% TFA before equilibration by two aliquots of 0.1% TFA. The samples were loaded and the flow-through re-applied. Salts were washed away by the addition of 1 mL 0.1% TFA prior to elution with two 250 µL aliquots of 70% ACN, 0.1% TFA. The results were evaluated by untargeted mass spectrometry.

The depletion strategy assessment demonstrated that *Top12* enrichment was superior to *Top2* and acetone/TCA precipitation. *Top12* enrichment resulted in 27% more protein hits than *Top2* and 17% more than acetone/TCA precipitation (Figure 3-2).

Figure 3-2. Comparison of three conditions for the de-complexing of plasma prior to proteomic analysis. *The vertical axis is normalised to the percentage of the maximum number of proteins detected. (a) and (c) show the results of antibody depletion removing the top two proteins, albumin, and IgG (a) and the top 12 proteins detected in plasma (c). (b) shows the results of the standard acetone/TCA precipitation of all proteins with no specific depletion. Top12 depletion (c) demonstrates 17% more hits than acetone/TCA precipitation (b) and 27% more hits than removal of albumin and IgG alone (a).*

3.2.1.2 *Assessment of further improvements to depletion strategy*

To evaluate if it was possible to modify and further improve the optimal depletion strategy using *Top12* columns, a combination of depletion and precipitation procedures was also attempted. Aliquots from a pooled plasma sample were used in all experiments. Depletion by *Top2* and *Top12* columns were performed as per section 3.2.1.1. The following experiments were performed:

(a) *Top12* followed by acetone precipitation. Plasma (10 µL) was depleted using a *Top12* column. The eluate from the column was precipitated with acetone and left to incubate at -20 °C for 16 hours before centrifugation at +4 °C, 16900 x g for 10 minutes. Supernatant was discarded.

(b) Two *Top12* columns in series. Plasma (10 µL) was depleted using a *Top12* column and the eluate was added to yet another *Top12* column.

(c) *Top2* and *Top12* in series. Depletion was performed using 10 µL plasma on a *Top2* column, the eluate was added to a Top12 column and incubated whilst end-over-end mixing for 60 minutes.

(d) *Top12* reapplied to the same cartridge. Plasma (10 µL) was depleted using a *Top12* column and reapplied to the same column before 60 minutes incubation whilst end-over-end mixing.

(e) *Top12*, no alterations to protocol. Plasma (10 µL) was depleted using a *Top12* column with no alterations or additions to the manufacturer's protocol.

The final eluate from experiments (a) - (e) were all freeze dried before digestion and cleaned up as described in Chapter 2, section 2.2.3.1. The results were evaluated by mass spectrometry.

The results (Figure 3-3) demonstrate that the normal protocol without alterations (e) results in 12% more hits than when reapplying the eluate to the same column (d), 18% more hits than when using *Top2* and *Top12* in series (c), 32% more hits than when applying two *Top12* columns in series, and 73% more hits than when performing a *Top12* depletion followed by acetone precipitation. Therefore, the Top12 column using the manufacturer's instructions was the optimum method for the purification of low-abundant proteins present in plasma and analysis by MS.

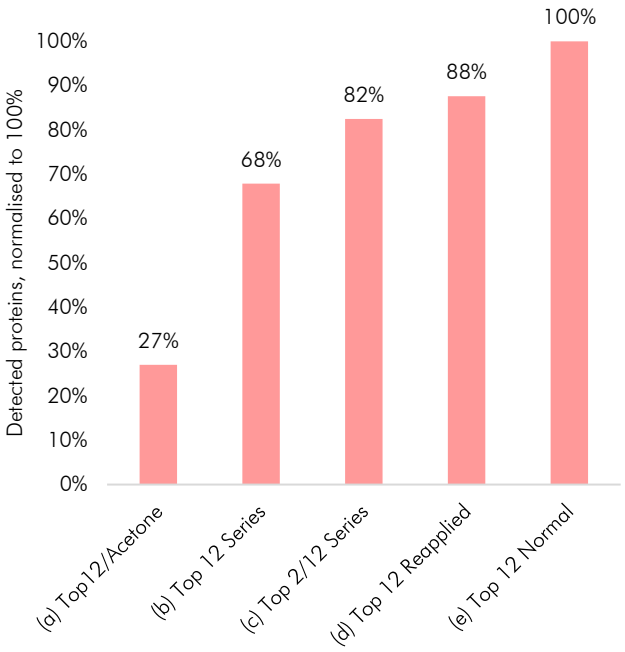


Figure 3-3. Optimising the depletion strategy. *The vertical axis is normalised to the percentage of the maximum number of proteins detected. Top12 depletion combined with acetone precipitation (a), using two Top12 columns in series (b), using a Top2 column in series with a Top12 column (c), reapplying the eluate from a Top12 column to the same column (d) and using a Top12 column according to the manufacturer's instructions without any alterations (e). The normal protocol resulted in the greatest number of protein hits.*

3.2.1.3 *Evaluation of the optimum digestion proteases for the analysis of low-abundant proteins in plasma*

The efficiency of sequencing grade trypsin, mass spectrometry grade trypsin and combined trypsin/Lys-C was evaluated by digesting equal volumes of plasma following the protocol referred to above.

The results (Figure 3-4) showed that MS grade trypsin gold produced 24% more hits than trypsin/Lys-C and 50% more hits than sequencing grade trypsin. The rationale is most likely that MS grade trypsin has higher activity and purity, and more stringent specificity in its cleavage of proteins and thereby leading to fewer random, non-specifically cleaved peptides not recognised by the protein database used for identification.

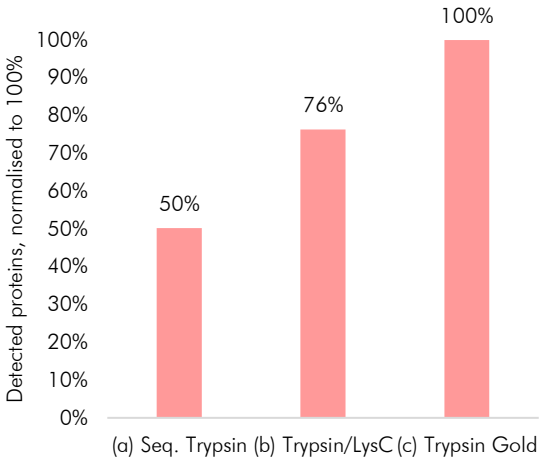


Figure 3-4. Relative number of protein hits after digestion of plasma using sequencing grade trypsin (a), a combination of trypsin and Lys-C (b) and using MS grade trypsin gold (c). *The vertical axis is normalised to the percentage of the maximum number of proteins detected. MS grade trypsin gold renders 24% more hits than combined trypsin and Lys-C, and 50% more hits than sequencing grade trypsin.*

3.2.1.4 *Optimising the digestion time*

The ideal time required for sufficient digestion was evaluated. The efficiency of the tryptic digestion over time was determined by quantitating the production of albumin peptides. The rationale behind choosing albumin rather than a less abundant protein was that it was desired to monitor the largest and most high-abundant protein possible to represent the sample matrix. A pooled plasma sample was prepared and split into six different aliquots. The six aliquots were incubated in a +37 °C water bath for 1, 2, 3, 4, 5 and 16 hours, respectively. The digestion process was halted by snap-freezing the samples on dry ice after their allocated digestion time. The samples were SPE-cleaned and analysed by targeted mass spectrometry. A linear production of peptides with increasing incubation time was observed, indicating that overnight or 16 hours digestion resulted in the most efficient protein digestion (Figure 3-5).

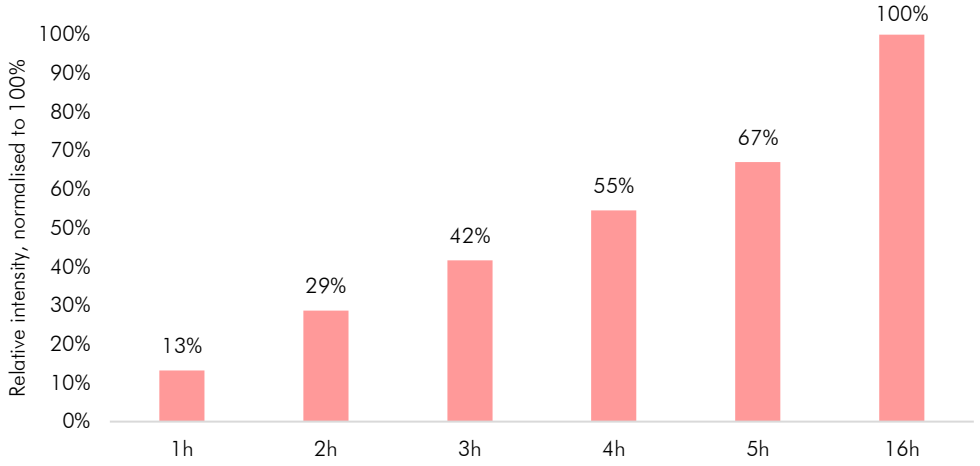


Figure 3-5. Relative intensities of albumin as a result from the digestion of a pooled plasma sample in timed intervals consisting of 1, 2, 3, 4, 5 and 16 hours. *Incubation for 16 hours showed the highest intensity*

3.2.1.5 *Comparison of top-down and bottom-up fractionation*

Fractionation strategies are often employed in plasma proteomics to reduce the complexity of the sample by limiting the dynamic range and subsequently increasing the availability of measurable proteins [140]. In our laboratory, fractionation has traditionally been performed as part of a bottom-up proteomics strategy, meaning that digested peptides are separated into fractions. However, top-down proteomic techniques have been described as an alternative and better strategy for obtaining greater protein coverage [141].

Comparing these two techniques experimentally, pooled and *Top12* depleted plasma was used. In the top-down methodology, an offline preparative isoelectric focusing (IEF) unit was utilised (GELFrEE, Expedeon, Cambridge, UK) to fractionate the intact plasma proteins based on their molecular weights. Table 3-1 shows the settings and timings for the different fractions. Ten fractions were opted for as this had previously been determined to be ideal in our laboratory (data not shown). After the proteins had been separated into molecular mass-based fractions, the separate fractions were digested and SPE cleaned.

Table 3-1. Settings for GELFrEE system in the fractionation of plasma proteins.

Fraction #	Load	1	2	3	4	5	6	7	8	9	10
Voltage (V)	50	50	50	100	100	100	100	100	100	100	100
Time (min)	16	41.5	7	2	2	3	5	7	10	15	20
Total time (min)	16	57.5	64.5	66.5	68.5	71.5	76.5	83.5	93.5	108.5	128.5

An additional sample was fractionated using the IEF unit and visualised on a 1D gel from BioRad (prepared according to Chapter 2, section 2.11.3), shown in Figure 3-6. It can be

noted that the masses of the proteins increase with increasing fraction number, thereby demonstrating the IEF unit's capacity of separating proteins by mass and reducing the sample complexity.

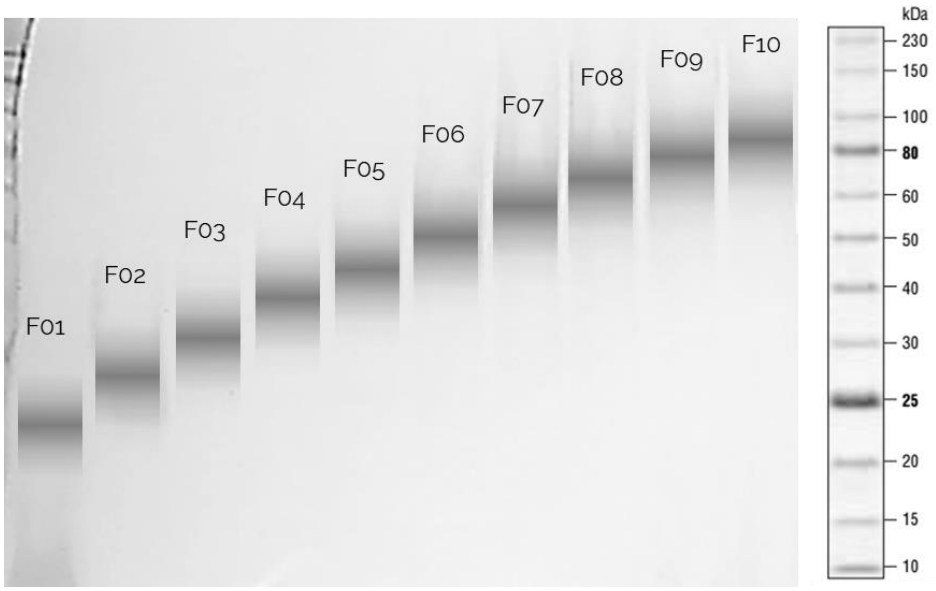


Figure 3-6. 1D gel of ten fractions from a Top12 depleted plasma sample, separated by mass on the GELFrEE system. *The figure shows graphically enhanced bands of the mass ranges of proteins present in each fraction obtained*

In the bottom-up methodology, the depleted *Top12* sample was digested and SPE-cleaned before chromatographic fractionation. The fractionation was performed utilising two LC pumps and two columns in parallel (known as two-dimensional liquid chromatography). It separated the peptides in the sample into ten fractions based on their isoelectric points before eluting each of the fractions onto the second analytical column for chromatographic separation prior to entering the MS. The percentages of organic solvent utilised for elution from the first dimensional high-pH column are displayed in Table 3-2.

Table 3-2. Percentages of eluent B in the ten online bottom-up LC fractions used to decomplex samples for untargeted proteomics analysis.

Fraction #	1	2	3	4	5	6	7	8	9	10
High-pH column [%B]	7.4	10.8	12.6	14	15.3	16.7	18.3	20.4	23.5	60

Both methodologies were evaluated by untargeted mass spectrometry. The comparison of top-down IEF fractionation of intact proteins and bottom-up two-dimensional LC fractionation of digested proteins demonstrated that 27% more hits could be detected in the bottom-up methodology than in the top-down methodology (Figure 3-7).

In conclusion, the bottom-up strategy resulted in a higher number of hits than the top-down strategy. The reason for this might simply be losses along the sample preparation as that the top-down methodology requires more steps. The bottom-up strategy was selected as the preferred option as it resulted in more hits and demanded fewer steps. However, fractionation of intact proteins certainly has its merits as a protein is confined to a unique fraction, thus allowing for fine-tuning of which proteins to capture in which fractions and which proteins to discard. Another advantage of the top-down strategy is that it renders greater flexibility in the instrumental analysis of the samples as each fraction can be loaded separately.

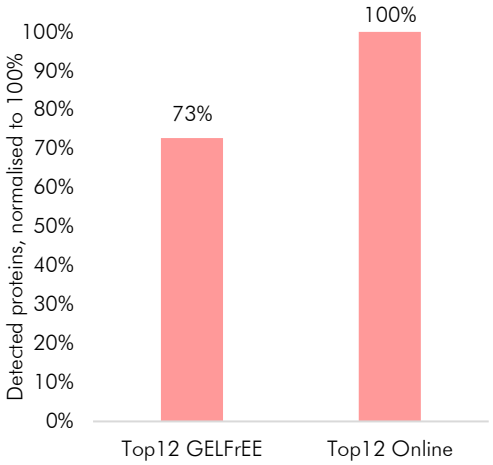


Figure 3-7. Top12 depleted samples fractionated as intact proteins by an IEF unit (GELFrEE) and as digested peptides by 2D-LC online fractionation, each into ten fractions. *The vertical axis is normalised to the percentage of the maximum number of proteins detected. The merged results from the ten fractions demonstrated that the online bottom-up fractionation resulted in 27% more protein hits.*

3.2.1.6 *Optimised workflow for proteomic biomarker discovery in plasma*

After optimising every critical step in the sample preparation of proteins purified from plasma/serum, a final protocol was decided upon for use in this study. This final and optimised sample preparation protocol producing the highest number of protein identifications was determined to be Top12 enrichment, followed by digestion using MS grade trypsin with 16 hours digestion and online bottom-up fractionation into ten fractions (Figure 3-8).

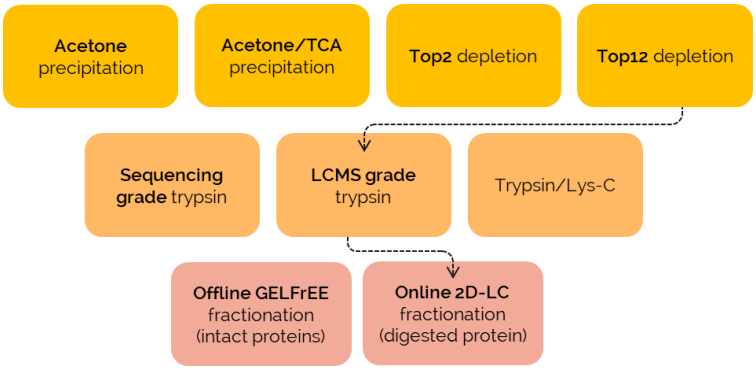


Figure 3-8. Optimised workflow for the preparation of plasma samples, including Top12 depletion, digestion by MS grade trypsin and online fractionation

The final protocol for untargeted discovery plasma proteomics sample preparation is presented in Chapter 2, section 2.3.

3.2.2 *Optimising the sample preparation of urine for untargeted discovery proteomics*

Urine is a less complex matrix for proteomics biomarker discovery than plasma, chiefly due to the fact that the dynamic range of protein concentrations is smaller, reportedly five orders of magnitude, thereby presenting a more even distribution of protein levels in a sample [142]. Urine does however provide challenges on its own as the amount of protein normally is very low; typically less than 150 mg of protein are excreted per day [143]. Urine is further rich in waste metabolites including creatinine at mM levels and salts, typically not well-tolerated by mass spectrometers. Moreover, the protein levels change with the urine's dilution, thus highlighting the importance of a robust normalising strategy between individuals for a non-biased comparison. Therefore, the challenge was to purify or enrich low amounts of protein from a solution containing very high amounts of salts and low molecular weight waste metabolites.

3.2.2.1 *Filtering and concentration of urinary proteins*

To assess strategy of concentrating the urinary proteins, two different molecular weight cut-off filters were evaluated - 3 kDa and 10 kDa, each capable of filtering 4 mL (Amicon Ultra-4, Merck, Darmstadt, Germany). The filters exclude any molecules with a mass smaller than 3 kDa or 10 kDa, respectively.

A pooled urine sample was used in the experiment. Prior to filtering, the urine was centrifuged at room temperature at 5000 x *g* for 30 minutes to separate sediment from solution, 2 mL of urine were thereafter applied to each filter and 2 mL of MilliQ water were added. The filters were centrifuged at room temperature at 5000 x *g* for 60 minutes before the concentrate was collected and freeze dried, followed by tryptic digestion and untargeted mass spectrometry analysis. The results demonstrated that the 10 kDa filter resulted in 22% more hits than the 3 kDa filter (Figure 3-9).

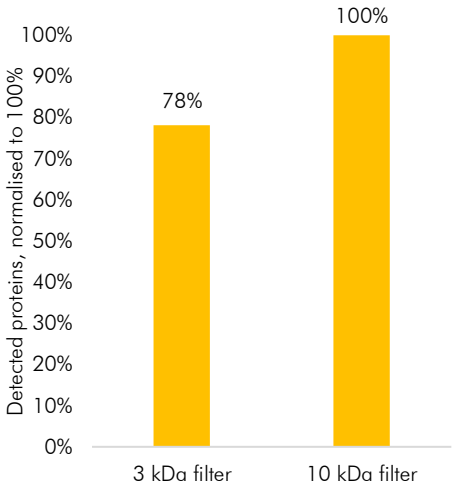


Figure 3-9. Evaluation of molecular weight cut-off filters for the preparation of urine for proteomic discovery mass spectrometry. *The vertical axis is normalised to the percentage of the maximum number of proteins detected. The 10 kDa filter resulted in 22% more protein hits than the 3 kDa filter.*

3.2.2.2 *Acetone precipitation to further decomplex and purify the sample*

It was hypothesised that an acetone precipitation step post-filtering would increase the coverage of proteins by further purifying the sample, extracting any residual salt or small molecules remaining. This was evaluated by filtering pooled urine according to section 3.2.2.1, using a 10 kDa filter. After filtration, the concentrate was split into two equal aliquots and four volumes of ice-cold acetone were added to one of the aliquots, which was then incubated at -20 °C for 16 hours. The acetone precipitated sample was centrifuged at +4 °C, 16900 x *g* for 10 minutes. The supernatant was discarded, and the both of the samples freeze dried before tryptic digestion and SPE-cleaning. The two samples were evaluated by targeted mass spectrometry, measuring the high-abundant urinary protein uromodulin. The results demonstrated that acetone precipitation rendered a 16% intensity increase compared to the filtered neat urine (Figure 3-10). The reason for the signal increase in the acetone precipitated sample is likely due to the removal of non-protein compounds, thereby resulting in a cleaner sample matrix with less interference and ion suppression.

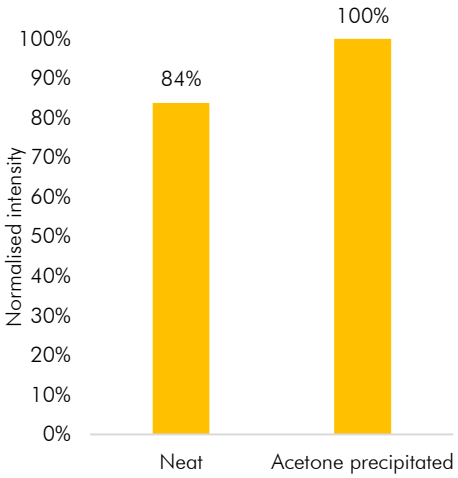


Figure 3-10. Comparison of uromodulin intensity in neat, filtered urine and filtered urine precipitated with acetone. *Acetone precipitation results in a 16% increase in signal intensity.*

3.2.2.3 *Digestion enzyme and digestion time*

The parameters of ideal digestion enzyme and digestion timing were extrapolated from the method development performed for plasma discovery proteomics, sections 3.2.1.3 and 3.2.1.4 of this chapter. It was consequently decided to utilise LCMS-grade trypsin for digestion with 16 hours incubation time.

3.2.2.4 *Chromatographic fractionation of urine*

Extrapolating from the plasma discovery proteomics method development, it was determined that bottom-up online fractionation would be utilised instead of top-down fraction of intact proteins as demonstrated in section 3.2.1.5. Taking the lesser complexity of urine compared to plasma into account, it was theorised that fewer fractions might be feasible to render sufficient protein coverage. This was evaluated by preparing a urine sample, filtered in a 10 kDa molecular weight cut-off filter followed by acetone precipitation, tryptic digestion for 16 hours and SPE-cleaning. The sample was thereafter

online fractionated into two, four, six and ten fractions and detected by untargeted mass spectrometry. The results demonstrated that ten fractions yielded 18% more hits than six fractions and 53% and 51% more than two and four fractions respectively (Figure 3-11).

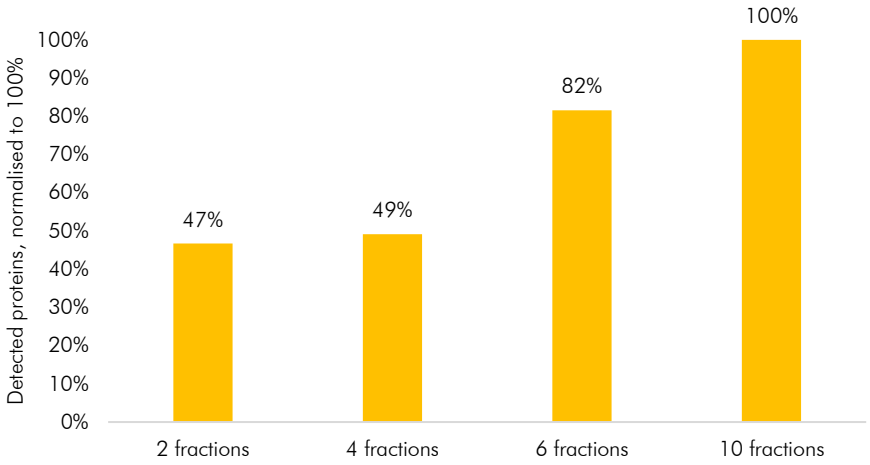


Figure 3-11. Detected proteins in urine prepared by online, bottom-up fractionation into two, four, six and ten fractions. *The vertical axis is normalised to the percentage of the maximum number of proteins detected. Ten fractions resulted in the greatest number of hits, 18% more than in six fractions and 53% and 51% more than in two and four fractions, respectively.*

The fractionation experiment demonstrated that an increased number of fractions indeed resulted in a greater protein coverage, it was therefore decided upon ten online fractions also for urine.

3.2.2.5 *Optimised workflow for proteomic biomarker discovery in urine*

Combining all the optimised steps, the final protocol for untargeted discovery urine proteomics was filtering of 2 mL urine using a 10 kDa MWCO filter, followed by acetone precipitation and digestion for 16 hours using LCMS-grade trypsin and fractionation into ten online fractions. The steps evaluated in the modules and chosen methods are illustrated in Figure 3-12. The final protocol for untargeted discovery proteomics urine sample preparation is presented in Chapter 2, section 2.4.

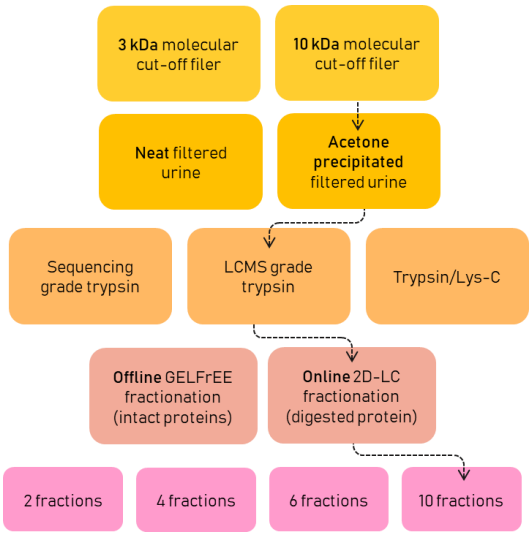


Figure 3-12. Optimised workflow for untargeted urine proteomics. *2 mL urine are concentrated using a 10 kDa MWCO filter and the concentrate acetone precipitates followed by freeze drying and digestion with LCMS-grade trypsin for 16 hours. In the instrumental analysis, the sample is fractionated into ten online fractions by 2D-LC.*

3.3 OPTIMISING THE INSTRUMENTAL PARAMETERS FOR UNTARGETED DISCOVERY MASS SPECTROMETRY PROTEOMICS ANALYSIS

The objective of the instrumental method development was to increase the number of detectable proteins in a sample. When performing discovery proteomics, it generally holds true that longer analysis-time will result in more protein identifications as sample de-complexion often involves extensive fractionation which reduces the number of peptides exiting the LC system and entering the MS simultaneously, thereby decreasing ion suppression and enabling more reliable detection of proteins as more peptides per protein will be measurable. The intent of the performed experiments in this section was to evaluate if it was possible to increase the number of detectable proteins and concurrently reduce the time of analysis.

In the setup utilised in the discovery experiments, the samples' analytes were fractionated into discrete fractions by a fractionation column and a high pH LC system, before being eluted on, one by one, onto the analytical column and a low pH LC system where the analytes were separated chromatographically prior to MS detection. Figure 3-13 gives a schematic illustration of the process.

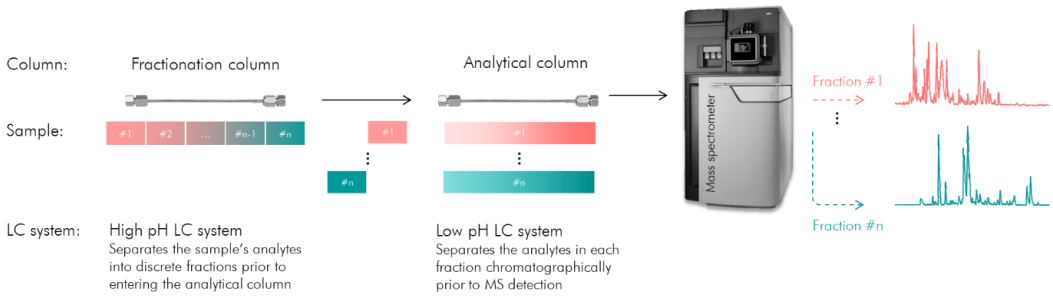


Figure 3-13. Set-up of the fractionation and chromatographic separation of a sample's analytes. *The analytes of a sample are fractionated into n fractions by the high pH LC system and a fractionation column. The discrete fractions thereafter sequentially enter the low pH LC system and analytical column where the analytes from each fraction are chromatographically separated prior to MS detection. After the first fraction has been analysed on the analytical column, the second fraction is eluted from the fractionation column and onto the analytical column. This repeats until the nth fraction has eluted off the analytical column and been detected by the mass spectrometer.*

The chromatographic separation and fractionation were optimised by an initial rough adjustment of the low pH analytical elution range (section 3.3.1), followed by an evaluation of the high pH fractionation and fine-tuning of the analytical elution parameters (section 3.3.2).

3.3.1 *Optimising the analytical chromatographic peptide elution range*

In the traditional LC-gradient set up utilised in our laboratory, peptides are eluted from the high-pH column at an increasing ACN percentage and separated on the analytical column with a linear gradient starting at 3% ACN and ending at 40% ACN. Generally, this leads to a relatively narrow elution range. It was hypothesised that by creating a gradient capable of better separating the peptides on the analytical column, ion suppression could be reduced and thereby, the number of detected proteins increased. Figure 3-14 demonstrates the traditional elution range when injecting 500 fmol of peptide standard 1 from Waters and the theoretically ideal elution range.

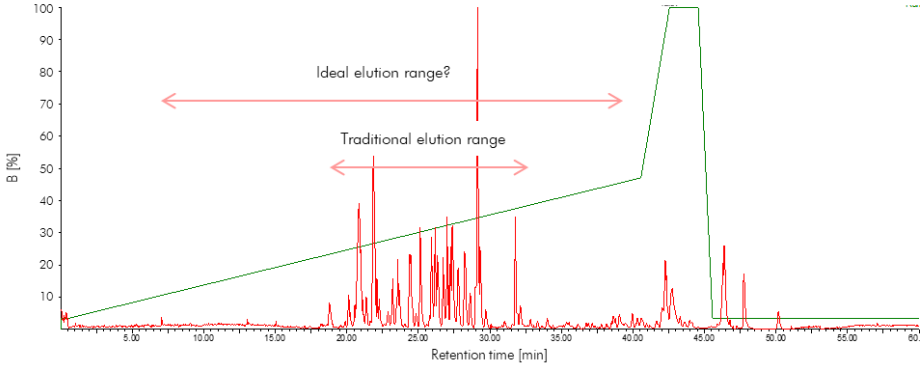


Figure 3-14. Elution profile on the analytical column, injecting 500 fmol of peptide standard, showing the actual elution range and the hypothetically ideal elution range.

Several different gradients were evaluated on the analytical LC-system, the high-pH system was bypassed and not utilised for the purposes of this experiment. The mobile phase in the analytical system consisted of A: water with 0.1% TFA and B: ACN with 5% DMSO and 0.1% TFA, pumped at a flowrate of 0.4 µL/min. The column was a 75 µm x 150 mm, 1.7 µm Peptide BEH C_{18} (Waters). The gradient most successful in separating the peptides was a curved rather than a linear profile and is illustrated in Figure 3-15. The chromatograms show that the curved gradient expands the elution range with approximately five minutes on each side. Several analytical column gradients were moreover evaluated in order to fine-tune the settings, these are described in the following section.

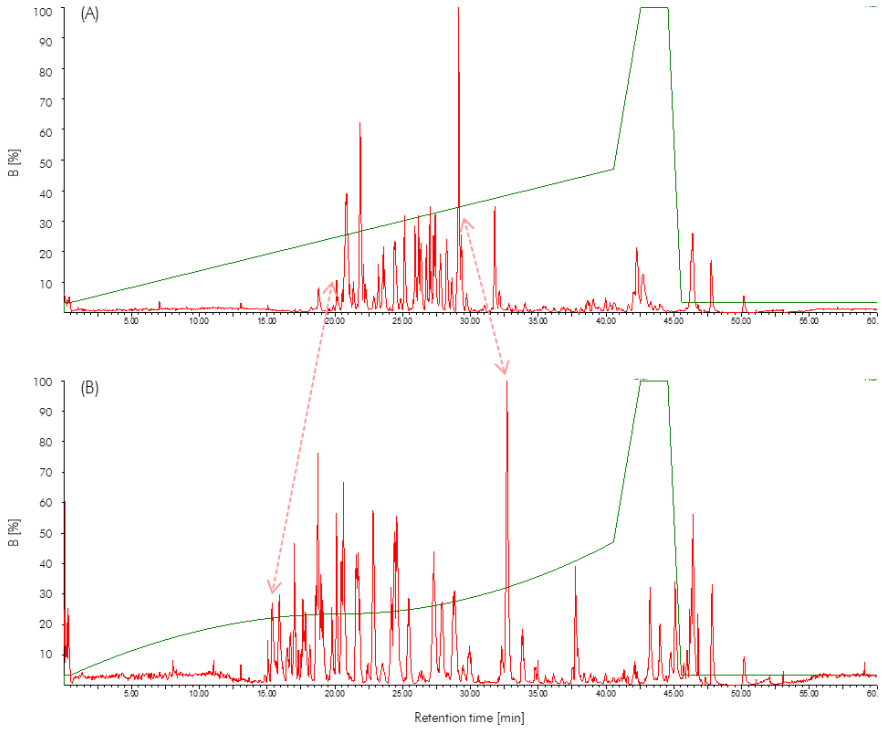


Figure 3-15. Gradient profile of 500 fmol peptide mixture comparing (A) the traditional linear gradient to (B) a curved gradient on the analytical column. *The curved gradient extended the elution range by approximately 5 minutes on each side of the traditional range.*

3.3.2 *Optimising the fractionation on the high pH column and fine-tuning of the analytical elution parameters*

Having extended the analytical elution range, thus improving separation of the peptides prior to entering the MS, attention was turned to the chromatographic fractionation system. As described in sections 3.2.1.5 and 3.2.2.4, ten fractions had been found optimal for both plasma and urine. However, it was considered that it might be possible to establish a more even distribution of the peptides eluting in each fraction. The following fractionation profiles were generated and evaluated on the high pH fractionation column and are described in detail as follows:

(i)	Profile # 1	Traditional percentage	Ten fractions
(ii)	Profile # 2	Even percentage	Ten fractions
(iii)	Profile # 3	Traditional percentage	Eight fractions
(iv)	Profile # 4	Even percentage	Seven fractions

The numbers of peptides eluting in the first (7.4% B) and second (10.8% B) fractions of the traditional gradient (*profile # 1*) are typically a great deal lower than in the subsequent fractions, thus it was hypothesized that it could be feasible to commence the fractionation at higher percentage of B and focus on improving the separation of peptides in later fractions. This strategy was implemented in *profile # 2* with ten fractions. It was further

considered if it would be possible to achieve an acceptable coverage with fewer fractions, thereby reducing the time of analysis. *Profile # 3* has the same final elution percentages as profile # 1 although the initial three fractions were merged thereby resulting in eight fractions. In *profile # 4*, the elution percentages of the first six fractions are identical to profile # 2 whereas the four last fractions have been merged, thus yielding seven fractions. The justification for merging the last rather than the first fractions in this experiment was that the elution would start at a percentage of B high enough to render a significant proportion of the peptides to elute in the first fraction, thereby aggravating rather than improving the separation. The percentages for the fractions in the different set-ups are listed in Table 3-3 and illustrated in Figure 3-16.

Table 3-3. Elution percentages and total analysis time of each fraction in the four different high-pH column profiles. *P = profile. The total time of analysis is given in hours.*

Fraction	1	2	3	4	5	6	7	8	9	10	Time
P # 1	7.4%	10.8%	12.6%	14.0%	15.3%	16.7%	18.3%	20.4%	23.5%	60.0%	13.8
P # 2	12.6%	14.9%	18.0%	21.3%	24.6%	29.7%	34.2%	39.3%	45.2%	60.0%	13.8
P # 3	12.6%	14.0%	15.3%	16.7%	18.3%	20.4%	23.5%	60.0%			11.0
P # 4	12.6%	14.9%	18.0%	21.3%	24.6%	29.7%	60.0%				9.7

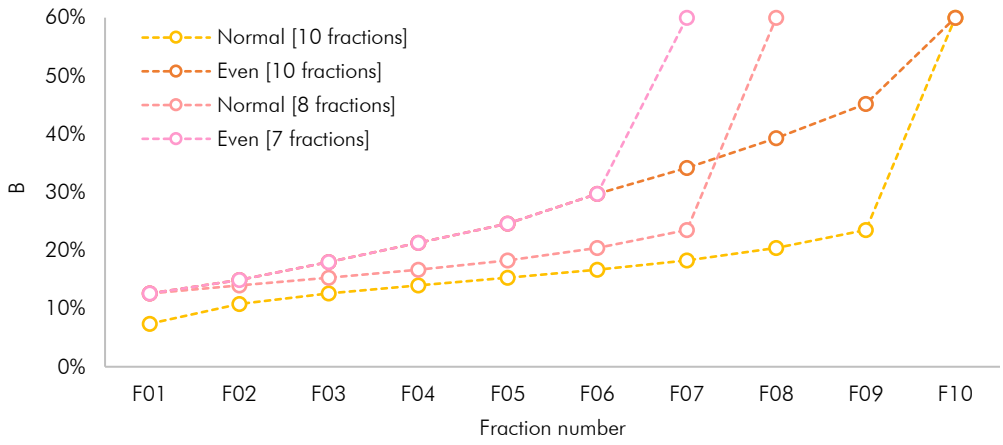


Figure 3-16. Percentages utilised in the different fractions to elute peptides from the high-pH column.

Fine-tuning the parameters for the analytical column set-up described in section 3.3.1, both the traditional linear gradient and two types of curved gradients were applied in conjunction with the high-pH column *profiles # 1 - # 4*. The linear gradient on the analytical column ran from 3% B to 40% B (except for the last fraction, running to 60% B) over 40 minutes. *Curve # 1* (identical to the set-up in section 3.3.1) was programmed to run from 3% B in a logarithmic profile up to the elution midpoint at 20 minutes, where the percentage of B reflected the percentage of the high-pH elution. Curve # 1 is equivalent to

the gradient developed in section 3.3.1, but altering the percentage at the midpoint to distribute the eluting peptides towards the beginning and end of the chromatogram. At the midpoint, the curve was set to an exponential profile running to 40% B for 20 minutes. *Curve # 2* had the same set-up as curve # 1, but with a lower percentage of B at the 20-minute midpoint. Figure 3-17 shows the different elution profiles on the analytical column.

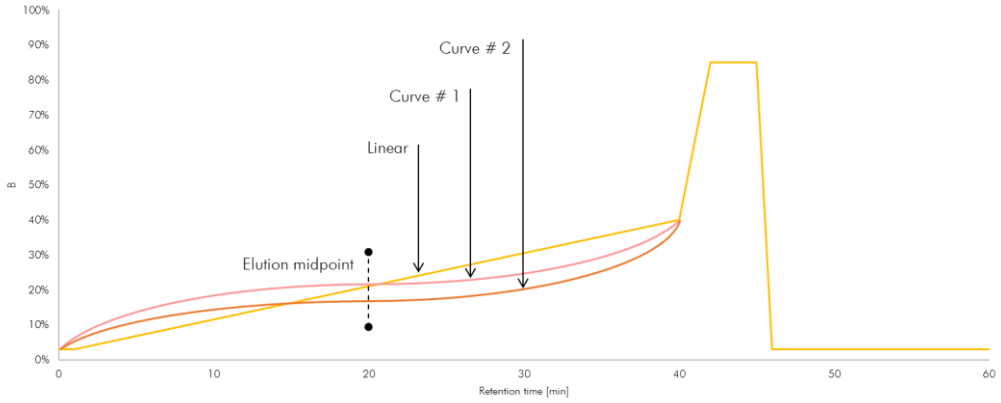


Figure 3-17. Gradients evaluated on the analytical column. *"Linear" is the traditional, linear gradient. "Curve #1" is a curved gradient utilising the same percentage of B at 20 minutes (mid-elution point (----)) as the high-pH column. "Curve #2" is a curved gradient with a mid-elution point percentage lower than the high-pH column*

Evaluating the fractionation experiment, a *Top12* depleted pooled plasma sample was utilised, prepared according to section 3.2.1.6. The high-pH fractionation mobile phase was composed of A: water with 10 mM NH_4OH (pH 9) and B: 100% ACN, pumped at a flowrate of 2 µL/min. The high-pH system was equipped with a 300 µm x 50 mm, 5 µm Peptide BEH C18 column (Waters). The analytical system's mobile phase was A: 0.1% TFA and B: ACN, 5% DMSO, 0.1% TFA, pumped at a flowrate of 0.4 µL/min. The column set up was a 180 µm x 20 mm, 5 µm Symmetry C_{18} trap column (Waters) trap in series with a 75 µm x 150 mm, 1.7 µm Peptide BEH C_{18} (Waters) operating at +45 °C.

The four different high-pH fractionation profiles were combined with the three different analytical profiles. In summary, the following 12 experiments were carried out:

(A) High-pH: *Profile # 1* combined with analytical column: *Linear/Curve # 1/Curve # 2*
(B) High-pH: *Profile #2* combined with analytical column: *Linear/Curve # 1/Curve #2*
(C) High-pH: *Profile #3* combined with analytical column: *Linear/Curve #1/Curve #2*
(D) High-pH: *Profile #4* combined with analytical column: *Linear/Curve #1/Curve #2*

The peptides were detected by untargeted mass spectrometry. The full results from the experiment are displayed in Figure 3-18. Experiments (A) demonstrated the typically observed pattern, with limited numbers of peptides in fractions # 1 and # 2, but other than that a relatively even distribution. It was clearly demonstrated that the peptide elution was left-skewed in experiments (B) where an even elution percentage was applied to the high-

pH column. The majority of the peptides in this experiment eluted in the early fractions, leaving few peptides left to elute in the final fractions. Experiments (C) demonstrated an overall even distribution. Experiments (D) demonstrated a relatively even distribution for curves #1 and #2 but a left-skewed profile for the linear analytical curve.

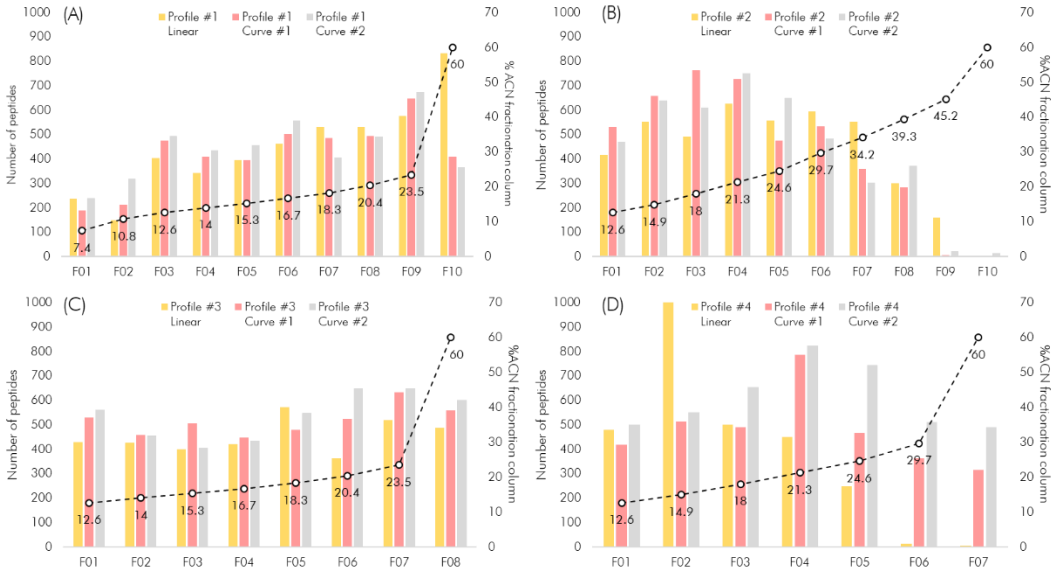


Figure 3-18. Number of identified peptides per fraction and condition in the evaluation of different 2D-LC gradients combined with three different gradients on the analytical column used to separate the peptides prior to MS-entry. *(A) shows the traditional gradient. (B) shows the evenly distributed ACN-increase profile. (C) shows the traditional gradient reduced to eight fractions by combining F01 - F03 of gradient (A). (D) shows the evenly distributed gradient from (B) reduced to seven fractions by combining F07 - F10.*

After merging the proteins identified in the individual fractions from each experiment (Figure 3-19), it was concluded that ten fractions with high-pH profile #1 in conjunction with analytical column curve #2 resulted in the largest number of identified proteins. The best compromise between number of hits and analysis time was determined to be eight fractions, high-pH profile #3 with analytical column curve #1. It was decided that the time saved by utilising eight fractions rather than ten (2.8 hours per sample) did not motivate the loss in identified proteins, thus the final protocol consisted of ten fractions, eluted from the high-pH column by profile #1 and separated on the analytical column applying curve #2.

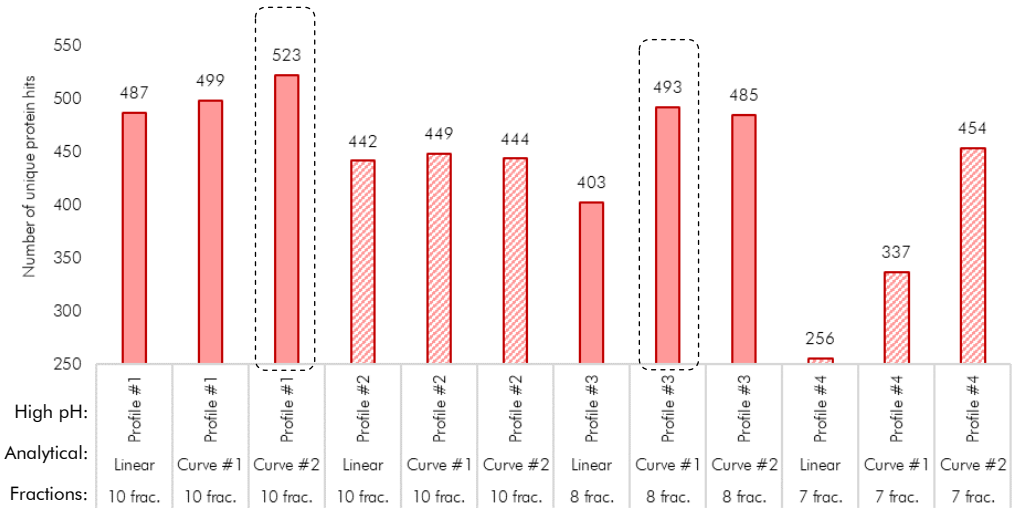


Figure 3-19. Number of total unique proteins in each of the fractionation experiments. *High pH profile #1 combined with analytical curve #2 resulted in the highest number of identifications while high pH profile #3 combined with analytical curve #1 was considered the best trade-off between analysis time and number of protein identifications*

3.3.3 *Final experimental set-up for untargeted discovery proteomics*

The final LC set-up for the untargeted discovery studies was 2D-LC fractionation according to profile #1, paired with curve #2 on the analytical column. In prior experiments carried out in our laboratory, it had been demonstrated that detecting the peptides with ion mobility enabled resulted in a greater coverage than without (data not shown), this was consequently included in the protocol.

The final instrumental settings for untargeted discovery proteomics are presented in Chapter 2, section 2.5.

3.4 METHOD DEVELOPMENT FOR TARGETED PROTEOMICS

As will be shown in chapters 4, 5 and 6, several proteins were found differentially expressed in the untargeted discovery studies of PD patients and centenarians. To confirm and validate the pathways affected and the putative proteomic biomarkers from the discovery studies, a targeted peptide assay was developed. It was decided to combine the putative biomarkers from the PD patients and the centenarians into one assay rather than two condition-specific assays; the rationale being that ageing is the greatest risk of developing PD, thus markers of delayed ageing from the centenarians could be of interest in the study of PD. Analogously, centenarians generally do not develop PD, therefore the absence of PD-markers expression would be of interest in the centenarian study.

The development of a targeted assay is a complex and time-consuming task, increasing in complexity and time-requirement with increasing number of peptides included. The development involves a number of steps that must be performed meticulously for the final assay to be functional and have sufficient sensitivity. The transfer from discovery to targeted proteomics also includes a change of instrumental platform, going from, in our case, an exceptionally sensitive high-resolution nano-LC-QTOF-MS set-up to a lower resolution UPLC-MS/MS instrument. Apart from the technical aspects involved, this transfer poses challenges as proteins detected in the discovery study may not be possible to detect in the targeted study. Figure 3-20 illustrates the targeted assay development process.

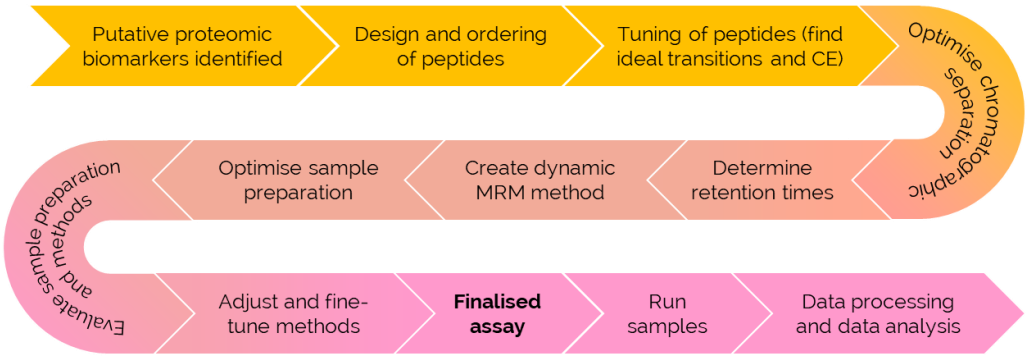


Figure 3-20. Workflow for developing a final targeted assay based on targets from discovery studies. *Identified putative proteomic biomarkers are developed into a targeted test by utilising synthetic peptide standards which are employed to develop ideal MRM and LC settings, the optimised and final assay can thereafter be applied to analysis of samples.*

3.4.1 *Construction of a targeted LC-MS/MS method*

The proteins in the targeted assay were selected from the discovery studies of Parkinson's disease and centenarians, described in Chapters 4, 5 and 6. Due to the suggested involvement of inflammation in neurodegenerative diseases and ageing, several known pro- and anti-inflammatory proteins identified from literature were also included in the multiplexed targeted assay. Targets were furthermore included from other related proteomic screening studies of plasma, urine and cerebrospinal fluid in Alzheimer's disease and other neurodegenerative conditions. This resulted in an assay monitoring putative markers of healthy ageing, inflammation, and neurodegeneration in general and specific to PD.

3.4.1.1 *Design of peptides for putative proteomic biomarkers*

The final targeted panel consisted of 127 proteins from discovery studies and literature. The following process was implemented designing the peptides:

- In-silico tryptic digestion of the target protein using Skyline [144]. This provides the theoretical peptides and peptide fragments from a protein
- Identification of potential peptide targets
- Amino acid sequence length preferentially 7 - 20 amino acids long
- Suitable amino acid sequence without repeated units of negatively charged amino acids (glutamic acid and aspartic acid)
- Peptide charge of $[M+H]^{2+}$ or $[M+H]^{3+}$ and suitable peptide fragments
- Peptide polarity, not too polar or apolar as this can cause peptides to elute in the LC void (no chromatographic retention) or cause the peptides to be retained too strongly on the column and not elute during the analytical LC range
- Check the peptide homology to other proteins using Basic Local Alignment Search Tool (BLAST) provided by UniProt [145]. The chosen peptides must be unique to the protein in question otherwise the measurement will be arbitrary.

Several of the proteins were represented by two peptides, leading to a total of 189 unique peptides. The selected peptides were purchased as synthetic peptide standards from GenScript (Amsterdam, Netherlands). Table 2-4, presented in Chapter 2, section 2.9, shows the amino acid sequences of the peptides included in the targeted assay.

3.4.1.2 *Determining peptide fragmentation and optimal collision energies*

The peptide standards were reconstituted in suitable solvents (for the vast majority this was 50% ACN) and a 1 pmol/µL pool containing all standards was created. Injections of 1 pmol peptide standard onto a Waters Acquity ultra performance liquid chromatography (UPLC) system coupled to a Waters Xevo-TQ-S triple quadrupole MS were performed and the two most high-abundant precursor-to-product ion transitions were selected for each peptide. Detection was performed in positive ESI mode. The capillary voltage was set to 2.8 kV, the source temperature to 150 °C, the desolvation temperature to 600 °C, the cone gas and desolvation gas flows to 150 and 1000 L/hour, respectively. The collision gas consisted of nitrogen and was set to 0.15 mL/min. The nebuliser operated at 7 bar. The peptide tuning was performed manually or using Skyline [144]. The optimal collision energies were determined by repeated injections, monitoring the most abundant transitions for each peptide, with altering collision energies. The voltages producing the most intense signals were chosen.

In the final assay, two transitions were selected for each peptide, one quantifier for concentration determination and one qualifier for identification, rendering a total of 378 analyte transitions.

3.4.1.3 *Chromatographic separation of peptides*

Chromatographic separation of the peptides was performed utilising a 1 x 100 mm, 1.7 µm ACQUITY UPLC® Peptide CSH C_{18} column (Waters). The mobile phase consisted of A: 0.1% formic acid and B: 0.1% formic acid in acetonitrile pumped at a flow rate of 0.2 mL/min. The gradient elution is described in Figure 3-21. The column temperature was set to +55 °C. Retention times of the peptides were determined by repeated injections of the peptide standards, 1 pmol on column.

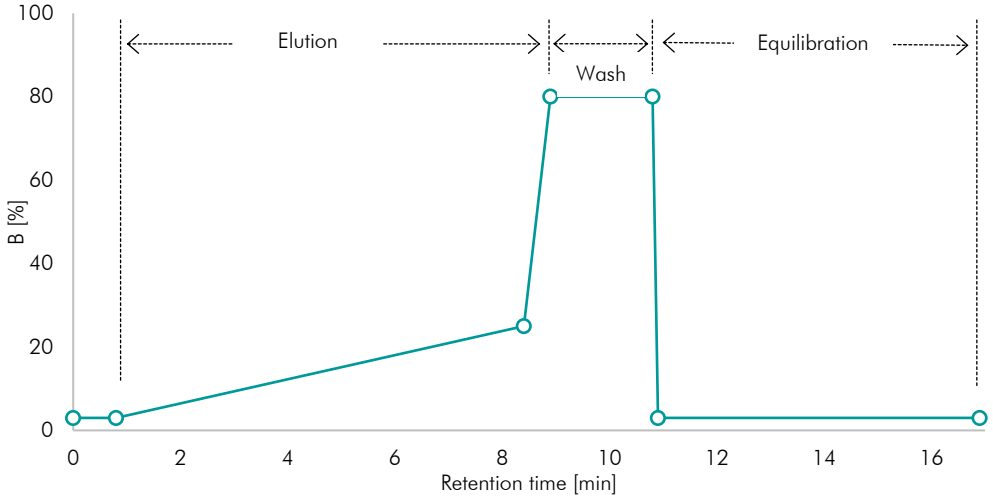


Figure 3-21. Gradient elution of targeted assay. *The starting conditions of 3% B were kept static for 0.8 minutes, before initialising the linear gradient, running for 7.6 minutes to 25% B, eluting the majority of the peptides. B was thereafter linearly increased to 80% over 0.5 minute and held for 1.9 minutes, eluting the most apolar peptides and washing the column, before returning to the initial conditions over 0.1 minute followed by equilibration for 6 minutes prior to the subsequent injection*

3.4.1.4 *Combining the LC and MS methods to a final assay*

The determined retention times were used to divide the peptides into timed segments containing maximum 24 peptides, measured by at least 12 points per peak and a dwell time of at least 100 milliseconds. Due to the large number of monitored analytes, the peptides were distributed over two MS methods. The peptides were sorted according to ascending retention time and split into timed acquisition segments. Figure 3-22 illustrates the strategy to allow maximum retention time windows for each segment. Dynamic MRM (dMRM) methods were created to only acquire data during the time of elution of the peptides in each segment. The LC flow was diverted to waste outside of the data acquisition interval.

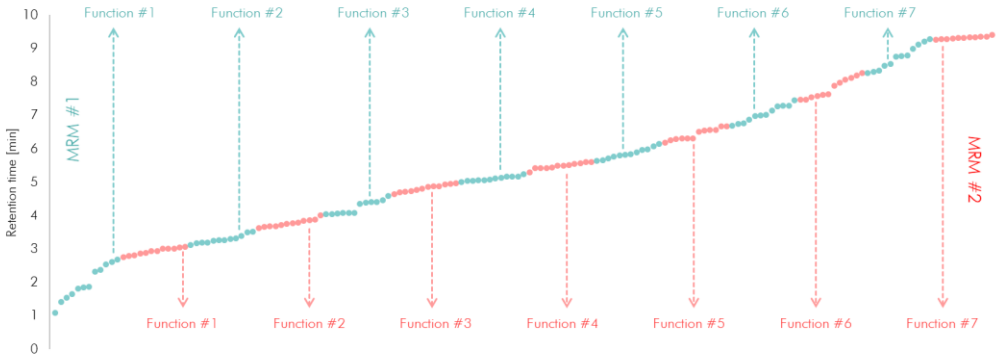


Figure 3-22. Illustration of the retention time-based segment division of the peptides included in the targeted assay for centenarians and PD. *The peptides were sorted by retention time and divided into acquisition segments allowing the largest possible windows for each segment*

Before the start of each targeted study, the retention time segments were adjusted and fine-tuned to ensure that all peptides could be detected. The elution profiles of the discovery and inflammatory markers upon injection of 1 pmol of peptide standards can be seen in Figure 3-23.

The final instrumental settings for targeted proteomics are presented in Chapter 2, section 2.9.

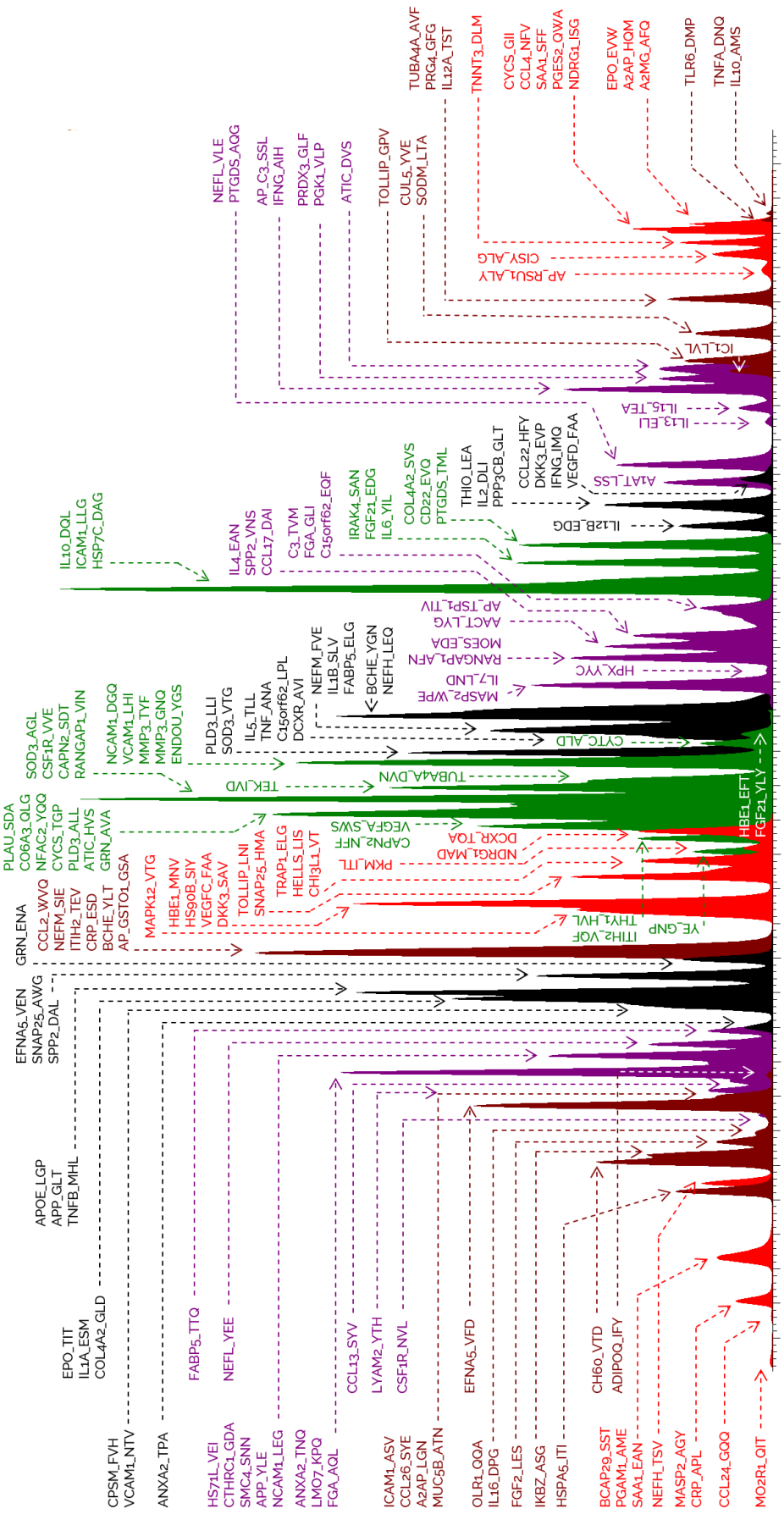


Figure 3-23 UPLC-MS/MS chromatogram showing the annotated peaks of the dynamic MRM functions of the targeted peptide assay. *The peptides are annotated by gene name or Uniprot ID, followed by the three first amino acids in the peptide sequence. Heavy isotope labelled internal standards are annotated by "AP" prior to the gene name.*

3.4.2 *Optimising sample preparation for targeted plasma proteomics*

As mentioned in section 3.2.1, plasma is a highly complex sample matrix and the sample preparation prior to targeted instrumental analysis must therefore be tailored to ensure that the analytes in the assay are detectable. In our case, with 127 unique proteins included in the assay, it was recognised that there would never be one sample preparation method ideal for all the proteins. Moreover, given the large number of samples that would be prepared for targeted analysis, the cost of sample preparation increasingly became a consideration to factor into the decision. The aim was consequently to develop a sample preparation method that allowed for detection of the maximum number of peptides at the lowest possible cost.

The following sample preparation procedures were evaluated:

(a) Neat plasma
(b) Precipitation by pure acetone
(c) Precipitation by acetone with 10% TCA
(d) Immunodepletion by *Top2*
(e) Immunodepletion by *Top12*

A pooled plasma sample was used for all preparations. The *Top2* and *Top12* samples were prepared as described in section 3.2.1.1. The acetone and acetone/TCA treated samples were precipitated overnight at -20 °C, centrifuged at +4 °C, 16900 x *g* for 10 minutes before the supernatant was pipetted off and discarded. All samples were freeze dried before tryptic digestion and solid phase extraction. The samples were reconstituted in 25 µL 3% ACN, 0.1% TFA per 10 µL plasma used. The injection volume was 5 µL.

In the comparison of the five different sample preparation methods, it was concluded that Top2 enabled detection of the highest number of compounds. In this method, 49% of the monitored analytes could be seen. Neat plasma, acetone precipitated plasma and Top12 depleted plasma resulted in a similar number of hits,

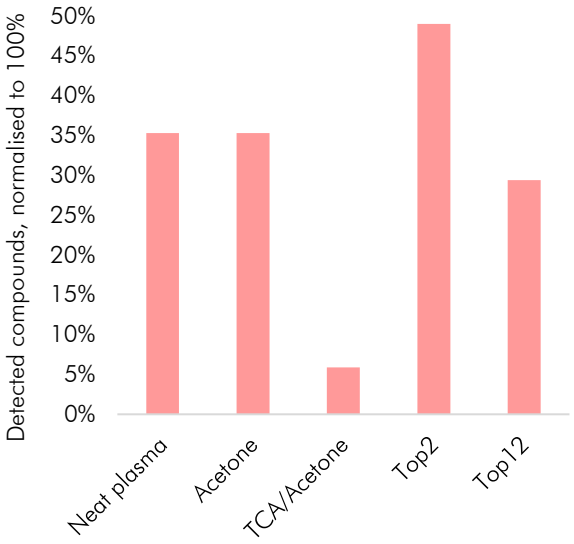


Figure 3-24. Sample preparation evaluation for targeted plasma proteomics. *Top2 resulted in detection of the largest number of compounds.*

while the TCA/acetone preparation under-performed compared to all other methods (Figure 3-24).

The results of the optimised preparation methodology of samples prior to targeted analysis demonstrated that Top2 depletion of 10 µL plasma allowed for the greatest number of proteins to be reliably detected. Although this resulted in an extra step and costs, it was decided that because of the extra sensitivity and reduction of matrix interference resulting from the removal of albumin and IgG by the Top2 column, that this would be the incorporated into the final targeted proteomics protocol.

The final protocol for targeted proteomics sample preparation of plasma is presented in Chapter 2, section 2.7.

3.4.3 *Sample preparation for targeted urine proteomics*

The preparation for targeted urine proteomics was kept identical as in the protocol developed for untargeted discovery urine proteomics in section 3.2.2, except for utilising 3 mL of urine instead of 2 mL to allow for a sufficient sample volume for the two injections required for the targeted assay.

The final protocol for targeted proteomics sample preparation of urine is presented in Chapter 2, section 2.8.

3.5 DEVELOPMENT OF TOOLBOXES FOR DATA ANALYSIS

After sample preparation and instrumental analysis, the mass spectrometric variables need to be identified, peak picked, integrated and quality controlled before statistical tests and biological interpretation can commence. In the case of untargeted proteomics, the process of peak picking, integration and identification is relatively streamlined as the software Progenesis QI-P guides the user through a set of well-defined steps. The analysis of targeted data is however more hands-on as the user has full control over all steps involved in turning raw data into a final table of results. The processing of targeted data can quickly turn into a time-consuming task, increasing in requirement with the number of samples and analytes as each peptide needs to be inspected in every sample.

Once the raw data have been processed and a table of results exported, a number of quality checks need to be performed to ensure that the data are of sufficient quality and that there are no confounding factors affecting the data before any actual interpretation of the

results can start. Common issues include instrumental drift and confounding effects of age and sex.

Aiming to expedite the targeted data processing and to develop a workflow for dealing with instrumental drift and confounders, a set of toolboxes were developed in the programming language Python.

3.5.1 *Development of a high-throughput peak picking application in Python for targeted proteomics assays*

Peak picking and quantitation are traditionally performed using vendor specific software. In the case of Waters instrumentation, the choice is normally TargetLynx, a quantitation application from the MassLynx package. The workflow of TargetLynx consists of creating compound specific methods utilising retention times and quantifier/qualifier transitions for identification, thereafter applying the methods to a run batch and adjusting poorly integrated peaks manually, sample by sample. Baseline corrections are often necessary, as peaks with fronting/tailing are managed differently depending on the size and shape of the front/tail. High-abundant peaks are in general well-handled as they in the majority of the cases are correctly integrated and require minimal manual correction. Quantitation using TargetLynx however becomes problematic if there is retention time drift, other peaks present in the chromatogram, poor baseline separation, poor peak shapes or a combination of some or all factors. In many cases, it is possible to adjust the quantitation methods so that they correctly identify the analyte peak, albeit often requiring a substantial amount of baseline re-drawing. Method adjustment is however in itself a time-consuming procedure including a significant amount of trial and error as there is not one ideal method adjustment applicable to all analytes. The result is a non-streamlined and time-consuming process with a substantial amount of user input in the form of manual adjustments and corrections.

The final targeted method included 189 unique peptides, each measured by two fragments, leading to 378 monitored transitions. The total number of samples in the targeted studies were approximately 600, meaning that around 220 000 peaks would need to be visually inspected, and a large proportion manually corrected. Assuming 5 seconds would be spent on each peak, it would take more than 300 hours to integrate the data. This was deemed not feasible.

To address this problem, a guided user interface (GUI) application was developed in Python, version 3.6 (Python Software Foundation, https://www.python.org/) and PyQt, version 5.6.0 (The Qt Company Ltd, Finland), aiming to create an user-friendly, clickable

software resulting in an accelerated data processing pipeline and a more robust procedure for peak integration. The GUI application allows for retention time alignment of peaks and simultaneous integration of all samples in a set. In essence, the time required to integrate one sample, or one thousand samples becomes the same. The application allows the user to specify the integration range, meaning that all peaks are integrated within an identical retention time window. A function with trapezoidal integration was utilised for determination of the peak areas.

In the initial step, the raw output files from the instrument are converted to text files using MSConvert from ProteoWizard [146]. The text files were thereafter imported to the application using the interface displayed in Figure 3-25. The interface allows the user to visualise the peaks generated by a transition in all samples simultaneously as an overlaid chromatogram.

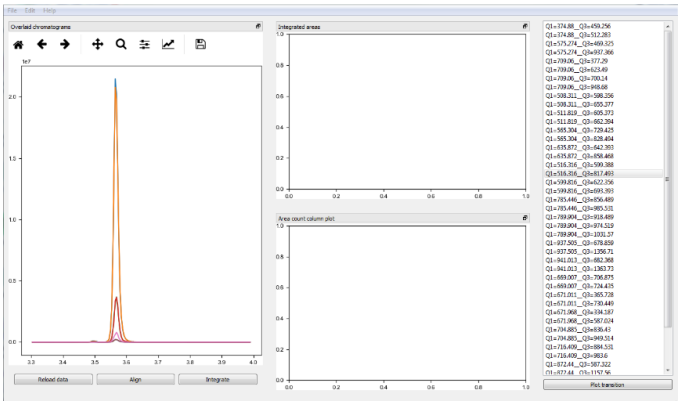


Figure 3-25. User guided interface of the targeted peak picking application. *The overlaid chromatogram for all imported samples is shown for the selected transition.*

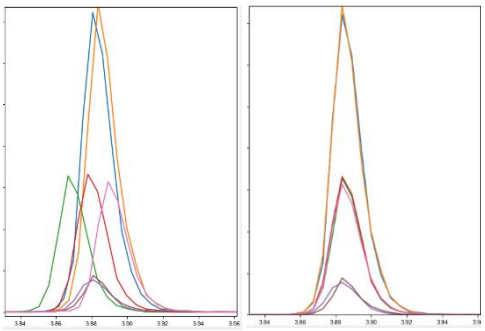


Figure 3-26. Example of a peptide demonstrating retention time drift. *The figure shows the chromatographic peaks before (left) and after (right) retention time alignment.*

If needed, the application allows the user to align peaks in the case of retention time drift. Figure 3-26 shows an overlaid chromatogram before and after alignment. The peak alignment function is implemented by centring on the apex of each peak within a given retention time span. As can be noted in the figure, all samples have been centred around the same retention time. In case of poor or failed alignment, the application allows the user to re-load the original, non-aligned data and redefine the retention time span.

Once aligned (if needed), the user can specify the integration range, upon which the integrated range of all peaks will be displayed and also the intensities for each sample

(Figure 3-27). After the integration has been completed, the results can be exported to a .csv file for further processing.

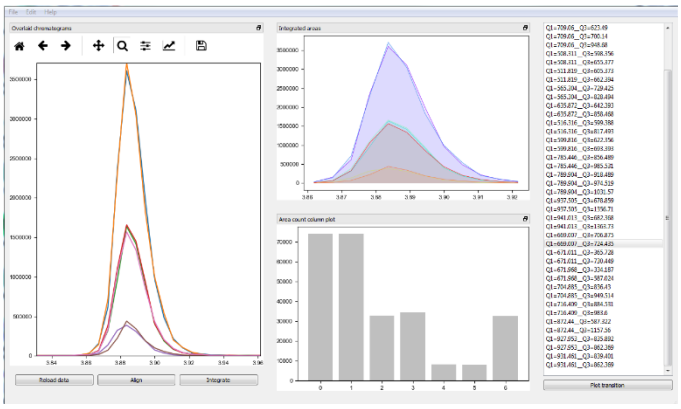


Figure 3-27. Resulting plots after integrating a MRM transition in the targeted peak picking application. *The integrated range and the resulting intensities are shown for the selected transition.*

To evaluate the performance of the GUI application, a sample set consisting of 176 samples was integrated using TargetLynx from Waters, and the GUI application.

The results from the integration of the quantifier and qualifier ions of peptides from Apolipoprotein E and Alpha-2-antiplasmin are shown in Figure 3-28. It was clearly demonstrated that the difference between the areas resulting from the two peak picking methods was minimal.

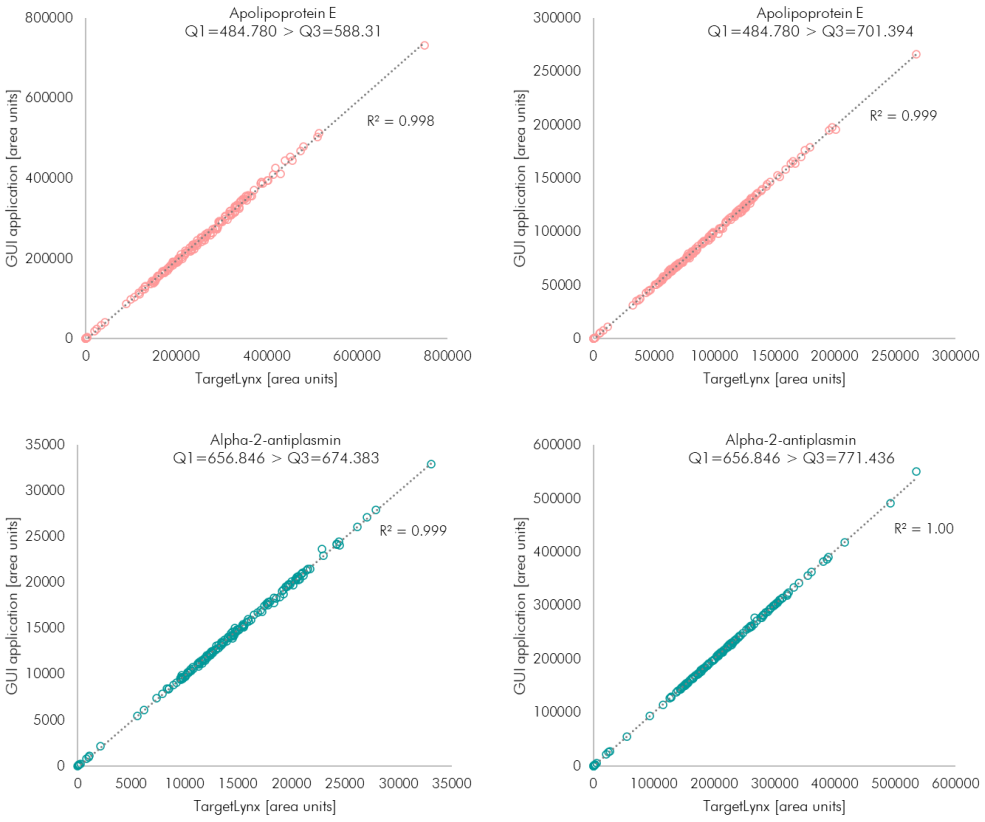


Figure 3-28. Comparison of integrated areas of peptides from Apolipoprotein E and Alpha-2-antiplasmin, produced by TargetLynx and an in-house GUI application. *The comparison shows the resulting area units after integration of two peptide fragments from each protein, where the precursor m/z is given by Q1 and the product m/z is given by Q3.*

The improvement in processing time and consistency in retention time windows rendered by the GUI application made this the primary strategy for peak picking. In a few cases, when peaks were of very low-abundance and/or it was concluded that each sample needed individual assessment, TargetLynx was utilised for quantitation.

In conclusion, an in-house Python-based application was created to automate the integration process, and greatly increased the throughput of data processing and analysis.

3.5.2 *Development of a strategy for correcting instrumental drift*

Instrumental drift can occur in both untargeted and targeted analyses. Generally, it is caused by parts of the instrument getting contaminated from contact with samples and leads to changes in sensitivity correlated with the run order of the samples. In targeted analyses, this effect can be mitigated by the use of internal standards. However, to fully correct the drift, each analyte would need to have an isotope-labelled internal standard, and this is rarely the case in large-scale multiplexed assays.

A script was developed in Python, utilising the function LOWESS from the package Statsmodels version 0.13.0 [147]. The samples were sorted according to run order and a Locally Weighted Scatterplot Smoothing (LOWESS) curve was individually fitted to each variable, a smooth curve thus capturing the overall drift for the variable by, in brief, fitting multiple polynomials to the data using weighted least squares [148]. The script was set to loop through all variables and return individual LOWESS points for each protein. The original data points were thereafter divided by the LOWESS points, resulting in individual drift correction of each variable. Figure 3-29 illustrates an example of a protein affected by run order drift, the fitted LOWESS curve, and the corrected data.

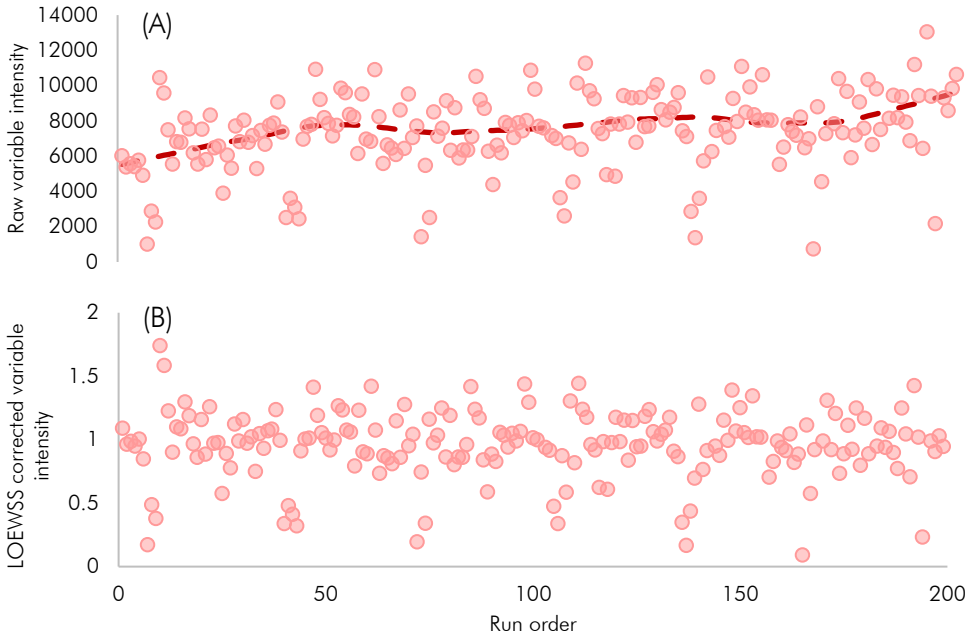


Figure 3-29. Example of a protein affected by run order drift. *(A) shows the raw data and the fitted LOWESS curve, (B) shows the protein after LOWESS drift-correction.*

3.5.3 *Development of a strategy and a script for outlier removal*

An outlier can be defined as any value deviating from the expected normal behaviour of a dataset. Outliers can create difficulties in data interpretation as they affect mean values, significance tests and fold-change calculations based on sample means. Outliers are moreover poorly handled by classification and machine learning models and may lead to non-robust models with unexpected behaviour [149]. For these reasons, it is necessary to adapt a strategy for dealing with outlier values. One option is to model the data by principal component analysis and delete any sample that falls outside the threshold set at a certain distance from the model's centre by Hotelling's T^2 distribution [66]. Although useful, this strategy encompasses that a whole sample is deleted. A more lenient strategy is to only target the extreme individual values and keep the remaining ones, thus avoiding the deletion of whole samples.

Traditionally, a certain number of standard deviations from the mean have often been utilised for outlier detection. This approach is however not without problems, since both mean and standard deviation are sensitive to outliers, meaning that the method for detecting outliers is itself affected by outliers. A more robust option is the median absolute deviation (MAD) as the median is less sensitive to extreme values [150].

For a feature variable $x_1, x_2, x_3, ..., x_n$, MAD can be defined as the median of the absolute deviations from the data's median, written as (3-1) [151, 152]:

$$MAD = M_i(|x_i - M_j(x_j)|) \quad (3\text{-}1)$$

where x_i is the initial observation and M_i the median of the series of initial observations subtracted by the median of the absolute deviation ($M_j(x_j)$).

One important consideration in outlier analysis is that the thresholds must not be set too conservatively. As the data are biological and expected to demonstrate inter-person variation, the outlier detection needs to take this into account and not filter out outliers too strictly. Taking this in consideration, the threshold was set to detect data points deviating more than ten MADs, thus only excluding the most extreme outliers and keeping the less severe ones. A script was developed in Python, going through the data variable by variable, and replacing outliers by missing values. The script moreover plotted the data before and after outlier removal, allowing for a visual overview of the process. Figure 3-30 illustrates example data from five variables before and after outlier correction.

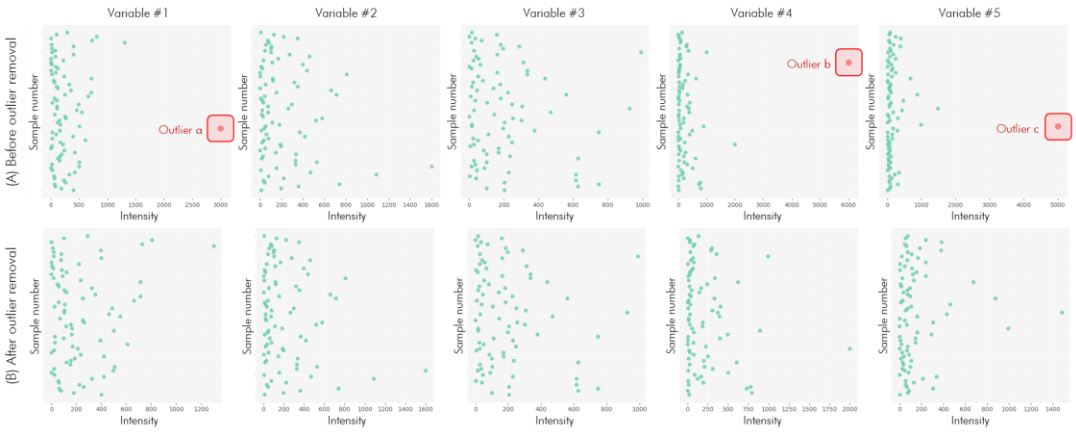


Figure 3-30. Example of data before and after outlier removal at a threshold of 10 median absolute deviations (MADs). *The figure shows five variables with outliers (a, b and c) detected at 10 MADs in variables #01, #04 and #05. (A) shows the data points for the variables before outlier removal, with the outliers highlighted in pink. (B) shows the same variables after outlier removal*

3.5.4 *Development of a script for age and sex adjustment*

The age and sex of a person can have an influence on their protein expression as described in several studies and also demonstrated in this thesis [153]. Age and sex can be used as informative variables when modelling data, but can also act as confounders, influencing both the dependent and independent variables, thus biasing the data interpretation. Therefore, it may in some situations be desirable to adjust the data for age and/or sex.

A script was written in Python, where each protein was set as the dependent variable Y and modelled with age and sex as independent variables using linear regression. The

protein values for each sample were thereafter predicted in the model and the residuals calculated as:

$$y_{j[res]} = y_{j[obs]} - y_{j[pred]} \quad (3\text{-}2)$$

where $y_{j[res]}$ is the residual, $y_{j[obs]}$ is the observed value for the point j and $y_{j[pred]}$ is the predicted value for point j (illustrated graphically in Figure 3-31). The residual vector is the proportion of the protein expression not related to the independent variable and thus contains the adjusted data. The script was run individually for all the proteins and the residuals extracted.

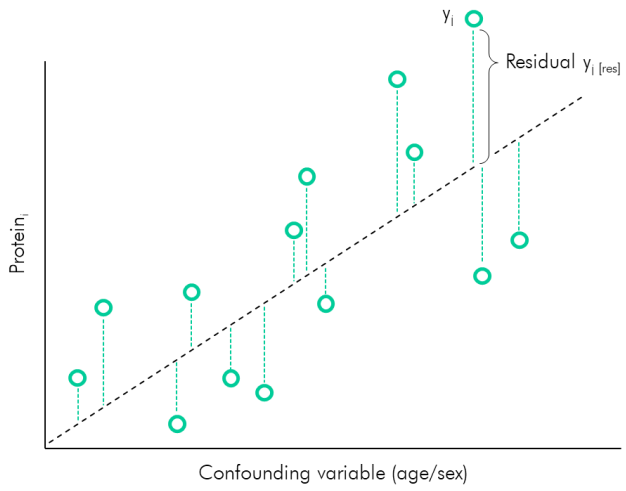


Figure 3-31. Illustration of the adjustment of data affected by age or sex. *The scatter plot shows the protein expression for protein$_i$ as a function of a confounding factor, x. The regression line represents the predicted y-values and y_j the observed value for observation j. The distance between the observed and predicted values is the residual and the proportion of the vector y not related to the confounding variable x.*

The adjustment was evaluated in several datasets and proved efficient in removing age and sex effects. Figure 3-32 shows an example of data corrected for age and sex. The initial data proved highly dependent on both age and sex; this was evaluated through multivariate analysis. An OPLS model with age set as the dependent variable y was significant ($p = 8E^{-7}$), as was an OPLS-DA model of males versus females ($p = 7E^{-24}$). After age and sex correction, both models exhibited insignificant relationships with age and sex, thus demonstrating the correction was successful. Moreover, the correction did not significantly affect the multivariate models related to disease.

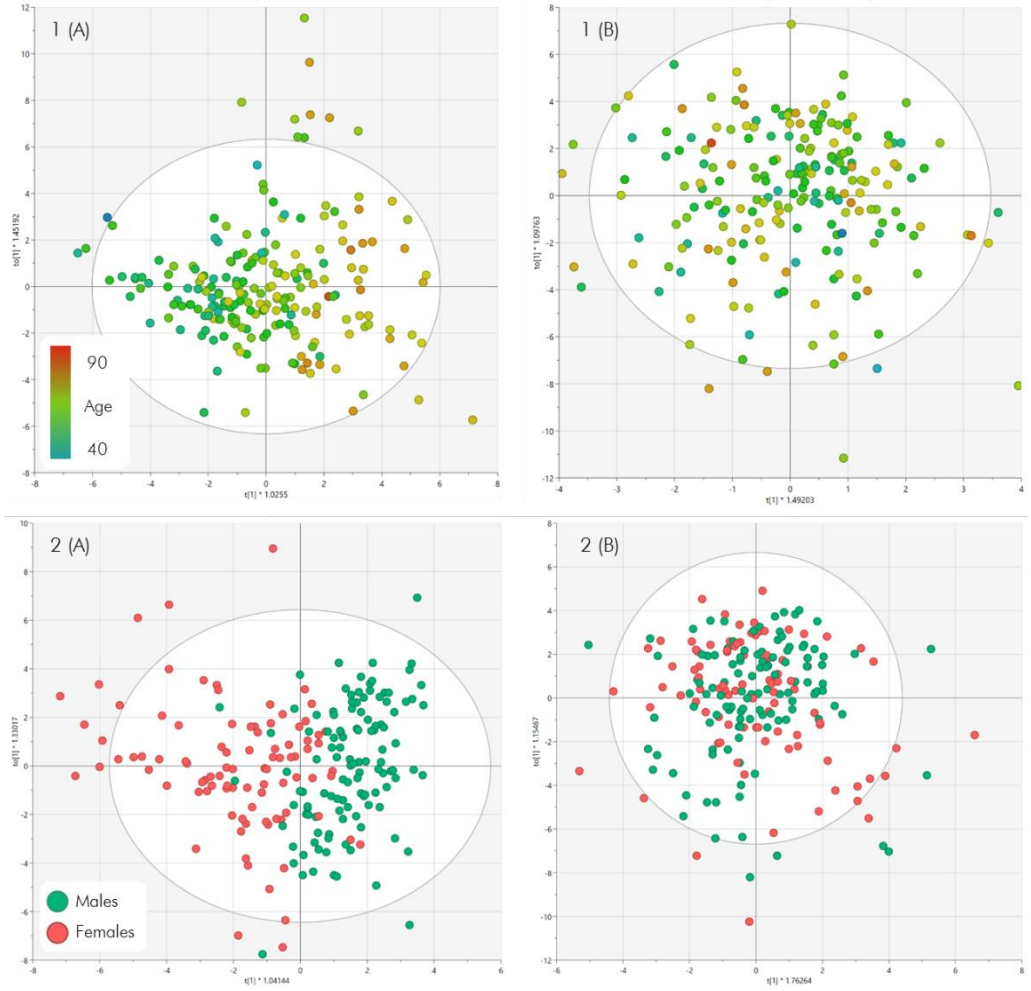


Figure 3-32. Age and sex adjustment of a dataset highly affected by both variables. *1 (A) shows an OPLS model of the data with age set as Y, where a trend in age can be seen along the x-axis. 1 (B) shows the same model after successful age-correction. 2 (A) shows an OPLS-DA model of males versus females, where a discrimination is possible between the two sexes. 2 (B) shows the same model after successful sex-correction. In all figures, t[1] represents the predictive principal component and to[1] represents the orthogonal component.*

3.5.5 *Development of a script for comparing machine learning prediction models*

In the targeted datasets presented in Chapters 4, 5 and 6, it was desired to create predictive machine learning models to determine if Parkinson's disease/centenarianism could be modelled from the measured proteins. OPLS-DA was modelled using the commercial software Simca (Umetrics). However, other machine learning models (described in Chapter 1, section 1.4) were implemented using functions and scripts written in Python. As mentioned, there exists a vast number of supervised machine learning algorithms to choose between for predictions, and as they are built on different mathematical foundations, a model that was suitable for one dataset may not be ideal for the next. It is in other words a matter of trial and error to find the optimal model for a certain dataset, which can turn into a time-consuming exercise.

Addressing this, a script was written in Python including three different classification algorithms: support vector machine (SVM), linear discriminant analysis (LDA) and Ridge regression classifier. The script imports a two-condition dataset and performs the following tasks:

- Perform k-fold cross-validation, using five cross-validation splits
- Extract the scores for the goodness of model fit from the five cross-validation splits
- Split the data into two groups, one for model training and one for prediction
- From the training data, determine the ideal number of predictors to use in the SVM and LDA models using recursive feature elimination
- Determine optimal settings for the Ridge regression classifier based on the training data
- Optionally reduce the number of predictors in the Ridge regression classifier model
- Build and fit models from the training data
- Calculate model scores
- Predict the test set in the newly built models
- Optionally predict an additional class of samples to test the models' specificity for the tested condition
- Calculate, for each model, the proportion of correctly and incorrectly predicted samples, and the sensitivity (3-3), specificity (3-4), and accuracy (3-5) given by [154, 155]:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3-3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3-4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3-5}$$

where TP = true positives, FN = false negatives, TN = true negatives, and FP = false positives.

The development of this script allowed for rapid and streamlined determination of the ideal machine learning model to use for a specific dataset. The example illustrated in Figure 3-33 demonstrates that, in this case, the LDA model performed superior to the SVM model and the Ridge classifier, which can also be seen in Table 3-4 displaying the classification metrics of the three models.

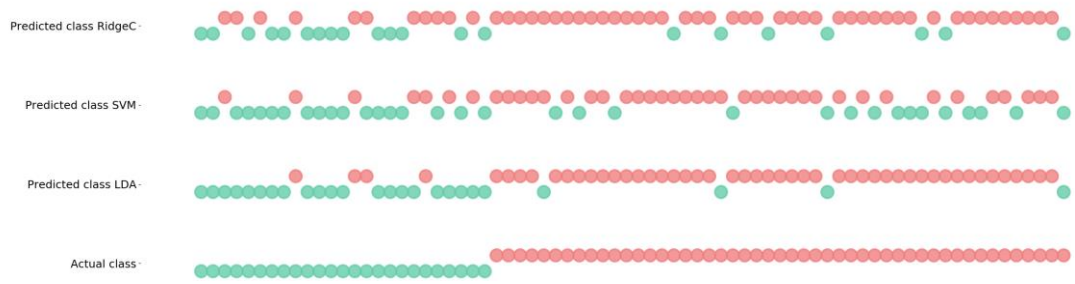


Figure 3-33. Prediction of the test dataset in SVM, LDA and Ridge classifier models. *The bottom row shows the actual classes of the samples where ■ control and ■ disease. The top three rows show the prediction results of the samples as control or disease from the three different classifying models. The LDA model performs superiorly when compared to the SVA and Ridge classifier models*

Table 3-4. Classification metrics from the prediction of a test dataset by the three machine learning models LDA, SVM and Ridge classifier. *The table shows the number of samples predicted as disease and control by the models, and the true number of samples in the classes. Accuracy, sensitivity, and specificity are reported for each model. LDA yields the best results.*

	LDA	SVM	RidgeC	Actual classes
Predicted as control	21	18	14	25
Predicted as disease	45	34	42	49
Accuracy	89%	70%	76%	
Sensitivity	92%	83%	79%	
Specificity	84%	55%	67%	

3.5.6 *Data processing and analysis workflow*

One important part of the data analysis development was to establish a pipeline that could be applied to all the datasets, ensuring consistency in both quality control and analysis. With the toolboxes developed and described in this section, the analysis workflow was streamlined, and the targeted data processing greatly expedited. Figure 3-34 illustrates the steps included in the processing and analysis of the targeted and untargeted data.

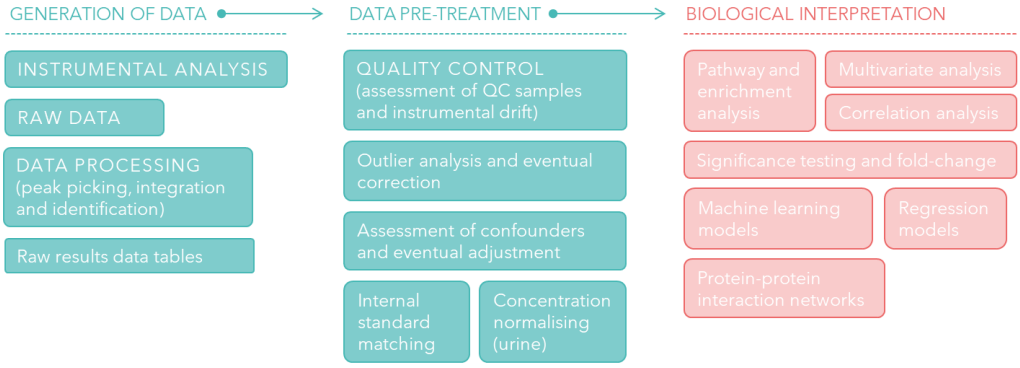


Figure 3-34. Data analysis pipeline for targeted and untargeted proteomics. *The raw data is first processed by Progenesis QI-P (untargeted) or an in-house Python application TargetLynx (targeted). Quality control checks are then carried out, evaluating instrumental drift, outliers, and confounders. Once the data have been pre-processed and the quality deemed adequate, several tools can be used for biological interpretation*

3.6 DISCUSSION

By carefully optimising the steps in the discovery proteomics sample preparation and analysis process, we increased the possibility of detecting lower abundant proteins and thereby improved the chances of identifying blood- and urine-based biomarkers. In the studies presented in this thesis, this was especially important as blood and urine were used to study Parkinson' disease, a condition usually researched by exploring CSF due to its proximity to the brain. Plasma and serum are not ideal biofluids for this purpose, in part because of their further distance from the brain but mainly because of their high abundance of house-keeping proteins, present at high concentrations and often considerable in size, making it difficult to detect small and/or low abundant proteins. Urine on the other hand, provides a less challenging matrix but is even further from the brain, thus emphasising the importance of proficient detection of proteins present at low concentrations. Comparing the number of detected proteins in a non-depleted and non-fractionated sample to a sample prepared using the optimised workflow, the number of detectable proteins was greatly increased.

With the development of such an extensive multiplexed targeted assay, it was apparent that quality control of the produced data was also a requirement for interpretation of the complex datasets. Therefore, method development was extended to include several data handling and analysis tools. Scripts were designed to optimise and extract information from the acquired data. The refinement of all aspects of the analyses were then applied to the cohorts of samples described in this thesis.

There are a few limitations to the method development experiments that need to be considered. The most important is the use of single replicates in the untargeted proteomics sample preparation development. The reason why single samples were used and not replicates was mainly due to time constraints as each fractionated sample takes between nine and 14 hours depending on the settings. Another consideration in the untargeted method development is that the entire chain of digestion parameters ideally should have been evaluated as continuous experiments rather than each step separately. In the fractionation of discovery samples, it would have been possible to fractionate the digested peptides offline in conjunction with the solid phase extraction and evaluate alongside the IEF and online fractions. In the experiment of trypsin efficiency as a function of digestion time, it is acknowledged that albumin may not be fully representative of the digestion efficiency and that several proteins ideally should have been evaluated to

fully capture the effect of the digestion time. Finally, in the targeted method development, it is recognised that also other sample preparation techniques could have been evaluated.

In the end, the methods developed for the untargeted discovery analyses of plasma and urine are time-consuming and low throughput, but it was argued that the gain in protein coverage justified the extra preparation and analysis time.

The use of proteomic techniques to study healthy ageing and identify markers of longevity in centenarians

4

Abstract. *Ageing is a complex and multi-factorial process, often leading to life-threatening diseases. Centenarians are individuals who, remarkably, have escaped or survived the most severe pathologies and reach an advanced age with their cognitive abilities intact at a higher frequency than the normal population. For this reason, centenarians can give us clues into what drives healthy ageing and longevity, and they may also help us understand the divergence between healthy ageing and the development of neuro-degenerative conditions.*

In this study, we applied an optimised discovery proteomics workflow to a cohort of samples consisting of healthy controls (n = 10) with a mean age of 68.8 years, and a group of cognitively healthy centenarians (n = 10) with a mean age of 103.8 years. The samples were profiled using label-free proteomics mass spectrometry with the aim of identifying markers of healthy ageing and longevity. Several proteins were differentially expressed between the groups and pathway analysis indicated that inflammatory pathways were upregulated in the centenarians.

Thirteen proteins found in the discovery phase were added to a targeted, mass spectrometric, MRM-based assay, where several pro-and anti-inflammatory proteins from literature were also included. The targeted assay was applied to a new and larger set of samples (n = 186) consisting of centenarians free of cognitive decline, healthy controls, and children of centenarians. The targeted analysis validated eight of the proteins from the discovery study; these proteins were A2M, ADIPOQ, CST3, PTGDS, PKM, SAA1 and SERPINA1.

Comparing our results with studies of ageing from the literature, it was hypothesised that the centenarians exhibited altered expression of several proteins and that some of these proteins likely exert protective functions, while others may indicate that they are closer to the end of their lives. We saw an increase of several inflammatory proteins, thus indicating an overall elevated state of inflammation in the centenarians. However, several anti-inflammatory acting proteins were also differentially expressed, possibly counteracting some of the detrimental inflammatory effects.

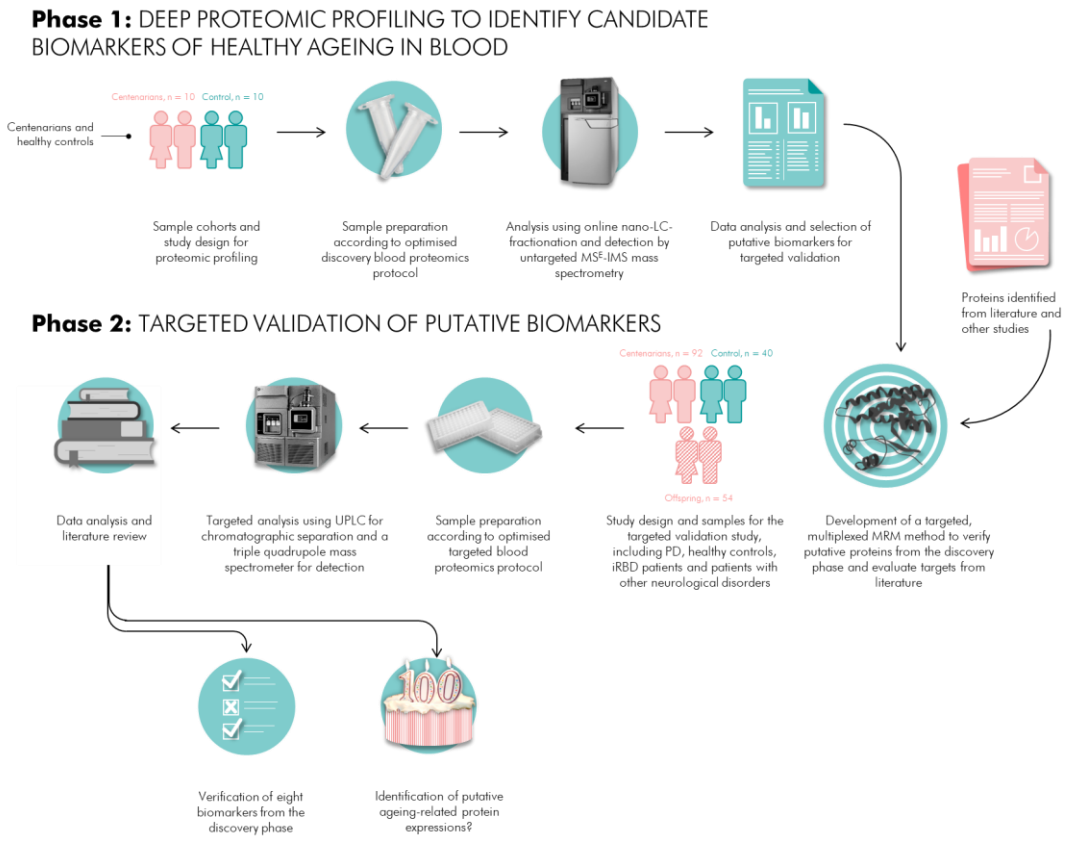


Figure 4-1. Graphical abstract of the blood-based discovery and validation study of centenarians presented in this chapter.

4.1 INTRODUCTION AND AIMS

Centenarians are an intriguing population with individuals who have remarkably avoided or survived events usually leading to mortality. To this day, we do not know for certain if their longevity is due to genetics or environment – or a combination of both. Studies of twins found that approximately 25% of the longevity could be explained by genetics while 75% appeared to be due to environment and chance [156]. Centenarians have been shown to develop neurodegenerative diseases at a lower rate than the normal population, and when they do, it is with a markedly delayed onset [133]. As age is the greatest risk factor for the two most common neurodegenerative conditions, Alzheimer's and Parkinson's diseases, this is highly interesting and it has been proposed that centenarians have protective mechanisms providing resilience in place [134]. Centenarians consequently make an ideal population to study healthy ageing. The comparison between centenarians and Parkinson's patients may provide additional information and insights relating to both groups. However, in this chapter, the focus will be solely on the centenarians; the comparison of centenarians and Parkinson's disease patients will be examined in Chapter 7.

In the experiments performed in this chapter, a carefully optimised plasma proteomics workflow was applied to a discovery cohort aiming to identify differentially expressed proteins between centenarians and controls. Putative biomarkers were developed into a targeted assay, also including pro- and anti-inflammatory proteins from literature, and applied to a new, larger set of samples to validate the discovery results.

The aims of the experiments performed in this chapter were:

- In the discovery phase, find novel targets and pathways related to the longevity observed in centenarians
- Develop the markers from the discovery phase into a targeted MRM assay and validate the discovery results
- Assess the expression of several pro- and anti-inflammatory proteins from literature to determine if the centenarians exhibit differences compared to controls

4.2 DISCOVERY PROTEOMICS OF CENTENARIANS AND CONTROLS TO IDENTIFY MARKERS OF HEALTHY AGEING

In the discovery phase of the centenarian study, the untargeted plasma proteomics workflow developed in Chapter 3 was applied. The purpose was to identify proteins with differential expression between centenarians and controls, aiming to find putative markers of healthy ageing and longevity, and to include these in a targeted assay to be confirmed in a new and larger set of samples.

4.2.1 *Materials and methods*

4.2.1.1 *Sample cohort*

The plasma samples from controls and centenarians were collected and provided by IRCCS Istituto delle Scienze Neurologiche di Bologna, Italy. The control group consisted of healthy individuals, and the centenarian group of elderly but cognitively healthy persons. The mean age of the control group was 68.8 years, and the mean age of the centenarians was 103.8 years, leaving an age difference of 35 years between the two groups. A summary of the characteristics of the sample groups is presented in Table 4-1.

Table 4-1. Sample characteristics of the discovery proteomics plasma samples from centenarians and controls. *The mean age of the controls was 68.8 years, and for the centenarians 103.8 years. In the control group, 60% of the samples were females while in the centenarian group, 80% were females*

Group	Number of samples	Percentage males / females	Age ± SD
Control	10	40% M | 60% F	68.8 (± 6.7)
Centenarian	10	20% M | 80% F	103.8 (± 2.9)

4.2.1.2 *Preparation of plasma samples for discovery proteomics*

The protocol for the preparation of plasma proteomics discovery samples is described in detail in Chapter 2, section 2.3. Briefly, ten microlitres of plasma were depleted from high-abundant proteins to allow for detection of medium- and low-abundant species. The proteins were digested into tryptic peptides before solid phase extraction.

4.2.1.3 *Instrumental analysis*

After sample preparation, the centenarian and control samples were online fractionated into ten fractions and analysed on a Waters Synapt-G2-S*i* quadrupole time of flight mass spectrometer utilising ion mobility. Detection was performed in positive MS^E mode as described in Chapter 2, section 2.5.

4.2.1.4 *Data analysis*

The acquired data were analysed as described in Chapter 2, section 2.6. In summary, the data were processed fraction-wise in Progenesis utilising the ion-accounting workflow. The fractions were merged, and the data were filtered according to the criteria described in Chapter 2, section 2.6. Run order drift was observed and LOWESS scaling was applied to correct for this as described in Chapter 3, section 3.5.2. After LOWESS scaling no run order drift was observed.

4.2.2 *Results from the discovery analysis of centenarians and controls*

Utilising the protocol developed for discovery plasma proteomics in Chapter 3, 875 proteins were detected in the analysis of centenarians and controls. Out of these identified proteins, 420 had a confidence score above 15 and two or more unique peptides. 74 of the proteins were found to be significantly different between the two groups on a nominal 95% confidence p-value significance level. Out of these significant proteins, 54 had a confidence score higher than 15 and two or more unique peptides.

The interpretation of the discovery results was divided into three parts; pathway analysis of the significantly differentially expressed proteins to identify any affected pathways, multivariate analysis to find protein covariation relating to centenarianism and/or age, and finally univariate analysis to explore the individual protein expressions and their difference between the groups and relationship with age in detail.

4.2.2.1 *Pathway analysis*

The proteins expressing a nominally significant difference ($p < 0.05$) between centenarians and controls were further investigated using Ingenuity Pathway Analysis (Qiagen). Supplementary table 1 shows the p-values and fold-changes utilised in the analysis. The pathway analysis demonstrated that inflammatory pathways were affected - acute phase signalling, coagulation system and complement system. In addition, pathways relating to lipid and bile acid homeostasis, LXR/RXR and FXR/RXR activation, were found to be significant. Figure 4-2 shows the significant IPA pathways.

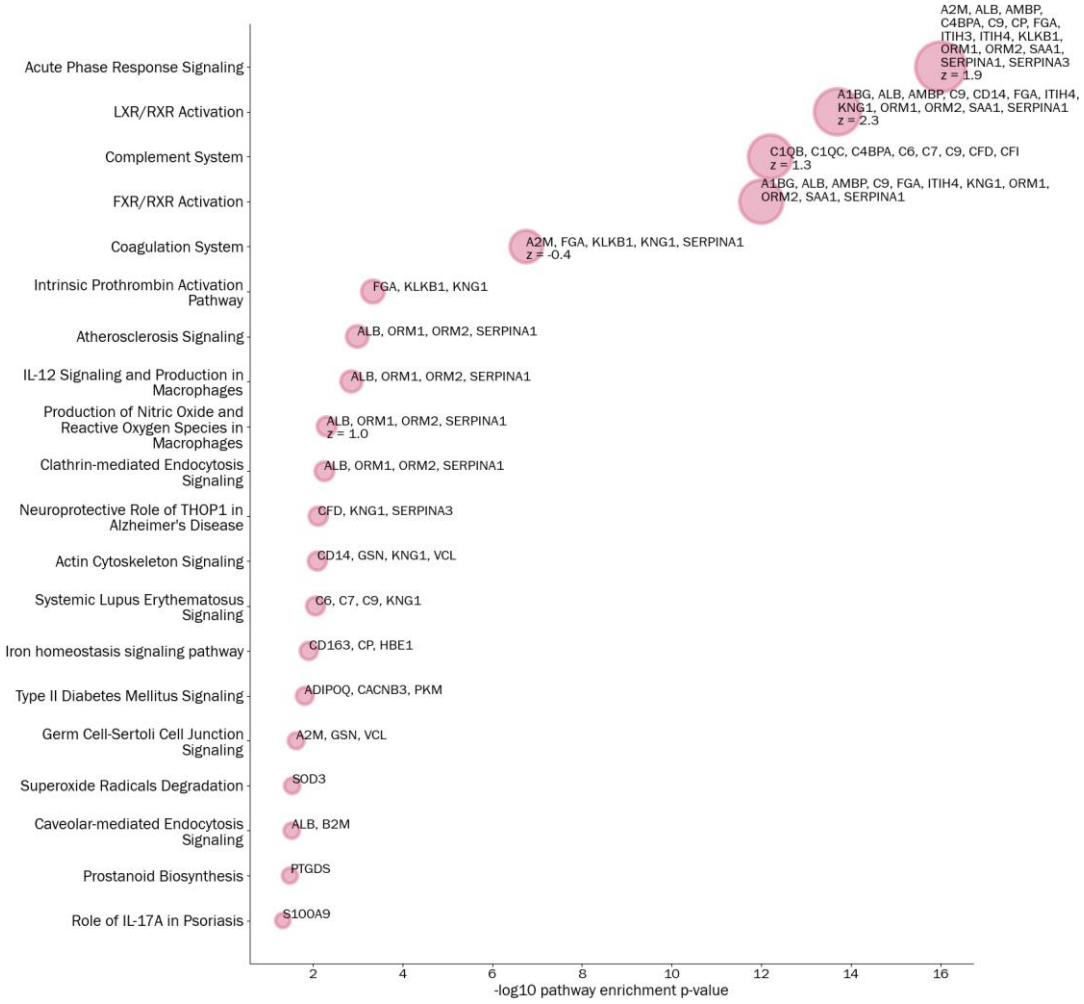


Figure 4-2. Statistically significant pathways from the Ingenuity pathway analysis of the discovery proteomics analysis of centenarians and controls. *The pathways are annotated by the respective proteins included in each. Circle radii indicate the significance of the pathway enrichment p-value, and z-scores denote a suggested up- or downregulated pathway, where z > 0 indicates up regulation and z < 0 indicates downregulation. The pathways where no up- or downregulation could be established lack z-score annotations.*

Below, the top five significant pathways are described in brief. The enrichment p-values are described, as well as the z-scores, which tell if a pathway is predicted to be up- or

downregulated. A z-score with an absolute value greater than 2 is deemed significant and negative values predict downregulation, while positive values predict upregulation.

Acute phase response. Out of the significant proteins, 15 were found included in the acute phase response pathway. The p-value was $1E^{-16}$ and the z-score 1.9, thereby indicating pathway activation. The acute phase response is a rapid, inflammatory response with an increase of inflammatory factors and positive acute phase plasma proteins, and a decrease of negative acute phase proteins. The purpose is to provide protection against microorganisms, although the response can also be triggered by injury, trauma and inflammation [157]. The production of acute phase proteins is predominantly stimulated by cytokines (small, intercellular signalling proteins), which are produced during inflammatory processes. Examples of cytokines are IL6, IL1B and TNF-alpha. IL6 is the primary activator of acute phase proteins, although other cytokines can affect subgroups of acute phase proteins [158]. The acute phase proteins are involved in defence against infections, homeostasis and take part in coagulation and transport of biomolecules. Certain acute-phase proteins can take part in initiation, amplification, attenuation and resolving of inflammation. They can also act as pro-or anti-inflammatory agents. Many of the complement proteins are additionally acute-phase proteins, with the ability to bring upon a pro-inflammatory response through activation of the complement cascade. Hemopexin on the other hand, acts as an anti-inflammatory protein, scavenging reactive oxygen species and protecting against oxidative stress, thereby dampening the inflammatory response [158].

LXR / RXR and FXR/RXR activation pathways. Twelve of the proteins were found in the LXR/RXR pathway. The p-value was $2E^{-14}$ and the z-score 2.3, thus indicating activation of the pathway. Liver X receptors (LXRs) are pivotal in the control of lipid metabolism. They are activated in response to raised levels of intracellular cholesterol and promote the expression of several genes involved in transport, excretion, and absorption. LXRs are moreover capable of affecting immune and inflammatory responses [159]. Eleven of the proteins were found in the FXR/RXR activation pathway. The p-value was $1E^{-12}$, however there was not enough congruency in the protein expression to suggest activation/deactivation of the pathway. Farnesoid X receptors (FXRs), also called bile acid receptors, react to intracellular bile acids and activation of these receptors can modulate triglyceride levels and glucose metabolism. LXRs and RXRs are intracellular and can regulate target genes directly in the nucleus. Both receptor families form heterodimer complexes with retinoid X receptors (RXRs), a nuclear receptor superfamily, and regulate a vast number of genes [160, 161].

Complement system. Eight complement proteins were identified. The enrichment p-value was $6.3E^{-13}$ and the z-score 1.3, suggesting that the pathway may be activated. The complement system is an inflammatory response to external pathogens and can be activated through three different routes - the alternative, the classical and the lectin pathway, all prompting the production of the central protein C3 and ultimately the complex C5b-9, MAC. The classical pathway is activated by the binding of antibody-antigen complexes to C1. The alternative pathway is triggered by bacteria and fungi. The lectin pathway is activated by the binding of mannose-binding lectin (MBL) to mannose residues on pathogen surfaces, which in turn activates MASP1 and MASP2. Complement activators include amyloid beta, lipofuscin constituents, CRP, cholesterol, immunoglobulin and advanced glycation products. Complement components have been reported to be upregulated in Alzheimer's disease and have been shown to be involved in microglia-mediated elimination of synapses. In mice, C3 production is upregulated by microglia and astrocytes. A study found that by inducing C3 deficiency, the mice were protected from cognitive impairment seen in normal ageing [162].

Coagulation cascade. Five proteins were identified in the coagulation cascade pathway. The p-value was $1.7E^{-7}$ and the z-score -0.4, thereby very weakly suggesting that the pathway may be deactivated. The coagulation pathway involves the defence mechanism of blood clotting, a process protecting the vascular system after tissue injury through the formation of a haemostatic plug, covering the site of injury. The proteins and molecules participating in coagulation are present in the blood under non-stressed conditions in their inactive form. The species are transformed to active enzymes through the cleavage of peptide bonds when stimulated. In addition, platelets and endothelial cells undergo biochemical changes to gain new properties aiding coagulation upon stimulation by agonists [163]. The coagulation cascade consists of two entry pathways; the intrinsic (three proteins found in this pathway: FGA, KLKB1 and KNG1, with a p-value of $4.6E^{-4}$ and no suggested up- or downregulation) and the extrinsic. Both pathways lead to the formation of fibrin, a protein that polymerises and together with platelets forms a clot covering the site of injury. In the extrinsic pathway, factor XII is first activated, followed by the activation of factors XI, IX and X and prothrombin. In the extrinsic pathway, a complex consisting of tissue factor and factor VII is initially constructed, followed by the activation of factors VII and X and prothrombin [164]. The extrinsic pathway, also known as the tissue factor pathway, is the primary initiator of blood clotting.

Although modelled as separate entities, there is evidence that the coagulation and complement cascades modulate each other's activity. Factor XIIa and kallikrein can cleave

C1s and thereby trigger the classical complement pathway. Thrombin can activate C3, C5, C6 and factor B. Kallikrein is further able to cleave C5 and factor B. Factor XIIa can cleave C3 [165].

Moreover, several proteins were predicted by Ingenuity as upstream regulators based on the observed proteins' expression in this dataset. Table 4-2 shows the three top proteins which on a significant level were predicted to be activated upstream regulators. No proteins were predicted to be deactivated on a significant level.

Table 4-2. Predicted activated upstream regulators and the target proteins in the IPA analysis of the discovery proteomics of centenarians and controls.

Upstream Regulator	p-value of overlap	Target molecules in dataset
IL6	$1.09E^{-9}$	A2M, ALB, CD14, CD163, CFD, CP, CST3, FGA, GFAP, HBE1, LRG1, ORM1, PPBP, S100A9, SAA1, SERPINA1, SERPINA3
CEBPB	$2.4E^{-8}$	ADIPOQ, ALB, CD14, CFD, CHIT1, CP, FBLN1, GFAP, NRP1, ORM1, SAA1, SERPINA1
STAT3	$3.74E^{-5}$	A2M, ADIPOQ, FGA, GFAP, NRP1, PROCR, S100A9, SAA1, SERPINA1, SERPINA3

These proteins are involved in inflammatory response and transcription. IL6 can induce acute phase response. STAT3 can bind to IL6 elements. CEBPB can regulate the expression of genes involved in inflammatory responses.

Overall, the pathway analysis suggests that the centenarians have increased levels of inflammation compared to controls. There is a large amount of overlap between the proteins seen in the different pathways as all the most relevant pathways are involved in inflammation, thus making it difficult to pinpoint the exact ongoing mechanisms. Proteins from the acute phase response signalling pathway are for example included to some extent in all the other pathways. Moreover, LXR/RXR and FXR/RXR overlap almost completely, with only one unique protein (CD14) found in LXR but not in FXR.

4.2.2.2 *Multivariate analysis*

To obtain an overview of the proteins mostly affected by age and to test if there was an age-related multivariate protein expression, an OPLS model with age as the dependent variable *y* was constructed. The model suggested a relationship with age for many of the proteins, although the model itself was not significant (ANOVA $p = 0.069$). This means that there is not enough age-related covariation between the proteins to build a model with adequate predictive ability for protein expression and age. Figure 4-3 shows the 40 proteins, which according to the OPLS model have the strongest positive and negative correlation with age.

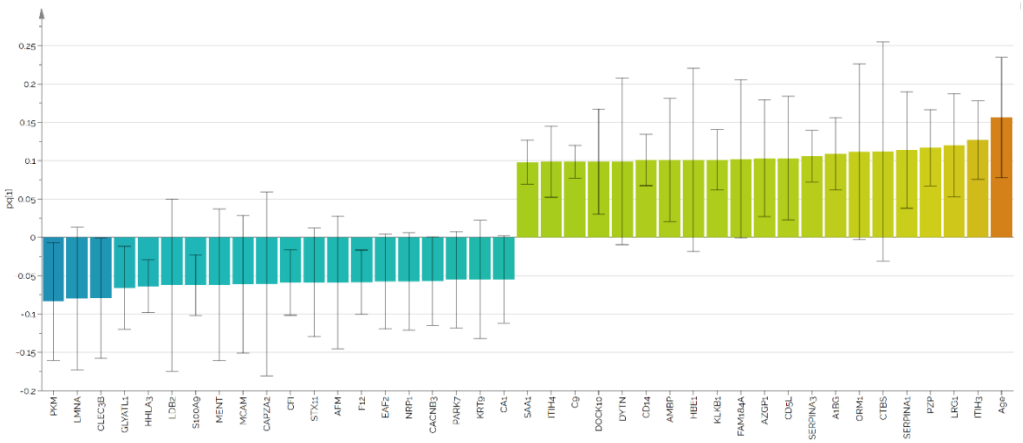


Figure 4-3. OPLS loadings of the results from discovery proteomics of centenarians and controls. *The figure shows the 40 most extreme proteins positively and negatively correlated with age. pq[1] on the y-axis represents the predictive loadings. The bars are coloured sequentially according to their correlation with age, where blue represents strong negative correlation and red strong positive correlation.*

An OPLS-DA model that compared centenarians and controls was created to assess group-specific protein expression. This model was not significant (ANOVA p = 0.16), again indicating that there is not sufficient co-variation between the proteins related to the discriminating factor of age in each group. An OPLS-DA analysis of males versus females was further performed and proved non-significant (ANOVA p = 0.18), thereby implying that no general gender-dependent difference in protein expression could be demonstrated in this study.

In summary, the multivariate analyses demonstrated that the protein expression was not gender dependent and, although the expressions of some individual proteins were related to age, no overall age-related multivariate protein expression could be established.

4.2.2.3 *Univariate analysis*

A univariate approach was undertaken to further investigate the age-correlation and differences between the groups for individual proteins. For an overview, the difference in protein expression between the centenarian and the control group was visualised in a Volcano plot (Figure 4-4). The Volcano plot shows that the overall distribution of up- and down-regulated proteins was relatively similar but that among the significant proteins, a majority were upregulated in the centenarians.

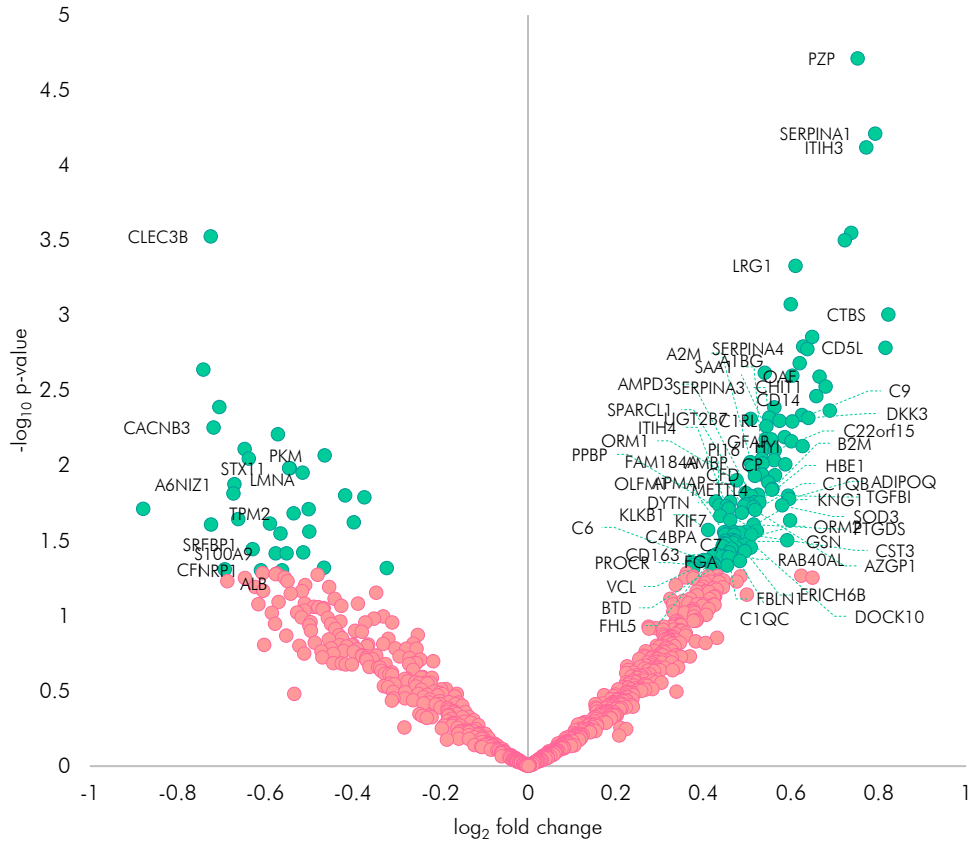


Figure 4-4. Volcano plot of the results from the discovery proteomics study of centenarians versus controls. *The x-axis shows the logarithm of the average fold change, base 2, and the y-axis shows the negative logarithm of the p-values, base 10.* ■ *significant proteins on a nominal p-value level of 0.05,* ■ *non-significant proteins. The significant proteins which were identified by two or more peptides and with an identification confidence score above 15 are annotated by their gene names.*

4.2.2.3.1 *Linear regression correlating protein expression with age*

The multivariate analysis of protein expression versus age that was performed in section 4.2.2.2 showed promise but did not produce significant models allowing for identification of a protein panel significantly correlated with age. To investigate the protein-age correlation in detail, separate linear regression models were created for each of the significantly different proteins. Controls and centenarians were modelled individually and together as one group. The slope of each model was examined for a significant difference from zero. In the combined analysis of centenarians and controls, 67 proteins had a slope significantly different from zero. In the regression analysis of centenarians only, 23 proteins demonstrated a slope significantly different from zero, and in the control group, only one protein had a significantly non-zero slope.

Table 4-3 shows the summary of the p-values in the group comparison between centenarians and controls, and results from the regression analysis. As expected, most of

the proteins differing in the group comparison also demonstrated a significantly non-zero slope in the combined control-centenarian model. The regression analysis demonstrated that none of the proteins had a non-zero slope in both the control and centenarian group, thereby indicating that no overall age-protein correlation could be established. The beta-coefficients of the models are provided in Supplementary table 2.

Table 4-3. Significance of univariate linear regression models of age versus protein expression. *The table shows the significance level of the proteins differentially expressed comparing the groups centenarians and controls, and the results from the non-zero slope significance test when modelling centenarians and control together, centenarians alone, and control alone. The proteins are denoted by their gene names. The significance levels are represented by asterisks, where ****$p < 0.0001$, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$ and NS $p > 0.05$*

	Significance level of t-test between centenarians and control	Significance level of age versus protein expression regression, testing if the slope is different from zero		
		Centenarians and control	Centenarians	Control
PZP	****	****	NS	NS
SERPINA1	****	***	NS	NS
ITIH3	****	***	*	NS
CLEC3B	***	***	NS	NS
LRG1	***	***	**	NS
CTBS	***	**	*	NS
CD5L	**	***	NS	NS
SERPINA4	**	**	NS	NS
OAF	**	**	NS	NS
CHIT1	**	**	NS	NS
CD14	**	**	NS	NS
DKK3	**	**	NS	NS
UGT2B7	**	**	NS	NS
C1RL	**	**	NS	NS
C9	**	**	NS	NS
CACNB3	**	**	NS	NS
A1BG	**	**	**	NS
GFAP	**	*	NS	NS
HYI	**	*	NS	NS
PI16	**	**	NS	NS
C22orf15	**	***	****	NS
PKM	**	**	*	NS
SAA1	**	**	NS	NS
AMBP	**	**	*	NS
CP	**	*	NS	NS
STX11	*	*	NS	NS
LMNA	*	**	NS	NS
CFD	*	*	NS	NS
SERPINA3	*	**	*	NS
A2M	*	*	NS	NS
ORM1	*	**	NS	NS
A6NIZ1	*	*	NS	NS
METTL4	*	**	*	NS

	Significance level of t-test between centenarians and control	Significance level of age versus protein expression regression, testing if the slope is different from zero		
		Centenarians and control	Centenarians	Control
TGFBI	*	*	NS	NS
AMPD3	*	*	NS	NS
ADIPOQ	*	**	*	NS
B2M	*	*	NS	NS
SPARCL1	*	*	NS	NS
AZGP1	*	**	**	NS
SOD3	*	**	***	NS
FAM184A	*	*	**	NS
ITIH4	*	*	**	NS
APMAP	*	*	NS	NS
HBE1	*	*	NS	NS
C1QB	*	*	NS	NS
TPM2	*	*	NS	NS
DYTN	*	**	*	NS
KIF7	*	*	**	NS
PTGDS	*	*	NS	NS
KLKB1	*	*	**	NS
PPBP	*	*	NS	NS
OLFM1	*	*	NS	NS
DOCK10	*	**	*	NS
CST3	*	*	*	NS
GSN	*	*	NS	NS
KNG1	*	*	**	NS
ORM2	*	*	**	NS
C7	*	NS	NS	NS
FHL5	*	*	NS	NS
C1QC	*	*	NS	NS
SRFBP1	*	NS	NS	NS
ERICH6B	*	NS	NS	NS
RAB40AL	*	*	NS	NS
S100A9	*	*	NS	NS
C4BPA	*	*	NS	NS
FBLN1	*	*	NS	NS
C6	*	*	*	NS
FGA	*	NS	NS	NS
PROCR	*	*	NS	NS
CD163	*	*	NS	NS
BTD	*	NS	NS	*
VCL	*	NS	**	NS
CFI	*	NS	NS	NS
ALB	*	NS	NS	NS
NRP1	*	*	NS	NS

There are differences in the levels of many proteins between controls and centenarians. Although the sample groups are small and thereby limiting the interpretability, none of the proteins demonstrate a linear relationship with age in both the control and centenarian group, thus indicating that there is no overall age-dependent protein expression in the studied groups among the significant proteins.

4.2.3 *Summary and conclusions from the discovery phase*

Utilising the optimised workflow for discovery plasma proteomics, 875 proteins were detected and 74 found significantly different between centenarians and controls. 11 of the proteins were downregulated in the centenarians and 63 were upregulated. Multivariate and linear regression analyses demonstrated that none of the proteins had a distinct trajectory with age. The reason for this may be that the sample size was too small, and that the age gap between the two groups was too large. To determine age-protein correlations, a larger sample set with a wide age spread should ideally have been studied.

Somewhat surprisingly, pathway analysis indicated an overall elevated inflammatory response in the centenarian group compared to the control group, suggesting activation of the LXR/RXR pathway, acute phase response signalling and, less confidently, of the complement system. The coagulation system was suggested to be deactivated, albeit on a non-significant level. Even though centenarians reach an advanced age with their health remarkably intact, they are nevertheless close to the endpoint of their lives, and this could be why evidence of inflammation was observed.

The following proteins were chosen from the discovery study to be included in a targeted assay, aiming to validate the discovery results: A2M, ADIPOQ, CST3, CTHRC1, FGA, HBE1, PTGDS, SOD3, CSF1R, DKK3, PKM, SAA1 and SERPINA1. The selection was based on significance testing, quality of the protein identification, and literature reviews.

4.3 TARGETED PROTEOMICS TO CONFIRM FINDINGS FROM THE DISCOVERY PHASE AND TO IDENTIFY INFLAMMATORY TARGETS FROM LITERATURE

In the targeted proteomics study, the proteins identified and selected in the discovery phase were evaluated in a larger set of samples to confirm the findings from the univariate and pathway analyses. Several pro- and anti-inflammatory proteins from the literature were moreover included in the assay. Samples from centenarians and controls were analysed, but also samples from children of centenarians (called “offspring” in this

chapter). The offspring samples were included in the study to evaluate if they possessed any similarity in protein expression to the centenarians, which could indicate that there might be a genetical component to centenarianism, passed on from parents to children.

4.3.1 *Materials and methods*

4.3.1.1 *Sample cohort*

The plasma samples for the targeted study were collected and provided by IRCCS Istituto delle Scienze Neurologiche di Bologna, Italy. The control group consisted of healthy individuals; the centenarian group consisted of elderly but healthy persons without cognitive decline; the offspring group consisted of healthy individuals with at least one centenarian parent. A total of 186 plasma samples were included in the targeted study. Table 4-4 shows the characteristics of the three sample groups and Figure 4-5 shows a histogram of the age distribution.

Table 4-4. Characteristics of samples in the targeted proteomic study of centenarians, offspring, and controls. *The table shows the number of samples, the percentages of males and females, and the mean age in each group*

Group	Number of samples	Percentage males / females	Age ± SD
Control	40	50% M | 50% F	70.0 (± 7.1)
Offspring	54	35% M | 65% F	71.0 (± 6.6)
Centenarian	92	23% M | 77% F	103.8 (± 3.0)

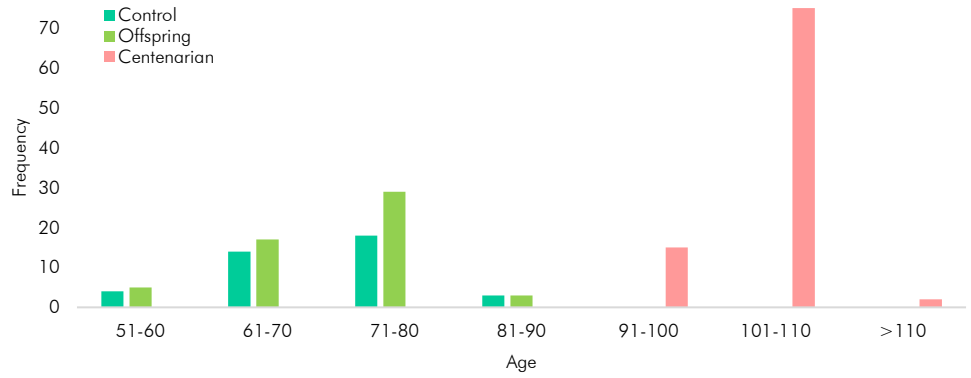


Figure 4-5. Histogram of age in the groups showing the frequencies of samples as 10-year bins

4.3.1.2 *Sample preparation for targeted proteomics*

The samples were prepared for targeted proteomics as described in Chapter 2, section 2.7. Briefly, 10 µL plasma was spiked with 150 ng yeast ENO1 (whole protein), and thereafter depleted of albumin and IgG using Pierce Top2 columns. The samples were tryptically digested and solid phase extracted. Quality control samples consisted of pooled, acetone

precipitated plasma. Calibration curves were prepared by spiking increasing amounts of peptide standards into blank and pooled quality control samples.

4.3.1.3 *Instrumental LC-MS analysis*

The settings and parameters for the instrumental analysis are described in detail in Chapter 2, section 2.9. In brief, the samples were reconstituted in 30 µL 3% acetonitrile, 0.1% trifluoroacetic acid containing 0.1 µM of isotope labelled internal standards. 5 µL was injected onto a UPLC system coupled to a triple quadrupole mass spectrometer. Two injections were made per sample, each with a different MRM method. In total, 189 peptides were monitored, representing 127 proteins.

4.3.1.4 *Peak picking, integration, and data pre-treatment*

Data analysis was performed as described in Chapter 2, section 2.10. In brief, the acquired data were peak-picked and integrated using an in-house software. The peptides were identified by the retention times given by blank and matrix calibration curves. Digestion efficiency was evaluated by monitoring the presence of yeast ENO1 in the samples. The analyte peak areas were normalised to the heavy isotope labelled peptide internal standard from the protein ALDOA. Run order drift was corrected by LOWESS curve fitting. Extreme outliers were detected at a threshold of 10 median absolute deviations and replaced by missing values. Outliers were identified in 14 samples, containing a total of 16 outlier values present in SERPINA3, FABP5, HSPA1L, PTGDS, SAA1, SOD3, TUBA4A and VCAM1. The outlier values were randomly distributed among the samples, apart from SAA1 where outliers were observed only in female centenarians. The table of outliers is presented in Supplementary table 3.

4.3.2 *Results from the targeted validation analysis*

The MRM-based proteomic assay with targets from the centenarian discovery study, studies of Alzheimer's and Parkinson's diseases, and pro- and anti-inflammatory proteins from literature, could reliably detect 29 individual endogenous proteins in the 186 samples. The data were investigated using several methods, and the results are presented in this section.

4.3.2.1 *Multivariate analysis*

Multivariate analysis was performed with two different objectives. Firstly, unsupervised PCA analysis to assess the data quality and to look for major trends in the data. Secondly, supervised OPLS and OPLS-DA analyses to determine relationships between the protein

expression and dependent variables such as age and sex, and to discriminate between the sample groups.

4.3.2.1.1 *Unsupervised Principal Component Analysis*

An unsupervised principal component analysis was performed to inspect the data quality post drift correction and outlier removal. The PCA demonstrated that there was a clear difference between centenarians and controls but not between offspring and controls (Figure 4-6).

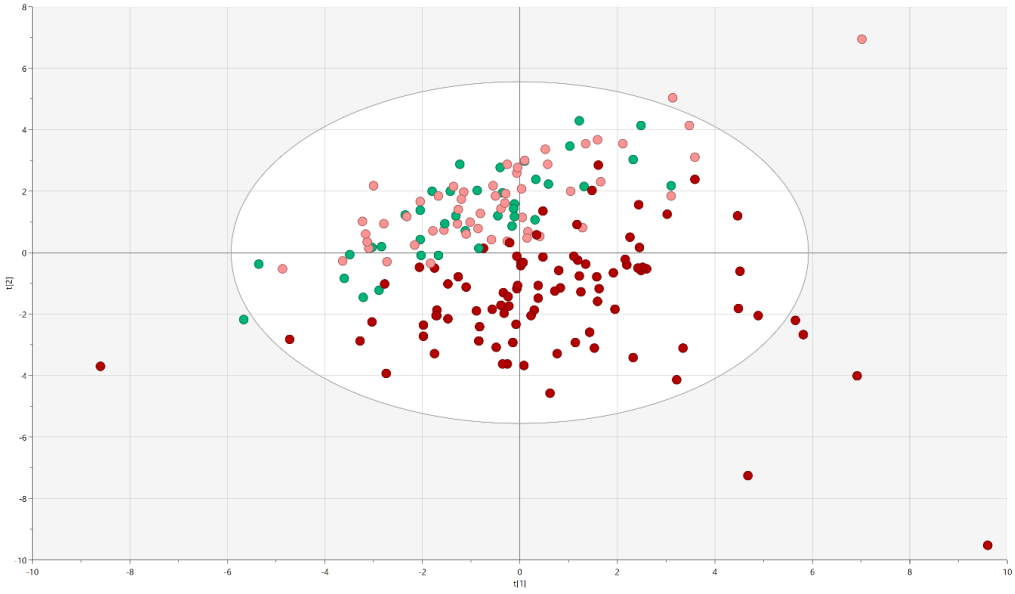


Figure 4-6. PCA of the targeted proteomics data from the study of centenarians, offspring, and control. *A clear separation could be observed between the centenarians and offspring/control; however, no difference could be discerned between the control and offspring groups.* ■ *centenarians,* ■ *controls, and* ■ *offspring. t[1] represents principal component 1, and t[2] represents principal component 2.*

4.3.2.1.2 *Supervised OPLS to relate age and protein expression*

To examine the relationship between the proteins and age, an OPLS regression model including all sample groups with age as the dependent variable *y* was generated (model scores shown in Figure 4-7). Model validation demonstrated that there was a strong relationship with age, with an ANOVA $p = 6\ E^{-37}$ and permutations $p << 0.001$. The model showed that there was a difference in protein expression between the two age groups. The loadings plot demonstrated that the difference was mainly caused by higher levels of PTDGS, VCAM1, CST3, ICAM1, A2M, ADIPOQ, SERPINA3, SAA1, SERPINA1 and FABP5, and lower levels of SERPINF2, BCHE, ITIH2, PKM, PRG4 and C3 in the centenarians.

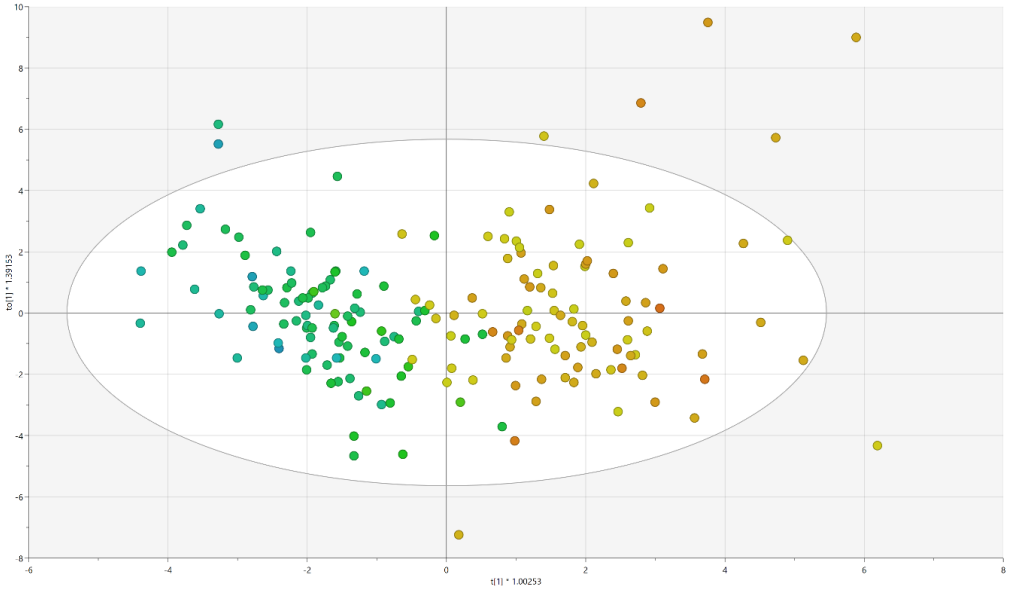


Figure 4-7. Scores from an OPLS model of the targeted data from the study of centenarians, offspring, and control with age as the dependent variable. *The score plot is coloured in a scale going from blue (youngest age) to red (oldest age). A distinct age-dependence can be observed. The x-axis represents the predictive component, and the y-axis the orthogonal component.*

An OPLS model of the control and offspring samples, with age as the dependent variable, was created but proved non-significant ($p = 0.4$), thereby suggesting that not enough covariation in the protein expression of these two groups exists to determine a multivariate relationship with age. The centenarians were moreover modelled in an OPLS, with age as the dependent variable. The model was weakly significant ($p = 0.02$), suggesting lower levels of SERPINF2, C3, HPX, ITIH2 and BCHE negatively correlated with age.

4.3.2.1.3 *OPLS-DA to discriminate between the sample groups*

Investigating the data further, discriminant OPLS-DA models were constructed, comparing all three groups to each other. The comparison between controls and offspring did not demonstrate a difference between the groups with both ANOVA and permutation $p >> 0.05$. The comparison between centenarians and controls showed a clear separation between the groups (ANOVA $p = 2E^{-12}$ and permutations $p << 0.001$), as did the comparison between centenarians and offspring ($1E^{-20}$ and permutations $p << 0.001$). Examining the relationship between the two discriminant centenarian models, a Shared and Unique Structures (SUS) plot was generated. This plot can be used to compare the results of two different models. The SUS plot in Figure 4-8 demonstrated that the loadings were indeed highly similar, only FABP5 differing slightly. This means that the discriminating proteins between centenarians and controls, and centenarians and offspring are the same. It moreover signifies that there is no discernible difference

between the offspring and control groups. The most discriminating proteins with higher levels in centenarians were found to be CST3, PTGDS, VCAM1, ADIPOQ and ICAM1. The most discriminating proteins with lower levels in centenarians were BCHE, A2AP, ITIH2, PKM, PRG4 and C3.

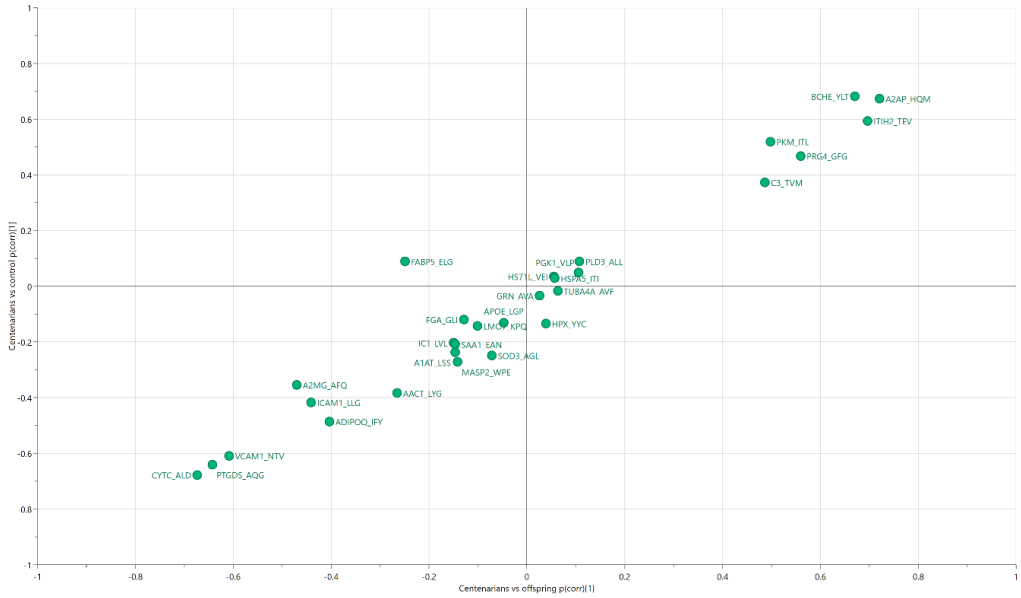


Figure 4-8. Shared and Unique Structures plot of the two OPLS-DA models centenarians versus offspring and centenarians versus control. *The SUS plot demonstrates a linear relationship between the models, thus illustrating that the discriminating proteins in both models are highly similar, only FABP5 differing slightly.*

Additionally, an OPLS-DA model of males versus females was constructed and proved significant ($p = 2.9E^{-6}$), thus indicating a protein expression correlation with sex. The model showed that SERPING1, APOE, HPX, ADIPOQ, SERPINA3, SOD3, A2M and SERPINA1 were upregulated in the females compared to the males.

4.3.2.1.4 *Conclusions from the multivariate analysis*

The multivariate analysis demonstrated that there is a strong relationship between age/centenarianism and protein expression when analysing all sample groups together. In summary, the proteins that exhibited altered expression with age (OPLS) or between the age-groups (OPLS-DA) were CST3, PTGDS, VCAM1 and ICAM1, all upregulated in the centenarians, and BCHE, ITIH2, PKM, PRG4 and C3, which were downregulated in the centenarians. Additionally, A2M, SERPINA3, SAA1, SERPINA1 and FABP5 demonstrated a relationship with age only in the OPLS model of all sample groups and SERPINF2 a difference between the centenarian and offspring/control groups only in the OPLS-DA models. The analysis further highlighted that no multivariate differences could be detected between the control and offspring groups, and that no correlation with age could be determined in these in these two groups either.

4.3.2.2 *Univariate analysis*

To examine the individual protein expressions and their relationship with the dependent variables, each protein was examined in detail by correlation analysis and group comparisons.

4.3.2.2.1 *Correlation analyses*

The relationship between the proteins, age and sex in the sample groups were examined by Pearson correlation, and correlation coefficients and p-values were extracted. All p-values were adjusted for multiple testing applying Benjamini-Hochberg false discovery rate with alpha = 0.05.

Correlation between protein expression and sex. Firstly, the relationship between the protein expression and sex was investigated and it was found that the following proteins were significantly correlated with sex, all downregulated in males compared to females: A2M, SERPINA3, ADIPOQ, APOE, HPX, ICAM1, SERPINA1, SERPING1, and SOD3.

Correlation between protein expression and age. Next, the age-protein correlation was examined in the following sample groups *(i)* control, *(ii)* offspring, *(iii)* control and offspring, *(iv)* centenarians, and *(v)* all samples. In the control, offspring and combined control/offspring groups, no significant correlations were identified. In the centenarians, the following proteins were significantly correlated (all negatively) with age: C3, HPX, and SERPINF2. In the correlation analysis of all samples, 19 proteins were found significantly correlated with age. These were:

- A2M
- SERPINA3
- ADIPOQ
- BCHE
- C3
- CST3
- FABP5
- ICAM1
- ITIH2
- LMO7
- MASP2
- PKM
- PRG4
- PTGDS
- SAA1
- SERPINA1
- SERPINF2
- SOD3
- VCAM1

Given the large number of proteins significantly correlated with sex, the data were adjusted for sex as per Chapter 3, section 3.5.4 and the same correlation analysis run again. The results from both analyses were compared and are summarised in Table 4-5. After sex-adjustment, the only protein changing from being significantly correlated with age to non-significant was SERPING1, all other proteins remained significant after adjustment.

Adjustment of age was not considered as this would remove the discriminating factor of what our study set out to explore.

Table 4-5. FDR-adjusted p-values and direction of correlation for Pearson correlation of protein levels and age in the groups control, offspring, combined control and offspring, centenarians, and all samples. *The table summarises the results from the analysis in data not adjusted for sex and sex-adjusted data. SERPING1 becomes non-significant in the group "All samples" after sex-adjustment. The arrows indicate positive or negative correlation. The correlation significance is represented by asterisks, where **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and NS $p > 0.05$*

	Not adjusted for sex					Sex-adjusted				
	Control	Offspring	Control/offspring	Centenarians	All samples	Control	Offspring	Control/offspring	Centenarians	All samples
A2M	NS	NS	NS	NS	**** (↑)	NS	NS	NS	NS	**** (↑)
ADIPOQ	NS	NS	NS	NS	**** (↑)	NS	NS	NS	NS	**** (↑)
APOE	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
BCHE	NS	NS	NS	NS	**** (↓)	NS	NS	NS	NS	**** (↓)
C3	NS	NS	NS	* (↓)	**** (↓)	NS	NS	NS	* (↓)	**** (↓)
CST3	NS	NS	NS	NS	**** (↑)	NS	NS	NS	NS	**** (↑)
FABP5	NS	NS	NS	NS	** (↑)	NS	NS	NS	NS	** (↑)
FGA	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
GRN	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
HPX	NS	NS	NS	* (↓)	NS	NS	NS	NS	* (↓)	NS
HSPA1L	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
HSPA5	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
ICAM1	NS	NS	NS	NS	**** (↑)	NS	NS	NS	NS	**** (↑)
ITIH2	NS	NS	NS	NS	**** (↓)	NS	NS	NS	NS	**** (↓)
LMO7	NS	NS	NS	NS	* (↑)	NS	NS	NS	NS	* (↑)
MASP2	NS	NS	NS	NS	** (↑)	NS	NS	NS	NS	* (↑)
PGK1	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
PKM	NS	NS	NS	NS	**** (↓)	NS	NS	NS	NS	**** (↓)
PLD3	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
PRG4	NS	NS	NS	NS	**** (↓)	NS	NS	NS	NS	**** (↓)
PTGDS	NS	NS	NS	NS	**** (↑)	NS	NS	NS	NS	**** (↑)
SAA1	NS	NS	NS	NS	*** (↑)	NS	NS	NS	NS	*** (↑)
SERPINA1	NS	NS	NS	NS	*** (↑)	NS	NS	NS	NS	** (↑)
SERPINA3	NS	NS	NS	NS	**** (↑)	NS	NS	NS	NS	**** (↑)
SERPINF2	NS	NS	NS	* (↓)	**** (↓)	NS	NS	NS	* (↓)	**** (↓)
SERPING1	NS	NS	NS	NS	* (↑)	NS	NS	NS	NS	NS
SOD3	NS	NS	NS	NS	* (↑)	NS	NS	NS	NS	* (↑)
TUBA4A	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS

4.3.2.2.2 *Group comparisons*

The sample groups control, offspring and centenarians were compared to each other, one by one, using a two-tailed Student's t-test. The p-values were adjusted for multiple testing by applying Benjamini-Hochberg FDR-correction. Both data adjusted for sex and non-adjusted data were evaluated.

Comparing centenarians to control, 21 proteins were significantly altered in the non-sex adjusted data and 16 in the sex-adjusted data. The proteins not significant post sex-correction were FABP5, LMO7, HPX, MASP2 and SERPING1. In the comparison of centenarians and offspring, the same 16 proteins were differentially expressed in the non-adjusted and the sex-adjusted data. Comparing centenarians to the combined group of control and offspring, 20 proteins were differentially expressed in the non-adjusted data and 18 in the sex-adjusted data. The proteins not significant after sex-adjustment were LMO7 and SERPING1. No significant differences in protein expression were identified in the comparison between control and offspring. The results are summarised in Table 4-6.

Table 4-6. Summary of the FDR adjusted results from the comparison of controls, offspring, and centenarians. *The significance levels are represented by asterisks, where **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and NS $p > 0.05$. Cent = centenarian, ctrl = control, and offsp = offspring.*

	Not adjusted for sex				Sex-adjusted			
	Cent. vs ctrl	Cent. vs offsp.	Cent. vs ctrl/offsp.	Ctrl vs offsp.	Cent. vs ctrl	Cent. vs offsp.	Cent. vs ctrl/offsp.	Ctrl vs offsp.
A2M	***	****	****	NS	**	****	****	NS
SERPINA3	****	***	****	NS	***	**	****	NS
ADIPOQ	***	***	****	NS	**	***	****	NS
APOE	NS	NS	NS	NS	NS	NS	NS	NS
BCHE	****	****	****	NS	****	****	****	NS
C3	*	****	****	NS	*	****	****	NS
CST3	****	****	****	NS	****	****	****	NS
FABP5	*	**	**	NS	NS	**	**	NS
FGA	NS	NS	NS	NS	NS	NS	NS	NS
GRN	NS	NS	NS	NS	NS	NS	NS	NS
HPX	*	NS	NS	NS	NS	NS	NS	NS
HSPA1L	NS	NS	NS	NS	NS	NS	NS	NS
HSPA5	NS	NS	NS	NS	NS	NS	NS	NS
ICAM1	**	****	****	NS	*	***	****	NS
ITIH2	***	****	****	NS	***	****	****	NS
LMO7	*	NS	*	NS	NS	NS	NS	NS
MASP2	*	NS	**	NS	NS	NS	*	NS
PGK1	NS	NS	NS	NS	NS	NS	NS	NS
PKM	***	****	****	NS	***	****	****	NS
PLD3	NS	NS	NS	NS	NS	NS	NS	NS
PRG4	**	****	****	NS	**	****	****	NS
PTGDS	****	****	****	NS	****	****	****	NS
SAA1	**	**	****	NS	**	**	***	NS
SERPINA1	**	*	***	NS	*	*	**	NS
SERPINF2	****	****	****	NS	****	****	****	NS
SERPING1	*	NS	**	NS	NS	NS	NS	NS
SOD3	**	NS	**	NS	*	NS	*	NS
TUBA4A	NS	NS	NS	NS	NS	NS	NS	NS
VCAM1	****	****	****	NS	****	****	****	NS

Next, the difference in protein expression between males and females was investigated in detail to clarify the sex-effect on the different proteins. A two-tailed Student's t-test was performed between males and females in the following groups (i) all samples, (ii) control, (iii) offspring, (iv) centenarians, and (v) control and offspring combined. The p-values were adjusted to account for multiple testing using Benjamini-Hochberg FDR-correction and are summarised in Table 4-7. Importantly, it is shown that there is a significant difference in age between males and females. The majority of the proteins do not exhibit a sex-related difference in expression in the subgroups, however, when inspecting the samples as a whole, ten proteins demonstrate significant differences in expression between males and females. When relating these results to the significance testing shown in Table 4-6, only HPX and SERPING1 are affected to a degree that places them below the significance threshold in the sex-corrected data compared to the non-corrected.

Table 4-7. Comparison of the protein expression between males and females in the targeted data from centenarians, offspring, and control. *The table shows the significance levels of the FDR-adjusted p-values represented by asterisks, where **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and NS $p > 0.05$*

	All samples	Control	Offspring	Centenarians	Control/ offspring
Age	*	NS	NS	NS	NS
A2M	*	NS	NS	NS	NS
SERPINA3	**	*	NS	NS	NS
ADIPOQ	**	NS	*	NS	**
APOE	****	NS	*	NS	**
BCHE	NS	NS	NS	NS	NS
C3	NS	NS	NS	NS	NS
CST3	NS	NS	NS	NS	NS
FABP5	NS	NS	NS	NS	NS
FGA	NS	NS	NS	NS	NS
GRN	NS	NS	NS	NS	NS
HPX	**	NS	NS	NS	*
HSPA1L	NS	NS	NS	NS	NS
HSPA5	NS	NS	NS	NS	NS
ICAM1	*	NS	NS	NS	NS
ITIH2	NS	NS	NS	NS	*
LMO7	NS	NS	NS	NS	NS
MASP2	NS	NS	NS	NS	NS
PGK1	NS	NS	NS	NS	NS
PKM	NS	NS	NS	NS	NS
PLD3	NS	NS	NS	NS	NS
PRG4	NS	NS	NS	NS	NS
PTGDS	NS	NS	NS	NS	NS
SAA1	NS	NS	NS	NS	NS
SERPINA1	*	NS	NS	NS	NS
SERPINF2	NS	*	NS	NS	*
SERPING1	***	NS	NS	NS	**
SOD3	*	*	NS	NS	NS
TUBA4A	NS	NS	NS	NS	NS
VCAM1	*	NS	NS	NS	NS

It was determined that the sex-corrected data would be utilised for all further modelling. The results from the univariate analysis, comparing centenarians to control and offspring, are represented in a Volcano plot (Figure 4-9). It is clearly demonstrated that several proteins exhibit highly significant p-values.

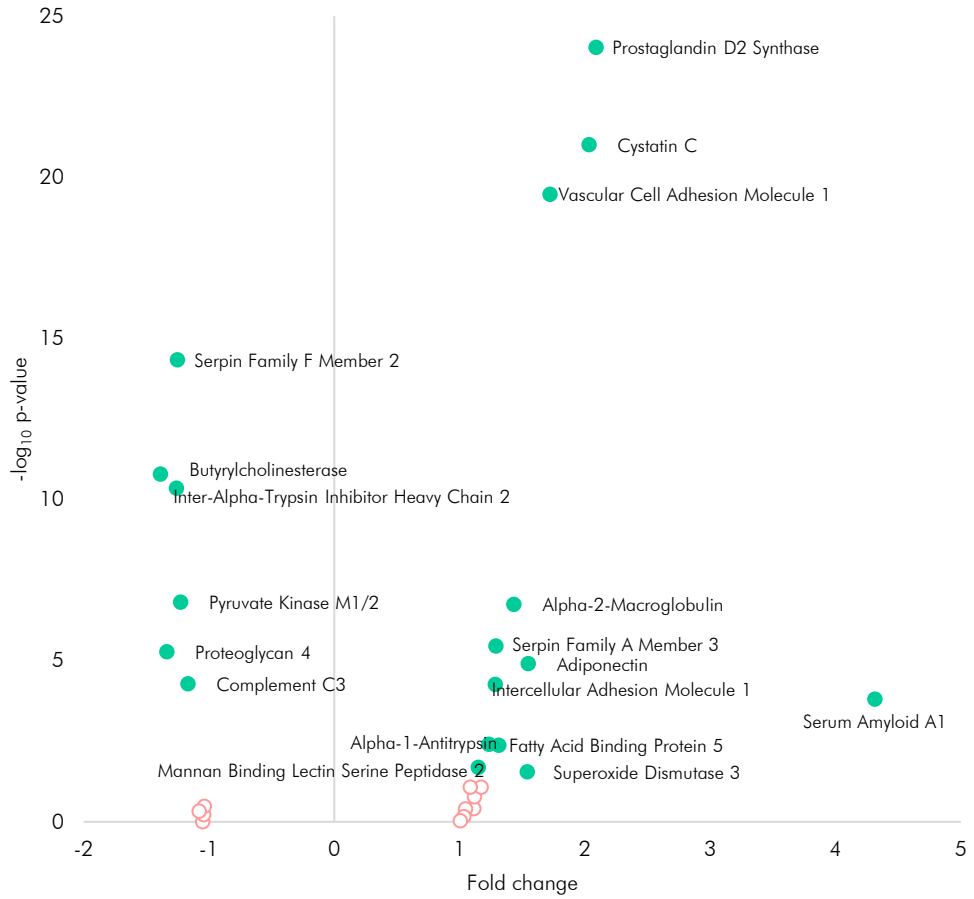


Figure 4-9. Volcano plot of the proteins from the targeted study of centenarians, offspring, and controls. *The data points are coloured according to their FDR-corrected p-value significance comparing centenarians to offspring and control, where* ● *p-value* > 0.05, ● *p-value* < 0.05.

The individual values of the samples in the significantly different proteins are shown in Figure 4-10. As demonstrated in the Volcano plot, a few proteins exhibit large differences between the centenarians and the control and offspring groups, especially PTGDS, CST3, VCAM1 and SERPINF2. The offspring samples with values closer to the distribution of the centenarians were examined in detail for the different proteins, but no congruency could be detected among these samples, therefore it was not possible to identify offspring samples with a centenarian-like profile.

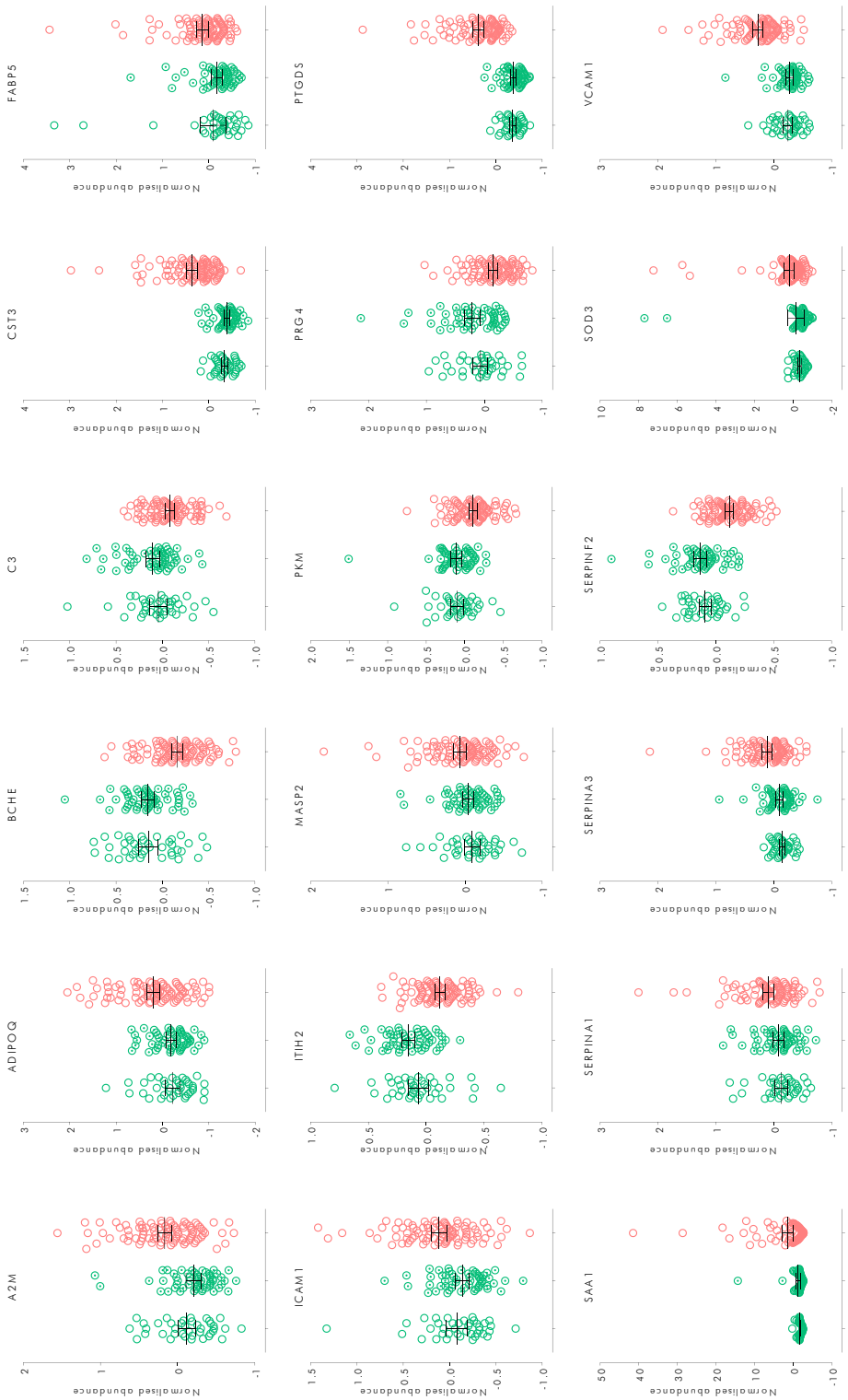


Figure 4-10. Significantly different proteins from the targeted proteomic study of centenarian, offspring, and control samples. *The error bars represent the 95% confidence interval. The offspring samples with protein levels close to the distribution of the centenarian samples were examined but no uniformity among these samples could be found. From left to right in each plot, the sample groups are represented by ○ controls, ⊖ offspring, and ○ centenarians*

4.3.2.2.3 *The targeted study validates eight proteins from the discovery phase*

Having analysed the targeted data thoroughly and now satisfied that bias had been accounted for, the results from the targeted study were compared to the results from the discovery phase. 13 proteins from the discovery phase had been selected for validation. Out of these 13 proteins, nine could be detected in the targeted validation study and eight demonstrated a significant difference between the centenarians and control and/or offspring samples. Comparing the results from the targeted study and the discovery study (Table 4-8), A2M, ADIPOQ, CST3, PTGDS, PKM, SAA1 and SERPINA1 were confirmed and demonstrated an expression matching the one seen in the discovery study. CTHRC1, HBE1, DKK3 and CSF1R could not be detected. FGA could not be confirmed and SOD3 could only be confirmed when comparing the centenarians to control, or control combined with offspring. In summary, the targeted study validated eight of the 13 proteins from the discovery phase.

Table 4-8. Comparison of results from the discovery study and the targeted study. *The table shows the significance levels of the proteins from the discovery study, which group they were highest in and the significance level in the targeted study for the comparisons centenarians versus control, offspring, and control and offspring combined and if the discovery results were replicated in the targeted study. A2M, ADIPOQ, CST3, PTGDS, PKM, SAA1 and SERPINA1 were confirmed. CTHRC1, HBE1, CSF1R and DKK3 could not be detected. FGA could not be confirmed. †SOD3 was confirmed, but not in centenarians vs offspring (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$, and NS $p > 0.05$).*

	Sig. level discovery study	Highest in	Sig. level targeted study			Targeted results
			Centenarians vs control	Centenarians vs offspring	Centenarians vs control/ offspring	
A2M	*	Centenarian	**	****	****	Confirmed
ADIPOQ	*	Centenarian	**	***	****	Confirmed
CST3	*	Centenarian	****	****	****	Confirmed
CTHRC1	*	Control				Not detected
FGA	*	Centenarian	NS	NS	NS	Not confirmed
HBE1	*	Centenarian				Not detected
PTGDS	*	Centenarian	****	****	****	Confirmed
SOD3	*	Centenarian	*	NS	*	Confirmed†
CSF1R	**	Centenarian				Not detected
DKK3	**	Centenarian				Not detected
PKM	**	Control	***	****	****	Confirmed
SAA1	**	Centenarian	**	**	***	Confirmed
SERPINA1	****	Centenarian	*	*	**	Confirmed

4.3.2.2.4 *Conclusions from the univariate analysis*

The univariate analysis concluded that there was a significant age difference between the males and females in the study, and moreover that some proteins exhibited altered expression between males and females. Therefore, the data was adjusted for sex to avoid bias. Correlation analysis demonstrated that there was no significant relationship

between age and protein expression in the control and offspring groups, not either when combining the groups. The proteins C3, HPX and SERPINF2 were significantly decreased with older age in the centenarians. A total of 18 proteins were differentially expressed between the centenarians and control/offspring. No differences in protein expression could be discerned between the control and offspring groups. Finally, eight of 13 proteins from the discovery phase were validated in the targeted analysis; these proteins were A2M, ADIPOQ, CST3, PTGDS, PKM, SAA1, SERPINA1 and SOD3.

4.3.2.3 *Machine learning for age prediction of centenarians*

The correlation analysis performed in section 4.3.2.2.1 could not establish an age-protein relationship in the control and offspring samples when analysing the individual expression of the proteins. To evaluate if more complex modelling could identify a protein-age dependency in the data, a machine learning approach was undertaken. The aim was to construct a regression model from the control and offspring samples and predict the centenarians to evaluate if their predicted age would be younger than their biological age.

Ridge regression was utilised to create a model of the control and offspring subjects, with age set as the dependent variable Y. The feature importance of the proteins in the model are shown in Figure 4-11.

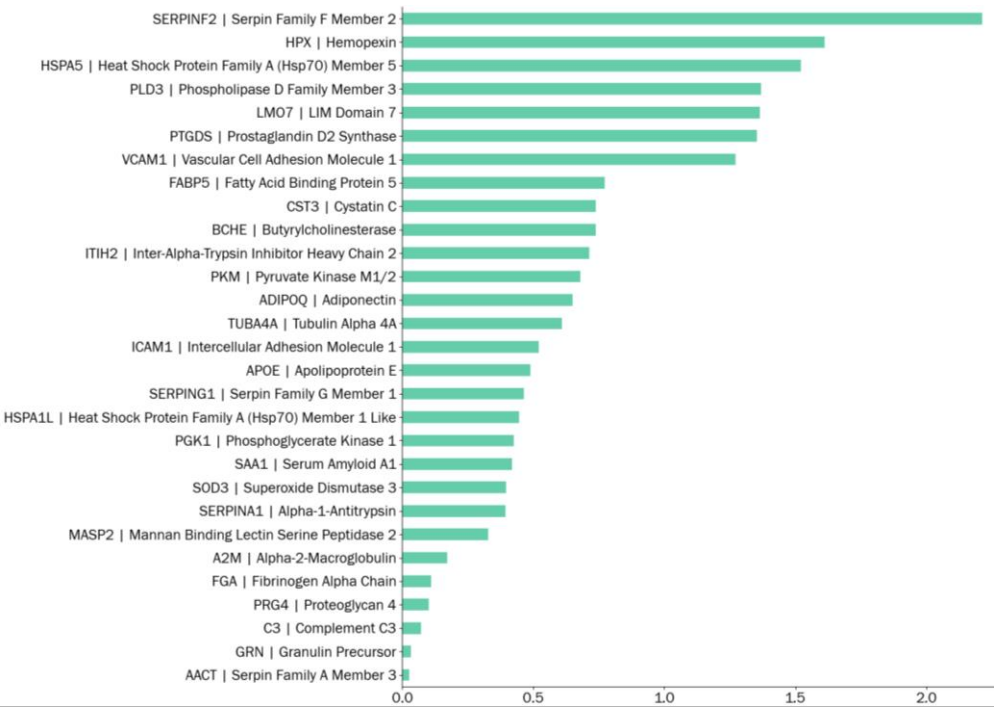


Figure 4-11. Feature importance of the proteins in a Ridge regression model of control and offspring individuals, relating protein expression to age.

Next, the ages of the centenarians were predicted in the Ridge regression model. The prediction resulted in the centenarians being predicted in the age span 70 to 90 years rather than their biological age of 100 to 112 years (Figure 4-12). Although initially promising, closer inspection concluded that the model fit for the training set alone (the control and offspring samples) was R2 = 0.22, thus demonstrating that the correspondence between the observed and predicted values was limited. Evaluating this further, the control and offspring samples were split randomly into two groups. One of the groups was used to build a new model and the second group was predicted in this model. The results of the age prediction were mostly random and did not correspond with the actual ages, signifying that the model had poor predictive ability and that no strong relationship with age could be determined and used for predictions of the centenarians' ages.

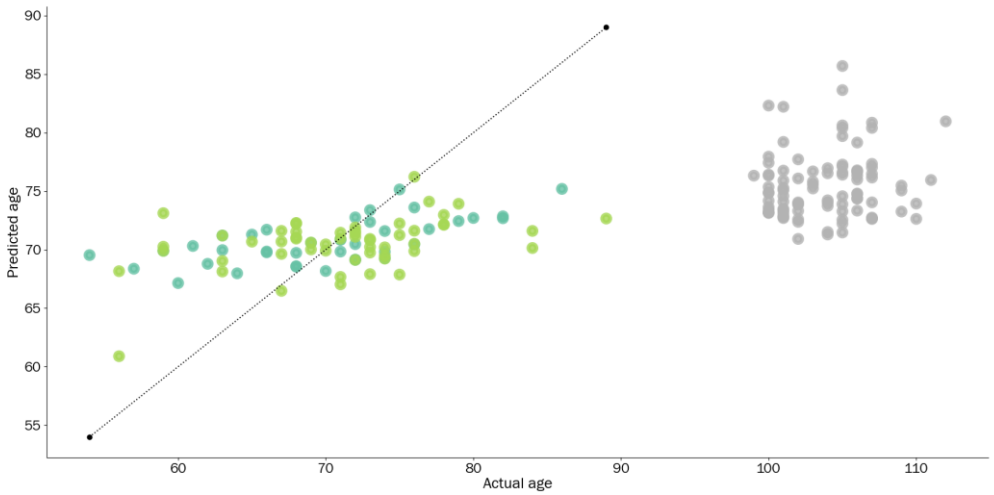


Figure 4-12. Prediction of the centenarians' ages in a Ridge regression model trained by control and offspring samples. *Although the centenarians are predicted in an age range approximately 30 years younger than their actual ages, so are the control and offspring individuals which were used to construct the model as demonstrated by the line expected in a perfect model (- - -) showing the minimum and maximum age of the control and offspring samples. The sample colours represent ■ controls, ■ offspring, and ■ centenarians.*

Other, non-linear, machine learning strategies such as neural network and random forest, both from Scikit Learn version 0.24, were attempted but also resulted in poor ability to correctly predict the ages in the training data. The consistent absence of predictive power in the training set consisting of offspring and control samples, regardless of model type, strongly indicates that there is insufficient correlation between the protein expression and age in the training data. The results from the machine learning regression consequently verifies what the univariate correlation analysis demonstrated - that the control and offspring groups do not exhibit any distinct age-related protein expression.

4.3.2.4 *Pathway analysis and literature studies*

A predicament in the study of centenarianism is that ageing in itself may alter the expression of some proteins regardless of centenarianism. There is a risk of confounding as many of the proteins whose expression correlate with normal ageing will also change in the centenarians. As mentioned in sections 4.3.2.2.1 and 4.3.2.3, the correlations between age and protein levels in offspring and controls were not significant, meaning that no information about the normal/expected protein changes over age could be acquired from this source and compared to the progression of the centenarians' protein expression. For this reason, the strategy to identify proteins relevant to centenarianism was based on the difference in protein expression between centenarians and offspring/controls in conjunction with pathway analysis and literature studies of the proteins' reported changes with age.

4.3.2.4.1 *Pathway analysis*

The significantly different proteins were analysed using Ingenuity Pathway Analysis (Qiagen). The pathway analysis (Figure 4-13) demonstrated that mainly inflammation-related pathways were affected, and the results largely overlapped with the significant pathways resulting from the discovery analysis (Section 4.2.2). The pathway with highest significance in both analyses was acute phase response signalling. Also, LXR/RXR and FXR/RXR activation, and complement and coagulation system were identified in both the discovery and targeted data. Although these results were expected as the targeted assay mainly contains inflammation-related proteins, it was reassuring that the protein expression from the targeted analysis closely matched the discovery phase. Given the limited number of proteins, the IPA software could not predict if the pathways were up- or downregulated and thus, no z-scores are provided.

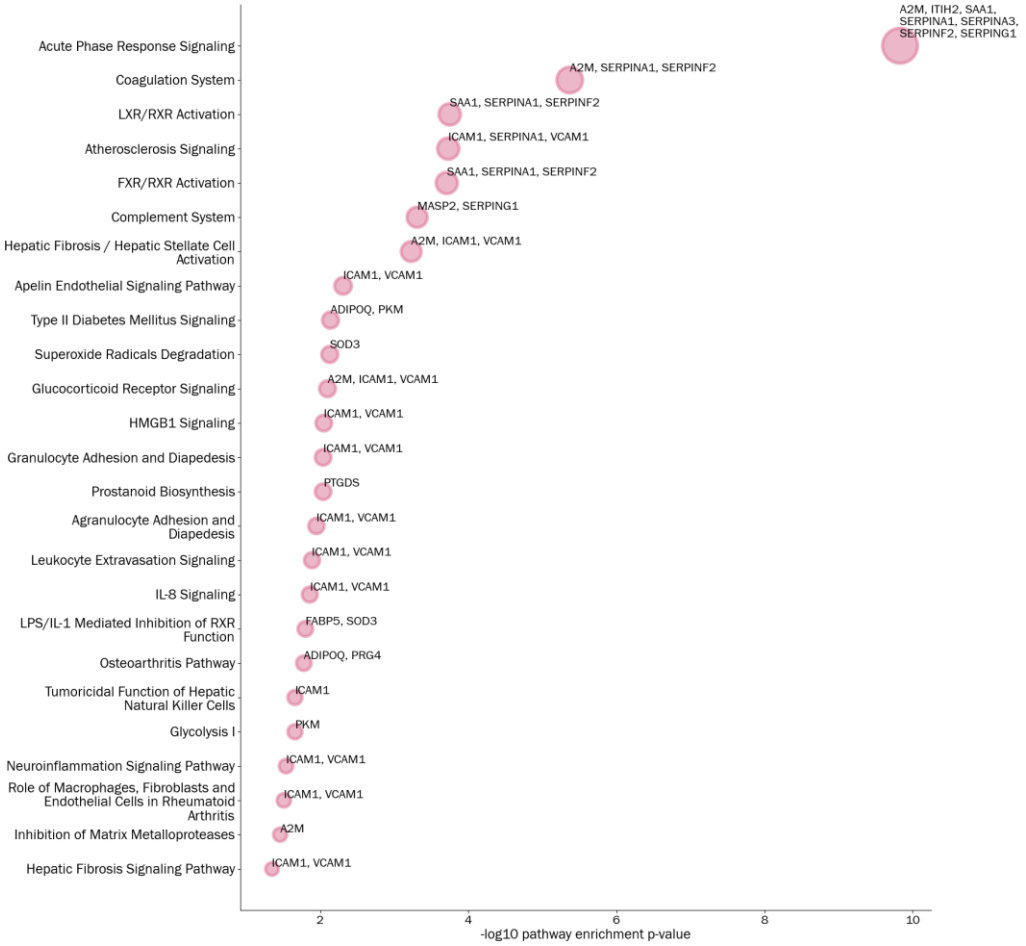


Figure 4-13. Pathway analysis in IPA of centenarians versus offspring/controls. *The pathways are annotated by the respective proteins included in each. Circle radii indicate the significance of the pathway enrichment p-value. The most significant pathway, acute phase signalling, was also the most significant pathway in the discovery phase. Additionally, the pathways LXR/RXR and FXR/RXR activation, and complement and coagulation system were identified in both the discovery and targeted data.*

4.3.2.4.2 *Literature studies*

Previous studies of ageing and centenarianism are a valuable tool in determining if a protein has been related to the normal process of ageing or if it is a potential marker of longevity. The strategy to identify proteins potentially involved in explaining centenarianism was composed of reviewing the significantly different proteins from our study but also the lack of significance in proteins that that have previously been described to correlate with age. The goal of the literature review was to separate proteins normally correlated with age (senescence) from markers that could otherwise explain the healthy ageing of centenarians. Figure 4-14 shows a schematic representation of how the proteins from the targeted study were classified with the help of previous published studies.

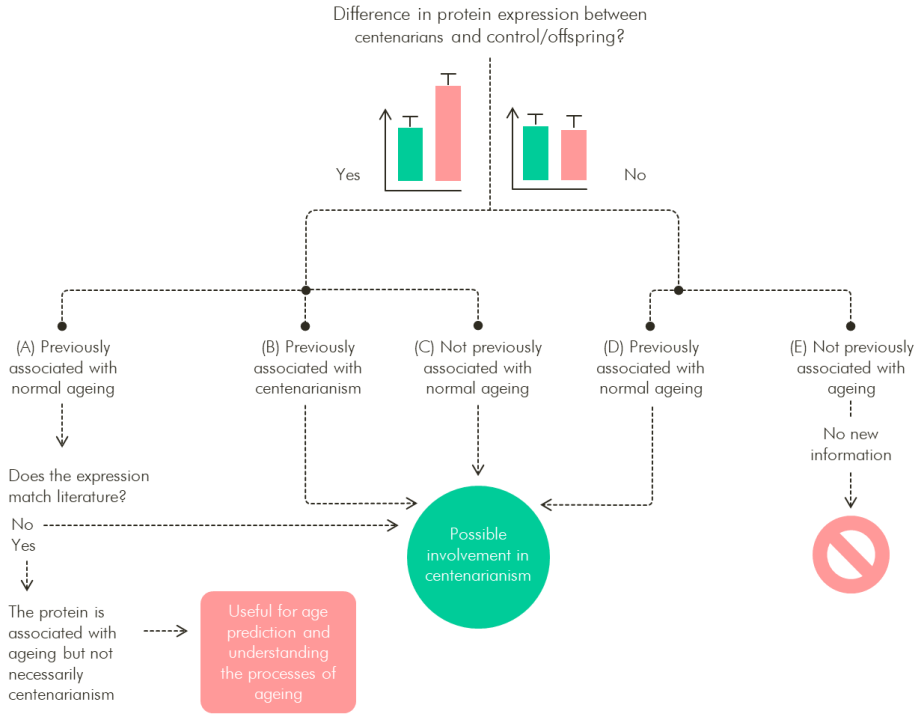


Figure 4-14. Schematic illustration of the workflow used when reviewing literature.

In a systematic review published by Johnson et al in 2020 [166], a large number of proteomic age studies were collated and the proteins most commonly reported to be associated with age listed. The authors highlighted several proteins found to have age-related levels in five or more studies, all detected in plasma and/or serum (hereafter called *"Tier A"* proteins). None of the most significant proteins from our study are included in this list of repeatedly reported age-associated proteins. The authors also compiled a list of proteins mentioned in two or more studies (called *"Tier B"* proteins), from this list of "lower-confidence" proteins, A2M, ADIPOQ, CST3, PTGDS, SERPINA3 and SERPINF2 are reported to alter expression with age. BCHE, ITIH2, PKM, PRG4 and VCAM1 are not mentioned.

A large-scale proteomic study published in 2019 by Lehallier et al describes undulating changes in the plasma proteome over age [167]. Comparing Lehallier's study to the results from this study, PTGDS, SERPINF2 and VCAM1 are also found in Lehallier's study where PTGDS demonstrated a drastic increase with age in the span 70 - 100 years and SERPINF2 decreased from 20 to 100 years. VCAM1 did not change with age.

The proteins from our study were compared with Lehallier's study and the review by Johnson et al. The proteins were also cross reviewed against other studies of ageing, longevity and centenarianism found through PubMed or Google Scholar. After reviewing literature, the proteins were classified as: (A) possible link to centenarianism, (B)

previously reported to correlate with normal ageing, and (C) no reported relationship with ageing. This resulted in a complex and intricate weave of potential risks and protective functions presented in Table 4-9.

Table 4-9. Literature compared with the significantly different proteins from the targeted study. *Based on previous reports from the literature, the proteins are classified as (A) possible link to centenarianism, (B) reported correlation with normal ageing, or (C) not previously reported linked to ageing*

Protein	Literature references	Classification
	Significantly differentially expressed between centenarians and control/offspring	
A2M (↑)	Not included in Lehallier's study. Tier B protein in Johnson's review. A2M has been reported to decrease with age [168]. In this study, A2M is upregulated in the centenarians.	A, B
ADIPOQ (↑)	Variant of ADIPOQ significantly positively correlated with longevity in a study of centenarians [169]. Reported to protect against age-related diseases [170].	A
BCHE (↓)	Hydrolyses Ghrelin of which low levels are associated with less obesity and possibly longevity [171]. High levels of BCHE are linked to lower mortality risk [172]	B
C3 (↓)	Found to increase with age in a microarray study of young and aged individuals, and AD patients [173]	A, B
CST3 (↑)	Involved in intra-cellular degradation of APP-Aβ [174]. Lower levels of CST3 were found associated with longevity (and late onset AD) in a proteomic study [175].	B
FABP5 (↑)	No changes in Lehallier or Johnson's studies. No compelling evidence from literature to support that FABP5 changes with age	A, C
ICAM1 (↑)	No changes in Lehallier or Johnson's studies. Reported to increase with age in an ageing study in rats [176]. Also reported to be upregulated in senescent cells [177].	B
ITIH2 (↓)	ITIH2 has been identified in ageing studies, downregulated with increasing age [178].	B
MASP2 (↑)	Tier B protein in Johnson's review but not found in Lehallier's study. Additional literature searches did not identify any results related to longevity or ageing.	A, C
PKM (↓)	Increased levels in older naked mole-rats compared to younger [179]. Otherwise, no evidence found of associations with age.	A, B
PRG4 (↓)	Not significant in Lehallier's or Johnson's publications. PRG4 has been found to decrease with advancing age in tendons and cartilage in animal studies [180, 181]	A, C
PTGDS (↑)	Up with age (Lehallier). Changes with age, tier B, Johnson.	B
SAA1 (↑)	No changes in Lehallier or Johnson's studies. No compelling evidence to support that SAA1 changes with age.	A, C
SERPINA1 (↑)	No changes in Lehallier's study, but a tier B protein in Johnson's review.	A, B
SERPINA3 (↑)	No changes in Lehallier or Johnson's studies. Upregulated with age in healthy controls in a study of prion disease [182]. Other studies have found no differences.	A, C
SERPINF2 (↓)	Downregulated with age in Lehallier's study and altered with age in Johnson's review (tier B). Upregulated in younger rats compared to older. Dietary intervention increased levels [183]	B
SOD3 (↑)	No changes in Lehallier or Johnson's studies. SOD3 is an antioxidant. No clear hits on ageing.	A, C
VCAM1 (↑)	Elevated levels are associated with ageing and have been linked to an increased inflammatory profile in endothelial brain cells [184]	B
	No significant difference between centenarians and control/offspring	
HPX	Altering with age in Johnson's review but not in Lehallier's study. It was also found to alter in a proteomics longevity study in men [175].	A, B
APOE	Alters with age in Johnson's review (tier B) but not in Lehallier's study. In ageing men, APOE changed with age [175]. Associated with increased lifespan in some studies but not others [185]	B/C
FGA	Altered with age in both Johnson's review (tier A) and Lehallier's study. It was also found to significantly correlate with age in a large-scale proteomics study [186].	A, B
GRN	Altered with age in both Johnson's review and Lehallier's study.	A, B
HSPA1L	Not found in Johnson's review or Lehallier's study. No compelling evidence to support an expression change with age.	C
HSPA5	Not found in Johnson's review or Lehallier's study. Increased levels were found to be associated with increased mortality in a study of longevity in men [175]	A, B
LMO7	Not included in Lehallier's study and not found in Johnson's review. No compelling evidence to support an expression change with age.	C
PGK1	Tier B protein in Johnson's review and changes with age in Lehallier's study.	A, B

Protein	Literature references	Classification
	Significantly differentially expressed between centenarians and control/offspring	
PLD3	Tier B protein in Johnson's review and weakly correlated with age in Lehallier's study. Variants are associated with AD, but no evidence that the protein is correlated with age or longevity.	B/C
SERPING1	No changes in Lehallier or Johnson's studies. No compelling evidence to support that SERPING1 changes with age from the literature.	C
TUBA4A	Not found in Johnson's review or Lehallier's study. No compelling evidence that the protein is correlated with ageing.	C

4.3.3 *Summary and conclusions from the targeted validation phase*

In the validation phase, 13 proteins from the centenarian discovery phase were included in an assay together with other proteins from neurodegenerative studies and several pro- and anti-inflammatory proteins from literature. The assay was applied to a cohort of samples consisting of centenarians, offspring, and controls. Eight proteins identified in the initial discovery study were confirmed and validated in the targeted study, these proteins were A2M, ADIPOQ, CST3, PTGDS, PKM, SAA1 and SERPINA1. In addition to these proteins, 11 other proteins were significantly differentially expressed between the groups, these were BCHE, C3, FABP5, ICAM1, ITIH2, MASP2, PRG4, SERPINA3, SERPINF2, SOD3 and VCAM1. No differences in protein expression could be discerned between the control and offspring samples.

Although limited in interpretation possibilities due to the targeted nature of the study, pathway analysis largely reflected the results from the discovery study, also demonstrating that inflammatory pathways were affected. Literature reviews aided in classifying the proteins into three groups: (i) proteins with a possible link to centenarianism, (ii) proteins correlated with normal ageing, and (iii) proteins with no known link to ageing or centenarianism. Based on these classifications and the observed protein expression, it was concluded that the centenarians likely exhibited both protective and risk protein characteristics. Several inflammatory proteins were upregulated in the centenarian group, including FABP5, MASP2, SERPINA3 and SAA1. Also, VCAM1 and ICAM1, both pro-inflammatory and indicators of oxidative stress. Interestingly, the central complement protein C3 was downregulated in the centenarian group, thereby indicating reduced complement response. The potent anti-inflammatory protein ADIPOQ was moreover upregulated in the centenarians.

In conclusion, it was hypothesised that the centenarians exhibited a protein expression both indicative of risk, likely due to being near the end of their lives, and of protective mechanisms which may have aided in keeping them healthy for such an extended amount of time.

4.4 DISCUSSION

The initial study performed in the discovery phase identified many proteins that differed in expression between centenarians and controls, and pathway analysis demonstrated that the centenarians had inflammatory pathways activated. From the discovery study, 13 proteins were selected for validation and added to an explorative targeted assay which also included targets from neurodegenerative disease discovery studies and known inflammatory proteins from literature. Eight of the 13 proteins were validated in a larger targeted study, and ten additional proteins from the augmented assay were differentially expressed in the centenarians.

The measured plasma proteins in the centenarians did show evidence of increased inflammation compared to controls and offspring. However, a number of proteins measured in the centenarians also show a pattern hypothesised to provide protective functions compared to offspring and controls. The pattern of elevated levels of inflammation corresponds with previous centenarian studies. Franceschi et al call this the "*paradox of the pro-inflammatory status of healthy centenarians*" in their seminal paper on inflammageing from 2000 [187]. There, they describe increased levels of pro-inflammatory agents in healthy ageing and centenarians and argue that individuals lacking what they call "*robust and anti-frail*" gene variants are more susceptible to age-related diseases. The authors suggest that inflammatory stimuli throughout life provide a biological background which may increase the susceptibility to inflammatory, age-related diseases such as Alzheimer's disease, diabetes and atherosclerosis, but that a lack of protective gene expression is likely required for these conditions to develop. They further rationalise that the inflammageing observed in centenarians may be a result of beneficial effects of the immune response at younger age, turning detrimental at older age. Our study may indicate that the higher levels of A2M, ADIPOQ and SOD3 provide protection against inflammation and oxidative stress. The lower level of the central complement cascade protein C3 observed in the centenarians may suggest inhibition of the complement cascade. The lower levels of PKM in the centenarians may reflect more efficient insulin-sensitivity.

Figure 4-15 shows a Venn diagram of the proteins divided into suggested protective and risk groups and their association and influence on ageing.

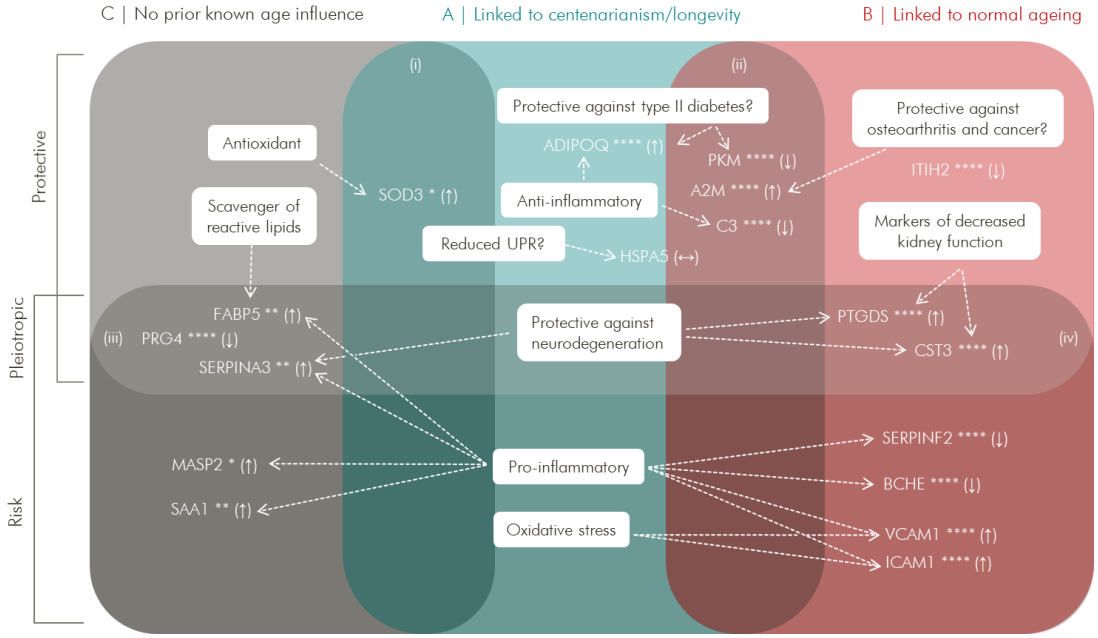


Figure 4-15. Classification of the significantly different proteins detected in the targeted study of centenarians. *The diagram is divided vertically into risk, protective and pleiotropic proteins. Horizontally, the diagram is divided into (A) proteins previously linked to centenarianism and longevity, (B) proteins linked to normal ageing and (C) proteins with no prior known age associations. Vertically, the proteins are divided into protective and risk with the overlapping area in between classified as pleiotropic. The overlapping area (i) represents no prior known age influence and a possible link to longevity. The overlapping area (ii) represents a link to both normal ageing and longevity. Area (iii) contain pleiotropic proteins with no prior link to age. area (iv) represents pleiotropic proteins with a prior ling to ageing.*

Risk protein expression. The proteins assigned to this group have been correlated with normal ageing in previous studies and are likely markers of increased mortality risk and/or being close to the end of life. Associated with increased risk of mortality are high levels of SAA1, MASP2, VCAM1 and ICAM1 and low levels of SERPINF2 and BCHE. These are all proteins indicative of inflammation. Serum Amyloid A1, SAA1, is a major acute phase protein whose blood levels increase drastically with inflammation. SAA1 has been shown to displace APOA1 in HDL during acute phase response, thereby suppressing HDL's anti-inflammatory properties [188]. It has also been demonstrated that SAA1 increases reactive oxygen species [189]. Some studies suggest that SAA1 may be pleiotropic, also acting as an anti-inflammatory agent [190]. Mannan Binding Lectin Serine Peptidase 2, MASP2, is an initiator of the lectin pathway of the coagulation cascade where it binds to sugar moieties on bacteria and pathogen particles to form the mannose-binding lectin complex. Activation of the lectin pathway is believed to contribute to brain inflammation as suggested in a study of traumatic brain injury [191]. Intercellular Adhesion Molecule 1, ICAM1, and Vascular Cell Adhesion Molecule 1, VCAM1, are both endothelial adhesion molecules. They can be activated by cytokines, complement proteins and reactive oxygen species and increase in concentration with endothelial activation [192]. VCAM1 is known to increase with age and has been linked to inflammation in endothelial brain cells. A study

showed that by transplanting blood from old mice to young, ageing was accelerated in the young mice's brains [184]. Serpin Family F Member 2, SERPINF2, is a serine protease inhibitor and a major inhibitor of plasmin, thereby regulating blood clotting [193]. SERPINF2 has been reported to decrease with age in several studies, however a study of rats demonstrated that SERPINF2 levels increased with dietary intervention and autophagy [183]. Butyrylcholinesterase, BCHE, is an esterase involved in detoxification of a number of compounds [194]. High levels of BCHE are linked to lower mortality risk [172] and low levels are associated with cardiovascular risk, mortality and systemic low-grade inflammation [195]. BCHE moreover hydrolyses ghrelin, a hunger hormone. Low levels of ghrelin are associated with less obesity and possibly longevity [171] and there are even suggestions of using BCHE gene therapy as obesity treatment [196]. In essence, the expression in this group of proteins indicates an overall increased state of inflammation, consistent with Franceschi's theory about inflammageing [197].

Protective protein expression. The protein expression proposed to provide protection against early death and promote longevity consists of high levels of A2M, ADIPOQ and SOD3 together with low levels of C3, PKM and ITIH2. These proteins have a heterogeneous set of functions. Alpha-2-Macroglobulin, A2M, is a protease inhibitor that can also inhibit cytokines, thereby mediating inflammatory responses [198]. It was found over-expressed in the longest-living rodent, the naked mole-rat, and is proposed as one of the causes for its resistance to developing cancers [199]. A2M can decrease the inflammatory response by "capturing" pro-inflammatory agents and has recently been suggested as a treatment to slow down osteoarthritis [200, 201]. In humans, A2M has been reported to decrease with age [199, 202] whereas our data show a 1.4-fold increase in the centenarians. Adiponectin, ADIPOQ, (1.5-fold increase in the centenarians) has been linked to longevity in several studies. Higher levels of ADIPOQ are beneficiary as it has anti-inflammatory, anti-oxidising, anti-diabetic and anti-apoptopic properties [203] and protects against age-related diseases [170]. Calorie restriction increases ADIPOQ and ADIPOQ-signalling further modulates the downstream AMPK and PPARα pathways [204]. Superoxide Dismutase 3, SOD3, (1.5-fold increase in the centenarians) is an antioxidant enzyme which catalyses the breakdown of superoxide. It is reported to have protective effects on various inflammatory diseases [205]. SOD3 was shown to influence NF-κB activity in a mouse study, thereby reducing the production of pro-inflammatory species [206]. Complement C3 (1.2-fold decrease in the centenarians) is a central protein in the inflammatory complement cascade. Although complement activation is part of the innate immune system and aims to clear pathogens and regulate the inflammatory

response, it is a response known to occasionally turn into a harmful promoter of inflammation. This "friendly fire" is reported to become more pronounced with advanced age [207]. Increased levels of complement have been associated with a number of detrimental conditions such as cardiovascular and neurodegenerative diseases [208]. Down-regulated levels of C3 have been associated with increased lifespan and negatively correlated with metabolic syndrome and abdominal obesity [209]. Pyruvate Kinase M, PKM, (1.2-fold decrease) is an enzyme that regulates the terminal step in glycolysis, transforming glucose to pyruvate and yielding one ATP. PKM is also involved in gluconeogenesis, the process of providing tissues with glucose in a fasting state [210]. Interestingly, recent research shows that PKM can initiate an alternative route to insulin secretion from ß-cells by locally increasing the ATP/ADP ratio in an oscillating manner, thus providing the energy required for insulin secretion [211]. A study of rats subjected to hyperglycaemic stress showed that the rats with type II diabetes exhibited higher levels of PKM compared to controls [212]. Another rat study examined the expression of PKM in response to a simple and a complex carbohydrate diet. The researchers found that PKM was significantly lower in the rats being fed complex carbohydrates [213]. The proposed protective mechanisms expressed in this group of proteins are heterogenous, however, the common theme among them is anti-inflammatory properties. The upregulation of A2M and ADIPOQ provides anti-inflammatory effects, while upregulation of SOD3 may indicate resilience against oxidative stress. Low levels of complement C3 implies that the complement response is not activated. The downregulation of PKM may be an indication of more efficient insulin-sensitivity.

Pleiotropic proteins. There are a number of proteins who appear to be pleiotropic, such as PTGDS and CST3, both markers of decreased kidney function but at the same time suggested to protect against neurodegeneration. Cystatin C, CST3, (2-fold increase in the centenarians) is a cysteine protease inhibitor. It is a marker of kidney dysfunction and high levels have been associated with increased risk of mortality [214]. Curiously, CST3 is also reported to have a neuroprotective role against AD and PD [215]. Prostaglandin D2 Synthase, PTGDS, (2-fold increase in the centenarians) is an inhibitor of platelet aggregation but has also been reported to increase with decreased kidney function [216]. PTGDS has also been suggested to enhance the anti-inflammatory function of astrocytes together with the PD gene DJ-1 [217]. Alpha-1-Antitrypsin, SERPINA3, (1.3-fold increase in the centenarians) is a positive acute phase protein, thus increasing with inflammation, but has also been suggested to act as a mediator of amyloid-β clearance in AD patients. Proteoglycan 4, PRG4, (1.3-fold decrease in the centenarians) is a joint-lubricating protein.

It has been proposed that PRG4 has a wider function and can bind to and effect inflammation cell surface receptors, thereby regulating the inflammatory response [218]. Fatty Acid Binding Protein 5, FABP5, (1.3-fold increase in the centenarians) is located in the cytosol and the plasma membrane [219], it is a carrier for lipids and fatty acids within the cell and is also involved in modulating inflammation [220]. It has been reported to scavenge reactive lipids and was found upregulated in mice exposed to Western diet [221]. Curiously, FABP5 was found downregulated in adipose tissue of insulin resistant subjects [222].

Non-significantly different proteins. The group of proteins previously described to correlate with age, or to be associated with longevity, but not differentially expressed in this study are HPX, APOE, FGA, GRN, HSPA5 and PGK1. This group of proteins can provide valuable information since the absence of expected change can give clues about pathways that are not affected in the centenarians. Hemopexin, HPX, is an acute phase reactant, increasing in response to inflammation. It also acts as an extracellular antioxidant, binding heme and thereby protecting molecules from heme oxidation [223]. Apolipoprotein E, APOE, has been studied with various outcomes in different ageing studies. When the different APOE alleles/genotypes have been researched, an influence on lifespan has been determined. The APOE ε4 genotype is for example strongly correlated with several detrimental conditions such as cardiovascular disease and Alzheimer's disease and therefore negatively correlated with extended lifespan, while the ε2/ε3 genotype is associated with longevity [224]. We measured total APOE and found no differences between the groups. Fibrinogen Alpha Chain, FGA, is a positive acute phase reactant and an element of fibrin, one of the major components of blood clots. Low FGA levels are associated with risk of bleeding due to decreased ability of clotting and high levels are linked to risk of cardiovascular disease [225, 226]. The lack of difference between the groups suggests that blood clotting is not affected in the centenarians. Granulin Precursor, GRN, is a pleiotropic protein. It is a regulator of lysosomal function and is involved in anti-inflammatory response where it can act as an inhibitor of cytokine release. It has also been found exerting proinflammatory effects in type II diabetes patients and obese individuals where it correlated with levels of CRP [227]. Mutations in GRN are moreover associated with frontotemporal lobe dementia. Heat Shock Protein Family A Member 5 (BiP), HSPA5, is a key regulator of protein folding and quality control in the ER and can initiate the unfolded protein response in cases of ER stress [228]. HSPA5 increases in response to ER stress and the lack of upregulation in the centenarians indicates that there is no ongoing stress in the endoplasmic reticulum. Phosphoglycerate Kinase 1,

PGK1, is an enzyme involved in catalysing the first ATP producing step in the glycolytic pathway [229].

In conclusion, we propose that the results from this study indicate that the centenarians have increased levels of inflammation and higher levels of several proteins normally associated with age, but also a protein expression that we suggest acts protectively, mainly by protecting against inflammation.

Study limitations. The greatest limitation we recognise in this study is the absence of suitable control samples. This is a problem hampering all studies of centenarianism and there is no easy way of amending the shortfall. Not only is the control issue related to comparing two groups who are at different ages, but also groups who have lived through different times. The centenarians have been exposed to an environment which may differ greatly from that of a 60 year old control [230]. One possible strategy to identify longevity-related proteins is to build regression models from control samples and predict centenarians. We attempted this in section 4.3.2.3 but as we were unable to build robust models from the training data, we could not explore this approach further. The reason for the non-robust models is likely that the control group was not large enough and that the majority of the samples were found in a relatively narrow age range, therefore not allowing for sufficient model training across the age range.

Blood-based discovery proteomics to find biomarkers of Parkinson's disease followed by targeted validation

5

Abstract. *There is a need for specific biomarkers of PD - for diagnosis, to follow the disease progression and to monitor new therapies as they are developed. A mass spectrometric blood-based test would be highly beneficial, as the sampling is relatively non-invasive, compared to for instance lumbar punctures, and because of the speed, accuracy, and cost-efficiency of targeted mass spectrometric assays. Aiming to identify biomarkers of Parkinson's disease, the first phase of our study was dedicated to discovery proteomics. Applying an optimised workflow to identify the maximum number of proteins in plasma and serum, two Parkinson's disease discovery cohorts were analysed by untargeted LC-MS. The cohorts consisted of newly diagnosed, treatment-naïve PD patients and controls, and of homozygous twins, discordant for developing PD a few years after the time of sampling. The analyses identified several putative proteomic biomarkers and pathway analysis indicated that inflammatory pathways were affected in the PD patients, along with unfolded protein response and Wnt signalling. A selection of the identified target proteins was added to an augmented targeted mass spectrometric, MRM-based assay, also including several pro- and anti-inflammatory proteins from literature. The targeted assay was applied to a larger cohort of newly diagnosed PD patients, iRBD patients, controls, and a positive control group consisting of non-PD neurological disorders. A total of 19 proteins from the targeted assay were differentially expressed between PD and controls. Five of the proteins from the discovery study were also differentially expressed in the targeted study, this included the Wnt signalling protein Dickkopf 3 and the unfolded protein response protein BiP. Discriminant multivariate modelling by OPLS-DA demonstrated that the control and PD groups could be differentiated based on the multivariate protein expression. The targeted data were further used to build machine learning models, and prediction of the data demonstrated that it was possible to discriminate Parkinson's disease from control with 100% accuracy based on the expression in a panel of nine proteins.*

Finally, the protein expression observed in the targeted analysis pointed towards the involvement of increased complement, increased unfolded protein response, and a reduction in Wnt/ß-catenin signalling in the PD patients. We postulate that the observed 1.8-fold downregulation of Dickkopf 3 in the PD patients may indicate that they have reduced protection of dopaminergic neurons.

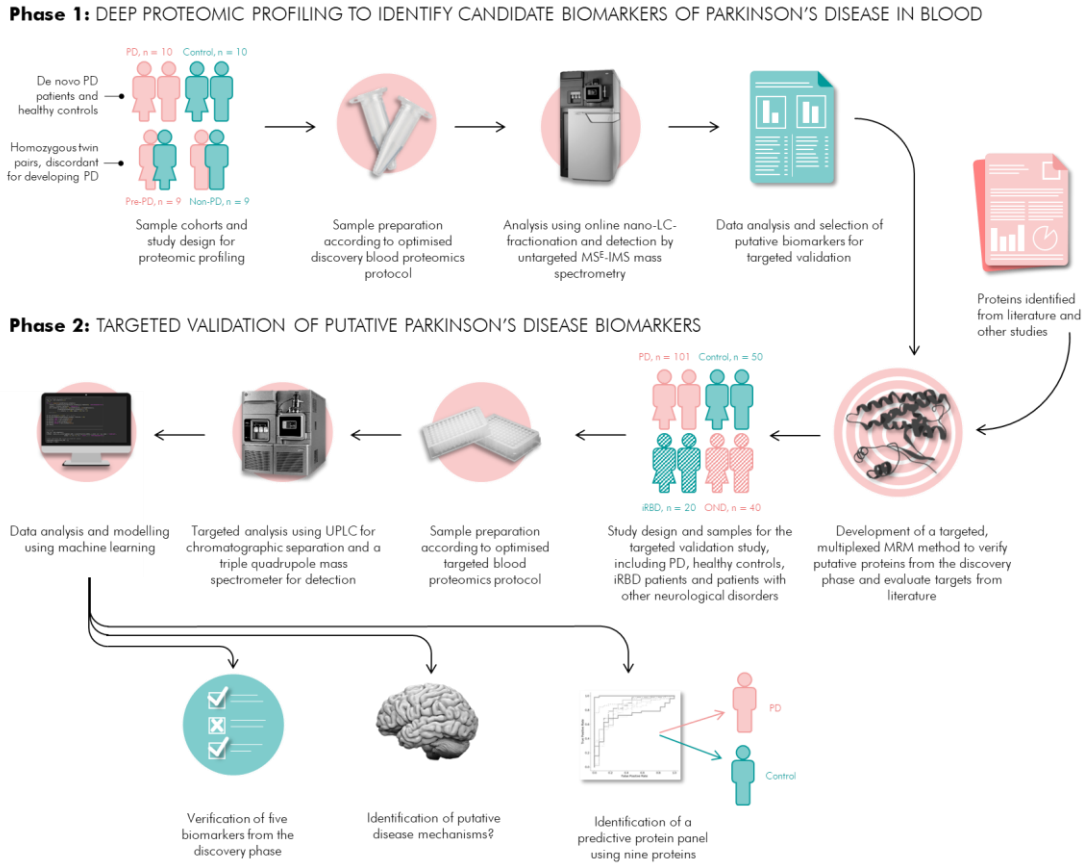


Figure 5-1. Graphical abstract of the blood-based discovery and validation biomarker study of Parkinson's disease presented in this chapter.

5.1 INTRODUCTION AND AIMS

Diagnosis of Parkinson's disease has traditionally been determined clinically and based on motor phenotype [231]. Evidence does however suggest that this diagnosis fails to capture the complexity of the disorder since motor symptoms appear years after the disease process has started and after approximately 60% of substantia nigra neurons have already been lost. Moreover, non-motor symptoms often occur prior to the manifestation of motor symptoms [232]. The accurate clinical identification of PD is consequently challenging and a meta-analysis from 2016, examining 13 clinicopathological studies, showed that no relevant improvement in the diagnostic accuracy had occurred over the past 25 years [233]. Therefore, identification of reliable PD biomarkers, ideally capturing the early stages of the disease and allowing for evaluation of disease progression, would be extremely beneficial. There have been several efforts to find PD biomarkers, but no single marker, or panel of markers, has yet been discovered to reliably discriminate between PD patients and healthy controls [234-236].

As described in Chapter 1, blood-based biomarker discovery is challenging due to the high concentration of house-keeping proteins such as albumin and immunoglobulins. These proteins are often of high molecular mass and consequently produce a large number of tryptic peptides thereby making it difficult to detect low- and medium-abundant proteins by mass spectrometry. Adding to that, plasma and serum are not ideal biofluids for studying diseases of the brain. The gold standard for examining neurological diseases is cerebrospinal fluid where neural cells secret proteins into the extracellular matrix which are then carried through the central nervous system by secreted cerebrospinal fluid [237, 238]. Blood and cerebrospinal fluid are not in direct contact but separated by the blood-brain-barrier. The blood-brain-barrier allows for selective transfer of molecules through different processes and some correlations between proteins in cerebrospinal fluid and plasma have been identified [239, 240]. Due to the lesser invasiveness, ease of sampling, and accessibility, blood-based biomarkers of neurodegenerative disease are desired and could more readily be used in clinical tests of neurodegeneration, thus making the effort of exploring blood-based biomarkers of Parkinson's disease worthwhile.

Extensive method development was performed in Chapter 3 to enable the detection of low- and medium-abundant proteins in blood-based discovery proteomics, and further to increase the number of identified proteins. In this chapter, the optimised method was applied to two different discovery cohorts and several putative biomarkers of Parkinson's disease were developed into a targeted test which was applied to a new, larger cohort of samples.

This chapter is dedicated to the search for new blood-based proteomic biomarkers of Parkinson's disease and goes through several stages, from the discovery phase – identifying potential targets, to a validation study of the putative biomarkers in a new and larger cohort of samples, and finally the development of a predictive panel of biomarkers using machine learning. The aims of the experiments performed in this chapter were to:

- In the discovery phase, identify putative blood-based proteomic biomarkers of Parkinson's disease by mass spectrometry
- Validate the targets from the discovery phase using a targeted, mass spectrometric MRM-based proteomic assay, applied to a new and larger cohort of samples
- Identify additional differentially expressed pro- and anti-inflammatory proteins from the literature
- Identify proteins, or a panel of a proteins, capable of distinguishing between Parkinson's disease and control

5.2 DISCOVERY PROTEOMICS IN PARKINSON'S DISEASE PLASMA AND SERUM TO IDENTIFY PUTATIVE PROTEOMIC BIOMARKERS

In the discovery phase of this study, the overall aim was to find blood-based biomarkers for Parkinson's disease that could be developed into a targeted test. The optimised discovery proteomics workflow described in Chapter 2 was applied to two sample cohorts: one used to explore the pre-symptomatic protein expression in Parkinson's disease, and one to identify putative biomarkers is newly diagnosed patients.

5.2.1 *Materials and methods*

5.2.1.1 *Discovery sample cohorts*

Two sample cohorts were analysed separately in the discovery phase; plasma from newly diagnosed and PD patients and serum from homozygous twins discordant for PD.

5.2.1.1.1 *De novo Parkinson's disease patients and controls*

Samples from de novo Parkinson's disease and controls were provided by Professor Brit Mollenhauer, Universitätsmedizin Göttingen-Klinik für Neurologie, Göttingen University, Germany. The samples were from the DeNoPa cohort and consisted of newly diagnosed, treatment-naïve PD patients and controls. The cohort has the potential to show proteins and pathways implicated in the early stages of the disease but not yet affected by symptom modifying treatment. The characteristics of the cohort are presented in Table 5-1.

Table 5-1. Characteristics of the samples included in the proteomic screening of de novo Parkinson's disease and control individuals. *The table shows the number of samples and percentages of males and females in each group. Time since diagnosis and age are not known. N/A = not available. SD = standard deviation.*

Group	Number of samples	Males/females	Years since PD onset ± SD	Age ± SD
De novo PD	10	80% M | 20% F	N/A	N/A
Control	10	40% M | 60% F	N/A	N/A

5.2.1.1.2 *Homozygous twins discordant for developing Parkinson's disease*

Parkinson's disease discordant twin pair samples from the Swedish Twin Registry were provided by the Karolinska Institute, Sweden, and consisted of serum from homozygous twins. One of the twins developed PD after sampling and one did not. This study benefits from having perfectly paired controls and has the potential to allow us to identify pre-symptomatic markers of idiopathic PD and the influence of lifestyle. The characteristics of the cohort are presented in Table 5-2.

Table 5-2. Characteristics of the samples included in the proteomic screening of pre-Parkinson's disease discordant twins. *The table shows the number of samples, percentages of males and females, number of years before PD onset in the pre-PD twins and average age in each group*

Group	Number of samples	Males/females	Years until PD onset ± SD	Age ± SD
Pre-PD	9	33% M | 67% F	4.6 (± 1.7)	64.8 (± 9.4)
Control	9	33% M | 67% F	N/A	64.8 (± 9.4)

5.2.1.2 *Sample preparation*

The samples from both cohorts were prepared according to the workflow described in detail in Chapter 2, section 2.3. In short, ten microlitres of plasma or serum were depleted of the twelve most abundant proteins using Pierce Top12 columns, before freeze drying, digestion and solid phase extraction. This allowed us to detect lower abundant proteins and increased the chances of identifying potential biomarkers and/or disease mechanisms for PD.

5.2.1.3 *Instrumental analysis*

All samples were analysed utilising a two-dimensional nano liquid chromatography system coupled to a Waters Synapt G2-S*i* time-of-flight mass spectrometer equipped with ion mobility separation. The samples were fractionated online into ten fractions and detection, using a label-free proteomics approach, was performed in positive MS^E mode. The detailed parameters of the instrumental analysis are described in Chapter 2, section 2.5.

5.2.1.4 *Data processing and analysis*

The acquired data were analysed as described in Chapter 2, section 2.6. In summary, the data were processed fraction-wise in Progenesis utilising the ion-accounting workflow. The fractions were subsequently merged, and the data were quality controlled. Run order drift was observed in the de novo PD dataset and LOWESS scaling was applied to correct it. Post LOWESS scaling, no run order drift was observed.

5.2.2 *Results from the study of de novo PD patients and controls*

Utilising the optimised plasma proteomics workflow, a total of 1238 proteins were identified and quantified in the study of newly diagnosed PD patients and controls. Of these, 696 proteins had an identification confidence score above 15 and were represented by two or more unique peptides.

5.2.2.1 *Univariate analysis*

A total of 47 proteins were found differentially expressed between de novo PD patients and controls on a nominal significance level of 95% using Student's t-test. Out of these, 35 proteins had two or more unique peptides and a confidence score larger than 15. The significance and direction of change are presented in a Volcano plot (Figure 5-2) and show that the number of proteins significantly up- or downregulated in the de novo PD and control groups are similar.

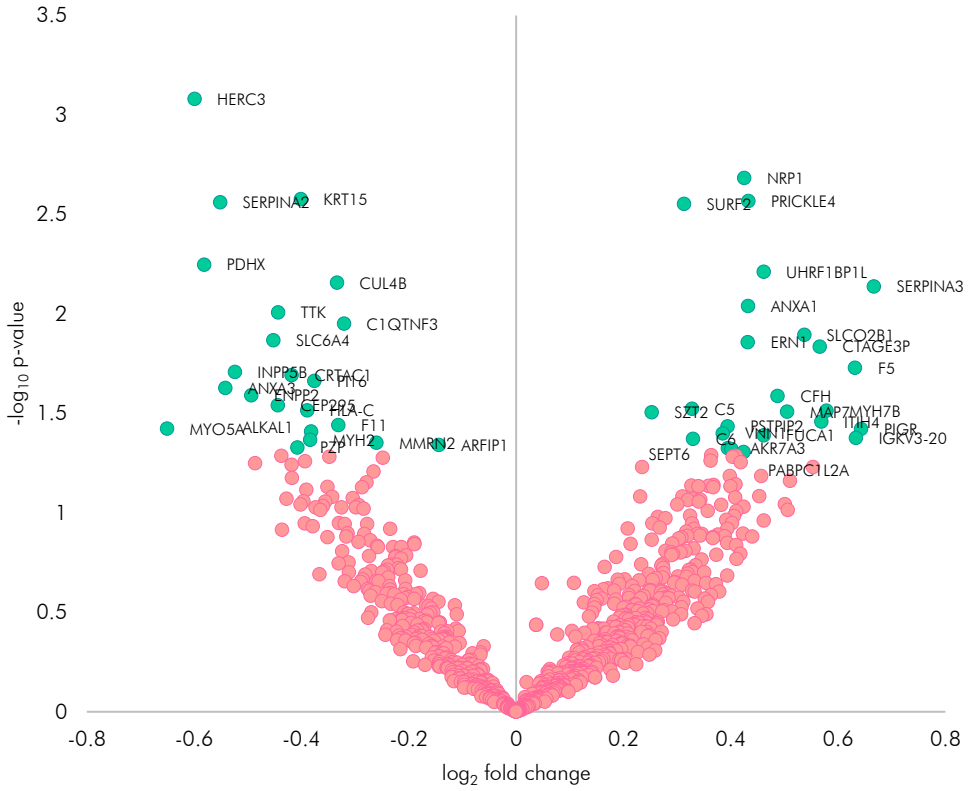


Figure 5-2. Volcano plot of all data from the discovery proteomics of de novo PD and control samples. *The horizontal axis shows log_2 of the average fold change and the vertical axis shows $-log_{10}$ of the p-values. The significantly different proteins are annotated by gene name.* ■ *significantly different,* ■ *not significantly different.*

Due to the skewed gender distribution in the sample set, an additional significance test was performed using males only. In this test, 63 proteins were significantly different between de novo PD patients and controls, 21 of which were also significant in the full dataset. Out of the significant proteins found in males, 52 were represented by at least two peptides and had an identification confidence score greater than 15, Figure 5-3 shows a Volcano plot of these results.

Figure 5-3. Volcano plot of male samples only from the discovery proteomics of de novo PD and control. *The horizontal axis shows log_2 of the average fold change and the vertical axis shows -log_{10} of the p-values. The significantly different proteins are annotated by gene name.* ■ *significantly different,* ■ *not significantly different.*

A detailed inspection of the similar and differentially expressed proteins discovered in the two comparisons, all data, and males only, demonstrated that all the proteins shared the same trend in direction of fold change regardless of being significant or not (Supplementary figure 1).

5.2.2.2 *Multivariate analysis*

A principal component analysis was performed to assess the data quality and to examine if any major trends or patterns could be detected in the data. The PCA did not detect any outlier samples and no run order drift post LOWESS scaling, therefore the data quality was deemed satisfactory. No clear separation could be observed between the de novo PD and control groups or between males and females. An OPLS-DA model of de novo PD versus control was constructed to assess if any discriminating protein expression could be discerned between the groups but proved non-significant (ANOVA p = 0.59). Due to the skewed sex distribution in the discovery set, an additional OPLS-DA model with males only was constructed, and this model also proved non-significant (ANOVA p = 0.26). The lack of significance in the discriminant models implies that there is not enough covariation in the protein expression in the two groups to separate them from each other.

In conclusion, multivariate analysis demonstrated that no large-scale differences induced by the instrumental analysis could be detected thus indicating that the data quality was satisfactory and that any disease-related changes in protein expression are likely subtle.

5.2.2.3 *Pathway analysis of the significant proteins in the de novo PD study*

Pathway analysis of the discovery proteomics results from the de novo PD study showed several pathways related to inflammation enriched in the newly diagnosed PD patients (*complement system, acute phase response signalling, prothrombin activation pathway* and *agranulocyte adhesion*), when viewing all the data as well as the males separately (Figure 5-4). It also showed enrichment of protein folding pathways (*ER stress pathway* and *unfolded protein response*). Below, brief descriptions of the inflammatory and protein folding pathways are provided.

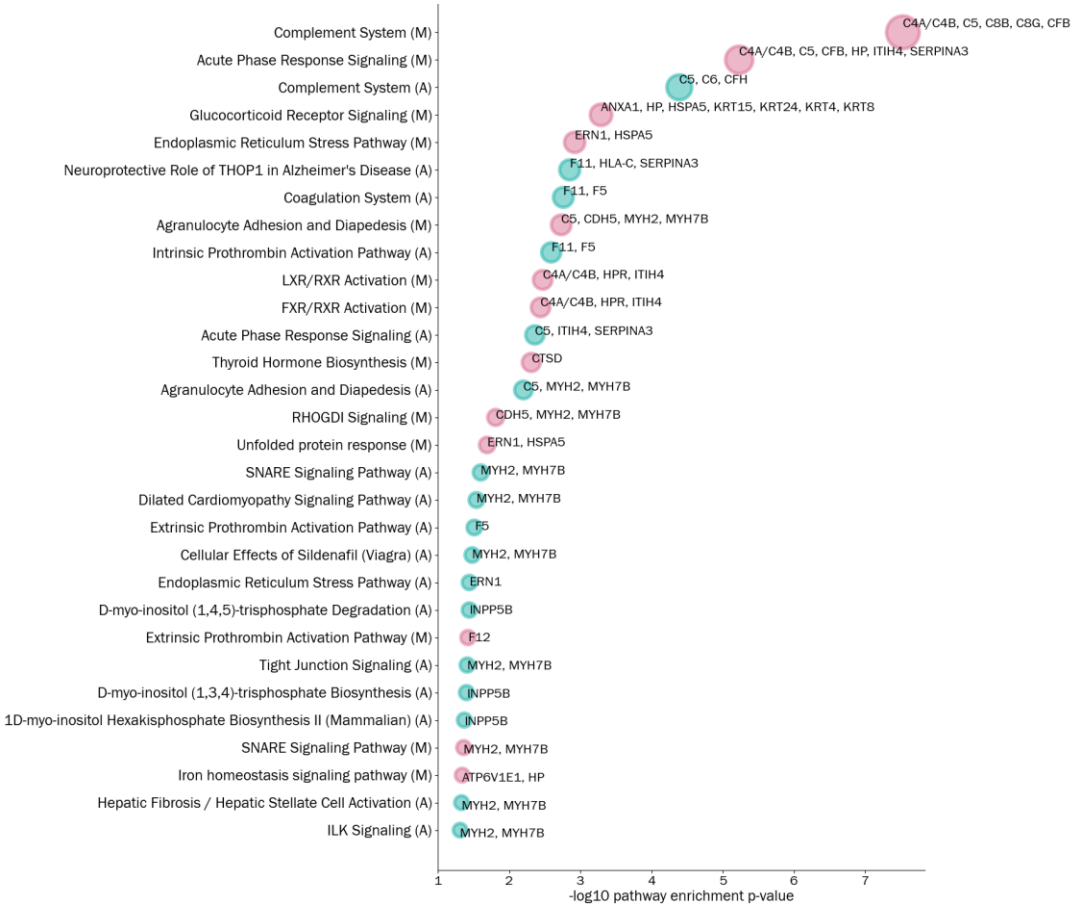


Figure 5-4. Significantly enriched pathways in the de novo PD and control discovery proteomics study of all samples (A) and of males only (M). *The pathways are coloured according to the set they were enriched in, where ■ all data, and ■ males only. Each pathway is annotated by the respective proteins included. Circle radii indicate the significance of the pathway enrichment p-value.*

Inflammation related pathways. The *complement system* demonstrated the highest level of significance in both comparisons. The complement system is part of the innate immune system and activates inflammation. *Acute phase response signalling* refers to a rapid inflammatory response which can be triggered by microorganisms, trauma, or immunological disorders. It results in an increase of pro-inflammatory compounds such as cytokines and can additionally activate the complement system [157]. *Agranulocyte adhesion and diapedesis* describes an inflammatory response to infection or injury and includes the process of certain molecules adhering to the surface of the endothelium [157]. Finally, the *extrinsic prothrombin activation pathway* is part of the coagulation cascade, serving to repair broken vessels via the production of thrombin which is needed for the conversion of fibrinogen to fibrin [157]. The pathway analysis therefore indicates that these inflammation- and clotting-related pathways may be dysregulated in the PD group.

Protein folding pathways. The endoplasmic reticulum manages the synthesis and folding of membrane and secretory proteins, ensuring that only correctly folded proteins exit the ER. The *endoplasmic reticulum stress pathway* describes the response of the cell under stressed ER conditions when handling extensive amounts of mis- or unfolded proteins [157]. *Unfolded protein response* describes a range of ER functions to alleviate stress induced by excessive handling of mis- or unfolded proteins. However, under extreme ER stress, the unfolded protein response can stimulate apoptosis [157]. Therefore, our results indicate that there is a significant change in the correct folding or function of the ER in the PD group.

5.2.3 *Results from the PD discordant twins study*

5.2.3.1 *Univariate analysis*

In the analysis of serum from PD discordant twin pairs, a total of 1141 proteins were detected and identified using the optimised workflow described in Chapter 2. Out of these proteins, 594 had a confidence score > 15 and were represented by two or more unique peptides per protein. The pre-PD and control twin pairs were compared using paired t-tests, which demonstrated that twelve proteins were differentially expressed utilising a nominal significance of 95% as threshold. These proteins were: PRG4, DKK3, C15orf62, SPP2, BCHE, TNNT3, CSF1R, MMP3, PTGDS, CDH1, ITIH2 and NCAM1. Figure 5-5 shows the fold change within each twin pair for the significant proteins.

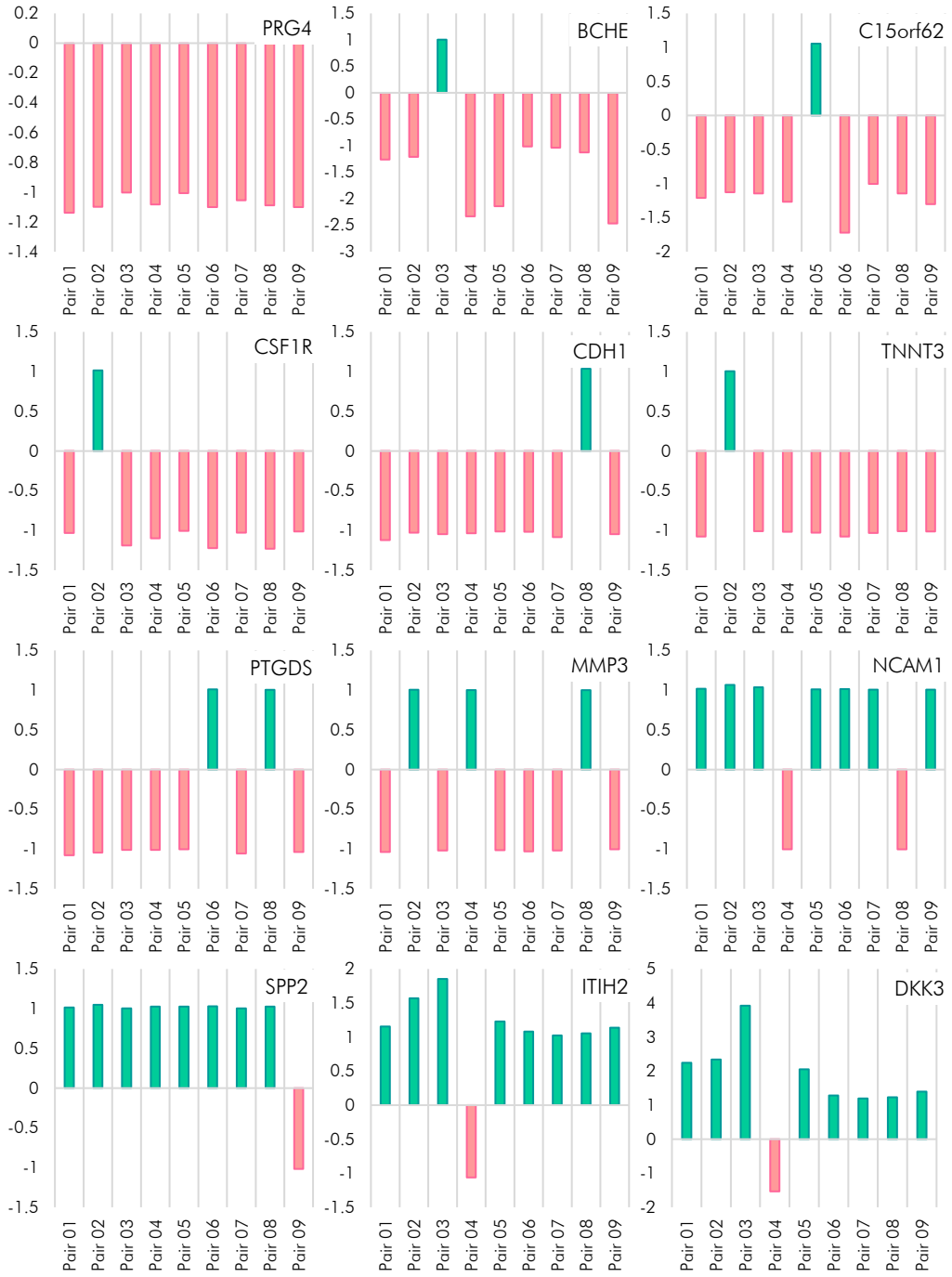


Figure 5-5. Fold changes in each pre-PD/control twin pair for the twelve significantly different proteins. *Positive values denote that the pre-PD twin demonstrated higher levels of a protein while negative values show that the pre-PD twin had lower levels of a protein.*

5.2.3.2 *Correlation between MS measured proteins and clinical data*

A large amount of questionnaire/clinical data had been collected for the twin pairs. The clinical variables consisted of alcohol dosage (ALC_DOSE), hypertension (HT, yes or no), diabetes (yes or no), body mass index (BMI), apolipoprotein A1 (APOA1), haemoglobin (HGB), glucose (GLU), total cholesterol (CHOLTOT), high density lipoprotein (HDL), low density lipoprotein (LDL), triglycerides (TRIG) and C reactive protein (CRP). A correlation matrix (Figure 5-6) was constructed from the provided clinical variables, the significantly different proteins, the PD related proteins PARK7 and BST1, and APOA1 and CRP measured by MS for comparison with clinically measured APOA1 and CRP.

The correlation analysis demonstrated that age was positively correlated with diabetes and the expression of TNNT3 and PTGDS. Males were negatively correlated with higher levels of HDL and the expression of CSF1R and BST1, and positively correlated with higher levels of LDL and the expression of NCAM1. The pre-PD individuals were negatively correlated with the expression of PTGDS and positively correlated with the expression of DKK3 and ITIH2. Alcohol dosage was positively correlated with HGB-levels. Hypertension was positively correlated with BMI. Diabetes was positively correlated with GLU and the expression of SPP2, TNNT3 and NCAM1. BMI was positively correlated with the expression of PRG4 and BCHE. Clinically measured APOA1-levels were positively correlated with HDL and MS-measured APOA1, and negatively correlated with the expression of NCAM1. HGB was negatively correlated with CRP and CSF1R and positively correlated with total cholesterol and LDL. Glucose was positively correlated with SPP2. Total cholesterol was negatively correlated with CSF1R and positively correlated with LDL and triglycerides. HDL was negatively correlated with BCHE. LDL was negatively correlated with CSF1R and positively correlated with triglycerides. Triglycerides were positively correlated with BCHE. Clinical CRP was positively correlated with MS-measured CRP and negatively correlated with SPP2.

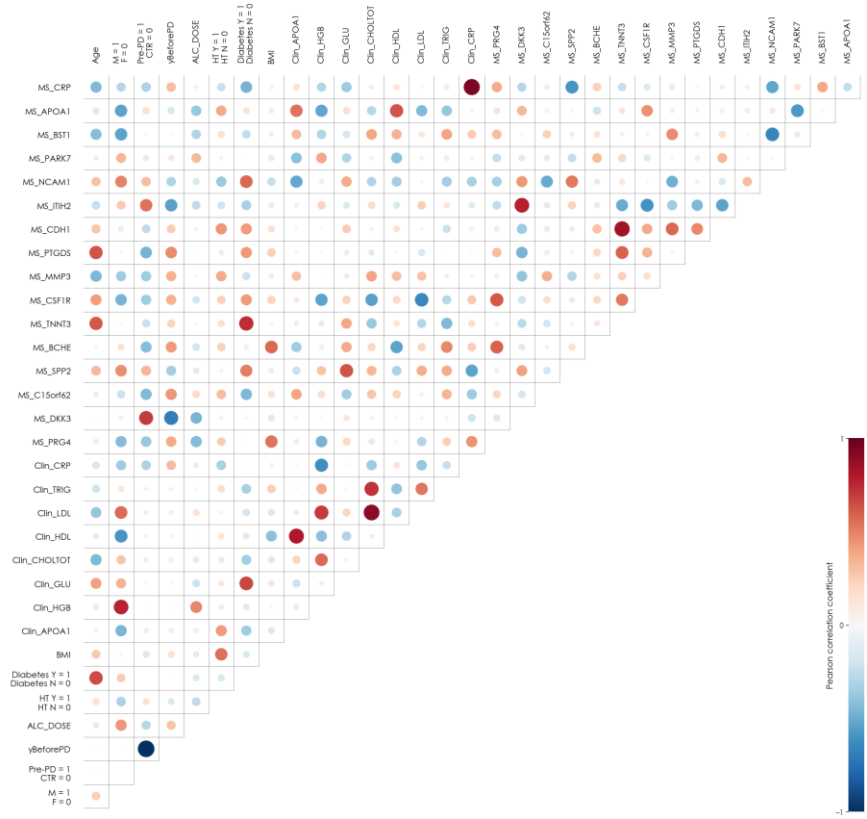


Figure 5-6. Correlation matrix of clinical variables, significantly different proteins, MS-measured APOA1 and CRP, and the PD related proteins PARK7 and BST1. *Blue represents negative correlation and red positive correlation. Circle radii represents the absolute Pearson correlation value. The prefix "Clin" denotes a clinically measured variable, while "MS" denotes a variable measured by mass spectrometry in this study. 1/0 after a variable describes the value given to the discrete class.*

5.2.3.3 *Multivariate analysis*

Principal component analysis was performed for quality control assessment and to investigate if any major patterns were present in the data. No correlation with instrumental drift or run order could be detected, thus indicating that the data was of sufficient quality. The PCA demonstrated that no clear pattern could be discerned between pre-PD and controls. When investigating the distribution of the twin pairs, the principal components PC3 and PC4 clearly demonstrated that the twin pairs clustered closely together (apart from two pairs) as shown in Figure 5-7. This suggests that the protein expression within each twin pair is highly conserved. Pairs 1 and 9 deviate from the clustering pattern and interrogation of the clinical data revealed that the twins in pair 1 were males and pair 9 were females; pair 1 was the oldest twin pair (81 years at sampling) and pair 9 was the youngest (49 years at time of sampling), both twins in pair 1 had diabetes. In both pairs, PD was diagnosed 6 years before sampling, a timescale similar to

the other twin pairs. PD-onset at the age of 43 for the pre-PD twin in pair 9 may indicate a genetic component.

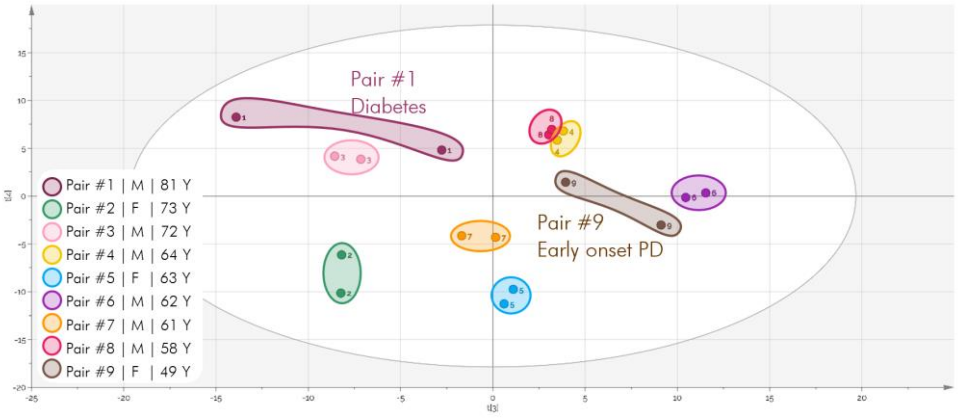


Figure 5-7. Principal component score scatter plot (PC4 vs PC3) of the proteomic results from the PD discordant twin study shows that the twin pairs group together, except for pair 1 and pair 9. *The legend describes if the twin pairs were males or females and their ages.*

The PCA model of the proteomic results also demonstrated that there was a difference between males and females. To investigate this further, a discriminant OPLS-DA model of males versus females was created. The model was significant (ANOVA p = 0.005, permutations p < 0.05) and showed that a number of proteins were influential for the difference between males and females (Figure 5-8).

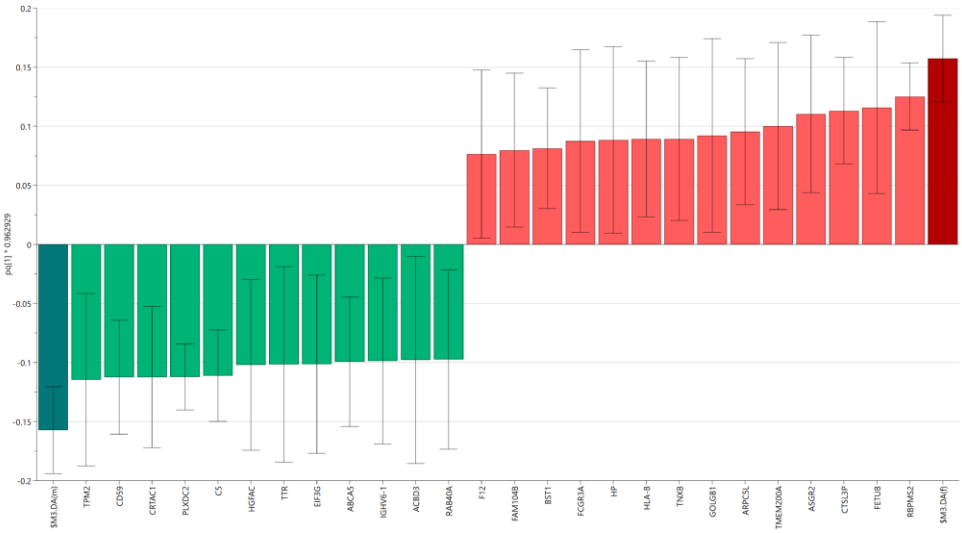


Figure 5-8. OPLS-DA loadings (pq[1]) from a model of males versus females showing the most influential proteins. *The proteins coloured in green were higher in males and the proteins in red were higher in females.*

The influence of age was evaluated in an OPLS model with age as the dependent variable Y. The model was significant (ANOVA p = 0.007, permutations p < 0.05) and demonstrated that the expression of a number of histone proteins were related to older age (Figure 5-9).

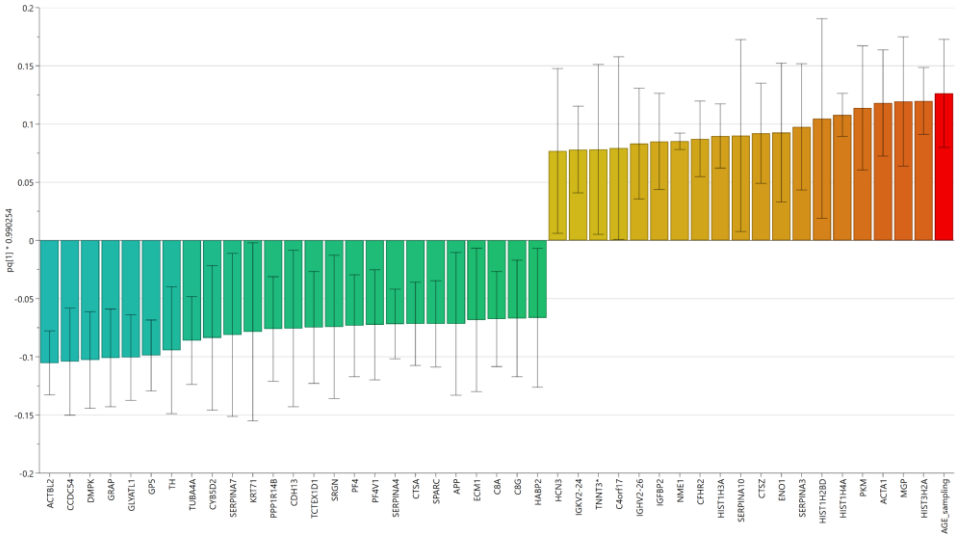


Figure 5-9. OPLS loadings (pq[1]) from a model with age as the dependent variable. *The loadings show the proteins most strongly related to age. The loadings are coloured continuously from blue (strong negative age-correlation) to red (strong positive age-correlation).*

An OPLS-DA model of pre-PD versus controls was created but proved non-significant, likely due to the highly similar protein expression within each pair. To maximise the separation between the twins, the pairs were centred around the average protein value, meaning that the twins had the same absolute value but were positive or negative depending on if the control or pre-PD twin had a higher initial value. An OPLS-DA model was constructed from the centred data but also proved non-significant. Although not enough covariance could be found between the proteins to build a significant model, the model's loadings do show which proteins are more strongly correlated with each group (Figure 5-10).

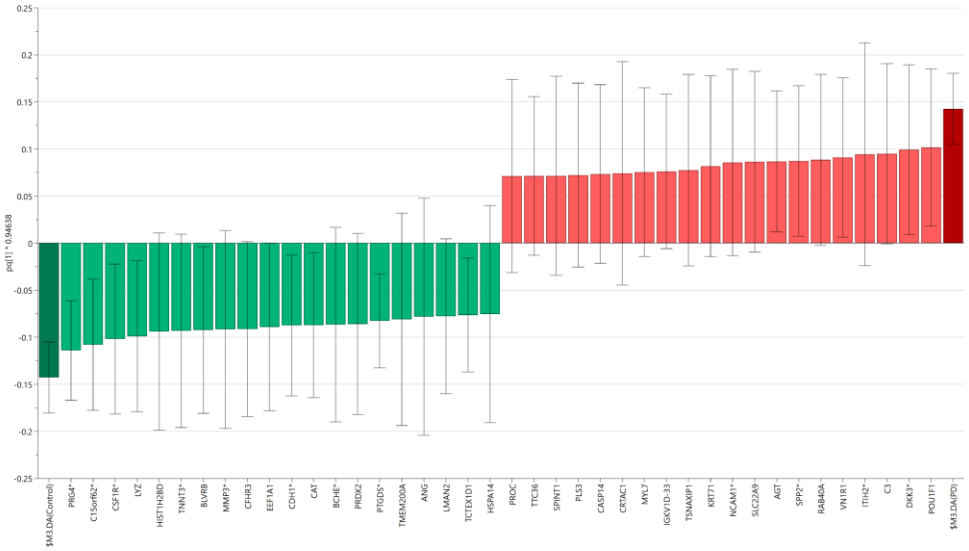


Figure 5-10. OPLS-DA loadings (pq[1]) from centred pre-PD and control pair analysis showing the 20 most influential proteins for each group. *The proteins in green are positively correlated with controls and the proteins in red are positively correlated with pre-PD. The proteins annotated with a * were significant in the univariate paired significance test.*

5.2.3.4 *Pathway analysis of the results from the PD twin study*

The significantly differentially expressed proteins between pre-PD and controls were included in a pathway analysis. Given the small number of significantly different proteins, the interpretability of the pathway analysis was limited and no z-scores indicating activation or deactivation were obtained for any of the pathways. The analysis did however demonstrate that neuroinflammation was one of the significant pathways. Pathways related to Wnt signalling and inflammation were also detected. The significant pathways are shown in Figure 5-11.

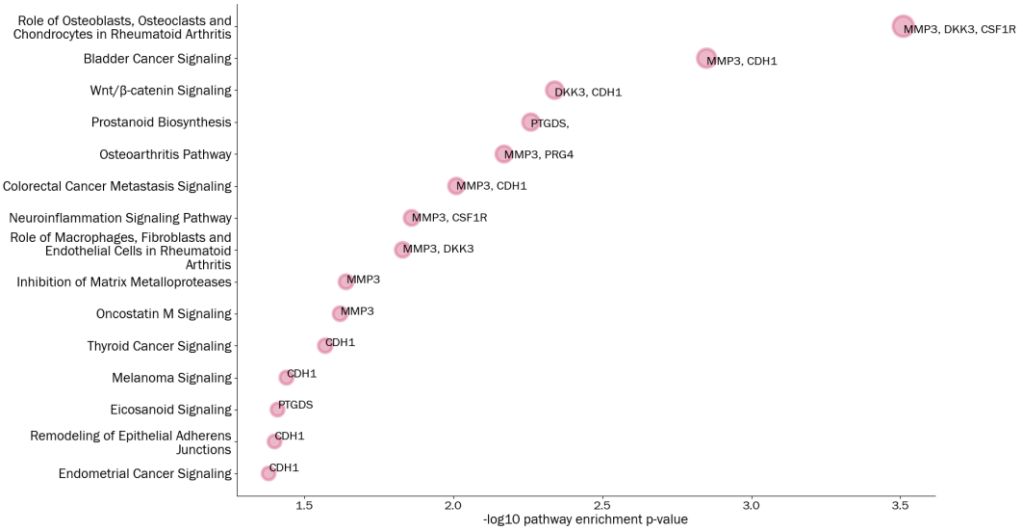


Figure 5-11. Pathway analysis results of the discovery proteomics of pre-PD and control twins. *The pathways are annotated by the respective proteins included in each. Circle radii indicate the significance of the pathway enrichment p-value.*

Neuroinflammation signalling pathway. Two of the significant proteins were part of the neuroinflammation signalling pathway (MMP3 and CSF1R), which had a p-value of 0.01 and indicate inflammation as a risk factor, even prior to PD onset. The neuroinflammation signalling pathway describes a range of inflammatory functions working to maintain the homeostasis of the central nervous system, clearing damaged tissue and targeting threats. However, the protective system can start producing an uncontrolled response leading to the destruction of healthy tissue and a state of chronic inflammation which eventually leads to necrosis of neurons and glial cells [157]. The role of MMP3 in the pathway is putative activation of NF-κB in microglia which can lead to production of proinflammatory proteins. CSF1R can be activated by IL34 and lead to microglial production [241].

Wnt/β-catenin signalling pathway. Two of the significant proteins were identified in the Wnt/ß-catenin signalling pathway (DKK3 and CDH1), with a p-value of 0.005. Wnt

proteins are involved in many different and complex aspects of cellular development, such as cell differentiation and proliferation. The Wnt/ß-catenin signalling pathway involves the low-density lipoprotein receptor-related family as co-receptors of Wnt ligands for the ß-catenin-dependent signal transduction, allowing it to translocate to the nucleus and expedite the transcription of Wnt target genes [157].

Inflammatory pathways. The pathway *Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis* describes the chronically inflammatory condition of the autoimmune disease Rheumatoid Arthritis. There is no clear evidence for a link between Rheumatoid Arthritis and PD and it possible that the pathway essentially signals inflammation. The same may also be observed in the *Osteoarthritis pathway*.

Cancer related pathways. Five cancer related pathways were found significant: bladder cancer, colorectal cancer metastasis, thyroid cancer, melanoma and endometrial cancer signalling. All cancer related pathways include CDH1 and in two of the cases also MMP3. CDH1 is indeed associated with hereditary cancer where a pathogenic variant of the gene is known to increase the risk [242]. CDH1 is also important for epithelial integrity and involved in Wnt signalling where it forms a complex with β-catenin and can modulate the pathway [243]. MMP3 is associated with cancer metastasis and tumour growth [244]. However, MMP3 has also been implicated in neurodegenerative disease through the activation of microglia [245]. Moreover, it is suggested to contribute to dopaminergic neuronal death mediated through oxidative stress [246] and to have the capability of disrupting the blood-brain barrier under certain conditions [247].

In summary, the number of statistically different proteins in the comparison of pre-PD and control twins was relatively low, thus limiting the power and validity of pathway analysis. There are nonetheless interesting proteins differentially expressed between the compared groups, especially DKK3, CSF1R and MMP3 due to their association with neurodegenerative diseases. DKK3, Dickkopf-related protein 3, is a glycoprotein belonging to the Dickkopf family, modulators of the Wnt signalling pathway. DKK3 has been seen downregulated in many cancer studies and was recently proposed to have a neuroprotective role [248]. It has been related to Alzheimer's disease in several studies and furthermore its expression is proposed to positively correlate with increased age [249]. CSF1R, Colony Stimulating Factor Receptor 1, is a cytokine acting as a receptor for CSF1 and IL34, promoting the release of pro-inflammatory chemokines. CSF1R has been associated with Parkinson's disease in gene studies [250]. CSF1R has also been suggested to eliminate microglia, possibly leading to loss of dopaminergic neurons [251].

5.2.4 *Conclusions from the discovery studies of newly diagnosed PD patients and pre-onset PD patients*

Several proteins were detected in the two discovery studies, many with differential expression between de novo PD/pre-PD and control. The results of the screening studies show that a number of interesting proteins in the circulation were up- or down-regulated in the serum/plasma of individuals at risk or patients recently diagnosed with Parkinson's disease.

Throughout the data sets, the presence of proteins associated with inflammation is reoccurring. In the pre-symptomatic PD discordant twins, neuroinflammation was indicated to be upregulated and in the de novo PD dataset the complement system was indicated to be activated. The complement system is part of the body's innate immune system and can be activated through three different pathways, all resulting in the formation of a membrane attack complex (MAC), inserting itself in the cell membranes and causing them to lyse. Studies have showed that inhibition of MAC after traumatic brain injury reduced neuronal damage and microglial activation [252, 253]. MAC is neurotoxic [254] and it is possible that the complement cascade may be involved in initiating the vicious cycle of microglial activation and production of reactive oxygen species suggested to occur in Parkinson's disease. A study from 2018 showed that complement activation is involved in the triggering of neuroinflammation and that complement-mediated neuroinflammation is associated with degeneration observed in traumatic brain injury [255]. Pathway analysis also implicated the involvement of protein misfolding in PD, thus another potential disease mechanism via activation of the unfolded protein response and subsequently neuroinflammation. Unfortunately, our data is not able to distinguish if endoplasmic reticulum stress is the driver of the suggested neuroinflammation observed in the data or vice-versa.

Overall, several promising proteins were identified as potential biomarkers and put forward for development into a targeted assay and validation in a larger set of patient samples, these proteins were: ANXA1, GOLM1, HSPA5, NRP1, UHRF1BP1L, SERPINA3, PRG4, DKK3, C15orf62, SPP2, BCHE, TNNT3, CSF1R, MMP3, PTGDS, ITIH2 and NCAM1.

5.3 TARGETED PLASMA PROTEOMICS TO VALIDATE THE PROTEINS IDENTIFIED IN THE DISCOVERY PHASE AND TO IDENTIFY INFLAMMATORY TARGETS FROM LITERATURE

In the targeted validation phase of this study, the putative biomarkers of Parkinson's disease which we identified in the discovery phase were evaluated in a new and larger set of samples. The aim was to confirm the discovery findings, and to evaluate if several pro- and anti-inflammatory proteins from the literature were differentially expressed. The samples analysed in the targeted validation study consisted of newly diagnosed PD patients, controls, patients with neurological non-PD disorders as a positive control group, and finally patients with rapid eye movement sleep disorder – a condition often preceding the development of Parkinson's disease.

5.3.1 *Materials and methods*

5.3.1.1 *Sample cohort*

The validation study consisted of plasma samples collected and provided by Universitätsmedizin Göttingen-Klinik für Neurologie, Göttingen University, Germany. The samples were from newly diagnosed Parkinson's patients, patients suffering from rapid eye movement sleep disturbance disorder (iRBD), controls and a heterogeneous group of non-Parkinson's disease but other neurological disorders (OND). The control and de novo PD samples were utilised to verify the findings from the discovery study. The other neurological diseases were included to assess the specificity of the peptides. The iRBD samples were evaluated for potential prediction of the iRBD patients who would converge to develop Parkinson's disease. A total of 211 samples were included in the targeted validation study. Table 5-3 shows the characteristics of the validation samples and Figure 5-12 shows a histogram of the age distribution in the groups.

Table 5-3. Characteristics of the samples included in the targeted validation study of Parkinson's disease. *The table describes the number of samples in each group, percentages of males and females, the mean age and mean years of symptom duration prior to diagnosis. SD = standard deviation.*

Group	Number of samples	Age ± SD	Percentages males/females	Symptom duration [Y] before diagnosis ± SD
Control	50	63.9 (± 7.1)	54% M | 46% F	
iRBD	20	66.9 (± 8.6)	55% M | 45% F	5.4 (± 4.1)
De novo PD	101	67.1 (± 10.6)	50% M | 50% F	2.3 (± 3.2)
OND	40	70.3 (± 8.8)	72% M | 28% F	2.2 (± 2.0)

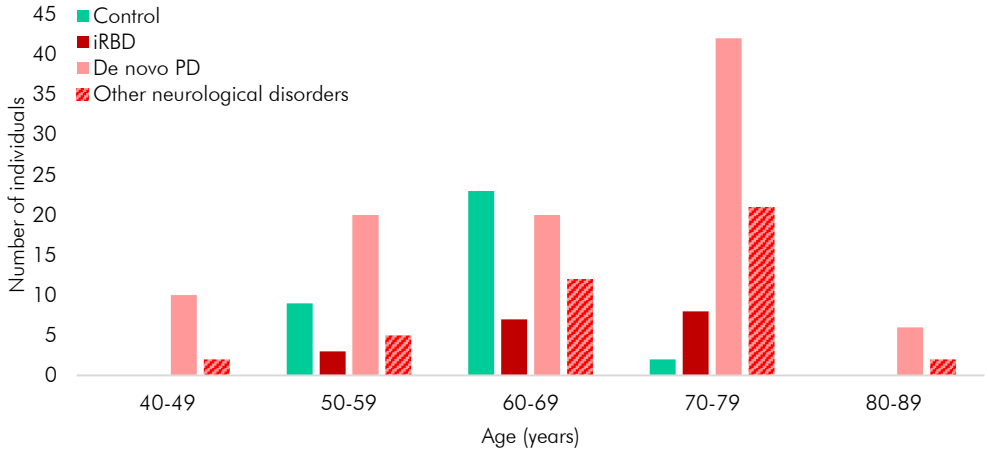


Figure 5-12. Age histogram showing the age distribution in the four different groups included in the targeted Parkinson's disease validation study - de novo PD, iRBD, control and other neurological disorders.

5.3.1.2 *Sample preparation for targeted proteomics*

The samples were prepared for targeted proteomics according to Chapter 2, section 2.7. Briefly, 10 µL plasma was spiked with 150 ng yeast ENO1 (whole protein) and depleted from albumin and IgG using Pierce Top 2 columns. The samples were digested and solid phase extracted. Pooled and acetone precipitated plasma were used as quality control samples. Calibration curves were prepared by spiking increasing amounts of peptide standards into blank and pooled samples.

5.3.1.3 *Instrumental analysis*

The settings and parameters for the instrumental analysis are described in detail in Chapter 2, section 2.9. In short, the samples were reconstituted in 30 µL 3% acetonitrile, 0.1% trifluoroacetic acid, containing 0.1 µM of isotope labelled internal standards and 5 µL was injected onto a UPLC system coupled to a triple quadrupole mass spectrometer. Two injections were made per sample, each with a different MRM method. In total, 189 peptides were monitored.

5.3.1.4 *Peak picking, integration, and data pre-treatment*

After acquisition, peak picking was performed utilising an in-house software written in Python or TargetLynx (Waters, UK). Peptide peaks were identified by blank and matrix calibration curves. The integrated peak areas were exported to Microsoft Excel where first the ratio between quantifier and qualifier peak areas was evaluated to ensure that the correct peaks had been integrated. The digestion efficiency was evaluated by monitoring the presence of yeast ENO1 in the samples, and the samples where yeast ENO1 was not detected were excluded from further analysis (16 samples out of which 12 controls, 2 iRBD,

and 2 PD), leaving a total of 195 samples in the final dataset. After the initial quality assessment, the quantifier area was divided by the area of an internal standard (heavy isotope labelled peptides from the proteins ALDOA or GSTO1) to yield a ratio used for the determination of relative concentrations. Any compound that also showed an intensity signal in the blank samples had the blank signal subtracted from the analyte peak intensity. Pooled plasma quality control samples were additionally evaluated to assess the robustness of the run. Run order drift was corrected by applying a LOWESS filter as per Chapter 3, section 3.5.2. Outliers at more than 10 median absolute deviations were removed and replaced by missing values as described in Chapter 3, section 3.5.3, the outliers are presented in Supplementary table 4.

5.3.2 *Results*

The MRM-based proteomic assay could reliably detect peptides from 32 unique endogenous proteins in the targeted sample set.

5.3.2.1 *Multivariate analysis*

Multivariate analysis was performed with two different objectives. Firstly, unsupervised PCA analysis to assess the data quality and to look for major trends in the data. Secondly, supervised OPLS and OPLS-DA analyses to determine relationships between the protein expression and dependent variables such as age and sex, and to discriminate between the sample groups.

5.3.2.1.1 *Unsupervised Principal Component Analysis*

For an overview of major patterns and groupings in the data, an unsupervised PCA was created. No run order issues could be detected in the LOWESS corrected data, nor any other instrumental or sample preparation bias. The analysis demonstrated that the control and DNP groups differed from each other (Figure 5-13). The iRBD group was situated in the middle of controls and DNP, and the OND group was distributed over the whole space with no evident clustering. The corresponding loadings of PC1 and PC2 demonstrated that the control samples contained higher levels of PPP3CB, DKK3, SELE and GRN, and lower levels of the majority of the other proteins.

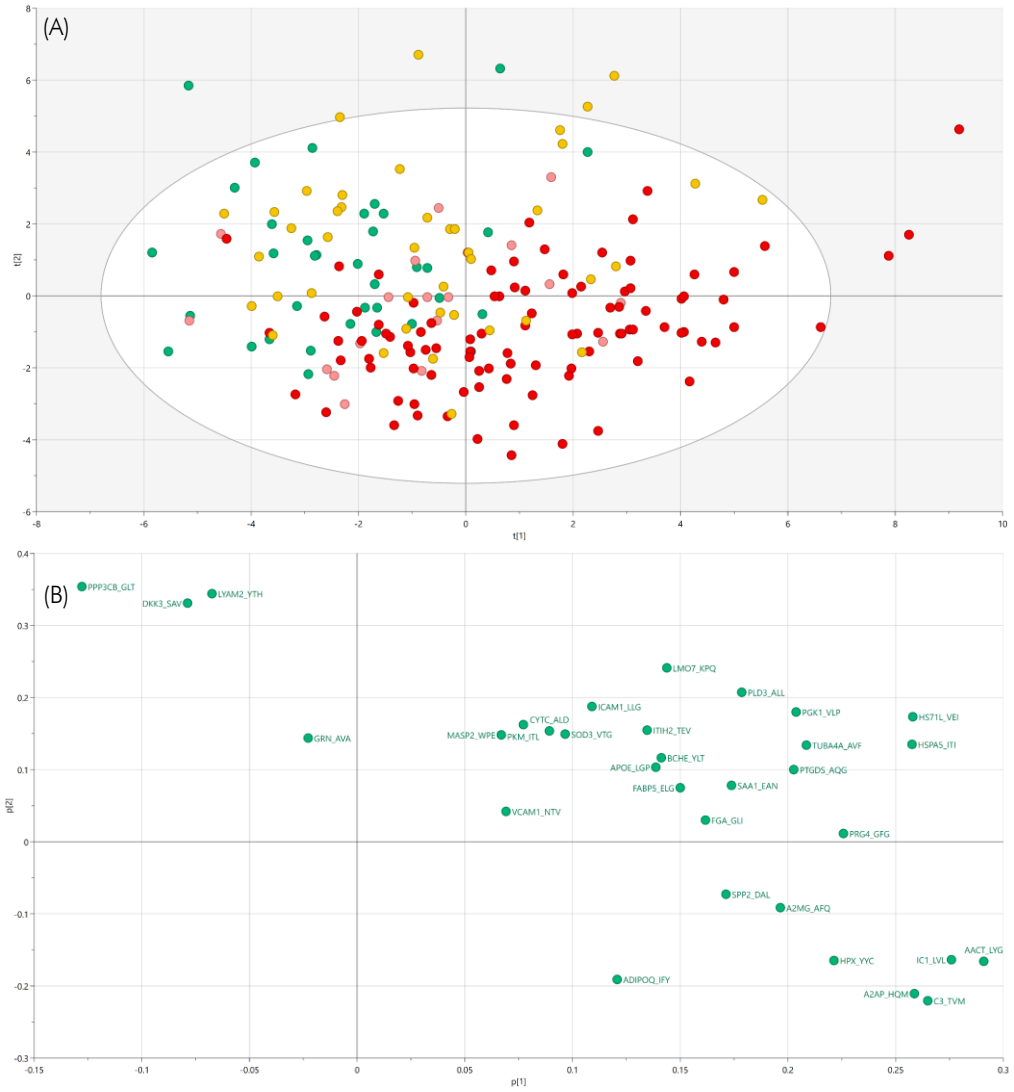


Figure 5-13. Principal component analysis of the outlier corrected targeted proteomics de novo PD data showing all groups. *(A) principal component 1 (t[1]) and principal component 2 (t[2]) show a separation between de novo PD and control. The iRBD samples distribute between de novo PD and control while other neurological disorders mainly cluster with control, indicating that some iRBD patients could be developing the early hallmarks of PD. (B) The corresponding loadings (p[1] and p[2]) demonstrate that de novo PD are correlated with lower levels of PPP3CB, DKK3, SELE and GRN.* ■ *de novo PD,* ■ *control,* ■ *iRBD and* ■ *other neurological disorders*

The OND and iRBD samples were excluded from the model to get a clearer view of the proteins responsible for the difference between de novo PD and controls. The two groups separated into two distinct clusters with DKK3, LYAM2, PPP3CB and GRN being higher in the controls, and IC1, A2AP, HPX, C3, and AACT higher in DNP (Figure 5-14).

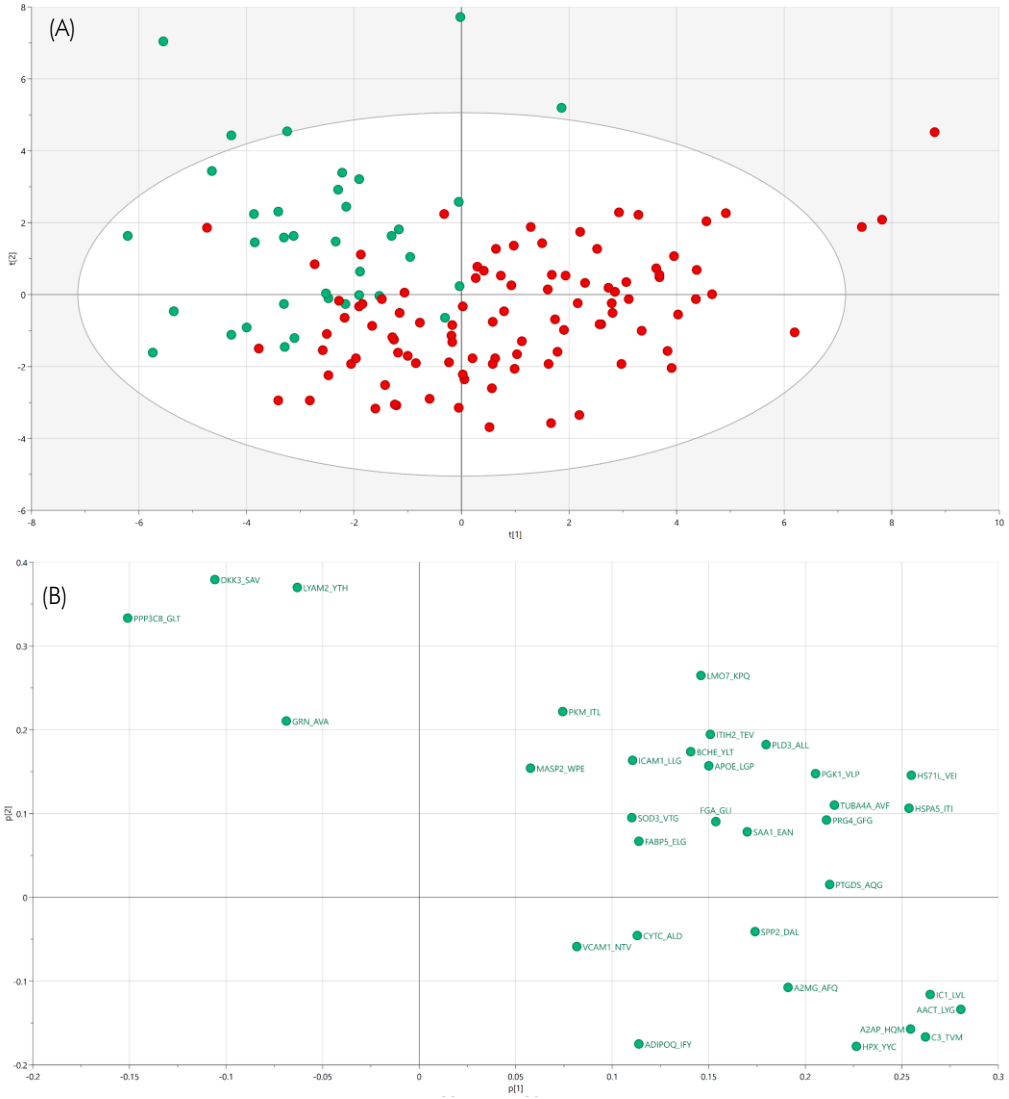


Figure 5-14. Principal component analysis of de novo PD and controls. *Scores (A) and loadings (B) of PC1 and PC2. The scores (t[1] versus t[2]) show a clear separation between the groups and the loadings (p[1] versus p[2]) demonstrate that it is caused by higher DKK3, LYAM2, PPP3CB and GRN in the controls and higher IC1, A2AP, HPX, C3, and AACT in de novo PD.* ■ *de novo PD and* ■ *control*

The PCA model was moreover evaluated for age and sex dependency with the proteins and suggested that there was a non-random protein expression related to both variables.

5.3.2.1.2 *Supervised OPLS and OPLS-DA models for evaluating the confounding effects of age and sex*

Investigating the age-protein expression dependency in detail, an OPLS model of all the samples with age set as the dependent variable Y was created. The model proved highly significant with ANOVA $p = 1.8\ E^{-14}$ and permutations $p << 0.001$. The model demonstrated that higher expression of CST3, PTGDS, VCAM1, A2MG and SOD3 were significantly positively correlated with older age (Figure 5-15).

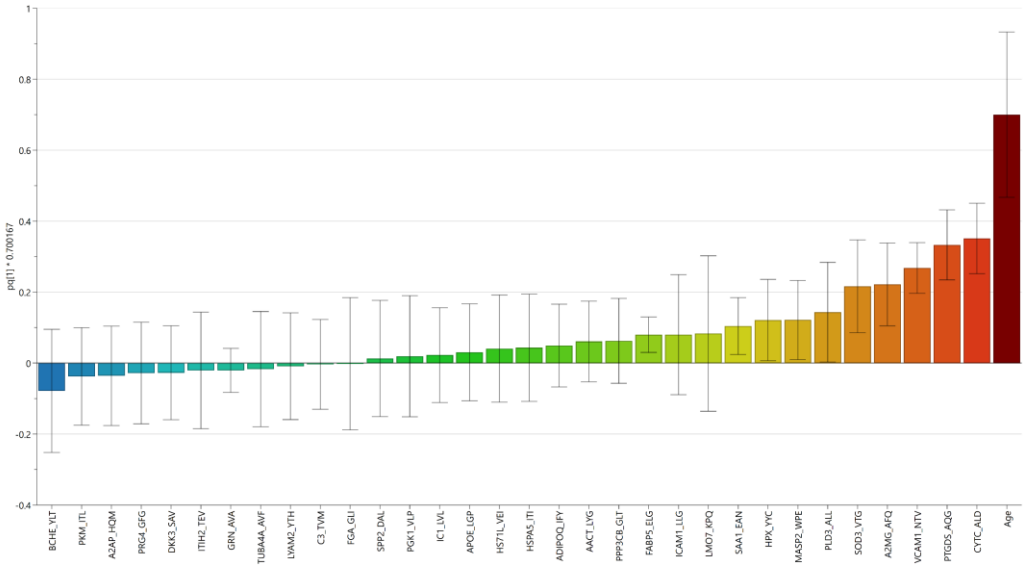


Figure 5-15. Loadings of the predictive component (pq[1]) from the OPLS model of all samples with age set as the dependent variable Y. *The proteins are coloured continuously according to their correlation with age, where blue represents negative correlation and red represents positive correlation. Higher levels of CST3, PTGDS, VCAM1, A2MG and SOD3 are significantly positively correlated with older age.*

The protein expression difference between males and females was further explored in an OPLS-DA model. Also this model was highly significant with ANOVA p = 6 E^{-17} and permutations p << 0.001. The model demonstrated that SPP2, ADIPOQ, SOD3, APOE, ITIH2 and PKM were significantly higher in females compared to males (Figure 5-16).

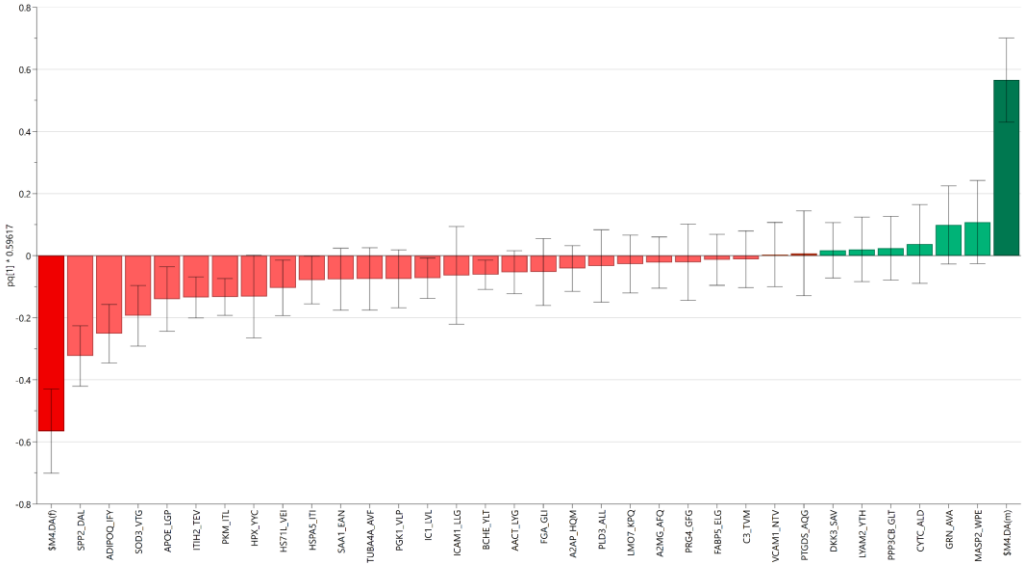


Figure 5-16. Loadings (pq[1]) from the predictive component of the discriminant OPLS-DA model of males versus females in all samples. *The proteins in red are higher in females and the proteins in green are higher in males. SPP2, ADIPOQ, SOD3, APOE, ITIH2 and PKM were significantly higher in females compared to males.*

The analyses demonstrated that both age and gender were significantly correlated with the expression of the measured proteins. As a result, it was determined that these factors

needed to be adjusted for, to ensure they did not have an impact on the discrimination between de novo PD and control, before progressing with further analyses.

Age and sex correction. Utilising multiple linear regression according to the methodology described in Chapter 3, section 3.5.4, the data were modelled with proteins set as dependent variables and sex and age as independent variables. The residuals (the part of the data not related to age and sex) were extracted and utilised in the subsequent models.

After age and sex correction, new supervised models were created. As previously, OPLS was utilised to determine the influence of age and OPLS-DA to determine discrimination between males and females. Both models were non-significant, thereby demonstrating that the correction was successful in removing age and sex dependency in the variables.

5.3.2.1.3 *Supervised OPLS-DA to discriminate between PD and controls*

Satisfied that confounding effects had been corrected for, an OPLS-DA model comparing controls and de novo PD was created from the age and sex corrected data (Figure 5-17). Model validation showed that the model was highly significant with ANOVA-p = $2\,E^{-28}$ and permutations $p << 0.001$. This indicates a high degree of covariation in the expression of the proteins in the two separate groups. The discriminating expression between de novo PD and controls was higher levels of GRN, DKK3, PPP3B and LYAM2 in controls, and higher levels of HPX, A2AP, C3, AACT, IC1, SPP2 and HSPA5 in de novo PD.

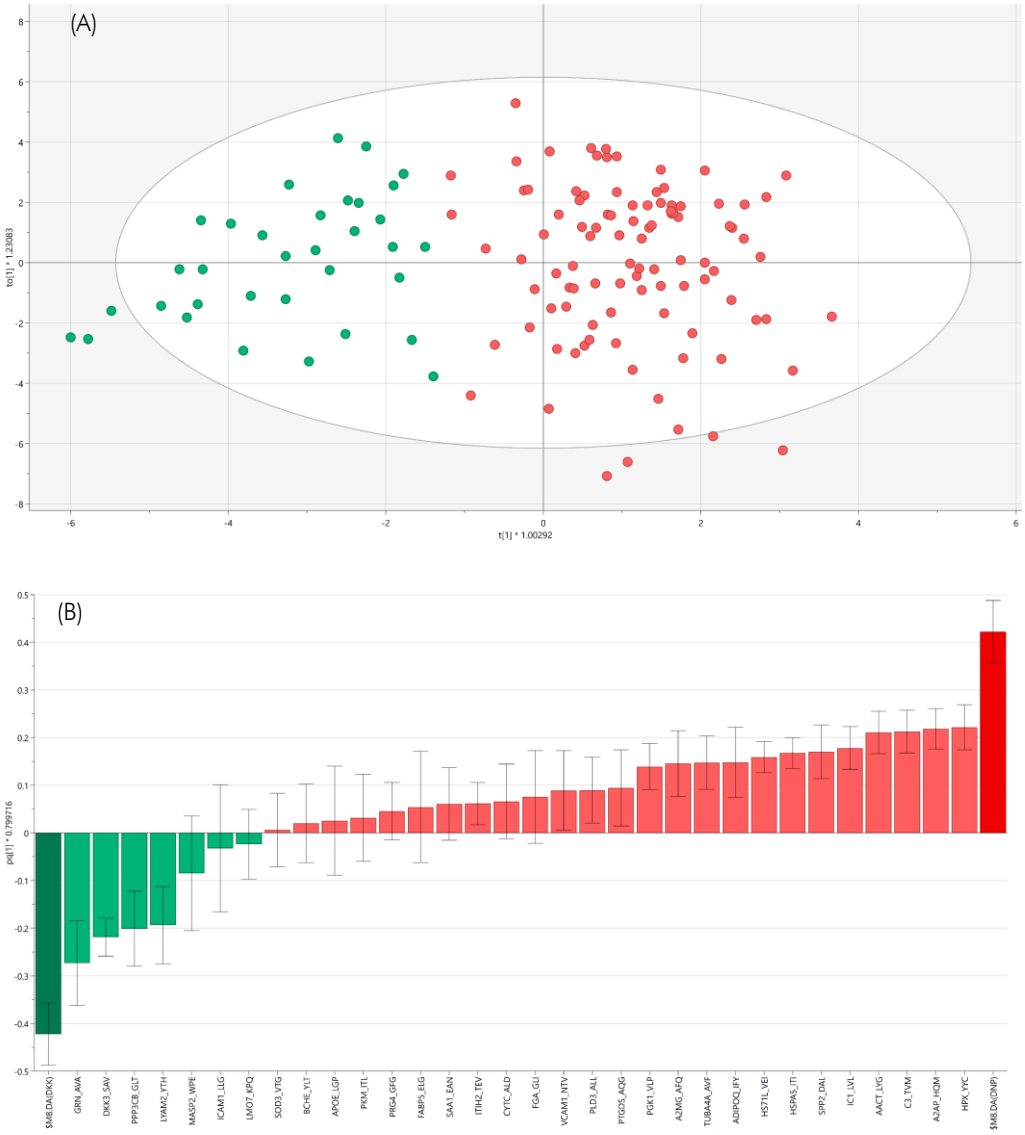


Figure 5-17. (A) Scores, t[1] versus to[1], and (B) loadings, pq[1], from the discriminant OPLS-DA model of de novo PD versus control after age and sex correction. *The model could separate the groups on a significant level. The proteins in the loading plot coloured in green are higher in the control group, while the proteins coloured in red are higher in the de novo PD group. Higher levels of GRN, DKK3, PPP3B and LYAM2 are found in controls, and higher levels of HPX, A2AP, C3, AACT, IC1, SPP2 and HSPA5 in de novo PD.* ■ *de novo PD and* ■ *control.*

The age and sex corrected OPLS-DA model of de novo PD versus control was compared to an identical model based on the non-corrected data in a SUS plot (Figure 5-18). The SUS plot demonstrated that the relationship between the two models was linear and that they were indeed noticeably similar. This signifies that the protein expression related to discrimination between de novo PD and control is largely unaffected by age and sex.

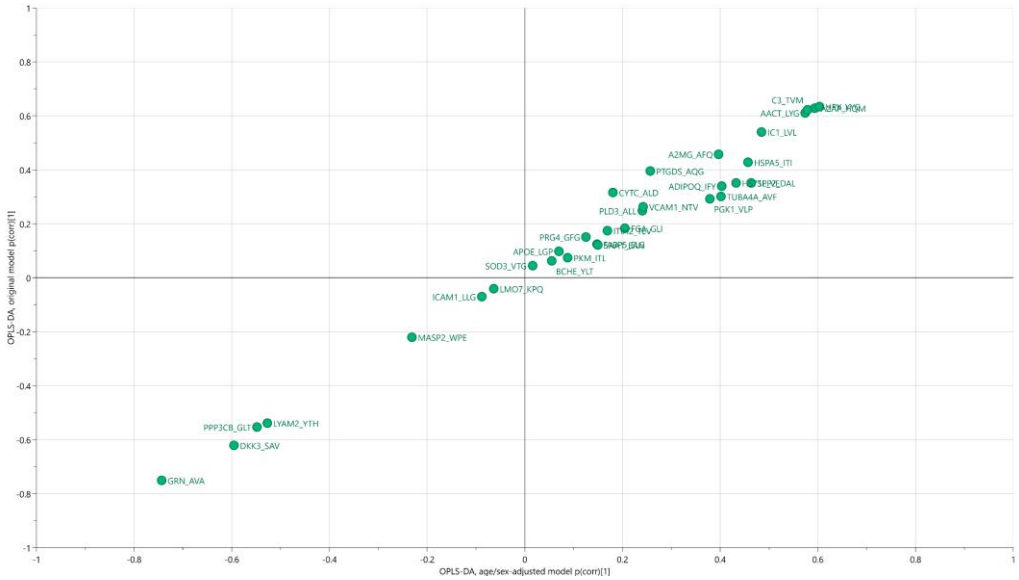


Figure 5-18. Shared and unique structures plot of the age and sex corrected OPLS-DA model versus the original OPLS-DA model of de novo PD versus control. *p(corr)1 represents the scaled predictive loadings. No major difference can be seen between the two models, signifying that the protein expression discriminating between de novo PD and control largely is unaffected by age and sex.*

5.3.2.1.4 *Prediction of iRBD and OND samples in the OPLS-DA model of de novo PD and controls*

To evaluate the specificity of the model and if the protein expression from the targeted experiment would allow for prediction of the iRBD samples more likely to develop PD, the OND and iRBD samples were predicted in the OPLS-DA model of de novo PD and controls (Figure 5-19). The OND prediction resulted in 52% of the samples classified as de novo PD and 48% as control, thus suggesting that the OPLS-DA model has a level of specificity for Parkinson's disease. The iRBD prediction resulted in 94% of the samples classified as de novo PD, thereby indicating that the variance of the most influential proteins in the OPLS-DA model is largely shared between the de novo PD and iRBD patients.

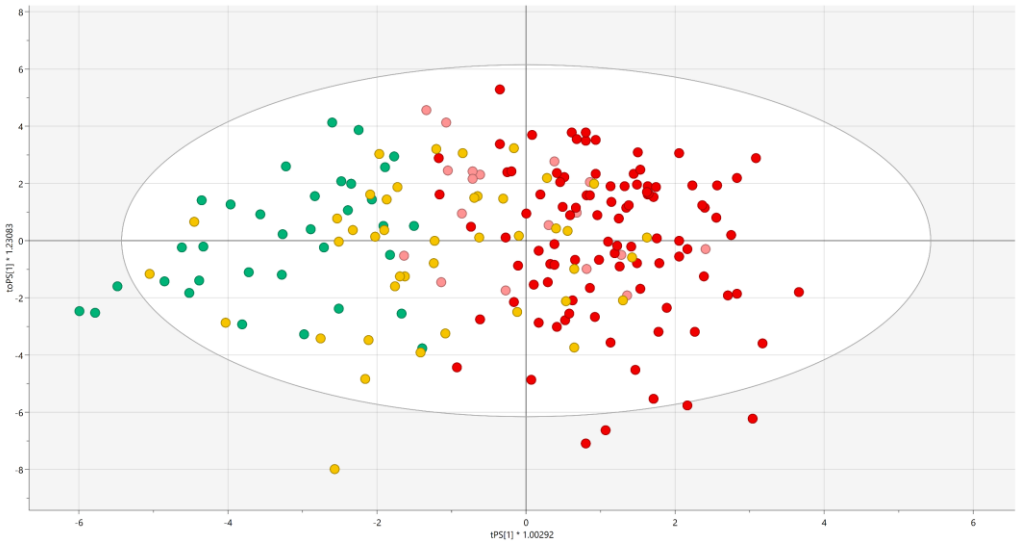


Figure 5-19. Prediction of OND and iRBD in the OPLS-DA model of de novo PD versus control. *The classification of the OND samples resulted in 52% predicted ad de novo PD and 48% as control. The iRBD classification resulted in 94% predicted as de novo PD.* ■ *de novo PD,* ■ *control,* ■ *iRBD (predicted) and* ■ *other neurological disorders (predicted). tPS[1] denotes the predictive component of the prediction, and toPS[1] the first orthogonal principal component of the prediction.*

5.3.2.1.5 *Conclusions from the multivariate analysis*

Multivariate analysis demonstrated that the protein expression was strongly influenced by age and sex. After adjusting for these covariates, the data were modelled by OPLS-DA, discriminating between de novo PD and controls, and produced a highly significant model that demonstrated that the de novo PD patients differed from controls by expressing lower levels of GRN, DKK3, PPP3CB and LYAM2, and higher levels of HPX, A2AP, C3, AACT, IC1, SPP2 and HSPA5. Prediction of the iRBD and OND groups in the PD-control model resulted in 94% of the iRBD samples and 52% of the OND samples predicted as PD.

5.3.2.2 *Univariate analysis*

For both the age and sex corrected data and the non-corrected data, the groups de novo PD, iRBD and other neurological disorders were compared to control using Student's t-test. Applying Benjamini-Hochberg FDR multiple testing correction with alpha = 0.05, 19 proteins were statistically different when comparing de novo PD and control in the age and sex corrected data and 21 in the non-corrected data. In the iRBD versus control comparison, five proteins were different in both the non-corrected and the age and sex corrected data. Comparing OND versus control, seven proteins were differentially expressed in the age and sex corrected data and eight in the non-corrected data. The FDR-adjusted p-values for the different comparisons are presented in Table 5-4.

Table 5-4. Summary of FDR-adjusted p-values from the comparison of de novo PD (DNP), iRBD and other neurological disorders versus control. *The table show results from age and sex corrected and non-corrected data, where* ****$p < 0.0001$, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$ *and NS* $p > 0.05$

	Age and sex corrected data			Non-corrected data		
	FDR adjusted p-value			FDR adjusted p-value		
	Control vs DNP	Control vs iRBD	Control vs OND	Control vs DNP	Control vs iRBD	Control vs OND
GRN	****	***	**	****	***	**
DKK3	****	NS	NS	****	NS	NS
SERPINF2	****	*	NS	****	*	NS
C3	****	*	NS	****	*	NS
HPX	****	NS	NS	****	NS	*
PPP3CB	****	NS	NS	****	NS	NS
SERPINA3	****	*	NS	****	*	NS
SERPING1	****	NS	NS	****	NS	NS
E-selectin (SELE)	****	NS	NS	****	NS	NS
HSPA5	****	NS	*	****	NS	*
A2M	****	NS	NS	****	NS	**
SPP2	****	NS	NS	***	NS	NS
HSPA1L	***	NS	*	***	NS	NS
ADIPOQ	***	NS	NS	***	NS	NS
TUBA4A	***	NS	*	**	NS	NS
PGK1	**	NS	*	**	NS	NS
PTGDS	**	NS	NS	***	NS	**
MASP2	*	****	NS	*	****	NS
PLD3	*	NS	*	*	NS	**
FGA	NS	NS	NS	NS	NS	NS
CST3	NS	NS	**	**	NS	****
VCAM1	NS	NS	NS	*	NS	**
PRG4	NS	NS	NS	NS	NS	NS
ITIH2	NS	NS	NS	NS	NS	NS
SAA1	NS	NS	NS	NS	NS	NS
FABP5	NS	NS	NS	NS	NS	NS
APOE	NS	NS	NS	NS	NS	NS
ICAM1	NS	NS	NS	NS	NS	NS
LMO7	NS	NS	NS	NS	NS	NS
BCHE	NS	NS	NS	NS	NS	NS
PKM	NS	NS	NS	NS	NS	NS
SOD3	NS	NS	NS	NS	NS	NS

The age and sex corrected data were used for all further analyses and plots. Figure 5-20 shows scatter plots of the 19 significantly different proteins comparing de novo PD patients and controls.

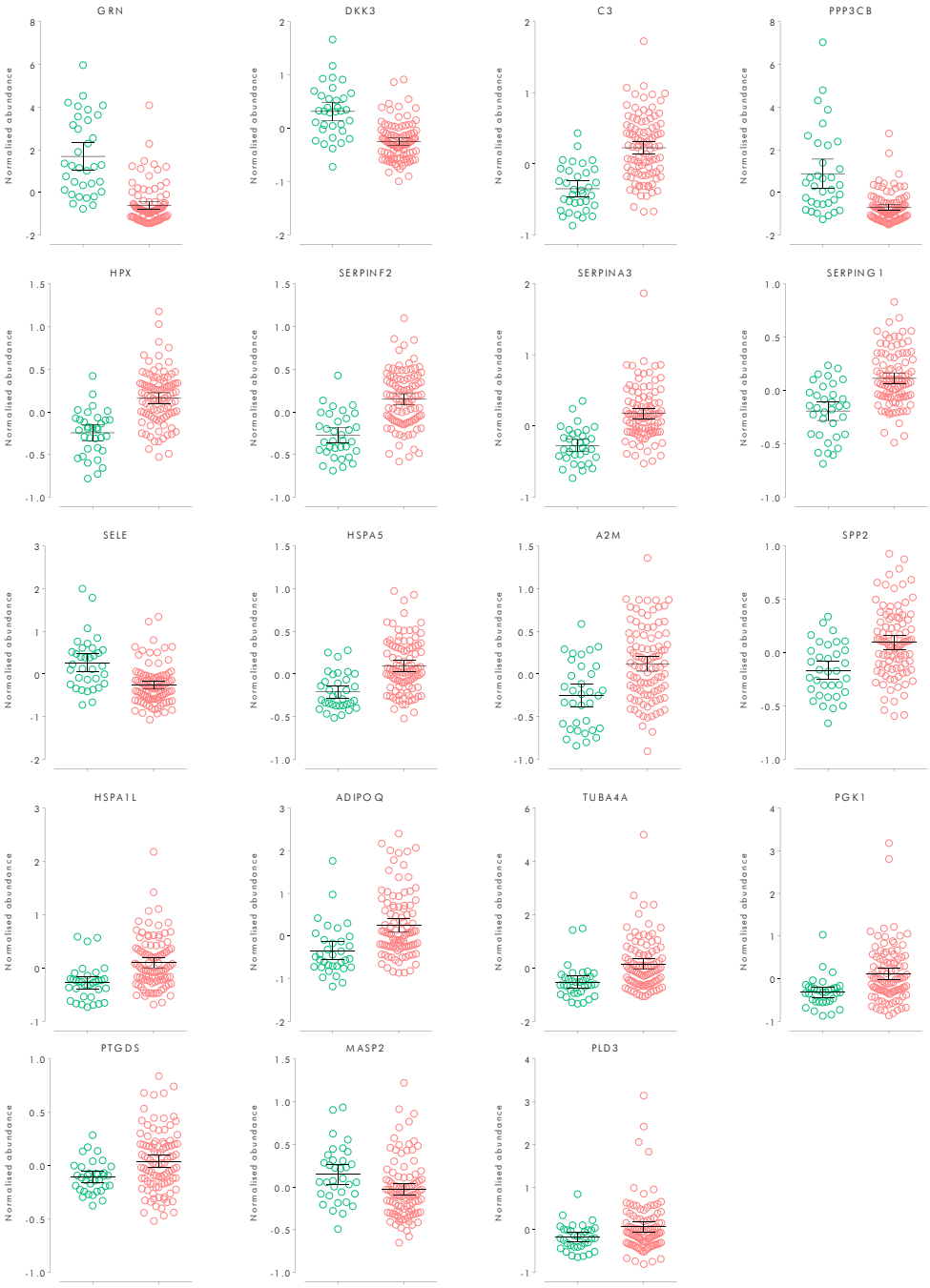


Figure 5-20. Significantly differentially expressed proteins in the comparison between de novo PD patients and control after FDR correction. *The proteins are expressed as normalised abundance. The error bars represent the 95% confidence interval. In total, 19 proteins demonstrated a significant difference between the two groups. The sample groups are represented by ■ controls, and ■ de novo PD.*

The five proteins significantly different between iRBD and control post p-value FDR-correction are shown as scatter plots in Figure 5-21. All the proteins differentially expressed in the iRBD patients were also differentially expressed in the de novo PD patients.

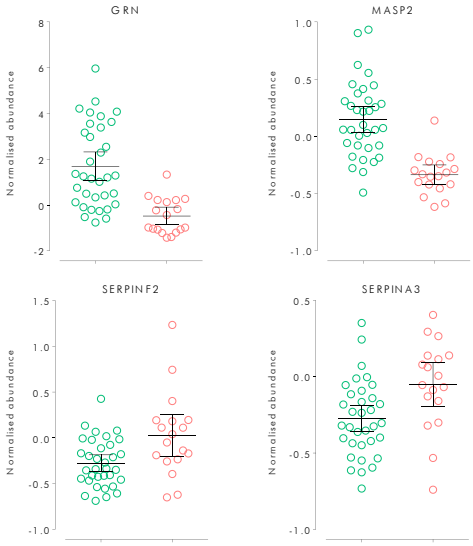


Figure 5-21. Significantly differentially expressed proteins in the comparison between iRBD and control after FDR correction. *The proteins are expressed as normalised abundance. The error bars represent the 95% confidence interval. In total, five proteins demonstrated a significant difference between the two groups. The sample groups are represented by ■ controls, and ■ iRBD.*

The seven proteins significantly differentially expressed between the OND group and control after p-value FDR-adjustment are shown in Figure 5-22. The OND proteins were also differentially expressed in the de novo PD group apart from CST3, which was uniquely found elevated in the OND group.

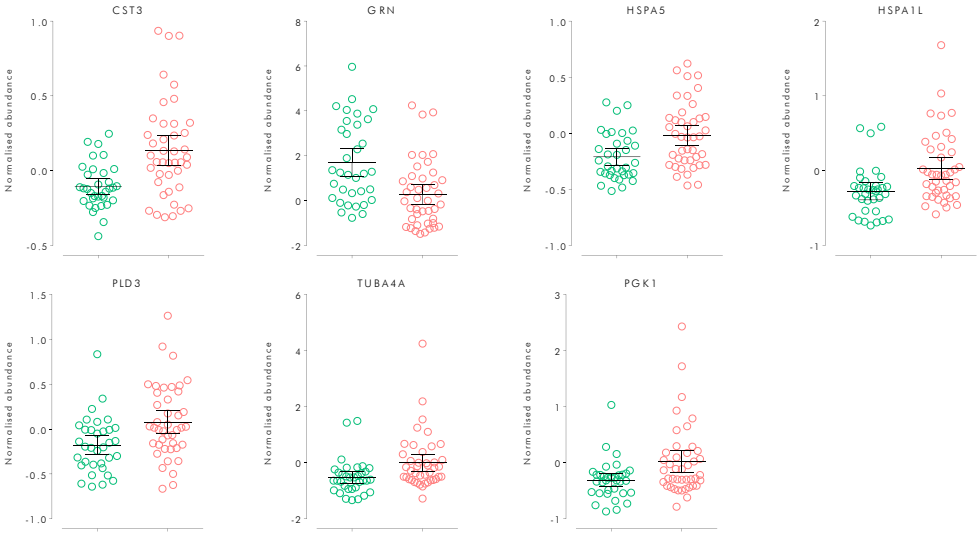


Figure 5-22. Significantly differentially expressed proteins in the comparison between other neurological disorders and control samples after FDR correction. *The proteins are expressed as normalised abundance. The error bars represent the 95% confidence interval. In total, seven proteins demonstrated a significant difference between the two groups. The sample groups are represented by ■ controls, and ■ other neurological disorders.*

5.3.2.3 *Roles of the proteins showing differential expression in de novo PD*

Many biomarker discovery experiments, functional assays and mouse models have been studied with the aim of understanding Parkinson's disease pathology and to find

predictive and diagnostic markers, therefore the scientific literature on the subject is plentiful. Any previously established link with Parkinson's disease and/or neurodegeneration was investigated by reviewing literature. The significantly different proteins from our study were cross reviewed against other studies of Parkinson's disease and neurodegenerative diseases using PubMed and Google Scholar and the findings are presented in Table 5-5.

Table 5-5. Description and reported PD links of the significantly different proteins between de novo PD patients and healthy controls.

Protein	Description and reported links to Parkinson's disease
Progranulin (GRN)	GRN was strongly downregulated in the de novo PD patients ($p = 1.2E^{-15}$). It was moreover downregulated in the iRBD patients ($p = 2.0E^{-4}$) and in the OND group ($p = 5.8E^{-3}$). Progranulin is the precursor of several granulins. It is suggested to act as a neurotrophic factor, promoting neuronal survival. It is also suggested to modulate lysosomal function together with granulin although the exact mechanism is not yet fully understood. Pathogenic mutations in the *GRN* gene are known to cause frontotemporal dementia [256]. Loss of progranulin has been linked to increased production of pro-inflammatory species such as TNF and IL6 in microglia. Moreover, a mouse study showed that *Grn*$^{-/-}$ mice had elevated levels of complement proteins, including C3, before neurodegeneration onset [257]. A study of GRN in neurodegenerative disease found it significantly downregulated in PD patients [258].
Dickkopf 3 (DKK3)	DKK3 was strongly downregulated in the de novo PD patients ($p = 5.5E^{-10}$). DKK3 is a glycoprotein belonging to the Dickkopf family, the majority of which are antagonists of the Wnt signalling pathway, although DKK3 is a modulator rather than an antagonist. DKK3 has been seen downregulated in many cancer studies and was recently proposed to have a neuroprotective role [248]. It has been related to Alzheimer's disease in several studies and furthermore proposed to positively correlate with increased age [249]. Interestingly, a mouse study found that DKK3 may protect dopaminergic neurons and proposed that DKK3 has potential as a pharmacological target for treatment of neurodegeneration [259]. Another study of mice and pluripotent stem cells showed that DKK3 is necessary for correct differentiation and survival of dopaminergic neurons [260].
Protein Phosphatase 3 Catalytic Subunit Beta (PPP3CB)	PPP3CB, also known as calcineurin A2, was strongly downregulated in the de novo PD patients ($p = 1.7E^{-9}$). It makes part of the calcineurin complex and is also component of the Wnt/Ca^{2+} pathway [261, 262]. PPP3CB was identified as a risk gene for AD in microarray studies in the early 2000s [263]. Calcineurin has been proposed to increase in response to accumulation of alpha-synuclein and to trigger both protective and toxic responses to maintain neuronal Ca^{2+} homeostasis [264].
Complement 3 (C3)	C3 was upregulated in the de novo PD patients ($p = 1.7E^{-9}$) and in the iRBD patients ($p = 3.1E^{-2}$). C3 is a central protein in the complement cascade, it is formed regardless of the initiating pathway branch. Complement activation has been linked to neurodegeneration in several studies.
Mannan binding serine peptidase 2 (MASP2)	MASP2 was downregulated in the de novo PD patients ($p = 1.9E^{-2}$) and in the iRBD patients ($p = 1.4E^{-5}$). MASP2 is an initiator of the lectin part of the complement cascade where it recognises certain sugar moieties on pathogens. MASP2 is moreover involved in the coagulation cascade, able to cleave prothrombin to thrombin.
Alpha-1-antichymotrypsin (SERPINA3)	SERPINA3 was upregulated in the de novo PD patients ($p = 4.5E^{-9}$) and in the iRBD patients ($p = 3.3E^{-2}$). SERPINA3's major target is cathepsin G although it can also inhibit other serine proteases. SERPINA3 has been associated with AD and proposed to mediate amyloid-beta clearance [265]. SERPINA3 was also found upregulated in studies of prion diseases and progressive MS [266, 267].
Alpha-2-antiplasmin (SERPINF2)	SERPINF2 was upregulated in the de novo PD patients ($p = 1.7E^{-9}$) and in the iRBD patients ($3.1E^{-2}$). SERPINF2 is a major regulator of the clotting pathway where it acts as an inhibitor of plasmin. SERPINF2 was upregulated in a recent proteomics study of platelet activation in AD patients [268]. Moreover, plasmin has been reported to cleave and degrade extracellular and aggregated alpha-synuclein [269].
Plasma protease C1 inhibitor (SERPING1)	SERPING1 was upregulated in the de novo PD patients ($p = 7.6E^{-8}$). SERPING1 is a modulator of the complement cascade where it inhibits C1r and C1s in the classical pathway and MASP1 and MASP2 in the lectin pathway. A recent study of medically PD induced mice demonstrated that there was an association between increased levels of SERPING1 and dopaminergic cell death in the substantia nigra [270].

Protein	Description and reported links to Parkinson's disease
Hemopexin (HPX)	HPX was upregulated in the de novo PD patients ($p = 1.7E^{-9}$). HPX is a glycoprotein with extraordinary binding affinity for heme, transporting it from the plasma to the liver [271] preventing the oxidative reactions of heme. Hemopexin is traditionally known as an anti-inflammatory protein, however, a study from 2017 described that nitration can occur on the second tyrosine (Tyr199) of the peptide YYCFQGNQFLR, thereby reducing its heme-binding capability [272]. This is the same peptide sequence as was used to identify Hemopexin in this experiment and it could be theorised that this is related to the observation of elevated levels of Hemopexin in the disease groups. Additionally, a recent study found that the neurotoxicity of haemoglobin increased with increased levels of hemopexin in absence of haptoglobin [273].
Alpha-2-macroglobulin (A2M).	A2M was upregulated in the de novo PD patients ($p = 5.7E^{-5}$). A2M is a protease inhibitor and a cytokine transporter. It is associated with AD where it contributes to the degradation of A-beta. A transcriptomics study found the gene significantly upregulated in PD patients [274].
E-selectin (SELE)	SELE was downregulated in the de novo PD patients ($p = 1.7E^{-6}$). SELE is an endothelial cell adhesion molecule and plays a role in the interaction between leukocytes and the endothelium. E-selectin is expressed by cytokines on inflamed endothelium surfaces [275]. It plays an important role in the recruitment of compounds to sites of inflammation [276].
Endoplasmic reticulum chaperone BiP (HSPA5)	HSPA5 was upregulated in the de novo PD patients ($p = 2.1E^{-6}$) and in the OND group ($p = 1.8E^{-2}$). HSPA5 a key regulator in protein folding and the degradation of misfolded proteins [277]. Several neurodegenerative diseases exhibit protein misfolding as part of the pathology, PD among them. When proteins fold incorrectly and/or aggregate, HSPA5 activates the unfolded protein response to alleviate endoplasmic reticulum stress [278].
Heat shock 70 kDa protein 1-like (HSPA1L)	HSPA1L was upregulated in the de novo patients ($p = 1.2E^{-4}$) and in the OND group ($p = 1.8E^{-2}$). It is a heat shock protein involved in the quality control system of the cell. Among its functions are folding and transport of newly synthesised polypeptides and re-folding or destruction of misfolded proteins [279, 280]. In a publication from 2005, overexpression of HSPA1L was suggested to reduce neurodegenerative symptoms in Parkinson's disease, Huntington's chorea and spinocerebellar ataxia [281].
Secreted phosphoprotein 2 (SPP2)	SPP2 was upregulated in the de novo PD patients ($p = 8.8E^{-5}$). SPP2 is associated with Gerstmann-Straussler disease, a rare inherited prion disease [282]. A study of kidney function associated SPP2 with biomarkers of bone and mineral disease, and also found an inverse relationship with Wnt antagonists [283].
Adiponectin (ADIPOQ)	ADIPOQ was upregulated in the de novo PD patients ($p = 1.5E^{-4}$). ADIPOQ has been linked to longevity in several studies. Higher levels of ADIPOQ are beneficiary as it is has anti-inflammatory, anti-oxidising, anti-diabetic and anti-apoptopic properties [203] and protects against age-related diseases [170]. Calorie restriction increases ADIPOQ and ADIPOQ signalling further modulates the downstream AMPK and PPARα pathways [204].
Phosphoglycerate kinase 1 (PGK1)	PGK1 was upregulated in the de novo PD patients ($p = 1.2E^{-3}$) and in the OND group ($p = 3.7E^{-2}$). PGK1 is an enzyme involved in the first ATP producing step of the glycolytic pathway [229]. Intriguingly, PGK1 was recently proposed as a drug target to slow down the progression of Parkinson's disease via the repurposing of the drug Terazosin which traditionally has been used to treat enlarged prostates. The drug is suggested to attenuate PGK1 activity and increase glycolysis thereby increasing oxidative phosphorylation, mitochondrial activity and ATP production and improving the parkinsonian phenotype [284].
Tubulin alpha 4A chain (TUBA4A)	TUBA4A was upregulated in the de novo PD patients ($p = 5.3E^{-4}$) and in the OND group ($p = 3.7E^{-2}$). TUBA4A is a major component of microtubules - core components in the cytoskeleton of the cell. A number of posttranslational modifications can modulate the tubulins' properties and PTMs have been reported to be especially abundant in the microtubules of neurons [285]. Mutations in the *TUBA4A* gene have been linked to amyotrophic lateral sclerosis and frontotemporal dementia. A case study from 2021 found that there may be a link between a *TUBA4A* mutant and selective neuronal loss in the substantia nigra [286].

5.3.2.4 *Pathway and enrichment analysis*

Given the targeted nature of the analysis and the limited number of proteins, the possible output from a pathway analysis is limited. Still, it can provide valuable information about

enrichment of biological processes and functions, and about the proteins' interactions with each other.

The significantly differentially expressed proteins from the comparison of de novo PD and control were analysed using DAVID Bioinformatics Resources 6.8 [287, 288]. The analysis showed that the KEGG pathway *Complement and coagulation cascades* was enriched (FDR p = 3.8E-4), with the proteins C3, SERPINF2, SERPING1, MASP2 and A2M included. Gene ontology analysis suggested that the biological processes *Platelet degranulation* (FDR p = 1.5E-5) and *Negative regulation of endopeptidase activity* (FDR p = 1.7E-7) were enriched. The molecular functions *Endopeptidase inhibitor activity* (FDR p = 4.8E-4), *Serine-type endopeptidase inhibitor activity* (FDR p = 0.003) and *Protein binding* (FDR p = 0.005) were moreover suggested to be enriched.

The interactions between the significantly different proteins were explored in STRING version 11.0 [289]. The resulting network (Figure 5-23) showed that the proteins had a significant number of known interactions and a protein-to-protein enrichment p-value of 1.2E-9. All proteins except PTGDS, PLD3, DKK3 and PPP3CB were connected in the interaction network.

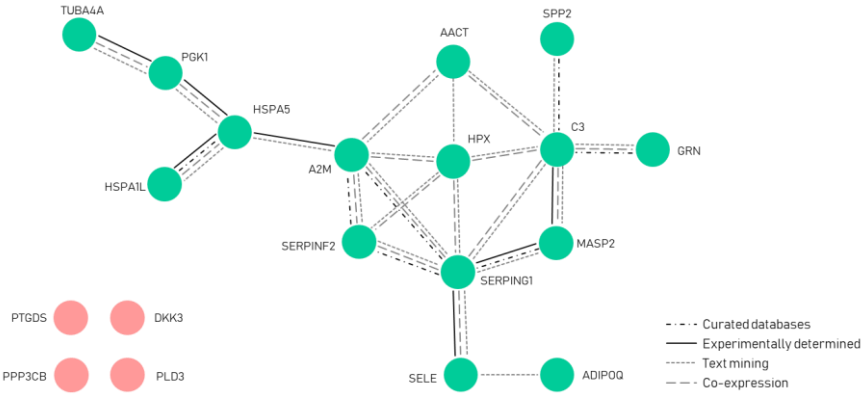


Figure 5-23. Protein-to-protein interaction network from STRING. *The enrichment p-value was 1.2E-9, demonstrating that there is a significant number of known interactions between the proteins*

5.3.2.5 *Comparison of the results from the discovery and the targeted studies*

Comparing the results from the discovery studies, where 17 proteins had been selected for validation, to the targeted results, eight proteins were detected and five demonstrated a significantly different expression. Diverging in expression between the two studies, DKK3 was upregulated in the control twins in the discovery study but significantly downregulated in de novo PD in the targeted study. PTGDS was downregulated in the discovery pre-PD twins but upregulated in de novo PD. HSPA5, SERPINA3 and SPP2

were upregulated in de novo PD/pre-PD in both the discovery and targeted studies. Nine of the discovery proteins could not be reliably quantitated in the targeted assay. Table 5-6 summarises the comparison between the discovery and the targeted study.

Table 5-6. Comparison of results from the discovery studies of de novo PD and controls, and pre-PD and control twin pairs with the results from the targeted study of de novo PD patients and controls. *Eight of the discovery proteins were detected in the targeted study and five demonstrated a significant difference between de novo PD and controls.* $^{****}p < 0.0001$, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, *NS* $p > 0.05$.

Discovered in	Gene	Significance level discovery (all/males only)	Discovery highest in	Significance level targeted study/highest in
de novo PD	ANXA1	** / *	de novo PD	Not detected
de novo PD	GOLM1	NS / *	de novo PD	Not detected
de novo PD	HSPA5	NS / *	de novo PD	**** / de novo PD
de novo PD	NRP1	** / *	de novo PD	Not detected
de novo PD	UHRF1BP1L	** / ***	de novo PD	Not detected
de novo PD	SERPINA3	** / *	de novo PD	****/ de novo PD
PD twins	PRG4	**	Control twin	NS
PD twins	DKK3	*	Pre-PD twin	****/ Control
PD twins	C15orf62	*	Control twin	Not detected
PD twins	SPP2	*	Pre-PD twin	****/ de novo PD
PD twins	BCHE	*	Control twin	NS
PD twins	TNNT3	*	Control twin	Not detected
PD twins	CSF1R	*	Control twin	Not detected
PD twins	MMP3	*	Control twin	Not detected
PD twins	PTGDS	*	Control twin	***/ de novo PD
PD twins	ITIH2	*	Pre-PD twin	NS
PD twins	NCAM1	*	Pre-PD twin	Not detected

5.3.2.6 *Prediction and machine learning models to classify samples as PD or control*

Artificial-intelligence and machine learning are increasingly used to model medical data, for example utilising prior knowledge to predict future outcomes [62]. They are powerful tools allowing for prediction and classification [61]. The multivariate OPLS-DA classification performed in section 5.3.2.1.3 showed promise in discriminating between PD patients and control subjects. Further exploring the prospect of discriminating between the two groups, other machine learning modelling strategies were investigated and are presented in this section.

5.3.2.6.1 *Receiver operating characteristic curve analysis*

To assess the classification ability of the individual proteins, a receiver operating characteristic (ROC) curve was generated from the de novo PD and control samples utilising the web based tool easyROC [290]. A ROC curve plots the true positive rate

(sensitivity) versus false positive rate (1 - specificity) at varied threshold settings and returns the area under the ROC curve (AUC) [291]. The area under the curve allows for an aggregated measure of a variable's performance across the full range of possible classification thresholds. Figure 5-24 shows the ROC curves of the significantly different proteins when comparing de novo PD and control, split over two graphs for the proteins upregulated in the de novo PD group and for the proteins downregulated in the de novo PD group.

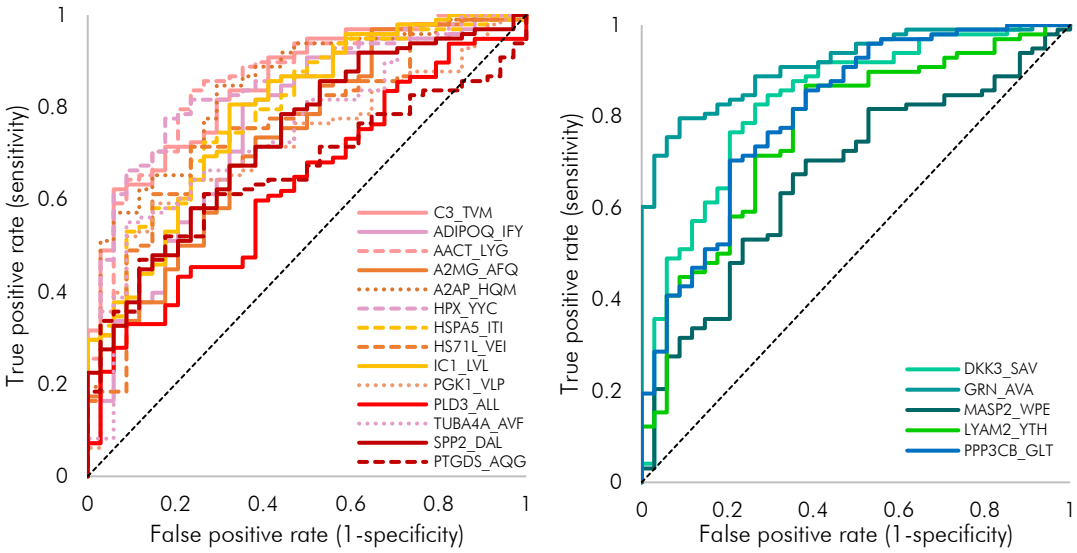


Figure 5-24. Receiver operating characteristic curves of all significantly different proteins in the comparison between de novo PD and control. *On the left: the proteins upregulated in the de novo PD group, and on the right: The proteins downregulated in the de novo PD group. The dashed diagonal lines represent an area under the curve of 0.5, a value at which there is no discrimination and the samples would be randomly classified as belonging to either group.*

Areas under the curve were extracted to assess the classification ability of each protein for the control and de novo PD groups (Figure 5-25). The largest AUC among the proteins upregulated in the de novo PD patients were AACT, C3, HPX and A2AP, while the downregulated proteins with largest AUCs were GRN, DKK3, and PPP3CB.

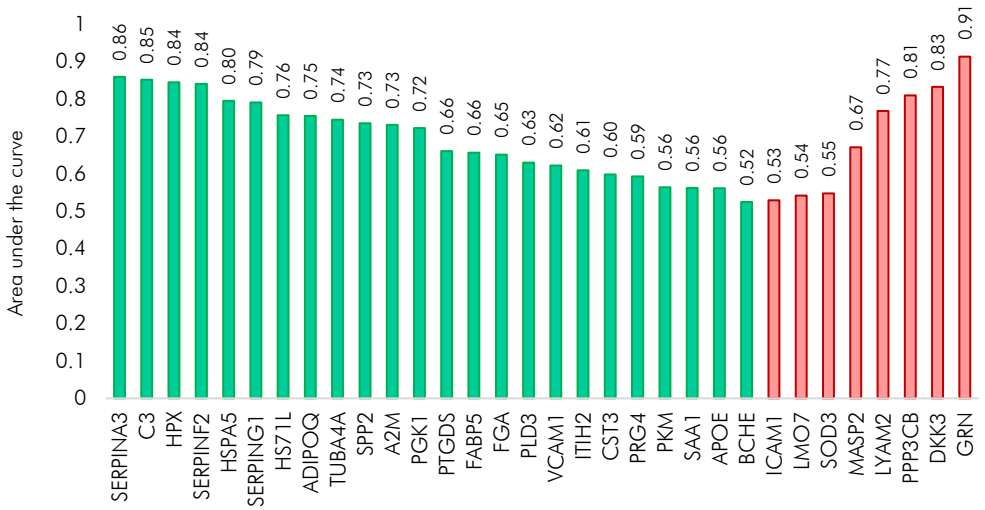


Figure 5-25. Area under the (ROC) curve showing all detected proteins from the targeted analysis of de novo PD and control. *The green bars denote proteins upregulated in de novo PD and the red bars denote proteins downregulated in de novo PD. The AUC value is displayed above each bar.*

5.3.2.6.2 *A machine learning model can predict who belongs to the de novo PD group and to the control group*

Next, the data-modelling workflow developed in Chapter 3, section 3.5.5 was applied, aiming to find the best model for discriminating between de novo PD patients and healthy controls.

Cross-validation to assess the overall quality of the data for prediction. The targeted proteomics de novo PD dataset was first filtered to only contain the de novo PD and control samples, thus excluding the iRBD and OND samples. The de novo PD/control dataset was thereafter divided into five different groups (called "test sets") for cross validation (CV), each group containing a proportion of different PD and control samples, using the function StratifiedKfold from Scikit Learn version 0.24.2 [292] with shuffled values and fixed random state. The remaining samples, not added to the test set, are called the "training set". Figure 5-26 shows the sample distribution in each cross-validation group, where it is demonstrated that in total across all five CV groups, all of the samples from the de novo PD/control dataset were included in a test set and thus subjected to cross-validation.

Figure 5-26. Cross-validation iteration groups. *The de novo PD and control samples are divided by the dashed vertical line, with de novo PD samples to the left and control samples to the right. In each of the five CV groups #1 - #5, the samples selected for the test set are coloured in green, while the remaining samples, the training set, are coloured in red. The test samples from all five CV groups together make up the full set of samples.*

The three different classifier algorithms linear discriminant analysis, support vector machine and Ridge classifier, all from Scikit Learn, were applied to build models from the training sets and predict the test sets for each cross-validation group/iteration. The model scores were extracted and are presented in Table 5-7. It can be noted that all three models performed exceedingly well in the cross-validation, with average scores of 1 or close to 1. This demonstrated that both model fitting and prediction was satisfactory in the models, regardless of which samples were picked for training and prediction.

Table 5-7. Cross-validation results. *The table shows the results from five cross-validation iterations, using three different models. The scores from the training data, and the prediction scores are presented for each model and iteration, and also the total average scores. SD = standard deviation.*

CV iteration	Linear discriminant analysis		Support vector machine		Ridge classifier	
	Training score	Test score	Training score	Test score	Training score	Test score
# 01	1.00	1.00	1.00	1.00	0.99	0.96
# 02	1.00	0.96	1.00	0.96	1.00	0.89
# 03	1.00	1.00	1.00	1.00	0.98	0.96
# 04	1.00	1.00	1.00	1.00	0.98	0.92
# 05	1.00	1.00	1.00	1.00	0.99	0.96
Average ± SD	1.00 (± 0.0)	0.99 (± 0.02)	1.00 (± 0.0)	0.99 (± 0.02)	0.99 (± 0.008)	0.94 (± 0.03)

Construction of models for predicting which samples belong to the de novo PD and control groups. The de novo PD/control dataset was anew split, this time into two equally large parts, each containing the same portion of control and de novo PD samples, using the function train-test split (Scikit Learn). One set, the training set, was used for building and training the models, and the other, the test set, for predictions. Given the superior performance of the LDA and SVM models in the initial assessment, the focus was on these. The most discriminating proteins, determined to be most apt for separating between de novo PD patients and controls, were chosen using recursive feature

elimination (RFECV, SciKit Learn) in the training set. Recursive feature elimination selected nine proteins for each of the models. These proteins were:

- **LDA**: BCHE, DKK3, GRN, HPX, HSPA5, ITIH2, MASP2, SERPING1 and SOD3
- **SVM**: C3, DKK3, GRN, HSPA5, ICAM1, ITIH2, MASP2, PGK1 and SERPING1

The importance of the proteins for discriminating between the PD and control samples in the SVM and LDA models are shown in Figure 5-27. The Ridge model's coefficients are presented in Supplementary figure 2.

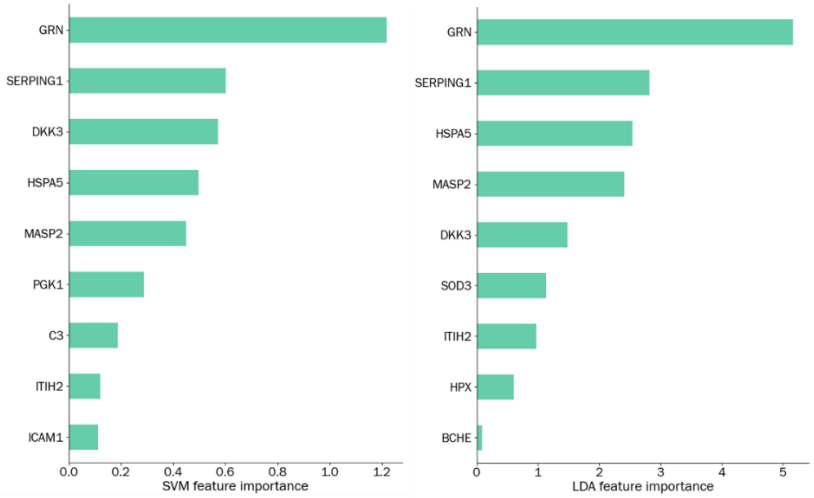


Figure 5-27. Selected features and feature importance of the LDA and SVM models. *The relative size of the bar of each protein shows how important it is to the model's ability to separate between de novo PD and controls in the training data. The SVM features are presented on the left and the LDA features on the right. In both models, GRN is selected as the most important feature, followed by SERPING1.*

Prediction of de novo PD patients and healthy controls. Having constructed models based on the training data - the test set, which consisted of the remaining samples that had not been used for model training and thus never seen by the machine learning models - was predicted in the models. The prediction resulted in the following, remarkable, outcome: in the LDA and SVM models 100% of the samples were correctly predicted as PD or control. In the Ridge classifier model, 97% of the samples were predicted correctly as PD or control. In the Ridge model, two of the control samples were incorrectly predicted in the PD group. The characteristics of the models are summarised below and in Figure 5-28.

- LDA: 100% sensitivity, 100% specificity, 100% accuracy
- SVM: 100% sensitivity, 100% specificity, 100% accuracy
- Ridge classifier: 88% sensitivity, 100% specificity, 97% accuracy

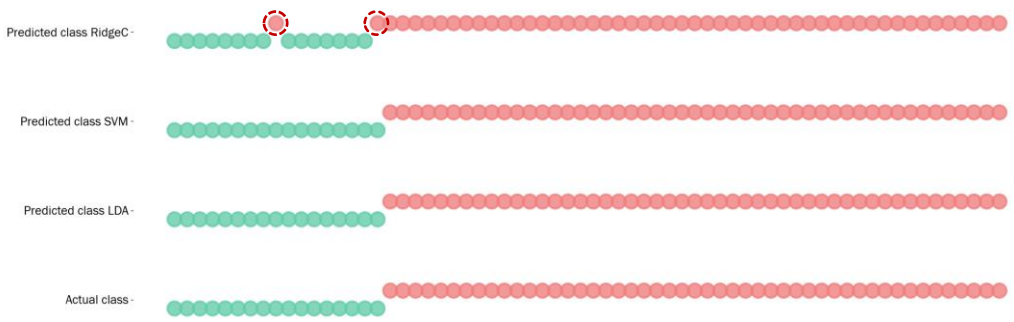


Figure 5-28. Prediction results of control and de novo PD patients in discriminant LDA, SVM and Ridge classifier models. *In the Ridge classifier model, two of the control samples (circled in red) were incorrectly predicted as PD. In the SVM and LDA models, all samples were predicted correctly as PD or control. The last row (labelled "Actual class") shows the actual classes of the samples.* ■ *predicted as control, and* ■ *predicted as PD.*

It was concluded that all three models were applicable to the data and that they all performed exceptionally well.

Specificity testing by prediction of iRBD and OND samples. To test the specificity of the models, the iRBD and OND samples were predicted in the three models. The iRBD samples' outcomes are at the time of writing unknown, therefore we do not know which individuals who will develop PD and could be used to test the models' ability to predict a pre-symptomatic PD state. For the model to show high specificity for Parkinson's disease compared to other neurological conditions, the OND samples would ideally not be predicted as PD but as controls, since they consist of a heterogenous group of other neurological disorders and not Parkinson's disease. The prediction of the 42 OND samples and 18 iRBD samples resulted in the numbers presented in Table 5-8. It is demonstrated that all iRBD samples were consistently predicted in the PD group, regardless of model. The OND samples were evenly predicted as PD or control in all models: the LDA model predicted 50% of the samples as PD, SVM predicted 57% as PD and Ridge classifier 52% as PD.

Table 5-8. Prediction of OND and iRBD samples in the discriminant control/PD linear discriminant analysis (LDA), support vector machine (SVM) and Ridge classifier models. *In the LDA model, 50% of the samples were predicted as PD, in the SVM model 57% as PD, and in the Ridge classifier model 52% as PD. All iRBD samples were predicted as PD in the three models*

	Predicted classes OND		Predicted classes iRBD	
	Parkinson's disease	Control	Parkinson's disease	Control
Linear discriminant analysis	21 (50%)	21 (50%)	18 (100%)	0
Support vector machine	24 (57%)	18 (43%)	18 (100%)	0
Ridge classifier	22 (52%)	20 (48%)	18 (100%)	0

The proportion of OND samples predicted as PD suggests that the model is not exclusively specific for Parkinson's disease. The OND group was highly heterogeneous, with 1–3 samples from each non-PD neurological disorder, meaning it was not possible to discern if any particular disorder overlapped with PD. It can be noted from Figure 5-29 that the prediction models classify largely the same samples as PD and control thereby signifying that these samples share the same characteristics.

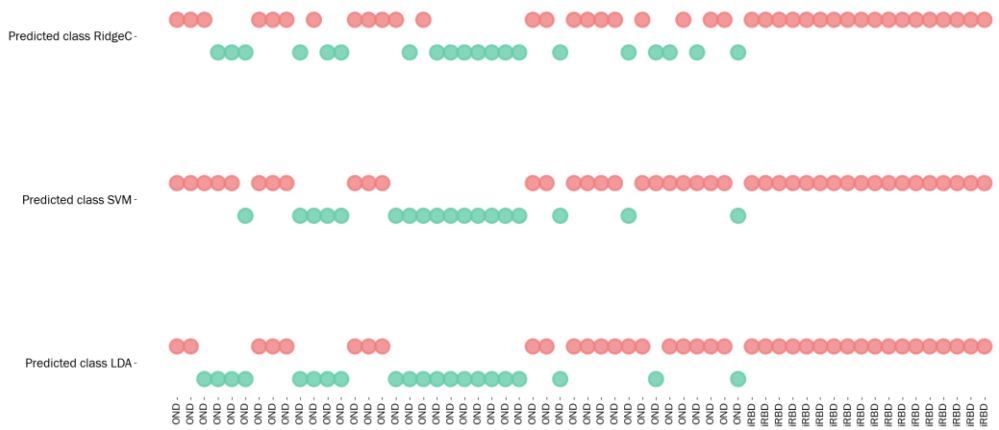


Figure 5-29. Individual results from the prediction of iRBD and OND samples in the discriminant LDA, SVM and Ridge control/PD classifier models. *All iRBD samples are predicted as PD. In all models, the OND samples are distributed evenly in their prediction as PD or control. The predicted class for the individual samples is largely conserved between the three models.* ■ *predicted as control,* ■ *predicted as PD*

5.3.3 *Summary and conclusions from the targeted validation phase*

Four sample groups were analysed utilising the targeted assay developed from discovery proteins and inflammatory proteins from the literature: treatment-naïve de novo PD patients, patients with iRBD, controls and a group of non-PD neurological disorders. The targeted assay included 189 proteins out of which 19 were differentially expressed between de novo PD and control, five between iRBD and control, and seven between OND and control. Eight of the proteins identified in the discovery phase could be detected in the targeted study due to low abundance or technical difficulties. Out of these proteins, five demonstrated a significantly different expression between de novo PD patients and controls. DKK3 and PTGDS diverged in expression when comparing the discovery and targeted studies. HSPA5, SERPINA3 and SPP2 were differentially expressed and demonstrated a protein expression matching the one observed in the discovery study. Literature studies highlighted that many of the proteins had links to PD and/or neurodegeneration and pathway analysis demonstrated that the complement and coagulation cascades were enriched. The ten proteins demonstrating the largest changes in the de novo PD patients were GRN, DKK3, SERPINF2, C3, HPX, PPP3CB, SERPINA3, SERPING1, SELE and HSPA5.

The targeted validation study indicated involvement of the unfolded protein response, complement mediated inflammation and Wnt signalling in the PD pathology.

A multivariate OPLS-DA model could separate de novo PD patients from control samples with high significance. Two machine learning models demonstrated that it was possible to classify samples as de novo PD or control with 100% accuracy based on the expression of BCHE, DKK3, GRN, HPX, HSPA5, SERPING1, ITIH2, MASP2 and SOD3 in a linear discriminant analysis model, and of C3, DKK3, GRN, HSPA5, SERPING1, ICAM1, ITIH2, MASP2 and PGK1 in a support vector machine model. The models were tested for specificity by predicting the OND group and demonstrated that they were not exclusively specific for PD as roughly half of the OND samples were classified as PD.

Finally, we identified a panel of proteins which could distinguish Parkinson's disease patients from healthy controls perfectly. Although exceptionally promising, these results need to be replicated in other studies to verify their validity. If the results would indeed replicate, this panel of proteins could be used as a screening assay to find patients in the early stages of Parkinson's disease.

5.4 DISCUSSION

Treatment-naïve de novo PD patients and pre-symptomatic PD discordant twin pairs were investigated by explorative, bottom-up proteomics. The analyses resulted in a large number of differentially expressed proteins, and demonstrated enrichment of inflammatory pathways (neuroinflammation, complement and clotting cascades), Wnt signalling and ER stress signalling.

Four sample groups were analysed utilising the targeted assay developed from discovery proteins and inflammatory proteins from the literature: treatment-naïve de novo PD patients, patients with iRBD, controls and a group of non-PD neurological disorders. The targeted assay included 127 proteins out of which 19 were differentially expressed between de novo PD and control, five between iRBD and control, and seven between OND and control. Literature studies highlighted that many of the proteins had links to PD and/or neurodegeneration and pathway analysis demonstrated that the complement and coagulation cascades were enriched. The ten proteins demonstrating the largest changes in the de novo PD patients were GRN, DKK3, SERPINF2, C3, HPX, PPP3CB, SERPINA3, SERPING1, SELE and HSPA5.

Taken together, the observed protein expression in the de novo PD patients points towards involvement of protein misfolding, inflammation and Wnt signalling. Protein misfolding and inflammation are well-known features of PD, but Wnt-signalling less so. This particular pathway is of specific interest as it involves the maintenance of dopaminergic neurons and could allow for new insights into the disease mechanisms of PD and suggest a direction for new treatments.

We propose that upregulation of the proteins HSPA5 and HSP1L indicate that the ER-associated degradation pathway and the unfolded protein response are activated. These are ultimately mechanisms aimed at protecting cells from endoplasmic reticulum stress and the grave consequences prolonged ER stress cause. Our study shows that the Wnt-related proteins DKK3 and PPP3CB are strongly downregulated in de novo PD. The downregulation suggests a disturbance of the Wnt signalling pathway. DKK3 is an activator of the canonical Wnt/ß-catenin branch and PPP3CB is a component of the non-canonical Wnt/Ca^{2+} signalling pathway. Wnt signalling is crucial for the development and maintenance of dopaminergic neurons [293] and it has been suggested that activation of the pathway could provide an important protective role in preventing the loss of dopaminergic neurons. Disturbance of the pathway may therefore disrupt one of the important defence mechanisms [294] and aggravate the pathology. A number of upregulated inflammatory proteins also suggest an increased level of inflammation in the PD patients. The strong upregulation of the protein C3 suggests that the complement cascade is activated. Augmented complement components have been linked to neurodegeneration in several studies. Notably, it has been suggested that complement is activated attempting to clear amyloid-beta plaques in Alzheimer's disease [295]. Moreover, the strong downregulation of GRN indicates loss of neuroprotection and increased susceptibility to neuroinflammation. These mechanisms are discussed in detail below and Figure 5-30 shows an illustration of the detrimental and protective mechanisms suggested to be taking place based on the protein expressions observed in this study.

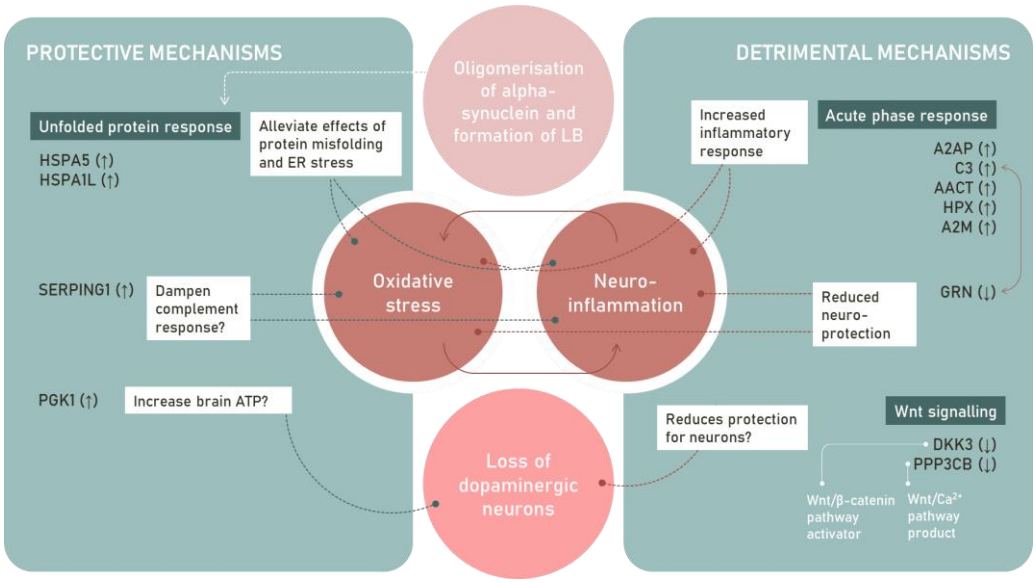


Figure 5-30. Suggested involvement of the proteins from the targeted validation study in Parkinson's disease. *It is suggested that the observed protein expression consists of protective and detrimental mechanisms. The upregulation of proteins from the unfolded protein response (HSPA5 and HSPA1L) are suggested to attempt to alleviate the effects of protein misfolding and ER stress, brought on by the oligomerisation of alpha-synuclein. SERPING1 may be upregulated attempting to dampen the complement response. The increase in PGK1 could be an indication of attempts to increase brain ATP and protect neurons. Upregulation of the inflammatory proteins A2AP, C3, AACT, HPX and A2M may indicate an overall increased inflammatory state. Downregulation of GRN indicated reduced neuroprotection and is modulated by increased levels of C3. Downregulation of the Wnt proteins DKK3 and PPP3CB may indicate that the Wnt signalling pathway is disrupted, thus reducing the protection of dopaminergic neurons.*

Neuroinflammation and complement. Neuroinflammation is involved in restoring homeostasis in the central nervous system, but it is a mixed blessing as it can cause detrimental damage if allowed to run uncontrolled. It is a feature in the pathology of several neurodegenerative conditions, Parkinson's disease included. There is an established link between neuroinflammation and complement. Complement components, receptors and regulators have been found in the central nervous system in several studies and many proteins have been shown to increase in expression during inflammation [296]. Augmented complement expression has also been linked to neurodegeneration in several studies. Complement factors have been observed at elevated levels with AD progression and it has been suggested that complement is initially activated attempting to clear amyloid-beta plaques [295]. In an AD mouse study, it was shown that C3 knockout was protective against synaptic loss [208]. Complement activation has been associated with the formation of alpha-synuclein and Lewy bodies in Parkinson's disease and a study from 2006 also found deposits of iC3b and C9 in Lewy bodies [297]. C3 is a central molecule in the complement cascade; it is formed regardless of the initiating pathway (classical, alternative or lectin) and ultimately leads to the formation of a membrane attack complex.

ER stress and the unfolded protein response. The endoplasmic reticulum is responsible for folding proteins into specific conformations, adding post translational modifications and sorting/exporting the proteins for their destination - secretion, the plasma membrane, or other organelles. Incorrectly folded proteins detected by the ER's quality control enter the ER-associated degradation pathway (ERAD) and are sent to the cytosol for proteasome degradation [298, 299]. Under circumstances leading to excessive amounts of misfolded proteins accumulating in the ER, the organelle enters a condition known as ER stress. Prolonged periods of ER stress can have grave consequences and lead to the collapse of a number of pathways [300]. Given the deleterious effects of unchecked ER stress, there is a defence in place - the unfolded protein response. The UPR attempts to restore homeostasis by reducing the influx of proteins to the ER and increasing the protein folding capacity [301]. If the UPR is unable to reduce ER stress, apoptosis is induced [302]. The UPR can be initiated through three different sensors, IRE1α, PERK and ATF6α, all activated by BiP-bound misfolded proteins [303]. Under non-stressed conditions, BiP interacts with the UPR sensors but when unfolded proteins accumulate in the ER, BiP will disassociate from the sensors and preferentially bind to the unfolded proteins instead, thereby allowing the cell to determine the level of ER stress by evaluating the levels of sensor-bound BiP [304].

Protein misfolding is a well-known component of Parkinson's disease pathology and believed to be the key-factor behind the oligomerisation and aggregation of alpha-synuclein ultimately leading to the formation of Lewy bodies [305]. ER stress has furthermore been linked to Parkinson's disease in a number of studies [306]. The higher expression of BiP (HSPA5) and HSPA1L in the DNP group does indeed point towards ER stress in the Parkinson's patients. The upregulation of BiP suggests activation of the UPR and the upregulation of HSPA7l, a part of the ERAD complex, indicates that also this pathway is activated.

Wnt signalling and dopaminergic neurons. Wnt signalling is a complex pathway, critical for several aspects of cell development in embryonic and adult tissues [307, 308]. In mice, knockout of the pathway, or of components in the pathway, led to embryonic lethality or severely affected phenotypes [309]. Wnt signalling is divided into three different branches; canonical Wnt/β-catenin, and non-canonical Wnt/planar cell polarity (PCP) and Wnt/Ca^{2+} [310]. Wnt proteins are consequently divided into two umbrella categories - Wnt1 for ligands of the canonical pathway and Wnt5a for ligands of the non-canonical pathways. Given the potency and significance of the pathway, Wnt signalling is tightly controlled by several enhancers and inhibitors. The inhibitors are separated into

two categories: secreted Frizzled-related proteins (sFRP) and Dickkopfs (DKK). Generally speaking, sFRPs exert inhibition by binding to Wnt proteins and can inhibit both the canonical and non-canonical pathways while DKKs can modulate the canonical pathway by binding to a Wnt receptor complex (LRP5/6). In the DKK class, DKK1, DKK2 and DKK4 act as antagonists while DKK3 is an activator [311]. In the adult human brain, Wnt signalling governs a range of crucial functions, including neuronal survival, synapse formation, neurogenesis, and regeneration. Adult neurogenesis is primarily governed by canonical Wnt/β-catenin signalling and new neurons can be produced in two areas of the brain, the subventricular zone and the subgranular zone [312]. It has been suggested that downregulation of Wnt signalling promotes dysfunction and/or death of dopaminergic neurons. Interestingly, restoration of dopaminergic neurons was shown in a mouse study where β-catenin was activated in situ [313]. Moreover, in a mouse study where neural stem cells were transplanted to the substantia nigra of medically PD-induced mice, re-expression of Wnt1 and repair of dopaminergic neurons could be seen [314]. The Wnt signalling pathway may have the potential for restoring dopaminergic neurons' function and it has been the focus of a number of studies in recent years [311].

In our study, the canonical pathway activator DKK3 was observed to be strongly downregulated in the DNP group. Also PPP3CB, a component of the non-canonical, β-catenin independent Wnt/Ca^{2+} signalling pathway was strongly downregulated. A downregulation or disturbance of Wnt signalling would appear to be happening in the DNP group, thereby removing an important line of defence against the detrimental loss of dopaminergic neurons.

Study limitations. One limitation of this study was the sample size in the discovery phase. However, due to the major complexity of plasma as a sample matrix, extensive fractionation and highly sensitive analysis were required to detect the low-abundant proteins, meaning that the sample size had to be limited to reduce the risk of irreparable instrumental drift. One additional limitation was the skewed sex distribution in the discovery de novo PD cohort. In the targeted analysis, the outcomes of the iRBD patients are still unknown, therefore we could not evaluate the accuracy of the predictive biomarker panels in their classifications of these samples. Moreover, as the positive OND group was highly heterogeneous, we did not succeed in determining which other neurodegenerative conditions were more likely to be confounded with PD.

Concluding remarks. After performing mass spectrometric proteomic studies in a discovery phase of newly diagnosed PD patients and pre-symptomatic PD patients, we

found several putative biomarker targets. We applied an augmented targeted proteomic assay to samples from newly diagnosed PD patients, iRBD patients, other (non-PD) neurological disorders and controls and identified several differentially expressed proteins between the sample groups. The most interesting targets were the C3, DKK3 and HSPA5. These proteins suggest involvement of complement, Wnt signalling and the unfolded protein response. The Wnt signalling-related protein DKK3 is especially interesting as it may point towards downregulation of the Wnt/β-catenin pathway, a so far relatively unexplored area in PD studies. The pathway may offer further insights into the mechanisms of the pathology and offer a route to develop new treatments. To our knowledge, this is the first time a proteomic study of Parkinson's disease finds DKK3 expression significantly downregulated in patients.

Exploring urine as a source of biomarkers for Parkinson's disease through discovery proteomics followed by targeted validation

6

Abstract. *In this study, urine was explored as a potential source of proteomic biomarkers for Parkinson's disease. Given its non-invasive and effortless collection, urine would be an ideal fluid for diagnostic screening. Our previous work on plasma from PD patients had showed considerable promise and we hypothesised that further targets could be discovered in urine.*

In the discovery phase, a cohort consisting of idiopathic PD patients (n = 9), symptomatic (n = 6) and asymptomatic (n = 6) LRRK2 mutation carriers and healthy controls (n = 10), were analysed by untargeted LC-MS proteomics. More than 2500 proteins were identified and several of the proteins expressed differences between the groups. Pathway analysis indicated that pathways related to PD were affected, among them neuroinflammation and PD signalling. A selection of the proteins identified in the discovery study were developed into a targeted test which also included several pro- and anti-inflammatory proteins identified from the literature.

The targeted assay was applied to a new and larger set of samples (n = 211), consisting of newly diagnosed PD patients, patients with REM sleep behaviour disorder, a positive control group of patients with non-PD neurological disorders, and healthy controls. A total of 23 proteins were differentially expressed between PD and healthy controls in the targeted analysis. Four of the proteins from the discovery phase were validated; these were MAPK12, PPP3CB, CAPN2 and NDRG1. DKK3 was confirmed to be differentially expressed, although with a diverging protein expression compared to the discovery study. The protein expression observed in the targeted assay pointed towards neuroinflammation and Wnt signalling. A discriminant machine learning model could differentiate between newly diagnosed PD patients and healthy controls with 85.1% accuracy using a panel of eight proteins.

In conclusion, we suggest that urine is a valuable biofluid for biomarker discovery and for exploring disease mechanisms.

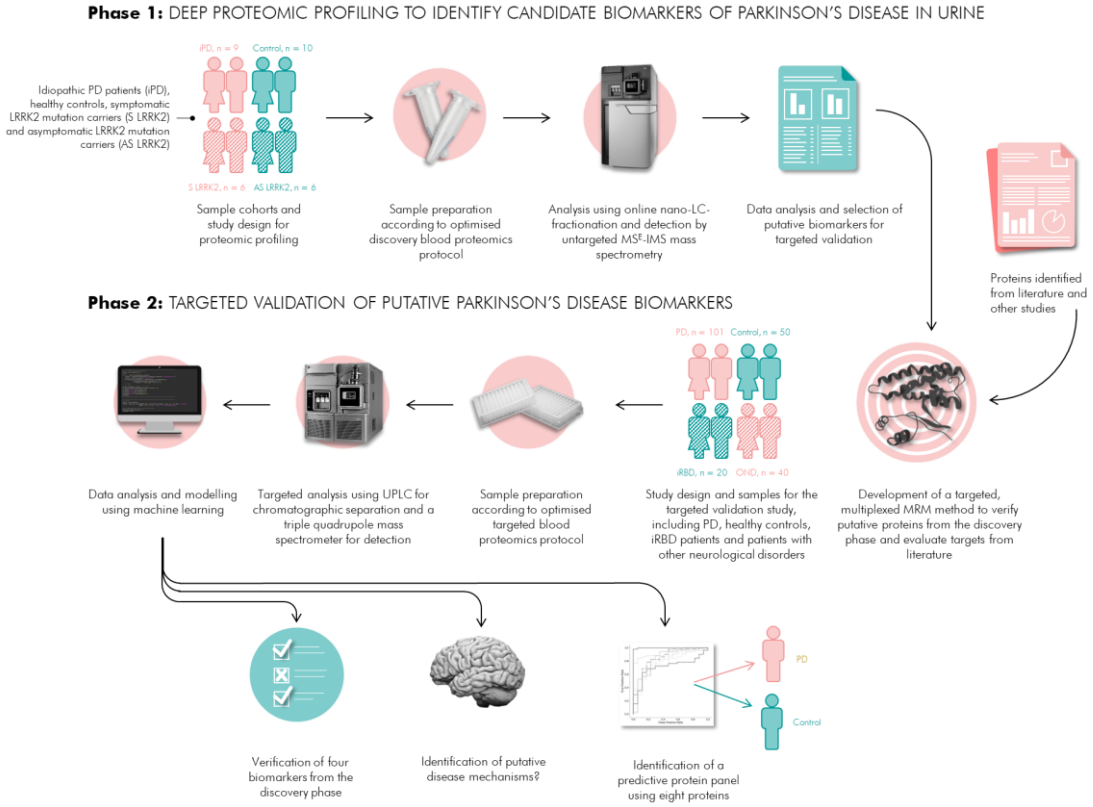


Figure 6-1. Graphical abstract of the urine-based discovery and validation biomarker study of Parkinson's disease presented in this chapter.

6.1 INTRODUCTION AND AIMS

Urine may not stand out as the obvious choice to study diseases of the brain. Anatomically, the brain and urine are not closely related, and the blood-brain barrier and the kidneys further provide filter barriers. For this reason, urine has been largely overlooked and there have been few biomarker studies of neurodegenerative disease utilising urine [315]. Urinary proteins originate from filtered blood, constituting roughly 30% of the proteins, and from the kidneys and urinary tract, making up approximately 70% [316, 317]. This means that markers discovered in blood may also be reflected in urine. Importantly, urine is not hampered by the same complexity and dynamic protein range as blood, thereby making it an attractive fluid for biomarker discovery. It has been demonstrated in several studies that a vast number of proteins can be detected in urine [318]. Exemplifying this - in our studies, we routinely detected approximately 1000 proteins in blood-based discovery experiments, whereas we detected nearly three times as many in urine.

Following the encouraging results from the previous chapter, using plasma to identify biomarkers of Parkinson's disease and to study potential disease mechanisms, we hypothesised that proteomic profiling of urine might also provide biomarker targets and

further insights into the disease process in PD. Moreover, urine provides an attractive biofluid choice because of its non-invasive sampling and abundant availability. It would thus be ideal for diagnostic screening.

Extensive method development was performed to maximise the number of detectable proteins, described in Chapter 3, and the final and optimised workflow was applied to a cohort of PD patients aiming to find urinary protein biomarkers. Targets were developed into a targeted assay, also including pro- and anti-inflammatory proteins from literature.

This chapter describes the process of identifying protein targets relevant to Parkinson's disease in urine. It outlines the steps involved in the discovery phase - where protein targets are identified, to the validation study in a new and larger set of samples, aiming to verify the discovery results.

The aims of the study performed in this chapter were the following:

- Utilising mass spectrometry, identify putative proteomic biomarkers of Parkinson's disease in urine
- Validate the targets from the discovery phase using a targeted, mass spectrometric MRM-based proteomic assay, applied to a new and larger cohort of samples
- Identify additional differentially expressed pro- and anti-inflammatory proteins from the literature
- Identify proteins, or a panel of a proteins, capable of distinguishing between Parkinson's disease and control

6.2 DISCOVERY PROTEOMICS TO IDENTIFY URINARY MARKERS OF PARKINSON'S DISEASE

In the discovery phase of this study, we explored the feasibility of identifying biomarkers of Parkinson's disease in urine. The optimised discovery proteomics workflow for urine, developed in Chapter 3, was applied to samples from idiopathic Parkinson's patients, controls, and LRRK2 mutation carriers (a common cause for developing familial late-onset PD [319]). The overall aim was to identify urinary proteomic biomarkers for PD and to validate these in a larger set of samples.

6.2.1 *Methods and materials*

6.2.1.1 *Discovery sample cohort*

Urine samples for deep proteomic profiling and biomarker discovery were provided by Professor Kailash Bhatia and had been collected from patients visiting the National Hospital of Neurosurgery and Neurology. The samples consisted of idiopathic PD

patients (iPD), symptomatic LRRK2 mutant carriers (LRRK2 S), asymptomatic LRRK2 mutant carriers (LRRK2 AS) and controls. The samples had been collected during routine visits and not post-fasting. The characteristics of the samples are described in Table 6-1.

Table 6-1. Characteristics of the urine samples analysed in the discovery proteomics study. *The sample groups include control, idiopathic PD patients (iPD), symptomatic LRRK2 mutant carriers (LRRK2 S), asymptomatic LRRK2 mutant carriers (LRRK2 AS) and controls. The distribution of males and females, the mean age and mean years of motor disease duration are presented. SD = standard deviation.*

Group	Number of samples	Percentage males/females	Age ± SD	Motor disease duration ± SD
Control	10	50% M | 50% F	69.8 (± 10.7)	
iPD	9	56% M | 44% F	68.3 (± 7.9)	11.7 (± 4.1)
LRRK2 AS	6	50% M | 50% F	66.8 (± 10.2)	
LRRK2 S	6	50% M | 50% F	65.7 (± 7.7)	19.0 (± 7.6)

6.2.1.2 *Sample preparation*

The sample preparation workflow is described in detail in Chapter 2, section 2.4. In summary, two millilitres of urine were filtered to remove low-molecular weight compounds and concentrate the proteins. The concentrated proteins were purified by acetone precipitation and freeze-dried before tryptic digestion and subsequent solid phase extraction. A colorimetric peptide assay was performed to determine the total peptide concentration in each sample. Before instrumental analysis, the peptide concentrations were normalised to 1000 ng/µL to allow for equal injection volumes.

6.2.1.3 *Instrumental analysis*

The instrumental discovery proteomics analysis was performed according to Chapter 2, section 2.5. In brief, 3000 ng of peptides were loaded onto a two-dimensional nano liquid chromatography system coupled to a Synapt-G2-S*i* mass spectrometer (Waters). The peptides were fractionated online into ten fractions on the first column and thereafter chromatographically separated on the second column. The mass spectrometer operated in positive electrospray ionisation mode with ion mobility separation.

6.2.1.4 *Data processing and analysis*

Data processing was performed as per Chapter 2, section 2.6. In brief, the ten discovery proteomics fractions were treated individually in the software Progenesis QI-P (Nonlinear, Waters), where the mass spectrometric data were searched against a database of the canonical human proteome. A false discovery rate of 4% was deemed acceptable for the identifications. At least two fragments per peptide, one peptide per protein and three

fragments per protein were set as ion matching thresholds. The individual fractions were merged in Progenesis to acquire the full protein expression in the samples.

6.2.2 *Results*

2640 proteins were detected and identified by at least one unique peptide. Out of these proteins, 1464 were identified with two or more unique peptides and demonstrated a confidence score above 15.

6.2.2.1 *Multivariate analysis*

The urine discovery data were modelled by PCA for an initial overview and quality control. No apparent clusters of the sample groups could be distinguished as demonstrated by Figure 6-2. The PCA did not demonstrate any run order issues, nor any major patterns of age or sex.

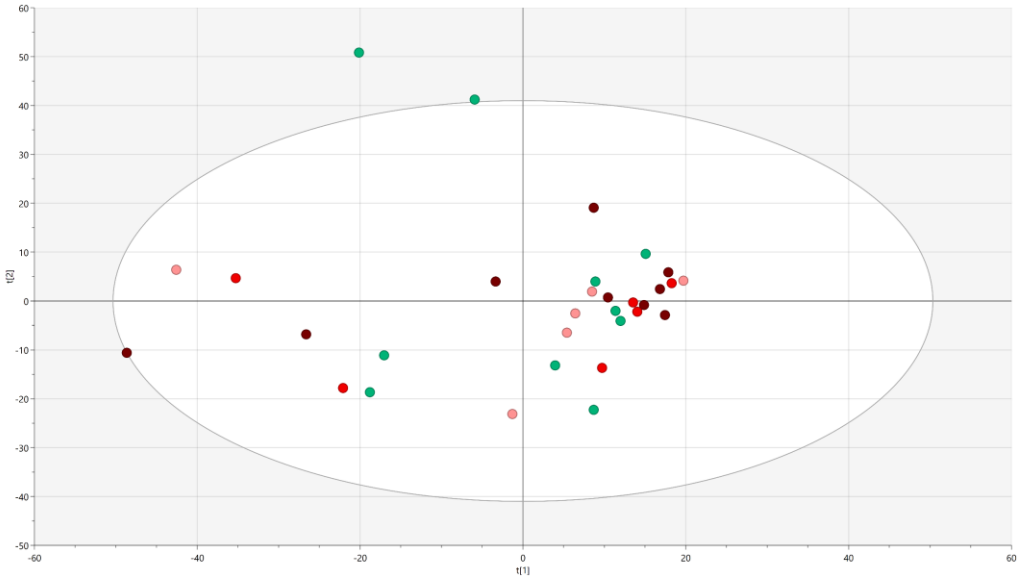


Figure 6-2. PCA of the iPD, LRRK2 mutation carriers and control urine discovery proteomics study. *Principal components 1 and 2 (t[1] versus t[2]) are shown. No apparent influence from instrumental drift, age, sex, or sample group clusters could be found. The samples are coloured according to ■ control, ■ AS LRRK2, ■ S LRRK2, ■ iPD.*

To confirm that no bias had been introduced from the instrumental analysis, an OPLS model with run order as the dependent variable Y was constructed. The model was insignificant with ANOVA $p > 0.05$. An OPLS model with age set as Y was created and proved non-significant. An OPLS-DA model of males versus females was also non-significant.

As the data quality control concluded satisfactory, comparisons of the different sample groups were performed utilising OPLS-DA. Neither of the OPLS-DA models were significant, thus indicating that there was not sufficient covariation between the proteins in the different sample groups to render multivariate class separation.

6.2.2.2 *Univariate analysis*

Using a nominal p-value threshold of 5% as cut off, the following numbers of proteins were significant when comparing the different groups by Student's t-test:

- 85 proteins differing between iPD and control
- 49 proteins differing between AS LRRK2 and control
- 25 proteins differing between S LRRK2 and control
- 23 proteins differing between AS/S LRRK2

The significance and the average fold change of the proteins identified by at least two unique peptides and a with an identification confidence score higher or equal to 15 in each comparison are demonstrated in the volcano plots in Figure 6-3 to Figure 6-6.

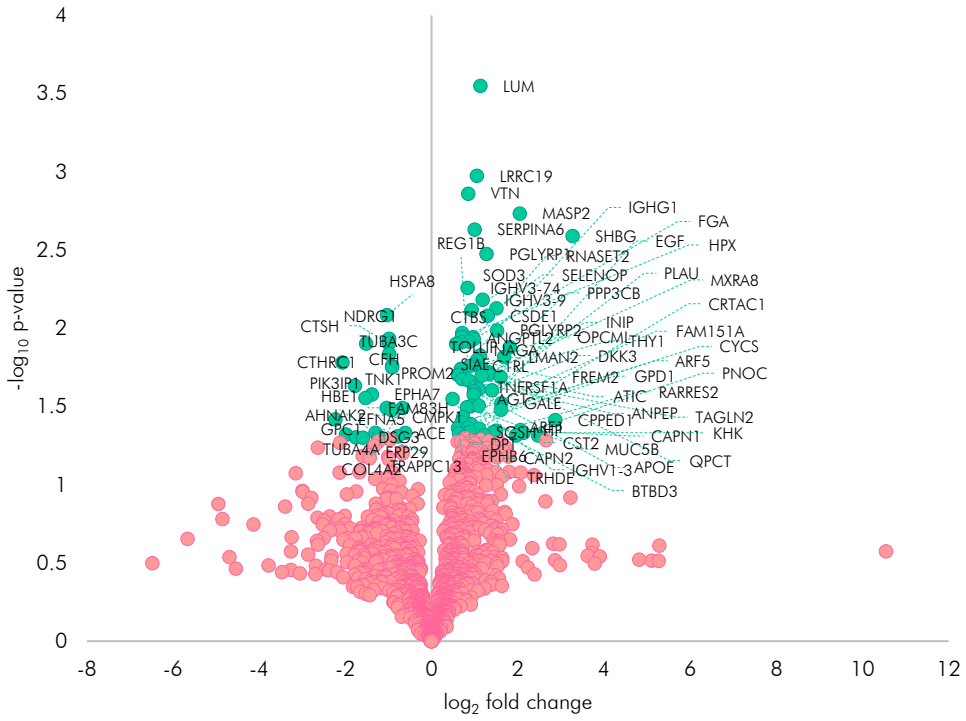


Figure 6-3. Volcano plot of iPD and control samples from the urine discovery proteomics study. *The proteins with negative fold change (to the left in the plot) were elevated in the control group and the proteins with positive fold change (to the right in the plot) in the iPD patients. The significantly different proteins at a nominal p-value threshold of 5% are denoted by their gene names.* ■ *significant, and* ■ *not significant.*

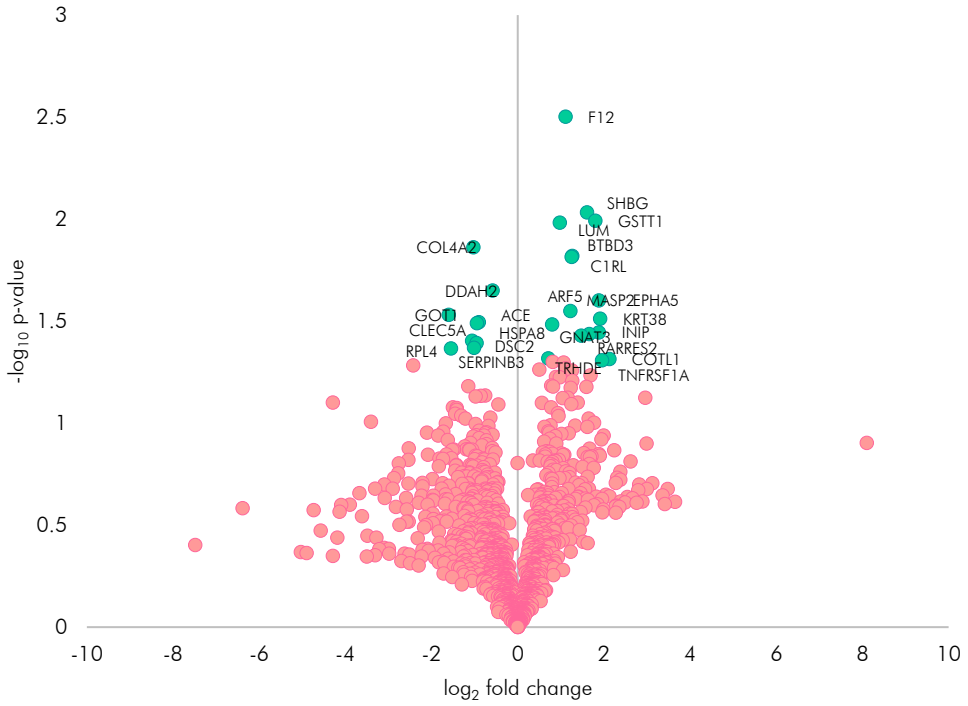


Figure 6-4. Volcano plot of symptomatic LRRK2 mutation carriers and control samples from the urine discovery proteomics study. *The proteins with negative fold change (to the left in the plot) were elevated in the control group and the proteins with positive fold change (to the right in the plot) in the symptomatic LRRK2 mutation carriers. The significantly different proteins at a nominal p-value threshold of 5% are denoted by their gene names. ■ significant, and ■ not significant.*

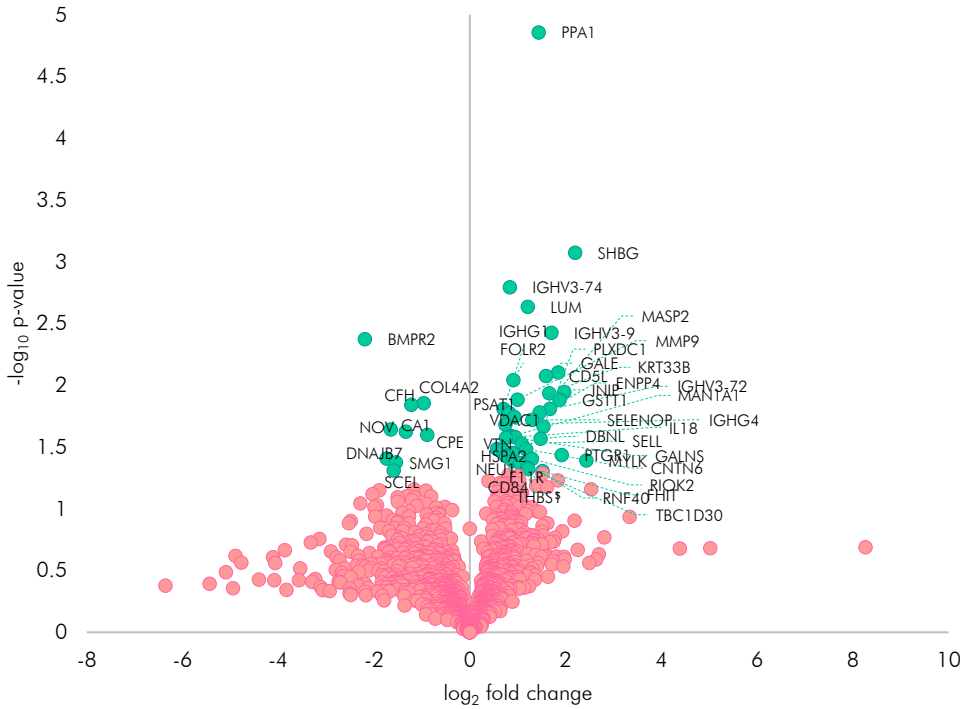


Figure 6-5. Volcano plot of asymptomatic LRRK2 mutation carriers and control samples from the urine discovery proteomics study. *The proteins with negative fold change (to the left in the plot) were elevated in the control group and the proteins with positive fold change (to the right in the plot) in the asymptomatic LRRK2 mutation carriers. The significantly different proteins at a nominal p-value threshold of 5% are denoted by their gene names. ■ significant, and ■ not significant.*

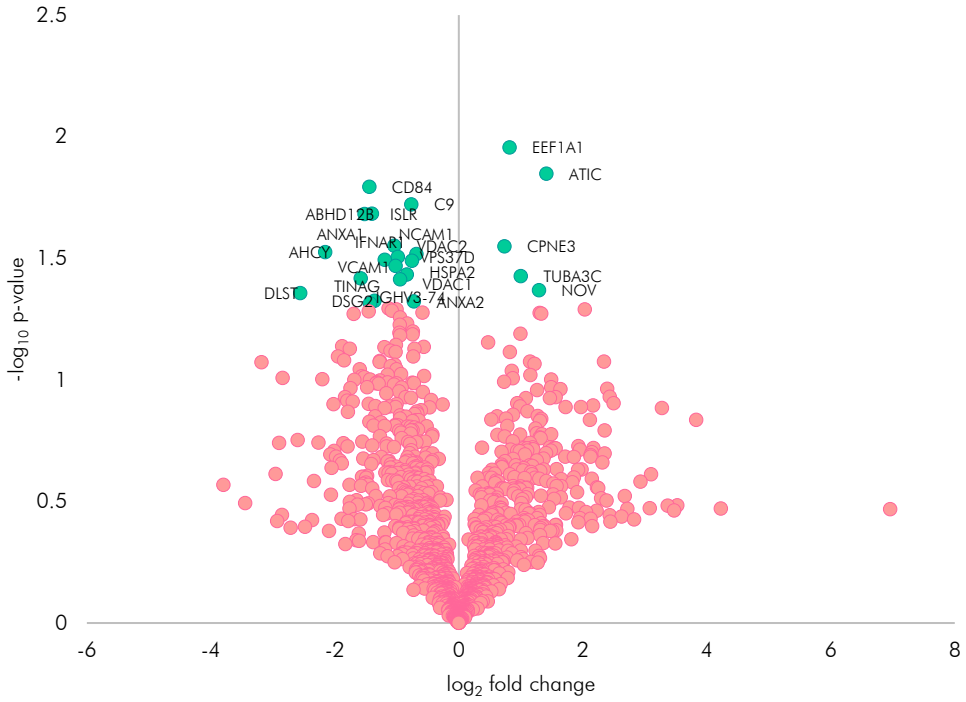


Figure 6-6. Volcano plot of symptomatic and asymptomatic LRRK2 mutation carriers from the urine discovery proteomics study. *The proteins with negative fold change (to the left in the plot) were elevated in the asymptomatic LRRK2 group and the proteins with positive fold change (to the right in the plot) in the symptomatic LRRK2 mutation carriers. The significantly different proteins at a nominal p-value threshold of 5% are denoted by their gene names.* ■ *significant, and* ■ *not significant.*

The proteins differentially expressed in both iPD patients and symptomatic LRRK2 mutant carriers when compared to controls were RARRES2, HSPA8, C1RL, SHBG, ACE, BTBD3, LUM, ARF5, TRHDE, TNFRSF1A, MASP2, COL4A2 and INIP. Out of these proteins, SHBG, LUM, MASP2, COL4A2 and INIP were differentially expressed also in the asymptomatic LRRK2 mutation group.

6.2.2.3 *Pathway analysis*

The proteins demonstrating a nominal significant difference between the groups at 95% confidence level from Student's t-test were moved forward for pathway analysis using Ingenuity Pathway Analysis (Qiagen). In the comparison between iPD and control, 71 pathways were significant. Comparing symptomatic LRRK2 and control, 18 pathways were significant and comparing asymptomatic LRRK2 and control, 43 pathways were significant. Figure 6-7 shows a selection of the significant pathways, based on significance and relevance to PD, from the comparison of iPD and control. The complete table of significantly enriched pathways from the analysis of iPD versus controls is presented in Supplementary table 5.

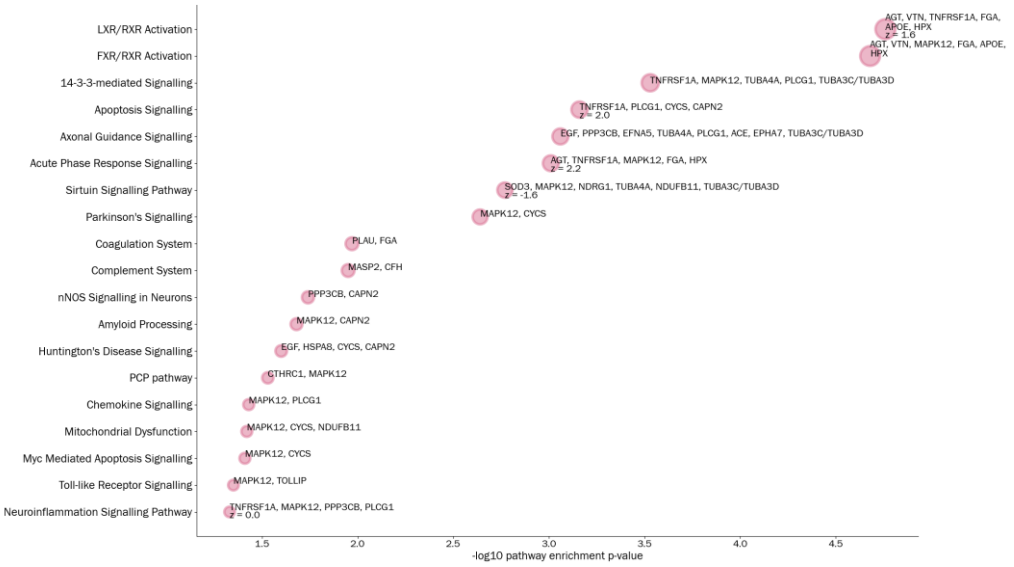


Figure 6-7. Pathway analysis results of iPD versus control. *The pathways are annotated by the respective proteins included in each. Circle radii indicate the significance of the pathway enrichment p-value, and z-scores denote a suggested up- or downregulated pathway, where z > 0 indicates up regulation and z < 0 indicates downregulation. The pathways where no up- or downregulation could be established lack z-score annotations. A selection of the significantly enriched pathways is shown.*

The pathways deemed most relevant to Parkinson's disease are described in brief in Table 6-2. The table provides a summary of the pathways, enrichment p-values, and z-scores which indicate if the pathways were predicted to be up- or downregulated. A z-score with an absolute value larger than 2 was deemed significant to determine up- or downregulation. In the cases where z-scores could not be established, IPA could not predict a direction of regulation. This means that although a pathway may be significantly enriched, the expression of the proteins cannot significantly determine its activation or deactivation.

Table 6-2. Selection of significantly enriched pathways from IPA in the comparison of iPD and control. *A brief description of the pathways is provided, and enrichment p-values, and z-scores which signify if the pathway was predicted to be up- or downregulated.*

Pathway	Pathway description
LXR/RXR and FXR/RXR activation	Six proteins were included in the LXR/RXR pathway which had an enrichment p-value of $1.7E^{-5}$. The z-score was 1.6, thereby weakly suggesting upregulation of the pathway. Six proteins were found in the FXR/RXR pathway, demonstrating an enrichment p-value of $2E^{-5}$. The LXR/RXR and FXR/RXR Activation pathways are involved in lipid metabolism, inflammation, and the catabolic conversion of cholesterol to bile acids. FXR plays an important role in the homeostasis of bile acids, lipoproteins, and lipids [157].
14-3-3-mediated Signalling	Five proteins overlapped with the 14-3-3-mediated signalling pathway, which demonstrated an enrichment p-value of $3.0E^{-4}$. 14-3-3 mediated signalling proteins are proposed to provide neuroprotective functions by inhibiting apoptopic processes and have been associated to Parkinson's disease where they are suggested to interact with alpha synuclein and LRRK2 [320].

Pathway	Pathway description
Apoptosis Signalling	Four proteins were included in the apoptosis signalling pathway with an enrichment p-value of $6.9E^{-4}$ and a z-score of 2.0, thus indicating upregulation. The apoptosis signalling pathway is initiated by caspases (intrinsic or extrinsic) and trigger a cascade of events ultimately leading to programmed cell death.
Axonal Guidance Signalling	Eight proteins were identified in the axonal guidance signalling pathway, the enrichment p-value was $8.7E^{-4}$.The pathway describes the process of guiding axons extended by neurons to form functional neural circuits [321].
Acute Phase Response Signalling	Five proteins overlapped with the acute phase response signalling pathway which had an enrichment p-value of $8E^{-4}$ and a z-score of 2.2, thereby indicating upregulation of the pathway. The acute phase response is a rapid inflammatory response to protect against microorganisms. It can also be triggered by tissue damage. Causes an increase of inflammatory agents such as cytokines.
Sirtuin Signalling Pathway	Six proteins were found in the sirtuin signalling pathway. The enrichment p-value was $1.7E^{-3}$ and the z-score was -1.6, weakly suggesting downregulation of the pathway. Sirtuin signalling has been proposed to increase longevity by delaying cellular senescence [322].
Parkinson's Signalling	Two proteins were identified in the Parkinson's signalling pathway, MAPK12 and CYCS. The enrichment p-value was $2.3E^{-3}$. This pathway describes the known molecular processes occurring in PD, including accumulation of alpha-synuclein, formation of Lewy bodies and loss of dopaminergic neurons.
Coagulation System	Two proteins overlapped with the coagulation pathway; the enrichment p-value was $1.1E^{-2}$. The coagulation pathway consists of a series of events cascading into the formation of blood clots in response to injury. The coagulation system contains a number of proteins to regulate the formation and dissolving of clots.
Complement System	Two proteins were included in the complement system pathway, the enrichment p-value was $1.1E^{-2}$. The complement pathway describes a cascade of events that can be initiated via three different branches. The complement system can lyse foreign cells, activate inflammation, mediate antibody response, and clear immune complexes and apoptotic cells.
nNOS Signalling in Neurons	Two proteins overlapped with the nNOS signalling in neurons pathway, which had an enrichment p-value of $1.8E^{-2}$. This pathway is associated with neuronal apoptosis [323].
Amyloid Processing	Two proteins were identified in the amyloid processing pathway, MAPK12 and CYCS, the same proteins identified in the Parkinson's signalling pathway. The enrichment p-value was $2.1E^{-2}$. The amyloid signalling pathway describes the processes occurring upon accumulation of beta-amyloid including oxidative stress, membrane damage and neuronal death.
Huntington's Disease Signalling	Four proteins were found in the Huntington's disease signalling pathway, EGF, HSPA8, CYCS and CAPN2. The enrichment p-value was $2.5E^{-2}$. The pathway describes the known molecular processes occurring with Huntington's disease, an autosomal, dominant neurodegenerative disorder.
PCP pathway	Two proteins were included in the PCP pathway, the enrichment p-value was $3.0E^{-2}$. The planar cell polarity pathway is one of three known Wnt signalling pathways. The Wnt signalling pathways may have the potential of restoring dopaminergic neurons' function and it has been the focus of a number of studies in recent years [311].
Chemokine Signalling	Two proteins were found in the chemokine signalling pathway. The enrichment p-value was $3.7E^{-2}$. Chemokines are part of the proinflammatory family of cytokines. This pathway involves the recruitment of molecules to alleviate pathological processes [324]
Mitochondrial Dysfunction	Three proteins were identified in the mitochondrial dysfunction pathway with an enrichment p-value of $3.8E^{-2}$. Mitochondria are large consumers of oxygen in the cell. Through different redox processes, oxygen is transformed to radical superoxide. To avoid ROS damage to other cellular compartments, an efficient antioxidant system is also in place. Mitochondrial dysfunction describes the state where the amount of ROS is too great for the antioxidant system to handle. Mitochondrial dysfunction is a trait of, among others, neurodegenerative diseases and diabetes.
Neuroinflammation Signalling Pathway	Four proteins overlapped with the neuroinflammation signalling pathway, the enrichment p-value was $4.7E^{-2}$. This pathway includes the functions aimed at restoring homeostasis in the central nervous system. While initiated as a protective response to clear harmful agents and injured tissue, neuroinflammation can cause detrimental damage if uncontrolled. It is associated with a number of neurodegenerative disorders, including PD.

In summary, the pathway analysis revealed that many pathways related to PD were enriched; it also highlighted enrichment of inflammation-related pathways. The pathways suggested to be upregulated in the PD patients, based on the protein expression in the PD and control groups, were LXR/RXR activation, apoptosis signalling and acute phase response signalling. Sirtuin signalling was indicated to be downregulated in the PD patients. The other enriched pathways did not demonstrate a protein expression allowing the software to conclude if they were up- or downregulated.

6.2.3 *Summary and conclusions from the discovery study of Parkinson's disease in urine*

We identified more than 2500 proteins in urine utilising a carefully optimised workflow in the discovery phase exploring potential urinary biomarkers for Parkinson's disease. The study demonstrated that several proteins were differentially expressed comparing the disease groups to control. Pathway analysis highlighted that several pathways relevant to PD were enriched. Of note, Parkinson's signalling and neuroinflammation signalling were two of the identified pathways. The LXR/RXR pathway, apoptosis signalling, and the acute phase response were suggested to be upregulated in the Parkinson's patients. These three pathways have inflammation as a common denominator [325-327]. Sirtuin signalling was predicted to be downregulated in the iPD patients. This pathway is associated with increased longevity and has been suggested to have an anti-inflammatory role [328]. In conclusion, the discovery study of iPD patients and LRRK2 mutation carriers showed great promise in utilising urine for biomarker identification.

Next, a selection of the identified targets was moved forward for targeted analysis and validation in a larger set of samples. These proteins were: RANGAP1, TUBA4A, MAPK12, APOE, FGA, HSPA8, PPP3CB, PLAU, COL4A2, THY1, CYCS, CTHRC1, ATIC, CAPN2, DKK3, EFNA5, ENDOU, HBE1, MASP2, MUC5B, NDRG1, SOD3 and TOLLIP. The selection was based on significance testing, quality of the protein identification, literature reviews, and potential relevance to Parkinson's disease.

6.3 TARGETED URINE PROTEOMICS TO VALIDATE THE PUTATIVE BIOMARKERS FROM THE DISCOVERY PHASE AND TO IDENTIFY INFLAMMATORY PROTEINS FROM LITERATURE

The discovery phase of this study identified several potential protein targets in urine. These proteins were included in a targeted assay, also containing a number of pro- and anti-inflammatory proteins from the literature. The assay was applied to a new and larger

set of samples to validate the discovery findings and to identify new inflammatory targets. The samples in the validation phase consisted of urine from newly diagnosed PD patients, controls, patients with neurological non-PD disorders as a positive control group, and finally patients with rapid eye movement sleep disorder – a condition commonly observed in patients years before developing Parkinson's disease.

6.3.1 *Methods and materials*

6.3.1.1 *Targeted validation cohort*

The validation cohort for targeted urine proteomics was provided by Professor Brit Mollenhauer, Göttingen University, Germany. The samples belonged to the DeNoPa cohort and were from newly diagnosed Parkinson's patients, patients suffering from idiopathic rapid eye movement sleep disturbance disorder (iRBD), controls and a heterogeneous group of non-Parkinson's disease but other neurological disorders (OND). The urine samples were collected from the same individuals as the PD plasma validation cohort described in Chapter 5, section 5.3.1. A total of 211 samples were provided for the validation study. The control and de novo PD samples were utilised to verify and confirm the findings from the discovery study. The other neurological diseases were included to assess the specificity of the putative biomarkers. The iRBD samples were evaluated for potential prediction of the iRBD patients who have a high probability to go on to develop Parkinson's disease. The conversion rate of iRBD to neurodegenerative synucleinopathies, including Parkinson's disease, is roughly ~80% according to literature [93]. Table 6-3 shows the characteristics of the validation samples.

Table 6-3. Characteristics of the urine samples analysed by the targeted proteomics assay. *The numbers of samples, distribution of males and females, and the average age are listed for each of the sample groups in the cohort.*

Group	Number of samples	Age ± SD	Percentages males/females	Symptom duration [Y] before diagnosis ± SD
Control	50	63.9 (± 7.1)	54% M | 46% F	
iRBD	20	66.9 (± 8.6)	55% M | 45% F	5.4 (± 4.1)
De novo PD	101	67.1 (± 10.6)	50% M | 50% F	2.3 (± 3.2)
OND	40	70.3 (± 8.8)	72% M | 28% F	2.2 (± 2.0)

6.3.1.2 *Sample preparation*

6.3.1.2.1 *Extraction of peptides*

The sample preparation is described in detail in Chapter 2, section 2.8. In summary, 4 mL of urine was spiked with 150 ng yeast ENO1 (whole protein) and filtered to concentrate the proteins and exclude low-molecular species. The concentrate was acetone precipitated and freeze dried followed by tryptic digestion and solid phase extraction.

Quality control (QC) samples were prepared by pooling equal volumes from the validation samples. Calibration curves were prepared by spiking increasing amounts of peptide standards into blank and QC samples.

6.3.1.2.2 *Creatinine measurement*

The concentration of the metabolite creatinine was measured in the samples according to Chapter 2, section 2.11.1.

6.3.1.3 *Instrumental analysis*

The parameters of the instrumental analysis are described in detail in Chapter 2, section 2.9. In short, the samples were reconstituted in 50 µL 3% acetonitrile, 0.1% trifluoroacetic acid, containing 0.1 µM of isotope labelled internal standards and 5 µL was injected onto a UPLC system coupled to a triple quadrupole mass spectrometer. Two injections were performed per sample, each with a different MRM method. In total, 189 peptides were monitored.

6.3.1.4 *Peak picking, integration, and data pre-treatment*

After acquisition, peak picking was performed utilising an in-house Python application or TargetLynx (Waters). Peptide peaks were identified by blank and matrix calibration curves. The digestion efficiency was evaluated by monitoring the presence of yeast ENO1 in the samples, all samples demonstrated a signal from the peptide thereby demonstrating that the digestion was efficient for all samples. The integrated peak areas were exported to Microsoft Excel. The urine concentration between the samples was normalised (see section 6.3.1.4.1). In the final data, outliers were detected utilising a script written in Python with ten absolute deviations set as threshold and replaced by missing values. The outliers are presented in Supplementary table 6. Pooled urine quality control samples were additionally evaluated to assess the robustness of the run.

6.3.1.4.1 *Normalising the urine concentration between samples*

The urine concentrations in the samples were harmonised by two different methods and evaluated: (i) normalising to creatinine concentration and (ii) probabilistic quotient normalisation (PQN). PQN is a mathematical method of adjusting the dilution within a set of samples. It compares each sample to a reference sample and calculates the relative dilution in comparison to the reference, thus normalising the concentration within the sample set [329].

Firstly, the raw and non-normalised data were visualised by principal component analysis (Figure 6-8). PCA clearly demonstrated that principal component 1, which accounted for the majority of the variation in the model (35%), represented the peptide intensity in the samples and thus also the urine concentration.

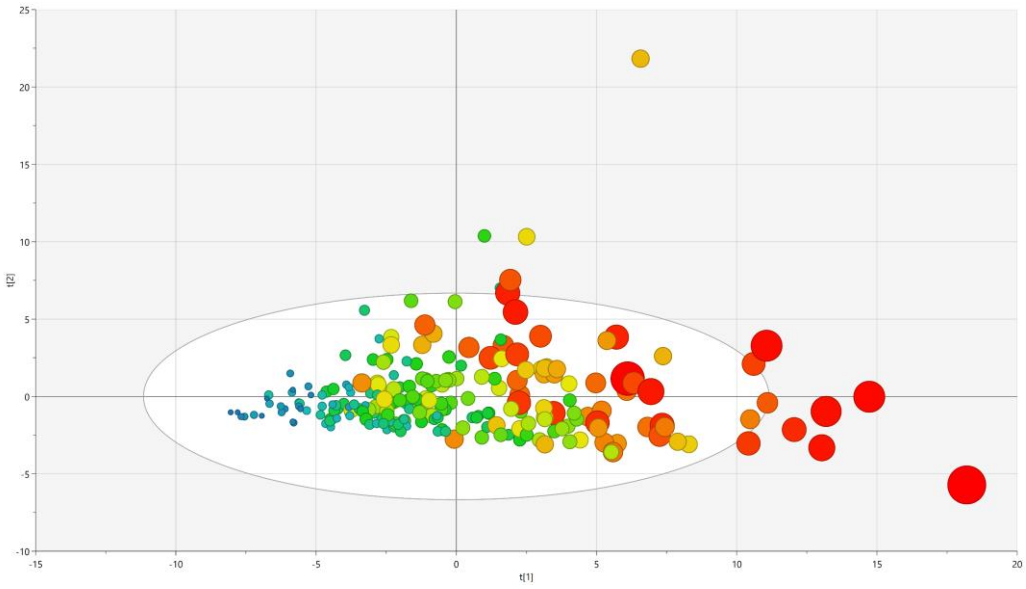


Figure 6-8. Principal component analysis of raw and non-normalised data from the targeted analysis of the de novo PD urine cohort. *The plot shows the first and second principal components (t[1] versus t[2]). The samples are continuously coloured and sized according to the average peptide intensity in each sample, where the lowest intensity is represented by ■ blue and highest intensity is represented by ■ red. It is clearly demonstrated that principal component 1, accounting for 35% of the model's variation, models peptide intensity with lower intensity in the samples on the left and higher intensity in the samples on the right. PC1 thus represents the peptide intensity in the samples and consequently the urine concentration.*

Next, the peptides were individually normalised to creatinine, and collectively normalised by using PQN. The two normalising strategies were evaluated by PCA scores, where the models were inspected for bias in the distribution of the average sample intensities, and by the distribution of the proteins in the PCA loadings.

Creatinine normalisation. Figure 6-9 shows PCA scores and loadings from the creatinine normalising. The scores show that there is a clear, non-random, direction of average intensity in the samples along principal component 1, accounting for 35% of the model's variation also in this model. The loadings plot further demonstrates that all proteins are distributed in the right-hand side of the loading space, meaning that PC1 is driven mainly by sample concentration and thereby demonstrating that the creatinine normalising was inefficient in equalising the urine concentrations between the samples. Moreover, the creatinine normalising produced six extreme outliers (Supplementary figure 3) which had to be excluded from the PCA model.

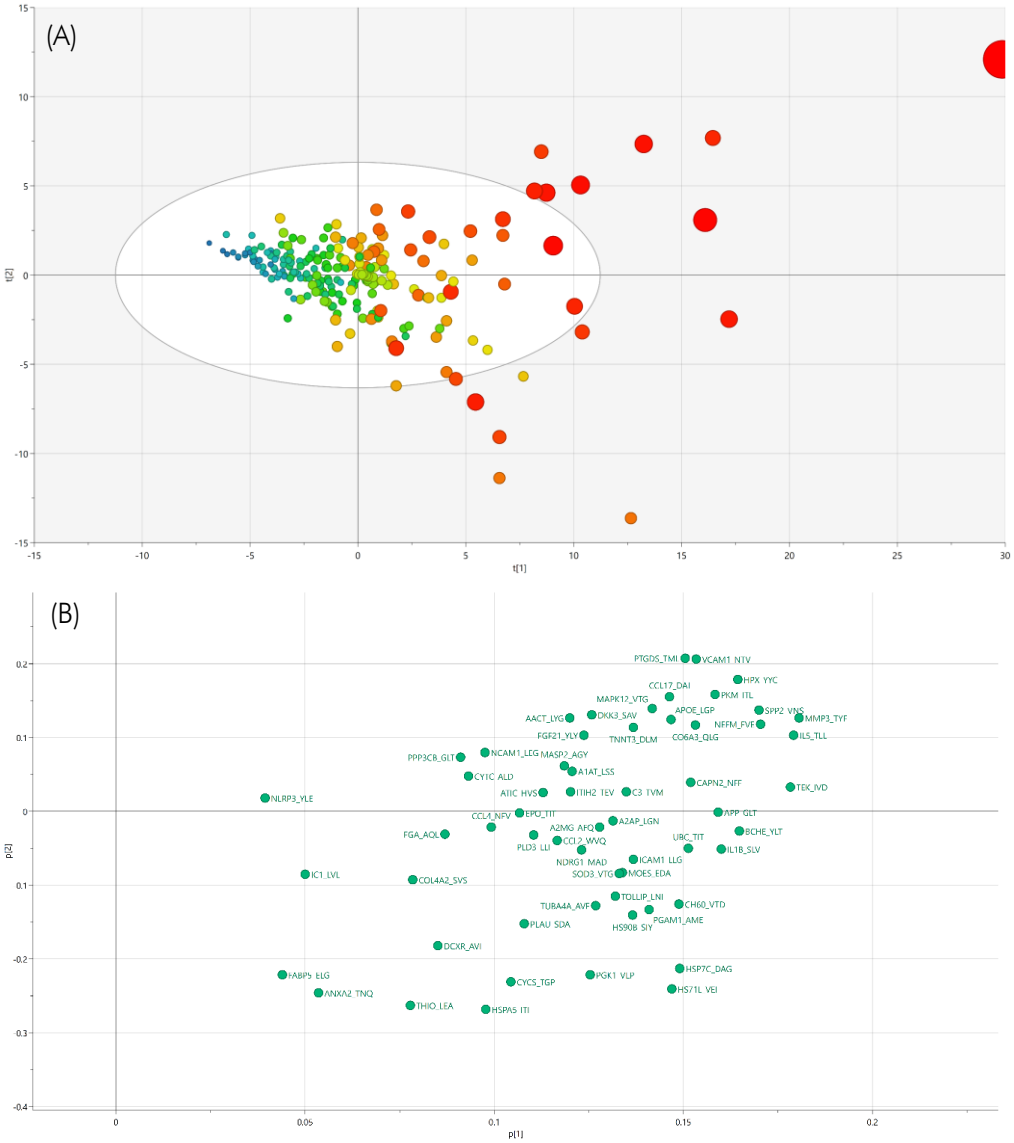


Figure 6-9. (A) PCA scores and (B) loadings from the normalising of urine concentrations by creatinine. *The scores (t[1] versus t[2]) are continuously coloured and sized by average sample intensity, where ■ lowest intensity, and ■ highest intensity. The score plot demonstrates that there is an evident direction of intensity – lower to the left and increasing towards the right of the plot. The corresponding loading (p[1] versus p[2]) show that samples towards the right indeed have higher levels of all peptides. It is thus demonstrated that creatinine normalising does not correct for the urine protein concentration efficiently.*

Probabilistic quotient normalisation. Next, the protein concentrations in the samples were normalised by PQN. Figure 6-10 shows the PCA scores and loadings of the results from this normalising strategy. The scores show that the sample intensities have been largely equalised and that no systematic variation related to sample intensity can be detected. The loadings demonstrate that the distribution of the proteins in the loading space is not heavily skewed to either side.

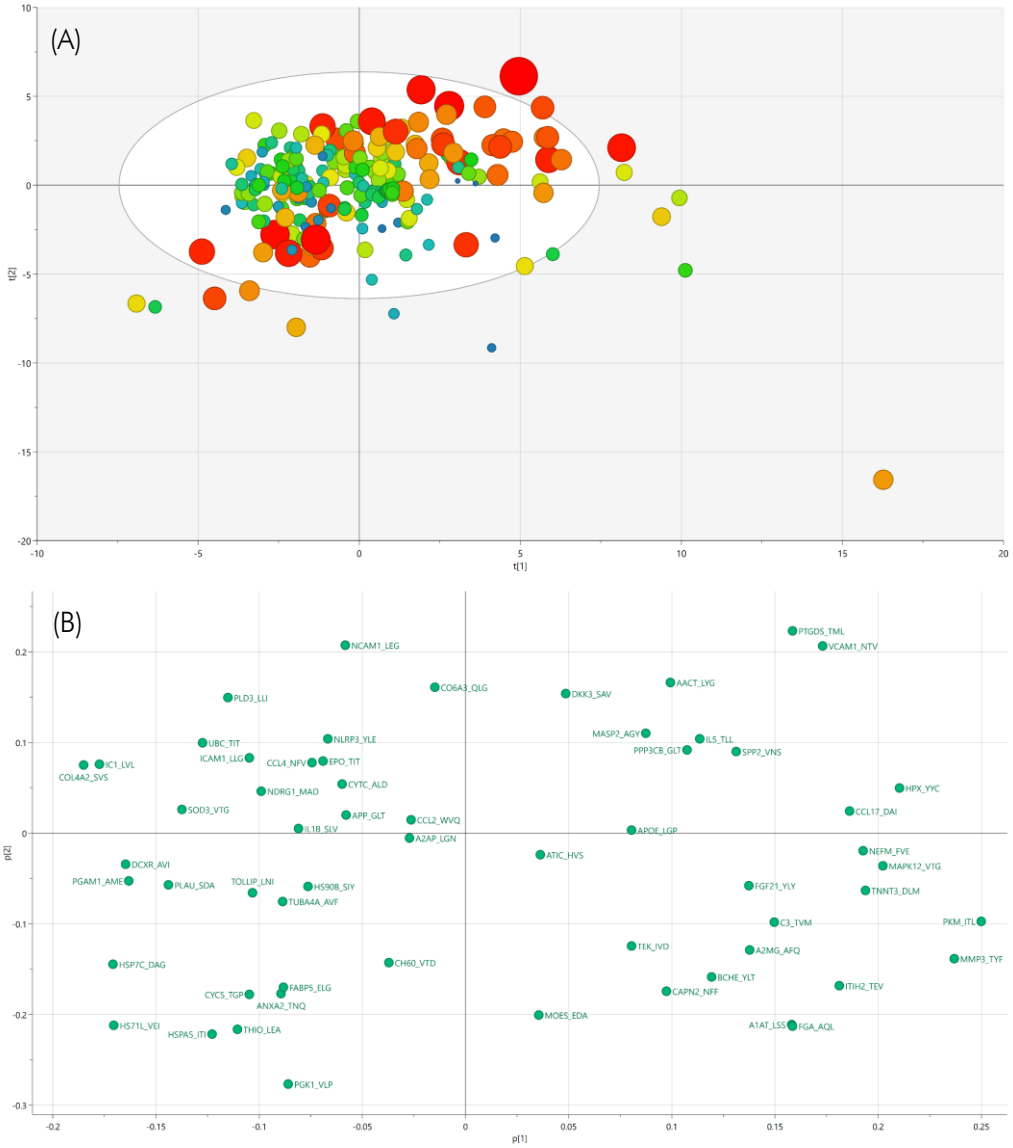


Figure 6-10. (A) PCA scores and (B) loadings from the normalising of urine concentrations by PQN. *The scores (t[1] versus t[2]) are continuously coloured and sized by average sample intensity, where ■ lowest intensity, and ■ highest intensity. The score plot shows that a continuous intensity trend is absent. The corresponding loadings plot (p[1] versus p[2]) demonstrates that the peptides are evenly distributed and that samples are not distributing in the score space due to a systematic intensity difference.*

The evaluation of the two normalising strategies concluded that PQN normalising was more efficient in equalising the urine concentrations between the samples and was therefore selected as the preferred normalising method.

6.3.2 *Results*

6.3.2.1 *Evaluation of repeated quality control samples*

The pooled urine samples utilised as quality control were evaluated by the coefficient of variation resulting from each protein. In total, 16 QC injections were performed. Figure 6-11 shows the coefficient of variation and average intensity of each protein. Out of the

measured proteins, all but four manifested a coefficient of variation of 20% or less. DKK3 had variation of 23%, CAPN2 had variation of 28%, and A2M and FGF21 displayed variations of 30% and 31%, respectively. As expected, the variation increased with lower intensities, the four proteins with largest variation having the lowest intensities.

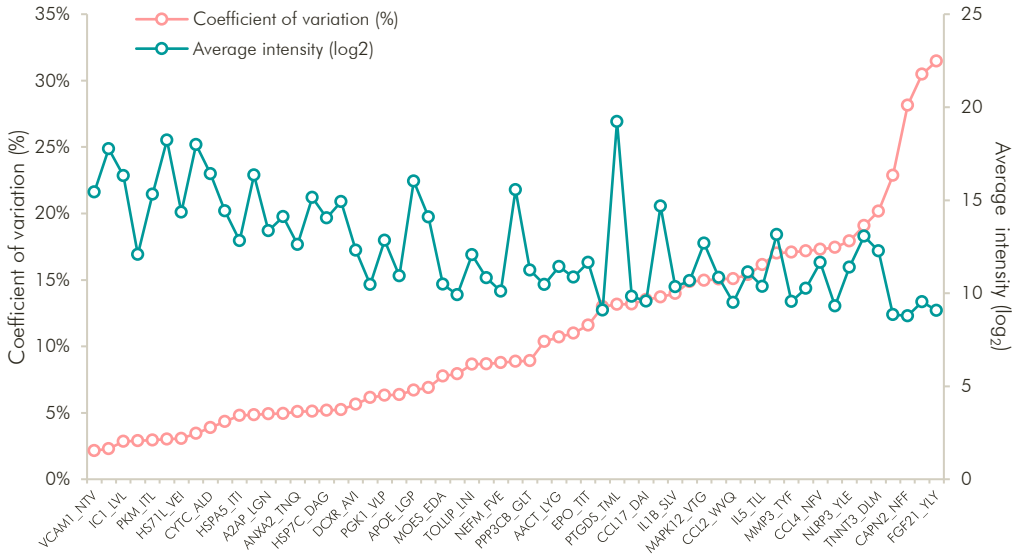


Figure 6-11. Characteristics of the pooled urine quality control samples from the targeted de novo PD analysis. *The average intensities (log_2) of the proteins in the QC samples and the consistency, represented as coefficient of variation, are shown. All but four proteins (DKK3, CAPN2, A2M and FGF21) demonstrated a coefficient of variation of 20% or less. The four proteins with highest coefficient of variation had the lowest average intensities.*

In summary, PQN was performed to equalise the urine concentrations, outliers at more than ten median average deviations were removed and the QC analysis proved satisfactory, thereby concluding the pre-treatment and initial quality assessment of the data.

6.3.2.2 *Multivariate analysis*

6.3.2.2.1 *Unsupervised Principal Component Analysis for quality assessment*

A PCA was performed to obtain an overview of major patterns in the data. No clear clusters of the different disease and control sample groups could be detected in any of the components. The first component demonstrated a weak age dependency and the second component modelled sex. Figure 6-12 shows a summary of the results.

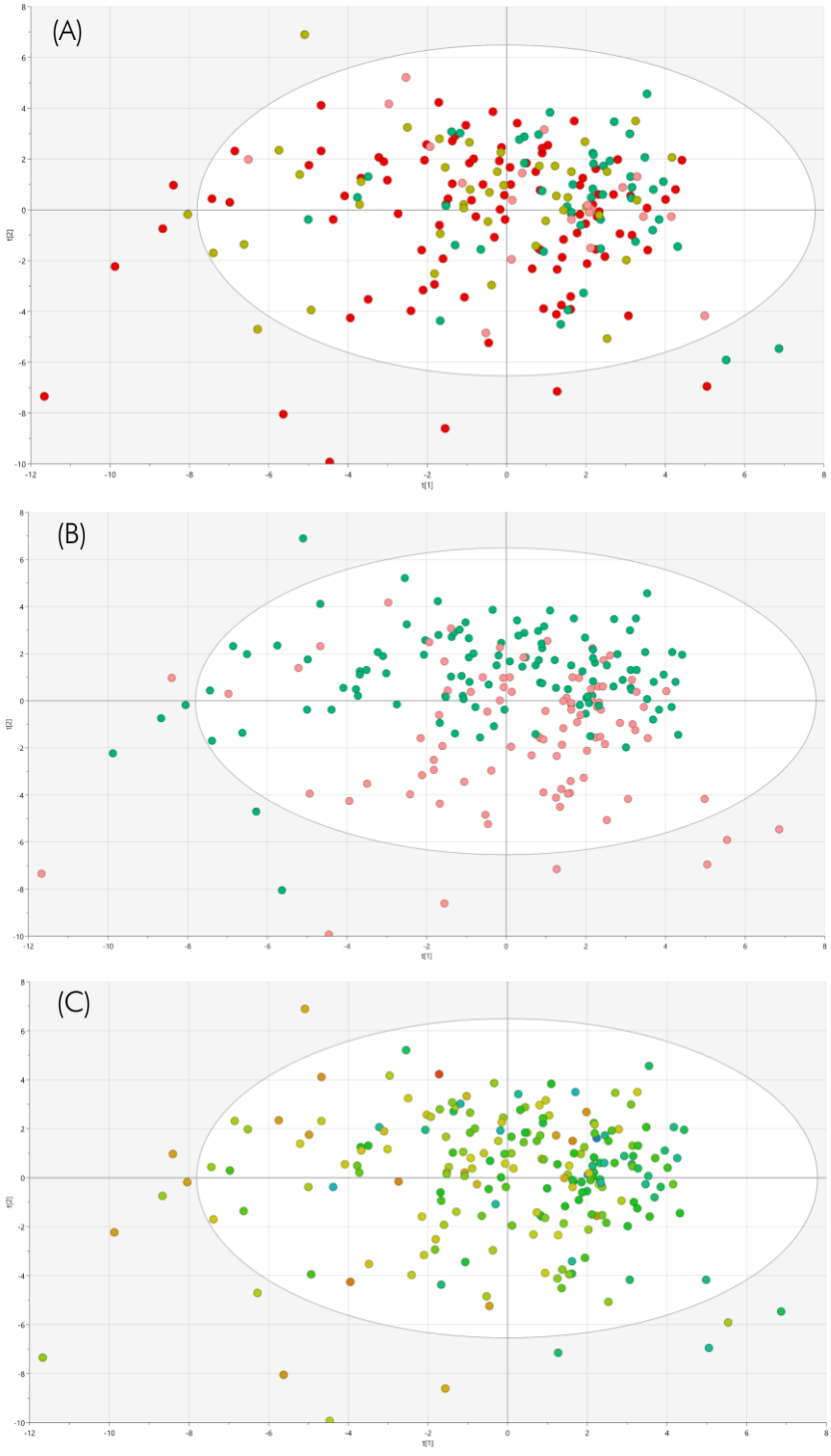


Figure 6-12. Overview PCA analysis of the targeted de novo PD urine cohort. *The first principal components (t[1] versus t[2]) are shown. (A) Data coloured according to sample groups, where ■ control, ■ de novo PD, ■ iRBD, and ■ other neurological disorders. No clear clusters of the groups can be discerned. (B) Data coloured by males/females, where ■ males and, ■ females. A clear discrimination can be detected vertically in principal component 2. (C) Data coloured continuously according to age, where ■ represents youngest (41 years) and ■ oldest (87 years). A weak age-dependency can be noted horizontally in principal component 1.*

6.3.2.2.2 *Supervised models exploring the confounding effects of age and sex*

Investigating the influence of age further, an OPLS model with age set as the dependent variable Y was created. The model was significant (ANOVA p = 7.8 E^{-7}, permutations $p << 0.001$) thereby confirming that there was an age-protein dependency in the data.

The discrimination between males and females was explored in an OPLS-DA model. The model proved highly significant (ANOVA p = 6.7 E^{-24}, permutations $p << 0.001$) thus verifying that the protein expression in the dataset differs between males and females. Figure 6-13 shows the proteins related to age and Figure 6-14 shows the proteins related to sex.

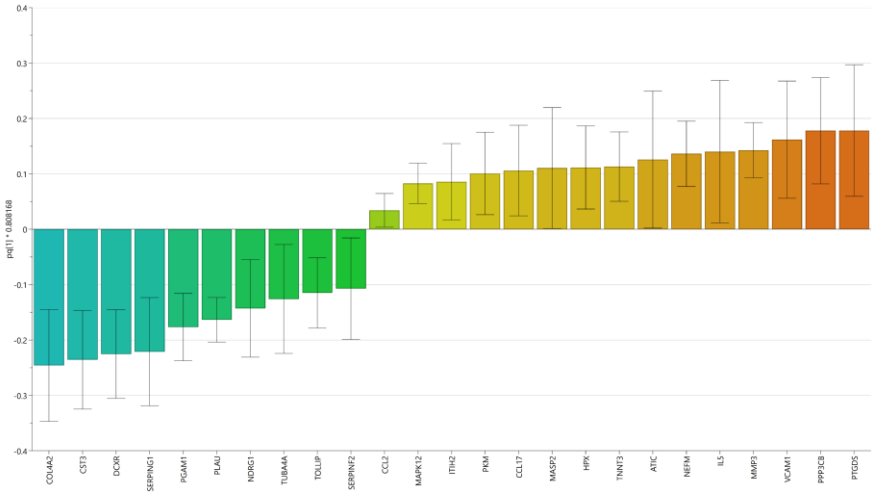


Figure 6-13. Predictive loadings from the OPLS model of the targeted proteomic analysis of the urine de novo PD cohort and age. *The proteins are coloured continuously according to their correlation with age, where blue represents negative correlation and red positive correlation. pq[1] represents the predictive loadings. The error bars represent the standard error.*

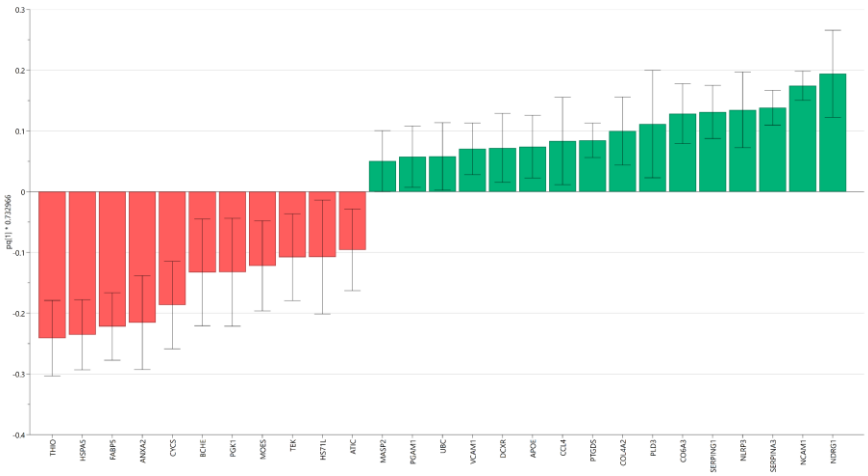


Figure 6-14. Predictive loadings from the OPLS-DA model of males versus females in the targeted proteomic analysis of the de novo PD cohort. *The red bars represent the proteins which were positively correlated with females and the green bars the proteins positively correlated with males. pq[1] represents the predictive loadings. The error bars represent the standard error.*

Age and sex adjustment. The data were adjusted for age and sex utilising linear regression, as described in Chapter 3, section 3.5.4, and modelled again. Both the OPLS model with age set as Y, and the discriminant OPLS-DA model of males versus females were non-significant thus demonstrating that the age and sex correction was efficient.

6.3.2.2.3 *Discriminant analysis of control versus de novo PD*

A discriminant OPLS-DA model of control and de novo PD was constructed from the age and sex corrected data. The model was highly significant, demonstrating an ANOVA p of 1.1 E^{-16} and permutations $p << 0.001$, thereby suggesting that there is a large amount of covariation in the protein expression related to the two groups. According to the model, the most discriminating variables were SPP2, CAPN2, TNNT3 and FGF21, upregulated in de novo PD patients, and SERPINF2, UBC, CCL4 and NCAM1, downregulated in PD (Figure 6-15).

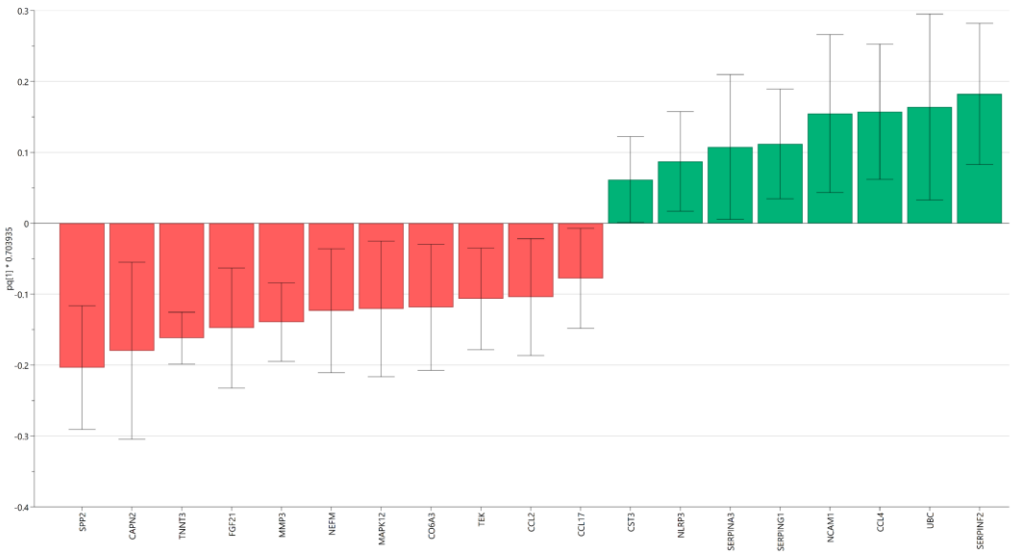


Figure 6-15. Predictive loadings from the discriminant OPLS-DA model of de novo PD patients versus controls in the targeted proteomics analysis of urine. *The red bars represent the proteins higher in de novo PD and the green bars the proteins higher in control. pq[1] represents the predictive loadings. The error bars represent the standard error.*

6.3.2.2.4 *Prediction of the OND and iRBD samples in the discriminant PD/control OPLS-DA model*

The iRBD and other neurological disorder samples were moreover predicted in the model and resulted in 55% of the iRBD samples and 86% of the other neurological disorders being predicted as de novo PD (Figure 6-16). The high rate of other neurological disorder samples predicted as de novo PD suggests that the specificity of the model has room for improvement as these samples largely act as a positive control group and should ideally not be predicted as de novo PD.

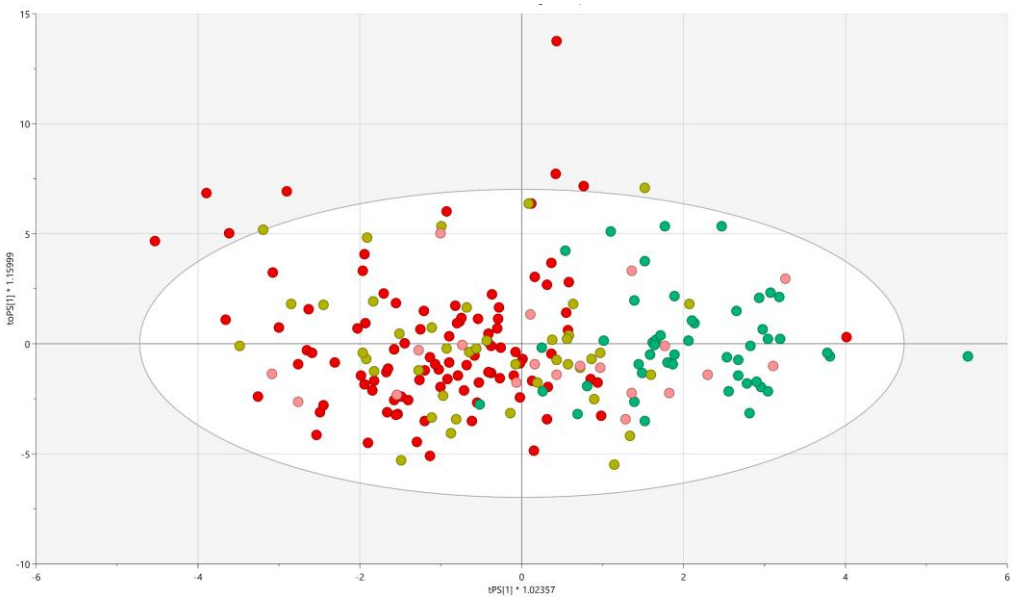


Figure 6-16. Prediction of OND and iRBD in the OPLS-DA model of de novo PD versus control. *The classification of the OND samples resulted in 86% predicted ad de novo PD. The iRBD classification resulted in 55% predicted as de novo PD.* ■ *de novo PD,* ■ *control,* ■ *iRBD (predicted) and* ■ *other neurological disorders (predicted). tPS[1] denotes the predictive component of the prediction, and toPS[1] the first orthogonal principal component of the prediction.*

In conclusion, the multivariate analysis demonstrated that the protein expression in the dataset was influenced by age and sex. After correcting for these factors, a discriminant analysis of de novo PD versus control was performed and proved highly significant, thereby demonstrating that there was a signature protein expression differentiating between de novo PD and control. The model however lacked specificity as demonstrated by the prediction of a positive control group of other, non-PD, neurological disorders.

6.3.2.3 *Univariate analysis*

The samples in the groups de novo PD, iRBD and other neurological disorders were compared to the control group using Student's t-test. Benjamini-Hochberg multiple testing correction at 95% significance level was applied. Between de novo PD and control, 23 proteins were significantly different. Comparing iRBD to control, four proteins were significantly different, all but one also differentially expressed in de novo PD. Between other neurological disorders and control, 11 proteins were significantly different, here too all but one also differentially expressed in de novo PD. Table 6-4 shows the summary of the FDR-corrected p-values from the three analyses.

Table 6-4. Benjamini-Hochberg FDR-adjusted p-value summary from the comparison of control to de novo PD, iRBD and other, non-PD, neurological disorders. **** = p < 0.001, ** = p < 0.01, * p < 0.05 and NS = not significant*

Protein (gene)	De novo PD vs control	iRBD vs control	OND vs control
Intercellular adhesion molecule 1 (ICAM1)	***	NS	NS
C-C motif chemokine 4 (CCL4)	***	NS	NS
Troponin T3, fast skeletal type (TNNT3)	**	NS	**
Secreted Phosphoprotein 2 (SPP2)	**	*	**
Collagen alpha-3(VI) chain (COL6A3)	**	NS	NS
Neurofilament medium polypeptide (NEFM)	**	NS	*
Polyubiquitin-C (UBC)	**	**	*
Phosphoglycerate mutase 1 (PGAM1)	**	NS	NS
Alpha-2-antiplasmin (SERPINF2)	**	NS	***
Calpain-2 catalytic subunit (CAPN2)	**	NS	**
Fibroblast growth factor 21 (FGF21)	**	NS	NS
Neural cell adhesion molecule 1 (NCAM1)	**	NS	NS
N-myc downstream regulated 1 (NDRG1)	**	NS	*
Matrix metallopeptidase 3 (MMP3)	**	NS	NS
Mitogen-activated protein kinase 12 (MAPK12)	**	NS	**
C-C motif chemokine 2 (CCL2)	*	**	*
Heat shock protein HSP 90-beta (HSP90AB1)	*	NS	NS
Serine/threonine-protein phosphatase 2B catalytic subunit (PPP3CB)	*	NS	NS
Cholinesterase (BCHE)	*	NS	NS
Interleukin-5 (IL5)	*	NS	NS
Alpha-1-antichymotrypsin (SERPINA3)	*	NS	NS
Dickkopf WNT signaling pathway inhibitor 3 (DKK3)	*	NS	NS
Heat shock 70 kDa protein 1-like (HSPA1L)	NS	NS	*
Plasma protease C1 inhibitor (SERPING1)	NS	NS	NS
Cystatin-C (CST3)	NS	NS	**
C-C motif chemokine 17 (CCL17)	NS	NS	NS
Heat shock cognate 71 kDa protein (HSPA8)	NS	NS	NS
Interleukin-1 beta (IL1B)	NS	NS	NS
Angiopoietin-1 receptor (TEK)	NS	NS	NS
NACHT, LRR and PYD domains-containing protein 3 (NLRP3)	NS	NS	NS
Amyloid-beta precursor protein (APP)	NS	**	NS
L-xylulose reductase (DCXR)	NS	NS	NS
Erythropoietin (EPO)	NS	NS	NS
Thioredoxin (TXN)	NS	NS	NS
Inter-alpha-trypsin inhibitor heavy chain 2 (ITIH2)	NS	NS	NS
Annexin A2 (ANXA2)	NS	NS	NS
Tubulin alpha-4A chain (TUBA4A)	NS	NS	NS
Hemopexin (HPX)	NS	NS	NS
Alpha-2-macroglobulin (A2M)	NS	NS	NS
Apolipoprotein E (APOE)	NS	NS	NS
60 kDa heat shock protein, mitochondrial (HSPD1)	NS	NS	NS
Collagen alpha-2(IV) chain (COL4A2)	NS	NS	NS
Prostaglandin-H2 D-isomerase (PTGDS)	NS	NS	NS
Pyruvate kinase M (PKM)	NS	NS	NS
Phospholipase D Family Member 3 (PLD3)	NS	NS	NS
Alpha-1-antitrypsin (SERPINA1)	NS	NS	NS
Extracellular superoxide dismutase [Cu-Zn] (SOD3)	NS	NS	NS
Fatty acid binding protein 5 (FABP5)	NS	NS	NS
Bifunctional purine biosynthesis protein PURH (ATIC)	NS	NS	NS
Phosphoglycerate kinase 1 (PGK1)	NS	NS	NS
Toll-interacting protein (TOLLIP)	NS	NS	NS
Endoplasmic reticulum chaperone BiP (HSPA5)	NS	NS	NS
Cytochrome C (CYCS)	NS	NS	NS
Urokinase-type plasminogen activator (PLAU)	NS	NS	NS
Mannan binding lectin serine peptidase 2 (MASP2)	NS	NS	NS

Protein (gene)	De novo PD vs control	iRBD vs control	OND vs control
Fibrinogen alpha chain (FGA)	NS	NS	NS
Complement C3 (C3)	NS	NS	NS
Vascular cell adhesion protein 1 (VCAM1)	NS	NS	NS
Moesin (MSN)	NS	NS	NS

The significantly different proteins between de novo PD and control are shown in Figure 6-17.

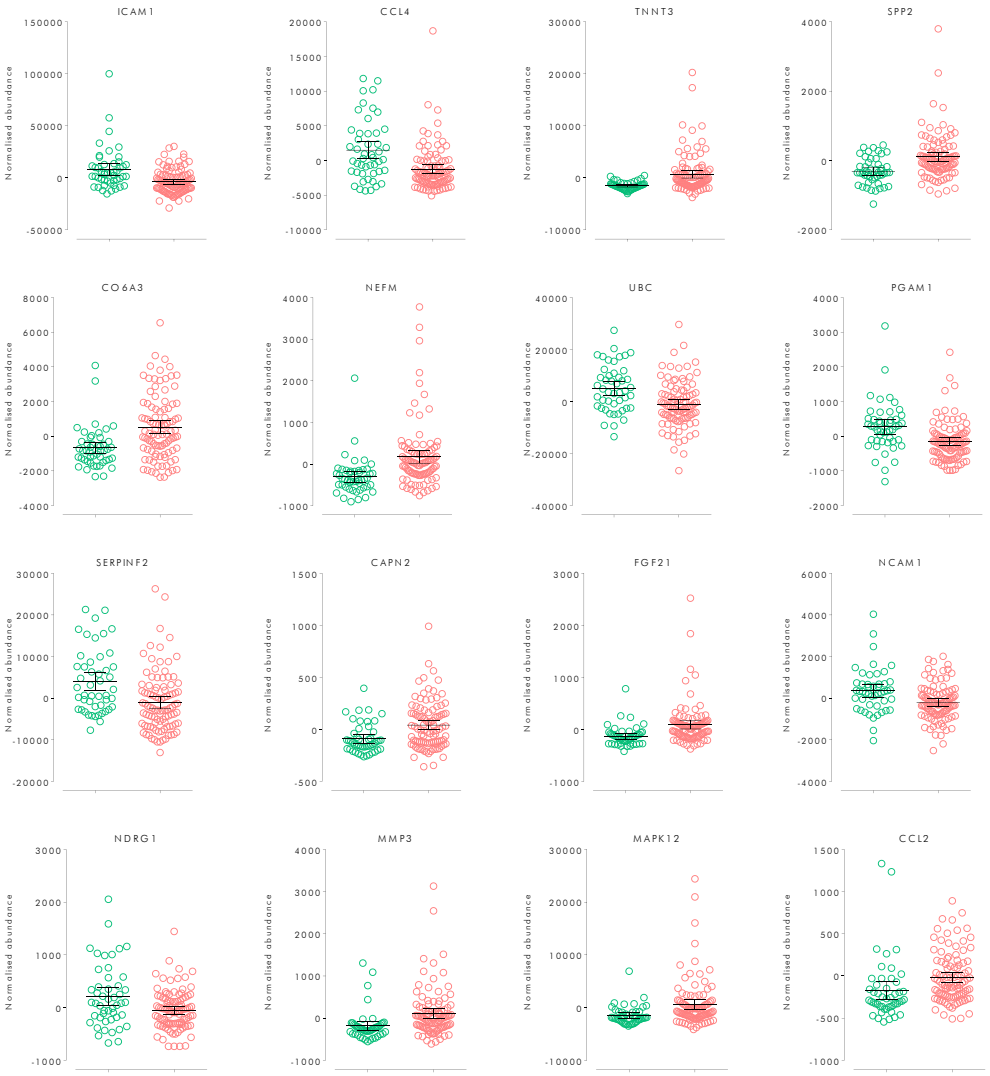


Figure 6-17. Significantly different proteins in the comparison of de novo PD and control from the targeted urine proteomics analysis after FDR correction. *The error bars represent the 95% confidence interval. In total, 23 proteins demonstrated a significant difference between the two groups. The plots are ordered according to level of significance, from most to least significant.* ■ *control, and* ■ *de novo PD.*

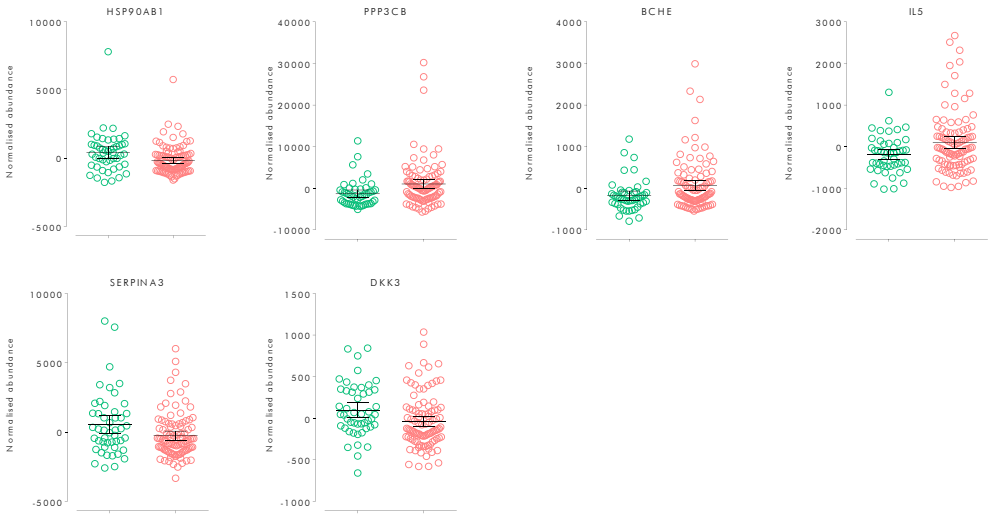


Figure 6-17. Significantly different proteins in the comparison of de novo PD and control from the targeted urine proteomics analysis after FDR correction. *The error bars represent the 95% confidence interval. In total, 23 proteins demonstrated a significant difference between the two groups. The plots are ordered according to level of significance, from most to least significant. (continued from previous page)*

The four proteins differentially expressed in the iRBD patients after FDR p-value adjustments are shown in Figure 6-18. Out of the four proteins, CCL2, UBC and SPP2 were also differentially expressed in the de novo PD patients and in the other neurological disorders group.

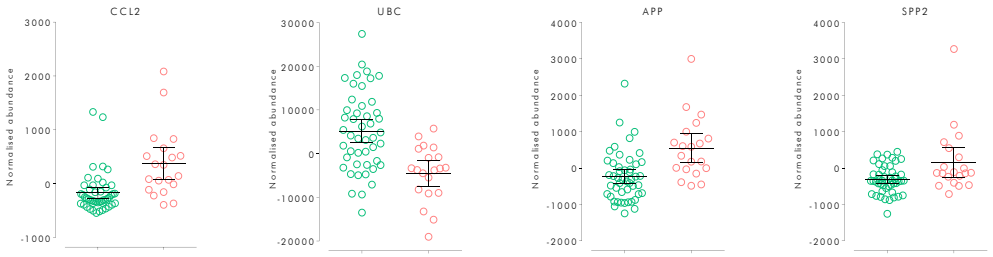


Figure 6-18. Significantly different proteins in the comparison of iRBD and control from the targeted urine proteomics analysis after FDR correction. *The error bars represent the 95% confidence interval. In total, four proteins demonstrated a significant difference between the two groups. The plots are ordered according to level of significance, most significant first.* ■ *control, and* ■ *iRBD.*

In the other neurological disorders group, 11 proteins were differentially expressed compared to control post-FDR adjustment of the p-values. Apart from CST3, all the proteins were also differentially expressed in the de novo PD patients. The protein expressions are shown in Figure 6-19.

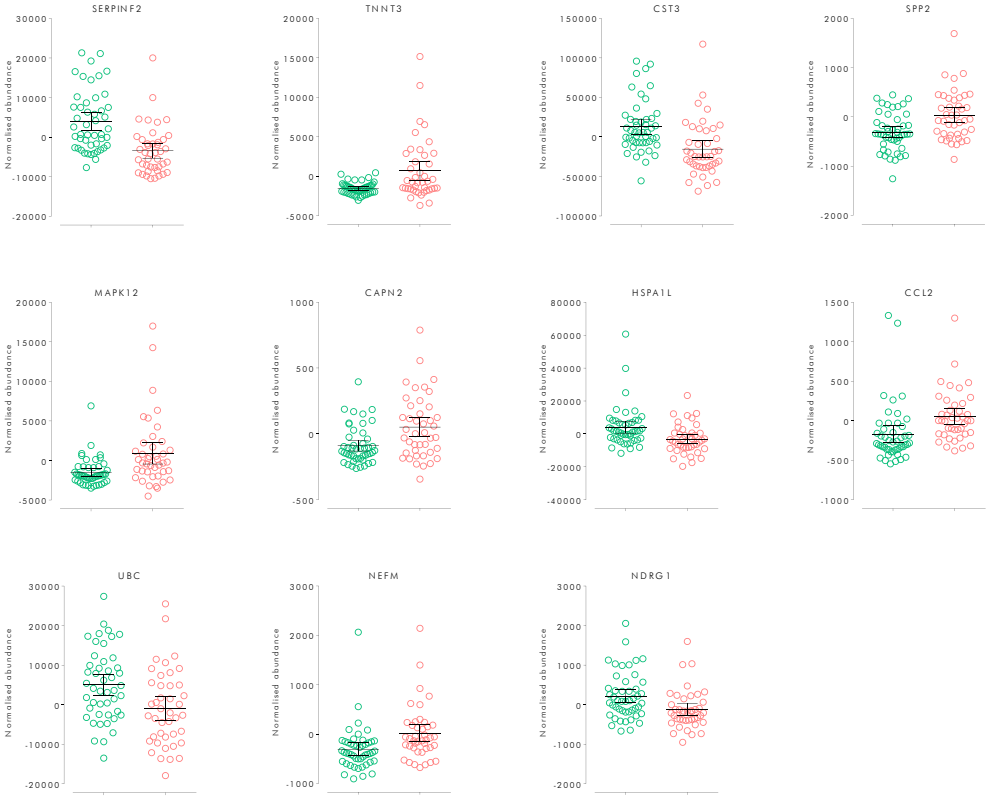


Figure 6-19. Significantly different proteins in the comparison between other, non-PD, neurological disorders, and control from the targeted urine proteomics analysis after FDR correction. *The error bars represent the 95% confidence interval. In total, 11 proteins demonstrated a significant difference between the two groups. The plots are ordered according to level of significance, most significant first.* ■ *control, and* ■ *other neurological disorders.*

Three of the measured proteins demonstrated a significant difference in all three disease groups; these were SPP2, UBC and CCL2 (Figure 6-20).

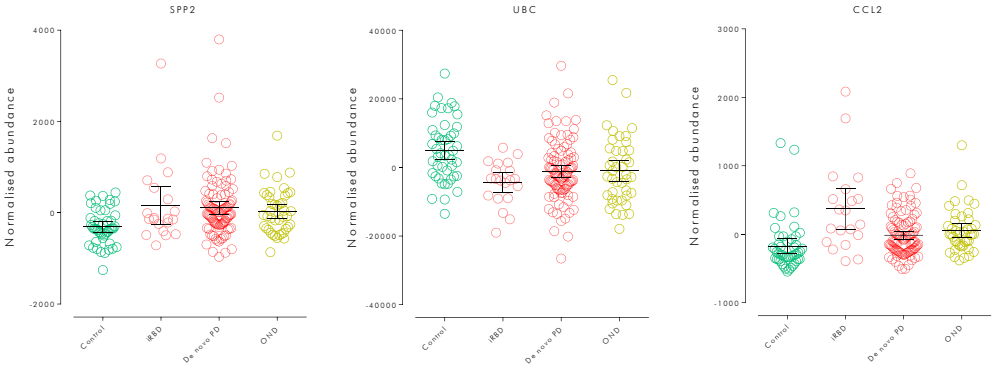


Figure 6-20. Proteins from the targeted study of urine which were differentially expressed in all three disease groups compared to control. *The error bars represent the 95% confidence interval*

The proteins that were differentially expressed in the three different groups were represented by a network created in Cytoscape version 3.8.0 [330]. The network allows for an accessible overview of the significant proteins in each group and further demonstrates which of the proteins were significantly expressed in more than one group, if they were up- or downregulated and their p-value significance. Figure 6-21 shows the proteins differentially expressed in de novo PD, iRBD and other neurological disorders. SERPINF2, NDRG1 and HSPA1L were significantly downregulated in both the de novo PD patients and the other neurological disorders group while TNNT3, CAPN2, NEFM and MAPK12 were significantly upregulated in both groups. UBC was significantly downregulated in all three disease groups, and CCL2 and SPP2 were significantly upregulated in the three disease groups. All other proteins were uniquely differentially expressed for only one of the conditions.

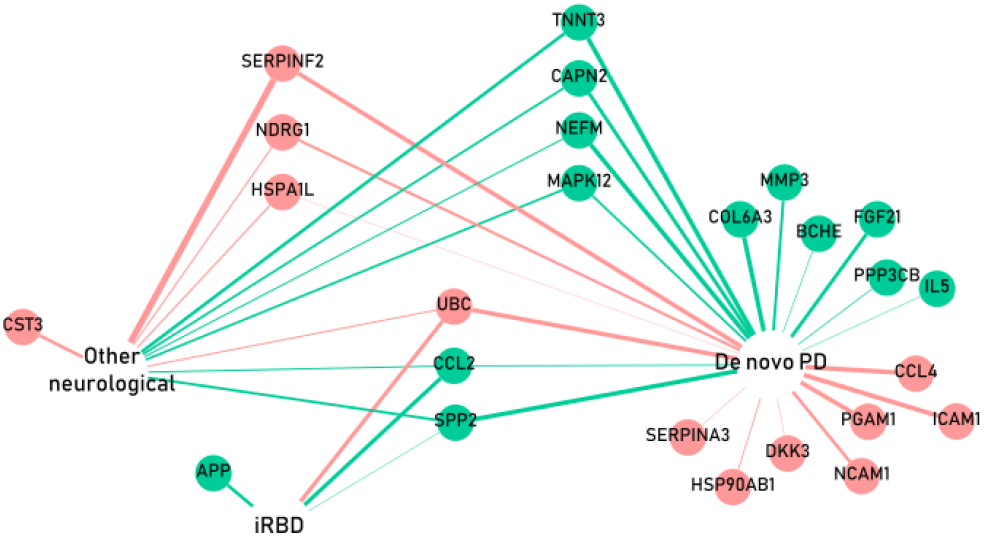


Figure 6-21. Network representation of the differentially expressed proteins in the groups de novo PD, iRBD and other neurological disorders compared to control. *The white circles represent the disease nodes. All proteins connected to a node were significantly differentially expressed in that group in relation to control. Pink edges denote downregulation of the proteins and green edges upregulation. The edge widths represent the p-value significances of the proteins in the different diseases, where wider edges mean more significant.* ■ *downregulated in disease group,* ■ *upregulated in disease group.*

In summary, a number of proteins were found differentially expressed in the disease groups post multiple testing correction; 23 proteins in the de novo PD patients, four in the iRBD patients and 11 in the other neurological disorders group. Out of these proteins, seven were found to overlap between de novo PD and other neurological disorders and three to overlap between all three disease groups.

6.3.2.4 *Comparison to the results from the discovery study*

Comparing the results from the targeted study with the discovery study, 16 of the proteins could be detected while six could not be detected reliably. Out of the 16 detected proteins, four were found significant in the targeted study with an expression matching the discovery study, these proteins were MAPK12, PPP3CB, CAPN2 and NDRG1. These proteins, apart from PPP3CB, were moreover differentially expressed also in the OND group. DKK3 was significantly different in both the discovery and targeted study, although upregulated in the discovery iPD patients and downregulated in the targeted study of de novo PD patients. Table 6-5 shows a summary of the results.

Table 6-5. Comparison the proteins discovered in the untargeted urine proteomics study and their expression in the targeted urine proteomics assay. *The table shows the significance level and direction of expression for the proteins in each study an if the discovery results were confirmed by the targeted study, where NS $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, ↑ upregulated, ↓ downregulated, ↔ no expression difference and ‡ expression discrepancy between the studies*

	Discovery study - iPD vs control		Targeted study - DNP vs control		Confirmed?
	Significance level	Expression in iPD	Significance level	Expression in de novo PD	
TUBA4A	*	iPD (↓)	NS	DNP (↔)	
MAPK12	**	iPD (↑)	**	DNP (↑) + OND (↑)	Y
APOE	*	iPD (↑)	NS	DNP (↔)	
FGA	*	iPD (↑)	NS	DNP (↔)	
HSPA8	**	iPD (↓)	NS	DNP (↔)	
PPP3CB	*	iPD (↑)	*	DNP (↑)	Y
PLAU	*	iPD (↑)	NS	DNP (↔)	
COL4A2	*	iPD (↓)	NS	DNP (↔)	
THY1	*	iPD (↑)		Not detected	
CYCS	*	iPD (↑)	NS	DNP (↔)	
CTHRC1	*	iPD (↓)		Not detected	
ATIC	*	iPD (↑)	NS	DNP (↔)	
CAPN2	*	iPD (↑)	**	DNP (↑) + OND (↑)	Y
DKK3	*	iPD (↑)	*	DNP (↓)	Y‡
EFNA5	*	iPD (↓)		Not detected	
ENDOU	**	iPD (↓)		Not detected	
HBE1	*	iPD (↓)		Not detected	
MASP2	**	iPD (↑)	NS	DNP (↔)	
MUC5B	*	iPD (↑)		Not detected	
NDRG1	*	iPD (↓)	**	DNP (↓) + OND (↓)	Y
SOD3	**	iPD (↑)	NS	DNP (↔)	
TOLLIP	*	iPD (↑)	NS	DNP (↔)	

6.3.2.5 *Literature studies and pathway/enrichment analysis*

6.3.2.5.1 *Literature review*

The proteins demonstrating a significant difference between de novo PD and control were reviewed for previously reported links to Parkinson's disease and neurodegeneration. The search engines Google Scholar and PubMed were used for disease associations and the portal GeneCards for the proteins' functions. Brief descriptions of any reported associations between the proteins and PD and/or neurodegeneration are presented in Table 6-6. A number of these proteins were also found in the targeted analysis of de novo PD plasma and are described in Chapter 5, section 5.3.2.3.

Table 6-6. Significantly different proteins and previously reported links to Parkinson's disease.

Protein	Description and reported link to Parkinson's disease
Intercellular adhesion molecule 1 (ICAM1)	ICAM was downregulated in the de novo PD patients ($p = 9.1E^{-4}$). ICAM1 is generally expressed on endothelial cells and is involved in inflammatory processes [331]. It has been proposed to play a role in the neuroinflammation associated with Parkinson's disease in a number of studies. One study suggests that ICAM1 interacts with T cells in the brain and modulates PD pathology [332]. ICAM1 has also been suggested to indicate astrocyte mediated inflammation in the brain, as proposed by a study of medically induced PD in monkeys, where ICAM1 was found upregulated in the substantia nigra [333].
C-C motif chemokine 4 (CCL4)	CCL4 was downregulated in the de novo PD patients ($p = 9.2E^{-4}$). CCL4 is a monokine involved in inflammation [334]. It has been suggested to be involved in neuroinflammation and a study of mice found the protein correlated with the age-related progression of amyloid-beta levels in the brain in a mouse study of Alzheimer's disease [335].
Troponin T3, fast skeletal type (TNNT3)	TNNT3 was upregulated in the de novo PD patients ($p = 1.1E^{-3}$) and in the OND group (3.4E-3). TNNT3 is involved in muscle contraction, where a process of binding Ca^{2+} and interaction with actin filaments leads to a muscle contraction [336]. No previously reported associations with PD or neurodegeneration could be found.
Secreted Phosphoprotein 2 (SPP2)	SPP2 was upregulated in the de novo PD patients ($p = 1.1E^{-3}$), iRBD ($p = 4.3E^{-2}$) and in the OND group ($p = 6.1E^{-3}$). SPP2 is part of the cystatin superfamily, a large group of proteins with a wide range of functions [337, 338]. As reported in Chapter 5, SPP2 is associated with Gerstmann-Straussler disease, a rare inherited prion disease [282] and a study of kidney function associated SPP2 with biomarkers of bone and mineral disease, and also found an inverse relationship with Wnt antagonists [283].
Collagen alpha-3(VI) chain (COL6A3)	COL6A3 was upregulated in the de novo PD patients ($p = 1.1E^{-3}$). COL6A3 is found in most connecting tissues and binds to extracellular matrix proteins where it organises matrix components [339]. Mutations in *COL6A3* have been implicated in muscular dystrophy [340]. A recently published genomic study of PD found that variants in the gene may increase the risk of developing PD [341].
Neurofilament medium polypeptide (NEFM)	NEFM was upregulated in the de novo PD patients ($p = 1.1E^{-3}$) and in the OND group ($p = 1.9E^{-2}$). NEFM makes up part of the axoskeleton and is involved in axonal maintenance [342]. NEFM is associated with a number of neurodegenerative conditions [343]. Mutations in the gene are additionally associated with familial PD [344]. It is used as a marker of neuronal damage.
Polyubiquitin-C (UBC)	UBC was downregulated in the de novo PD patients ($p = 1.1E^{-3}$), the OND group ($p = 1.9E^{-2}$) and in the iRBD patients ($p = 1.7E^{-3}$). UBC can be found in bound or free form; the bound form is involved in protein degradation, usually conjugated to lysine residues of target proteins whereas the free form can activate protein kinases [345]. UBC is involved in the ubiquitin-proteasome pathway and it has been suggested that alpha-synuclein is tagged by the ubiquitin system for degradation [346].

Protein	Description and reported link to Parkinson's disease
Phosphoglycerate mutase 1 (PGAM1)	PGAM1 was downregulated in the de novo PD patients ($p = 1.2E^{-3}$). PGAM1 is involved in glycolysis, where it catalyses the reversible transformation of 3-phosphoglycerate to 2-phosphoglycerate [347]. A study of ischemic damage in rabbits suggested that PGAM1 increases brain ATP and that the protein may reduce microglial activation and oxidative stress [348].
Alpha-2-antiplasmin (SERPINF2)	SERPINF2 was downregulated in the de novo PD patients ($p = 1.3E^{-3}$) and in the OND group ($p = 1.7E^{-4}$). As reported in Chapter 5, SERPINF2 is a major regulator of the clotting pathway where it acts as an inhibitor of plasmin. SERPINF2 was found upregulated in a recent proteomics study of platelet activation in AD patients [268]. Moreover, plasmin has been reported to cleave and degrade extracellular and aggregated alpha-synuclein [269].
Calpain-2 (CAPN2)	CAPN2 was upregulated in the de novo PD patients ($p = 2.4E^{-3}$) and in the OND group ($8.8E^{-3}$). CAPN2 is a calcium activated cysteine protease [349]. Calpain has been implicated in the pathogenesis of several neurodegenerative diseases. Calpain activity has been reported to be upregulated in AD as a consequence of malfunctioning Ca^{2+} homeostasis, leading to neurotoxicity and neuronal death [350]. In a study of medically PD-induced mice, calpain was increased in microglia, astrocytes and neurons, and neuronal cell death was moreover observed [351].
Fibroblast growth factor 21 (FGF21)	FGF21 was upregulated in the de novo PD patients ($p = 2.6E^{-3}$). FGF21 is involved in a number of molecular activities and is known to be a metabolic regulator and promoter of cell survival processes. It is also known to increase the uptake of glucose in adipose tissue [352]. FGF21 has been suggested to have protective functions against neurodegeneration by regulating the NF-κB and AMPKα/AKT pathways thereby reducing oxidative stress, neuroinflammation and protecting the mitochondria in neurons [353].
Neural cell adhesion molecule 1 (NCAM1)	NCAM1 was downregulated in the de novo PD patients ($p = 3.5E^{-3}$). NCAM1 is a member of the immunoglobulin family. It is involved in immune surveillance and interactions with the extracellular matrix [354].
N-myc downstream regulated 1 (NDRG1)	NDRG1 was downregulated in the de novo PD patients ($p = 5.3E^{-3}$) and in the OND group ($p = 2.0E^{-2}$). NDRG1 is involved in stress responses and is necessary for p53-induced apoptosis [355]. NDRG1 is implicated in Charcot-Marie-Tooth disease, a hereditary motor and sensory neuropathy. Mutations in the gene are known to lead to demyelination [356]. A computational study of PD risk genes and transcription factors identified NDRG1 as a hit [357].
Matrix metallopeptidase 3 (MMP3)	MMP3 was upregulated in the de novo PD patients ($p = 5.3E^{-3}$). MMP3 is involved in breaking down molecules from the extracellular matrix and bioactive compounds [358]. It has been implicated in neurodegenerative disease, activating microglia [245]. MMP3 is also suggested to contribute to dopaminergic neuronal death mediated through oxidative stress and to have the capability of disrupting the blood-brain barrier under certain conditions [246, 247].
Mitogen-activated protein kinase 12 (MAPK12)	MAPK12 was upregulated in the de novo PD patients ($p = 9.0E^{-3}$) and in the OND group ($p = 8.4E^{-3}$). MAPK12 is involved in the MAP kinase signal transduction pathway and the transduction of extracellular signals [359]. MAPK signalling has been suggested to play a role in the pathological processes of PD, including neuroinflammation, oxidative stress and neuronal death [360].
C-C motif chemokine 2 (CCL2)	CCL2 was upregulated in all the disease groups (de novo PD $p = 2.4E^{-2}$, iRBD $p = 1.7E^{-3}$, OND $p = 1.9E^{-2}$). CCL2 is a chemokine involved in immune regulation and inflammatory processes which bring upon a strong response and mobilise intracellular Ca^{2+} [361]. CCL2 has been suggested to contribute to the neuroinflammation observed in PD and to be involved in mediating neurodegeneration [362]. Variants in the gene have been associated with PD [363].
Heat shock protein HSP 90-beta (HSP90AB1)	HSP90AB1 was downregulated in the de novo PD patients ($p = 2.4E^{-2}$). HSP90AB1 is a member of the HSP 90 family, proteins involved in protein folding, degradation and signal transduction [364]. A cell study of PD found HSP90AB1 upregulated after poisoning the cells with MPP^+ [365]. HSP90 has been identified as a chaperone to alpha-synuclein and suggested to modulate its assembly [366, 367].
Serine/threonine-protein phosphatase 2B catalytic subunit (PPP3CB)	PPP3CB was upregulated in the de novo PD patients ($p = 3.0E^{-2}$). As reported in Chapter 5, PPP3CB makes part of the calcineurin complex and is also component of the Wnt/Ca^{2+} pathway [261, 262]. PPP3CB was identified as a risk gene for AD in microarray studies in the early 2000s [263]. Calcineurin has been proposed to increase in response to accumulation of alpha-synuclein and to trigger both protective and toxic responses to maintain neuronal Ca^{2+} homeostasis [264].
Cholinesterase (BCHE)	BCHE was upregulated in the de novo PD patients ($p = 3.6E^{-2}$). BCHE is an esterase involved in the detoxification of a number of compounds [194]. High levels of BCHE are

Protein	Description and reported link to Parkinson's disease
	linked to lower risk of mortality [172] and low levels are associated with cardiovascular risk, mortality and systemic low-grade inflammation [195]. BCHE activity was found decreased in a study of serum from PD patients [368].
Interleukin-5 (IL5)	IL5 was upregulated in the de novo PD patients ($p = 4.3E^{-2}$). IL5 is a cytokine involved in mediating immune response and can differentiate B-cells to immunoglobulin secreting cells [369]. It has been reported that the neuroinflammation in PD causes an increased release of inflammatory mediators, including IL5 [370].
Alpha-1-antichymotrypsin (SERPINA3)	SERPINA3 was downregulated in the de novo PD patients ($p = 4.4E^{-2}$). As reported in Chapter 5, SERPINA3's major target is cathepsin G although it can also inhibit other serine proteases. SERPINA3 has been associated with AD and proposed to mediate amyloid-beta clearance [265]. SERPINA3 was found upregulated in studies of prion diseases and progressive MS [266, 267].
Dickkopf 3 (DKK3)	DKK3 was downregulated in the de novo PD patients ($p = 4.4E^{-2}$). As reported in Chapter 5, DKK3 is a glycoprotein belonging to the Dickkopf family, the majority of which are antagonists of the Wnt signalling pathway, although DKK3 is a modulator rather than an antagonist. DKK3 has been seen downregulated in many cancer studies and was recently proposed to have a neuroprotective role [248]. It has been related to Alzheimer's disease in several studies and furthermore proposed to positively correlate with increased age [249]. Interestingly, a mouse study found that DKK3 may protect dopaminergic neurons and proposed that DKK3 has potential as a pharmacological target for treatment of neurodegeneration [259]. Another study of mice and pluripotent stem cells showed that DKK3 is necessary for correct differentiation and survival of dopaminergic neurons [260].
Heat shock 70 kDa protein 1-like (HSPA1L)	HSPA1L was downregulated in the de novo PD patients ($p = 5.0E^{-2}$) and in the OND group ($p = 1.7E^{-2}$). As reported in Chapter 5. HSPA1L is a heat shock protein involved in the quality control system of the cell. Among its functions are folding and transport of newly synthesised polypeptides and re-folding or destruction of misfolded proteins [279, 280]. In a publication from 2005, overexpression of HSPA1L was suggested to reduce neurodegenerative symptoms in Parkinson's disease, Huntington's chorea and spinocerebellar ataxia [281]. May be involved in translocation of PRKN [371].

6.3.2.5.2 *Protein-protein interactions*

Examining known interactions between the significant proteins, the online network tool STRING, version 11.0 was utilised [289]. The protein-protein interaction enrichment p-value was 0.03, thus demonstrating that there was a significant number of known interactions between the proteins. Twelve of the proteins were connected in the protein-protein interaction network while ten were not (Figure 6-22).

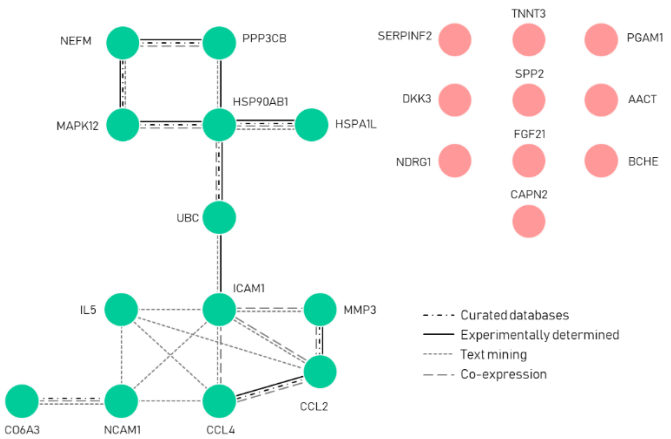


Figure 6-22. Protein-to-protein interaction network from STRING. *The enrichment p-value was 0.03, demonstrating that there is a significant number of interactions between the proteins.* ■ *connected in network, and* ■ *not connected in network.*

6.3.2.5.3 *Pathway and enrichment analysis*

As previously stated, the acquirable information from pathway analyses of targeted data is limited due to the nature of the data. The targeted proteomics data are acquired in a biased mode since the monitored proteins have been pre-selected – this is in contrast to untargeted proteomic data where the data are collected in an unbiased manner without any prior selection - and thus, the results from a pathway analysis will reflect this pre-selection of proteins. Pathway analysis of targeted data can however provide insights into enriched functions, processes, and the proteins' interactions with each other.

The proteins differentially expressed between de novo PD and control were analysed by DAVID Bioinformatics Resources 6.8 [287, 288]. The analysis resulted in eight KEGG pathways - *TNF signalling pathway, Amyotrophic lateral sclerosis, Influenza A, NOD-like receptor signalling pathway, Rheumatoid arthritis, MAPK signalling pathway, T cell receptor signalling pathway* and *Protein processing in endoplasmic reticulum*. Although nominally significant, none of the pathways was significant after multiple testing correction. Among the highlighted gene ontology (GO) molecular functions were *serine-type endopeptidase inhibitor activity, endopeptidase inhibitor activity, heat shock protein binding, protein binding* and *chemokine activity*, all nominally significant but not passing multiple testing correction. In the GO biological processes analysis, four processes were suggested to be enriched; *cellular response to drug* (FDR p = $6.3E^{-4}$), *positive regulation of ERK1 and ERK2 cascade* (FDR p = 0.01), *MAPK cascade* (FDR p = 0.04) and *negative regulation of endopeptidase activity* (FDR p = 0.04).

In summary, the literature review of the significantly different proteins demonstrated that several had a previously reported link to Parkinson's disease. An analysis of protein-protein interactions showed that there was a significant number of interactions between the proteins. Pathway analysis did not identify any significantly enriched pathways but suggested that the GO processes *cellular response to drug, positive regulation of ERK1 and ERK2 cascade, MAPK cascade* and *negative regulation of endopeptidase activity* were enriched.

6.3.2.6 *Prediction and machine learning models*

Univariate analysis showed much promise in identifying biomarkers that could be used to distinguish PD from control. Therefore, we attempted to use a machine learning approach to develop models where a panel of proteins could be used together to increase the discriminating power and help to further differentiate PD from controls. We also wanted to explore if there was a panel of proteins that would allow us to predict which of the iRBD

patients that were more likely to develop PD in the future, as iRBD is a strong risk factor for Parkinson's disease.

6.3.2.6.1 *Receiver operating characteristic curve analysis*

To assess the classification ability of the individual proteins, a receiver operating characteristic (ROC) curve was generated from the de novo PD and control samples utilising the web based tool easyROC [290]. A ROC curve plots the true positive rate (sensitivity) versus false positive rate (1 - specificity) at varied threshold settings [291]. Figure 6-23 shows the ROC curves of the significantly different proteins when comparing de novo PD and control, split over two graphs for the proteins upregulated in the de novo PD group and for the proteins downregulated in the de novo PD group.

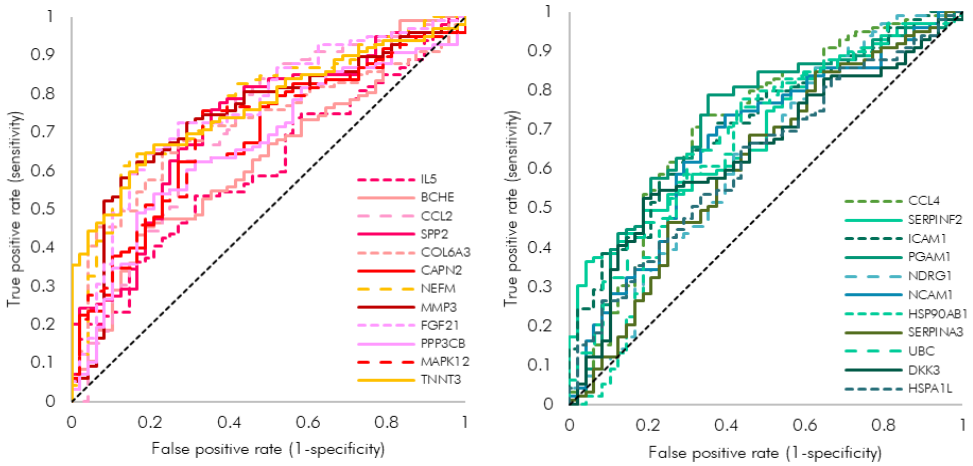


Figure 6-23. ROC curves of the significantly different proteins in the comparison of de novo PD and control from the targeted urine proteomics analysis. *On the left: the proteins upregulated in the de novo PD group, and on the right: The proteins downregulated in the de novo PD group. The dashed diagonal lines represent an area under the curve of 0.5, a value at which there is no discrimination and the samples would be randomly classified as belonging to either group.*

Areas under the curve (AUC) were extracted for the most promising proteins and are shown in Figure 6-24. The downregulated proteins with strongest expression were PGAM1, ICAM1 and CCL4, all attributed with AUCs of more than 0.7. In the upregulated proteins, the strongest expressions were found in NEFM, TNNT3, FGF21, MMMP3, SPP2, CO6A2 and CCL2, also these with AUCs greater than 0.7.

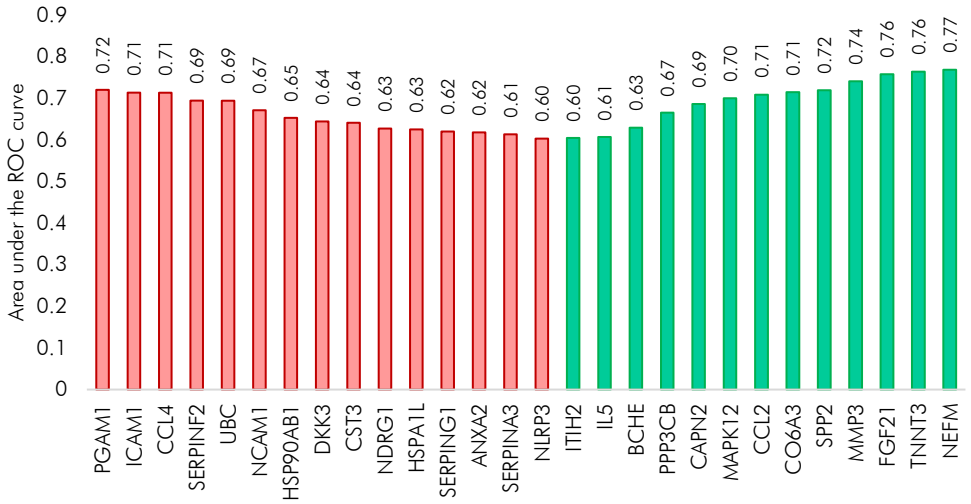


Figure 6-24. Area under the curve from the ROC analysis of targeted urine proteomics of de novo PD patients and controls. *The plot shows all significant proteins with an AUC > 0.6. The green bars denote proteins upregulated in de novo PD and the red bars denote proteins downregulated in de novo PD. The AUC value is displayed above each bar.*

6.3.2.6.2 *Machine learning for classification and prediction*

The strategy employed in Chapter 5, section 5.3.2.6 was also utilised in this analysis. Three different classifier algorithms were evaluated to determine which could best discriminate between PD and controls: linear discriminant analysis (LDA) with singular value decomposition (SVD) as solver, support vector machine (SVM), and Ridge classifier.

Initial cross-validation and comparison of machine learning models. The targeted proteomics dataset from the analysis of urine from controls, PD patients, iRBD patients and individuals with other neurological disorders was filtered to contain only the PD patients and the control samples. As an initial data quality overview, the data were divided into five groups (test sets), selecting different samples for each group while maintaining the proportion of de novo PD and controls, using the function StratifiedKfold from Scikit Learn version 0.24.2 with shuffled values and fixed random state. Each cross-validation iteration was performed by building models from the data remaining after the test set (called the training set) and predicting the classes of the samples in the test set. Figure 6-25 shows the sample distribution in the five cross validation iterations.

Figure 6-25. Cross-validation iteration groups in the targeted urine proteomics de novo PD and control samples. *The de novo PD and control samples are divided by the dashed vertical line, with de novo PD samples to the left and control samples to the right. In each of the five CV groups #1 - #5, the samples selected for the test set are coloured in green, while the remaining samples, the training set, are coloured in red. The test samples from all five CV groups together make up the full set of samples.* ■ *training set,* ■ *test set.*

The cross-validation test sets were predicted in models created from the training sets for the three different algorithms LDA, SVM and Ridge classifier. The model scores for the training and prediction sets were extracted and are shown in Table 6-7. As demonstrated by the CV scores, all models performed well with training model scores consistently above 0.8. The test scores, the prediction of the samples extracted for cross validation, were likewise attributed with scores of good magnitudes.

The initial quality check signified that it was possible to build well-performing models regardless of how the data were split for training and prediction.

Table 6-7. Cross validation summary of linear discriminant analysis, support vector machine and Ridge classifier from the five-fold split of the de novo PD and control urine samples measured by targeted proteomics. *The individual scores from each iteration fitting the training set and test set and the average scores are presented . SD = standard deviation.*

	Linear discriminant analysis		Support vector machine		Ridge classifier	
CV iteration	Training score	Test score	Training score	Test score	Training score	Test score
# 01	0.87	0.73	0.93	0.73	1.00	0.83
# 02	0.87	0.87	0.97	0.80	0.99	0.90
# 03	0.85	0.90	0.90	0.77	1.00	0.87
# 04	0.87	0.86	0.92	0.93	0.99	0.86
# 05	0.88	0.82	0.96	0.57	1.00	0.75
Average ± SD	0.87 (± 0.01)	0.84 (± 0.06)	0.94 (± 0.03)	0.76 (± 0.13)	1.00 (± 0.005)	0.84 (± 0.06)

Model building and variable selection. For model building and refinement, the PD/control samples data were split into two equally large parts, each containing the same proportion of control and de novo PD samples, using the function train-test split (Scikit Learn). One set was used for training and one for prediction. The optimal numbers of features to include in the LDA and SVM models were evaluated in the training set by

recursive feature elimination (RFECV, Scikit Learn). The ideal numbers of features selected by RFECV for the LDA algorithm were eight (NEFM, FGF21, CAPN2, NDRG1, CO6A3, CCL2, TEK and IL1B) and for the SVM algorithm 30 (PGAM1, NEFM, SPP2, CAPN2, NDRG1, CO6A3, MMP3, MAPK12, PPP3CB, CCL2, IL5, BCHE, TEK, SERPINA3, CCL17, HSP7C, DKK3, NLRP3, IL1B, A2M, APP, ANXA2, TOLLIP, PGK1, A1AT, CH60, DCXR, ATIC, C3 and MOES). The beta-coefficients of the proteins in the Ridge classifier model are illustrated in Figure 6-26. Many of the coefficients had values close to zero, thus demonstrating that their influence on separating the classes was limited.

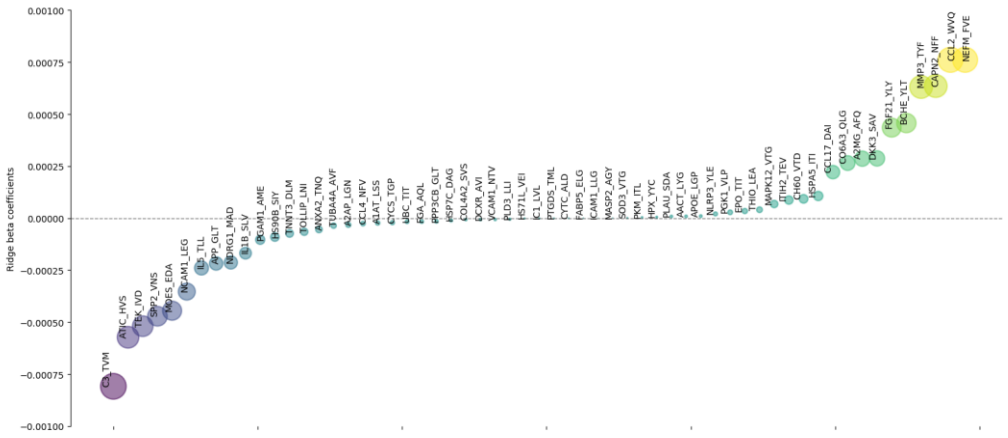


Figure 6-26. Ridge classifier beta coefficients in the de novo PD versus control training model based on the targeted urine proteomics. *A large proportion of the proteins have limited influence on the model as shown by the size of their beta coefficients.*

To reduce the complexity of the Ridge classifier model and to curtail the risk of overfitting, a new model was built, containing only the top 20 variables with the most influential beta coefficients (C3, NEFM, CCL2, CAPN2, MMP3, ATIC, TEK, SPP2, BCHE, MOES, FGF21, NCAM1, DKK3, A2M, CO6A3, IL5, CCL17, APP and NDRG1).

Prediction of de novo PD and healthy controls. The test dataset was predicted in the three final training models and resulted in the following outcomes: 85% of the total samples were correctly predicted in the LDA model, 79% of the total samples were correctly predicted in the SVM model and 82% of the total samples were correctly predicted in the Ridge classifier model. Figure 6-27 illustrates the individually predicted classes in the three models and the actual sample classes. In summary, the prediction models had the following characteristics:

- Linear discriminant analysis: 91.8% sensitivity, 72.0% specificity, 85.1% accuracy
- Support vector machine: 75.5% sensitivity, 88.0% specificity, 79.7% accuracy
- Ridge classifier: 91.8% sensitivity, 64.0% specificity, 82.4 % accuracy

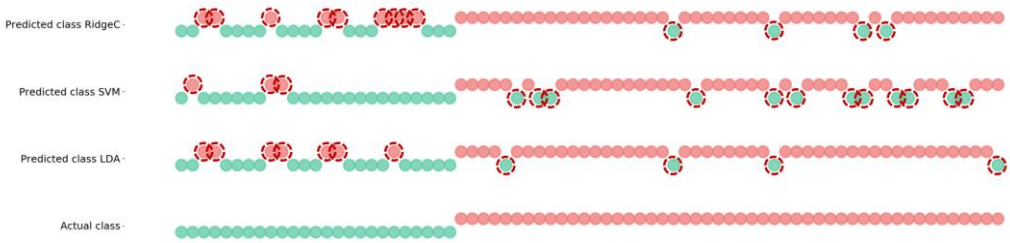


Figure 6-27. Prediction results of the test set in the LDA, SVM and Ridge classifier models. *85% of the samples were correctly predicted in the LDA model, 79% in the SVM model and 82% in the Ridge classifier model, resulting in a respective model accuracy of 85.1%, 79.7% and 82.4%. The samples which were incorrectly predicted are circled in red. The last row (labelled "Actual class") shows the actual classes of the samples.* ■ *predicted as control, and* ■ *predicted as PD.*

It was concluded that the LDA model had the most favourable characteristics with highest sensitivity, and overall accuracy.

Specificity testing by predicting OND and iRBD samples. Testing the specificity of the LDA model further, the OND and iRBD samples were predicted. The prediction resulted in 85.7% of the OND samples and 50% of the iRBD samples being predicted as de novo PD (Figure 6-28).

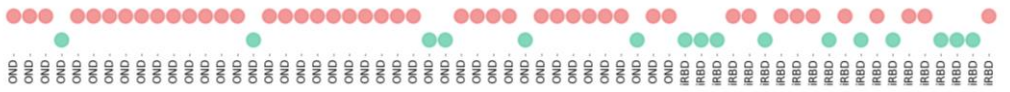


Figure 6-28. Prediction of other neurological disorders and iRBD samples in the LDA model. *The prediction resulted in 85.7% of OND samples and 50% of iRBD samples predicted as de novo PD.* ■ *sample predicted as de novo PD,* ■ *sample predicted as control.*

The high rate of OND samples being predicted as de novo PD echoes what the univariate analysis demonstrated - that many of the differentially expressed proteins are shared between the two groups. The LDA model is consequently not well-suited to separate between PD and the other neurological disorders included in the heterogeneous OND group, but rather between control and neurological disorders, including PD.

6.3.3 *Summary and conclusions from the targeted study*

In the targeted urine proteomics study, an augmented MRM assay was applied to a cohort of 211 samples from de novo PD patients, controls, iRBD patients, and patients with other non-PD neurological disorders. Five of the proteins selected from the discovery phase were differentially expressed between de novo PD and controls, these proteins were

MAPK12, PPP3CB, CAPN2, NDRG1 and DKK3. All apart from DKK3 had an expression matching the discovery study. Another 18 proteins from the targeted assay, not from the urine discovery phase, were differentially expressed between de novo PD patients and controls. Literature studies demonstrated that many of these proteins had previously been linked to, or had functions relevant for, PD pathology. A discriminant LDA machine learning model, based on the expression of eight proteins, could separate de novo PD patients from controls with a high level of accuracy. However, the LDA model was unable to distinguish between PD and other neurological disorders.

In conclusion, the study of Parkinson's disease using urine proteomics showed promising results and allowed us to measure several proteins that we could not detect in plasma. The use of urine as a screening tool would be extremely beneficial due to its non-invasive collection. Our panel of biomarkers can indeed distinguish between PD and control with good accuracy in this set of samples but would need to be evaluated in a much larger study to test its robustness and utility. If successful, the panel could be utilised as an entry screen to help determine which patients would need further and more specific clinical assessment.

6.4 DISCUSSION

Urine from idiopathic PD patients, symptomatic and asymptomatic LRRK2 carriers and controls were investigated in a discovery, bottom-up proteomics study, aimed at finding urinary biomarkers for Parkinson's disease. The study identified many targets from which a selection was developed into a targeted test. The targeted test we developed also included several inflammation-related proteins from the literature and blood-based putative PD and ageing biomarkers identified in studies of PD and centenarians. Four sample groups were analysed by the targeted assay: urine from treatment-naïve de novo PD patients, patients with iRBD, controls and a group of non-PD neurological disorders. The targeted assay included 127 proteins out of which 23 were differentially expressed between de novo PD and control, four between iRBD and control, and 11 between OND and control. Seven of the proteins were moreover found to overlap between de novo PD and other neurological disorders and three to overlap between all three disease groups. Five of the proteins from the PD discovery study demonstrated a significant difference between de novo PD and control. An LDA machine learning model demonstrated that it was possible to classify samples as de novo PD or control with 91.8% sensitivity and 72% specificity based on the expression of NEFM, FGF21, CAPN2, NDRG1, CO6A3, CCL2, TEK and IL1B. The model was tested for PD specificity by predicting the OND group but

demonstrated limited unique applicability for this as the majority of the samples were predicted in the de novo PD class. Still, the model has value as it is capable of differentiating between control and neurological disorders in urine.

The reoccurring functions and processes associated with the significantly, differentially expressed proteins from the targeted analysis included:

- Interactions with extracellular matrix
- MAPK signalling
- Inflammation/immune system

The extracellular matrix (ECM) surrounds cells in a tissue. It provides physical structure and is involved in a range of biochemical processes, including differentiation and homeostasis. The ECM is largely made up collagen, enzymes and glycoproteins [372]. Changes in the brain's ECM have been associated with neurodegeneration and cell-ECM interactions and have been found to regulate transcription factors promoting apoptosis. ECM matrix metallopeptidases have been implicated in increasing the inflammatory response of microglia by activation of chemokines [373]. Specifically in Parkinson's disease, the ECM constituents glycosaminoglycans (GAGs) have been found in Lewy bodies, thus suggesting a role in the accumulation of alpha-synuclein [374]. It has been proposed that the GAGs may impede the degradation of alpha-synuclein by binding to proteases that otherwise might have been purposed to degrade it [375, 376].

Mitogen-activated protein kinase (MAPK) cascades are involved in a number of cellular processes, including differentiation, inflammatory responses and apoptosis [377]. MAPK signalling has been implicated in neurodegenerative disease in a number of studies, primarily suggested to contribute to neuroinflammation induced by astrocytes and microglia [378].

In essence, the proteins detected in the targeted urine proteomics study point towards increased levels of inflammation in the de novo PD patients. Neuroinflammation is a well-known characteristic of the PD pathology where it is suggested to be part of a vicious and self-feeding cycle. As mentioned in the previous PD plasma proteomics chapter, neuroinflammation is a protective response aiming to restore homeostasis in the central nervous system. The response includes a range of inflammatory functions working to restore/maintain homeostasis, clearing damaged tissue and targeting threats. However, the protective system can start producing an uncontrolled response leading to the destruction of healthy tissue and a state of chronic inflammation which eventually leads to

necrosis of neurons and glial cells [157]. In animal models of Parkinson's disease, neuroinflammation has been demonstrated to exert an important role in the disease progression [379]. Links between neuroinflammation and other pathological PD mechanisms, including mitochondrial dysfunction and oxidative stress, have also been found [380]. There are even hypotheses suggesting that neuroinflammation is the cause of neurodegeneration [362]. Whether or not this will prove to be true remains to be seen, but what is undisputable is that neuroinflammation plays a crucially important role in the PD pathology.

In this study, we found a number of inflammatory markers upregulated in the de novo PD patients. The most prominent being MAPK12, CCL2, MMP3 and IL5. MAPK12 is one of the p38 mitogen-activated protein kinases. The p38 pathway is activated by extracellular stimuli such as heat, osmotic stress, growth factors or inflammatory cytokines and there is a strong association between p38 and inflammation. The activated pathway brings upon the production of proinflammatory cytokines and is also involved in apoptosis [381]. Studies have shown that CCL2 levels increase during neuroinflammation [382] and, moreover, that CCL2 can activate the p38 pathway [383]. It has also been demonstrated that dying neurons release, among others, MMP3 causing microglia to secrete toxic compounds thus harming neighbouring cells [384]. IL5 has been reported to be released by inflamed dopaminergic neurons [385]. Put together, this paints a picture of a self-amplifying loop of inflammation in the de novo PD patients.

As described in the plasma proteomics PD chapter, DKK3 is believed to activate the Wnt/β-catenin signalling pathway and provide protection for dopaminergic neurons. It has been suggested that downregulation of Wnt signalling promotes dysfunction and/or death of dopaminergic neurons. Restoration of these neurons was shown in a mouse study where β-catenin was activated in situ [313]. A mouse study showed re-expression of Wnt1 and repair of dopaminergic neurons after neural stem cells were transplanted to the substantia nigra of medically PD-induced mice [314]. The Wnt signalling pathway may thus have the potential of restoring dopaminergic neurons' function [311]. Here, we are observing downregulation of DKK3 thus indicating reduced dopaminergic neuron protection through the Wnt signalling pathway.

Neurofilament medium is part of the Type IV intermediate filament family, also containing NEFL and NEFH. The neurofilaments are found in neurons and are part of the cytoskeleton. When damage occurs in the central nervous system, neurofilaments are released from axons and migrate to CSF from which they later circulate into blood [386].

For this reason, neurofilaments are commonly used as indicators of neuronal damage. Neurofilaments have been found elevated in Parkinson's disease and Parkinsonian disorders throughout a number of studies [344]. The observed upregulation of NEFM in our study may therefore reflect the central nervous system damage occurring in the de novo PD patients.

Figure 6-29 gives a summary of the proposed ongoing mechanisms in the de novo PD patients based on the urinary protein expression from the targeted study. In conclusion, these mechanisms consist of an inflammatory loop - continuously fuelling neuroinflammation, upregulation of NEFM - indicating neuronal damage, and finally downregulation of DKK3 - proposed to reduce the protection of dopaminergic neurons.

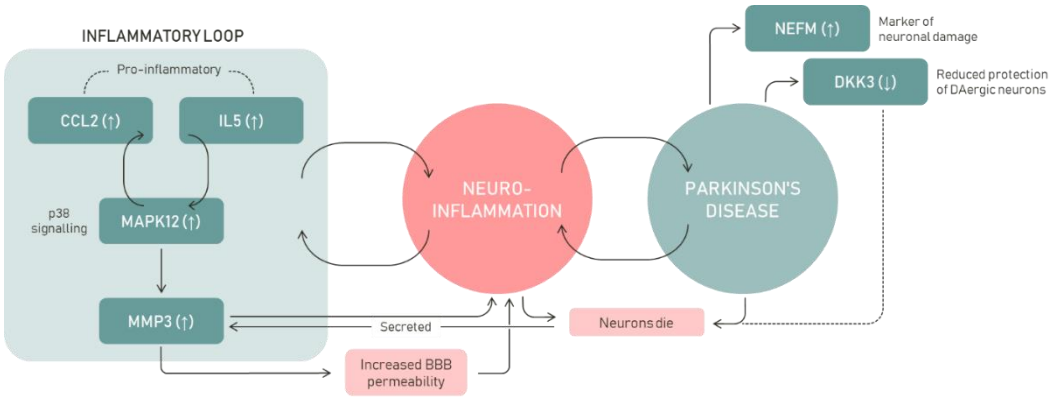


Figure 6-29. Proposed inflammatory mechanisms based on the protein expression observed in the targeted study of de novo PD urine. *The proteins are denoted by their gene names and the arrows indicate the direction of expression in the de novo PD patients compared to control. An inflammatory loop is suggested, with MAPK12 indicating activation of the p38 pathway, causing the release of CCL2 and IL5, which then feedback and continuously activate the inflammatory p38 pathway. MMP3 is secreted as neurons die, this increases the blood-brain barrier permeability and causes a locally toxic environment, thus increasing neuroinflammation and causing further neurons to die. The downregulation of DKK3 reduces the protection of dopaminergic neurons, thereby potentially accelerating the neuronal cell death. NEFM indicates neuronal damage.*

Lastly, there are a few limitations to consider in this experiment. Firstly, the sample size should ideally have been larger in the discovery phase to allow for more certainty in the interpretation of the results. However, it was argued in the discovery study that a deeper coverage, through extensive fractionation, of the urinary proteins was desired and therefore the number of samples had to be limited to curtail the risk of severe and irreparable instrumental drift. A second consideration is that the discovery study was performed on advanced PD patients while the targeted study's PD subjects had just been diagnosed, there may thus be discrepancies due to treatment or disease duration that are unaccounted for. Moreover, there were no LRRK2 mutation carriers in the targeted study, therefore the putative LRRK2-specific biomarkers could not be evaluated. The outcomes of the iRBD patients are not yet known, thus we do not know which of the iRBD patients

will develop Parkinson's disease in the future. This would have provided a valuable disease progression angle to the study and might also have allowed us to find a pre-symptomatic protein expression pattern.

As a final remark, although our study showed highly interesting and promising results, our findings will need to be evaluated in a larger study to test the reproducibility and to verify the protein expression we observed. It is however clear that urine holds merit as a biofluid in discovery studies of neurodegeneration and can be useful for exploring putative disease mechanisms.

How do the different proteomic studies relate to each other and what additional information can be gained by comparing their results?

7

7.1 CORRELATION OF PROTEIN EXPRESSION IN PLASMA AND URINE SAMPLES FROM PD PATIENTS

The plasma and urine samples which were included in the verification phases in Chapters 5 and 6 were paired, meaning that the individuals in the study provided both a plasma and a urine sample. These samples consisted of newly diagnosed Parkinson's disease patients, patients with iRBD, healthy controls, and a heterogenous group of individuals with other neurological disorders. The samples were analysed by the same explorative targeted assay and therefore provide an opportunity to directly compare the expression of the analysed proteins in both matrices.

Reiterating the conclusions from the targeted de novo PD urine and plasma studies – in the plasma study, the observed protein expression suggested that the de novo PD patients had increased inflammatory response, increased unfolded protein response, and disrupted Wnt signalling. In the urine study – the results suggested a self-amplifying inflammatory cycle and, also here, disrupted Wnt signalling in the de novo PD patients. Comparing the proteins detected in respective sample matrix of our targeted studies - 32 proteins were detected in plasma, and 59 proteins in urine. Out of these proteins, 26 overlapped between the matrices, leaving six proteins uniquely detected in plasma and 33 proteins uniquely detected in urine. The masses of the proteins in Dalton (Da) were extracted from UniProt [387]. The proteins are sorted by molecular weights and as can be noted, there is no systematic difference in the molecular weights of the proteins detected in the different matrices. Similar findings have been reported in studies of the molecular weights of proteins from plasma and urine from individuals without renal dysfunction [388]. Table 7-1 shows the proteins discovered uniquely in each matrix, and the proteins discovered in both plasma and urine.

Table 7-1. The proteins uniquely detected in the targeted analysis of de novo PD plasma and urine samples, and the proteins detected in both matrices. *The proteins are denoted by gene names and the molecular masses (Da) are given in parentheses after each protein.*

Unique in plasma	Common between plasma and urine		Unique in urine	
SAA1 (13532)	FABP5 (15164)	BCHE (68418)	CCL4 (10212)	PLAU (48507)
ADIPOQ (26414)	CST3 (15799)	HSPA1L (70375)	CCL17 (10507)	MMP3 (53977)
GRN (63544)	PTGDS (21029)	HSPA5 (72333)	CCL2 (11025)	HSPD1 (61055)
SELE (66655)	SPP2 (24338)	MASP2 (75702)	TXN (11737)	ATIC (64616)
PRG4 (151061)	SOD3 (25851)	VCAM1 (81276)	CYCS (11749)	MSN (67820)
LMO7 (192696)	APOE (36154)	FGA (94973)	IL5 (15238)	HSPA8 (70898)
	DKK3 (38390)	ITIH2 (106463)	EPO (21307)	UBC (77039)
	PGK1 (44615)	A2M (163291)	FGF21 (22300)	CAPN2 (79995)

Unique in plasma	Common between plasma and urine		Unique in urine	
	SERPINA3 (47651)	C3 (187148)	DCXR (25913)	HSP90AB1 (83264)
	TUBA4A (49924)		PGAM1 (28804)	APP (86943)
	HPX (51676)		TOLLIP (30282)	NCAM1 (94574)
	SERPINF2 (54566)		IL1B (30748)	NEFM (102472)
	PLD3 (54705)		TNNT3 (31825)	NLRP3 (118173)
	SERPING1 (55154)		ANXA2 (38604)	TEK (125830)
	ICAM1 (57825)		MAPK12 (41940)	COL4A2 (167553)
	PKM (57937)		NDRG1 (42835)	COL6A3 (343669)
	PPP3CB (59024)		SERPINA1 (46737)	

The proteins detected in both matrices were investigated for a linear relationship using Pearson correlation. Significant positive correlations were identified in MASP2, APOE, BCHE, PTGDS and ICAM1. CST3 demonstrated a significant negative correlation, the protein expression in urine decreasing with increasing expression in plasma. Figure 7-1 shows the Pearson correlation coefficients for the proteins identified in both plasma and urine.

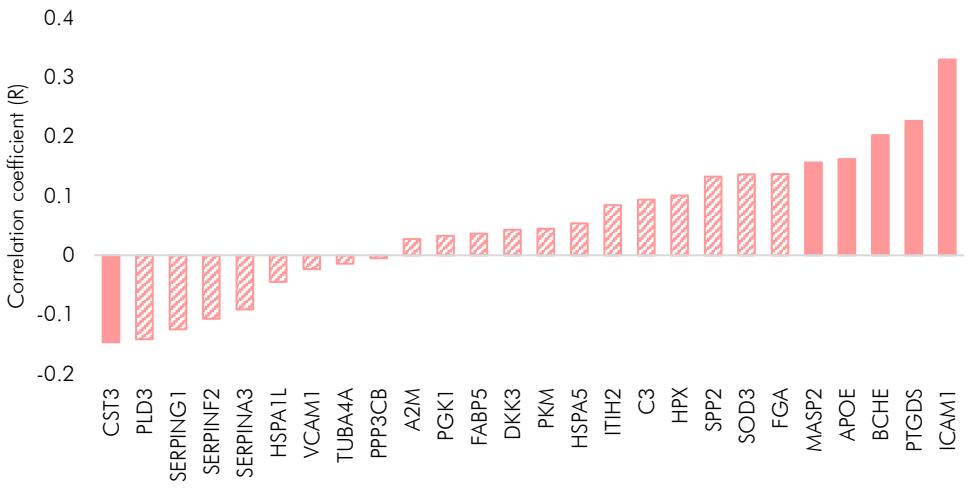


Figure 7-1. Pearson correlation between plasma and urine protein expression. The significant correlations are represented by filled bars and the non-significant by striped bars. *The proteins with negative correlation coefficients were negatively correlated between the matrices, while the proteins with positive correlation coefficients were positively correlated.*

As mentioned in Chapter 6, roughly 30% of the proteins found in urine originate from blood where they are filtered from plasma through the kidney [316, 317], however, there is limited scientific literature describing the correlation between the protein expression in plasma and urine. A study specifically investigating the levels of inflammatory cytokines in urine and plasma found poor correlations between the matrices and concluded that the protein levels in urine and plasma were largely independent of each other [389]. Another

study related urine and plasma proteomics and hypothesised that plasma, urine, and the kidney could be considered as a system where plasma is the input and urine the output with the kidney as the process midpoint. The authors suggested that the function of the kidney may be described as a black box where certain proteins are blocked, permitted to pass, or secreted [388]. Moreover, a study from 2014 of existing plasma, urine and kidney proteomic datasets found that the correlation between the plasma and kidney proteomes, and the urine and kidney proteomes, were greater than the correlation between the plasma and urine proteomes [390].

Investigating the cellular locations of the proteins identified in the matrices, the proteins uniquely discovered in plasma and urine, and the common proteins identified in both matrices, were analysed utilising Ingenuity Pathway Analysis and the cellular locations were extracted. The results from the IPA search are shown in Figure 7-2. The proteins uniquely identified in plasma were categorised as originating from the extracellular space, and one from the cytoplasm. The largest proportion of the proteins identified in both urine and plasma were classified as originating from the extracellular space, and to a lesser degree from the plasma membrane and cytoplasm. In the proteins uniquely detected in urine, the distribution of proteins suggested to originate from the extracellular space and the cytoplasm were relatively equally distributed.

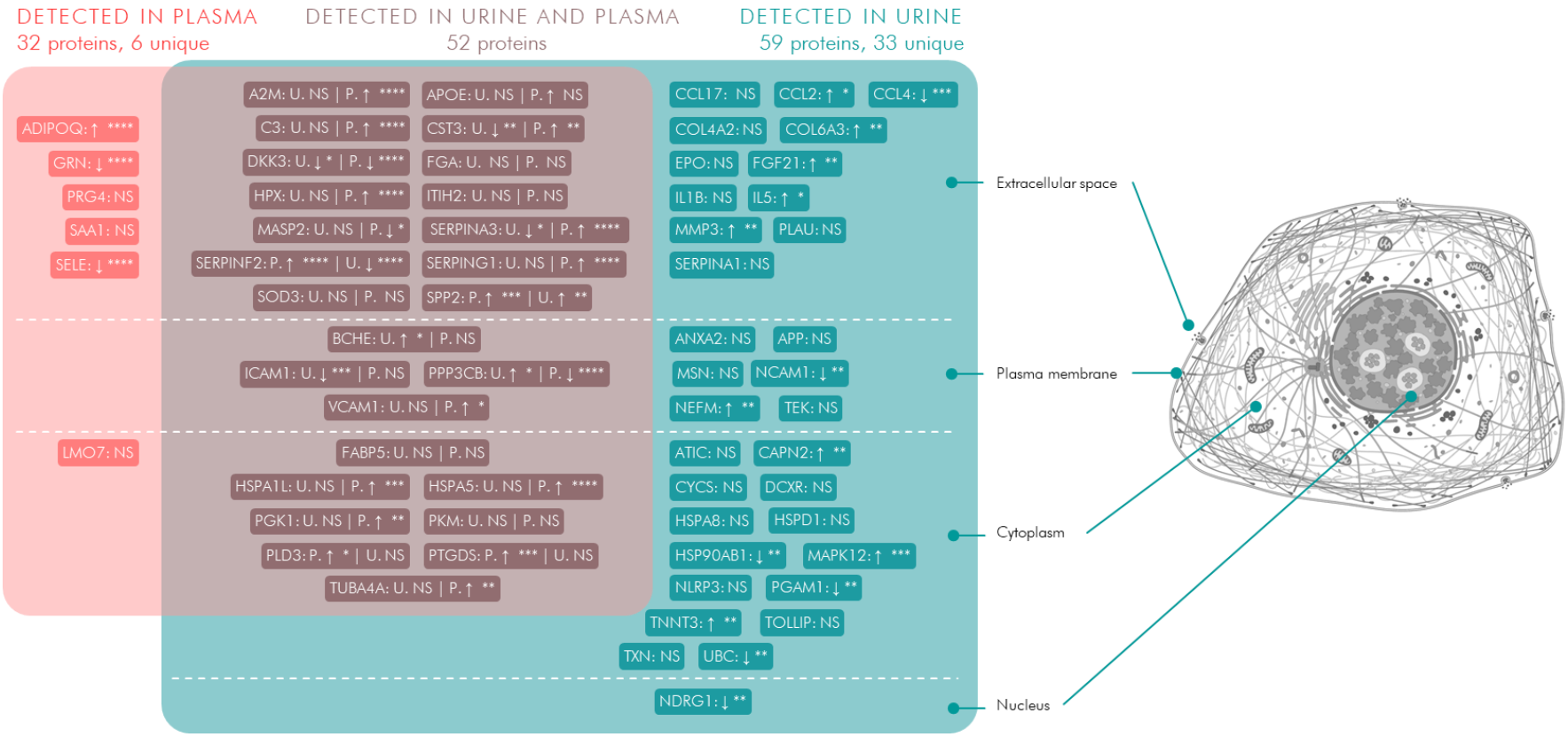


Figure 7-2. Venn diagram of the proteins detected in plasma and urine, and the proteins detected in both matrices. (■) *unique to plasma,* (■) *detected in both plasma and urine, and* (■) *unique to urine. The cellular locations of the proteins (represented by gene names) were generated form Ingenuity Pathway Analysis. The proteins were classified as originating from the extracellular space, the plasma membrane, the cytoplasm, or the nucleus. The arrows indicate if the protein was up- or down-regulated in de novo PD patients compared to healthy controls. FDR adjusted p-value significances are denoted by asterisks, where **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and NS = not significant.*

The significantly differentially expressed proteins from the targeted urine and plasma studies were visualised in a network created in Cytoscape [330] where the proteins were modelled according to p-value significance and fold changes, in their respective studies, for de novo PD, iRBD and other neurological disorders versus healthy controls. As demonstrated by the network in Figure 7-3, DKK3 and SPP2 had the same direction of expression change between controls and de novo PD patients in both the urine and plasma studies, with DKK3 being downregulated in the PD patients and SPP2 upregulated. Four other proteins were detected and significantly differentially expressed between PD patients and controls in the two matrices, with fold changes in opposite directions - these proteins were PPP3CB, HSPA1L, SERPINA3 and SERPINF2. In the comparison of patients with other neurological diseases and controls, CST3 was identified in both matrices – downregulated in OND patients' urine and upregulated in their plasma.

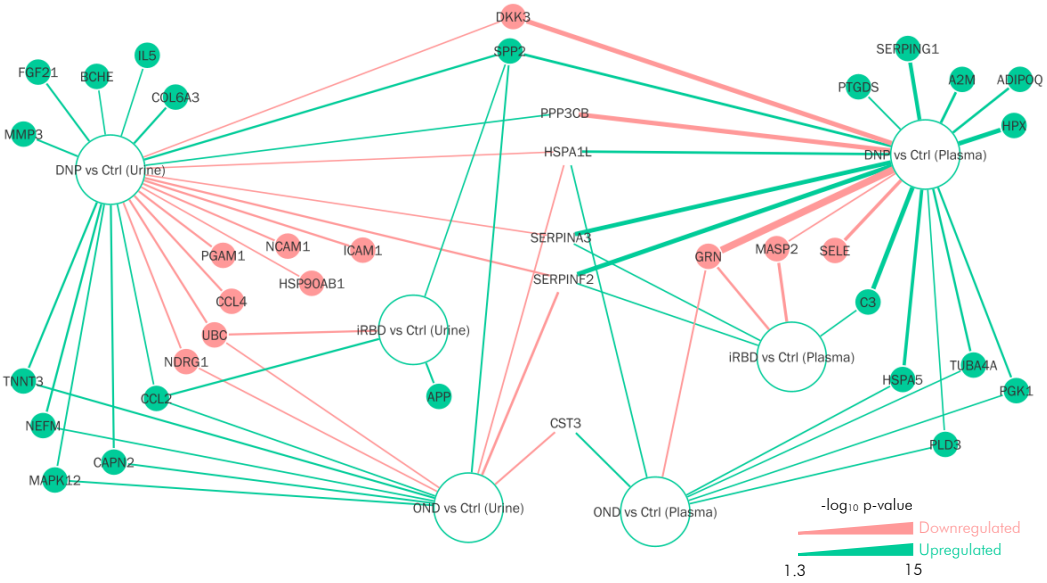


Figure 7-3. Network representation of the results from the targeted study of urine and plasma from the analysis of healthy controls, de novo PD, iRBD and other neurological disorders. *The network shows the proteins which were significantly different when comparing the patient groups to healthy controls in their respective matrix. The line width represents the p-value significance, wider line with higher significance, and line colour represents up- or downregulation in the patient groups, with ■ downregulated, and ■ upregulated.*

The reason for the expression differences between the two matrices is a complex question and further experiments, specifically aimed at disentangling this explicit query, would have been necessary to determine the reasons and the mechanisms responsible for the divergence. Cystatin C (CST3) has been utilised as a marker of renal damage and there is literature describing the relationship between its expression in plasma and urine. A study of acute kidney injury in rats found changes in urinary CST3, however this was not

replicated in plasma CST3 levels [391]. No specific studies relating the levels of the remaining four proteins in urine and plasma could be identified in the literature.

The main findings from the studies do not necessarily contradict each other, and it is of great value to complement the plasma study with the lower abundant proteins identified in urine. The identification of CCL2 and MAPK12 in urine make valuable contributions to the set of proteins identified in plasma and it is further interesting that downregulation of the Wnt signalling promoter DKK3, which was the major finding in the plasma study, was also observed in urine. Poor correlation between plasma and urine protein expression has been reported previously and thus does not invalidate the findings. The reason for the discrepancy between the two matrices may be due to the kidney's filtering system, however – this can only be speculated, and further studies would need to be performed to fully understand the inverse expression observed for the proteins CST3, PPP3CB, HSPA1L, SERPINA3 and SERPINF2 in the matrices. We suggest that both biofluids have merit and that they complement each other in terms of the proteins that can be detected in each matrix.

7.2 TARGETED DE NOVO PD PLASMA STUDY COMPARED TO TARGETED CENTENARIAN PLASMA STUDY

In this thesis, two targeted studies were performed utilising plasma samples and the same analytical mass spectrometric method described in Chapter 2, section 2.9. The centenarian study from Chapter 4 sought to find markers of healthy ageing and longevity by comparing centenarians to controls and centenarian offspring. In Chapter 5, the aim was to find blood-based biomarkers for Parkinson's disease by studying the protein expression in newly diagnosed PD patients, healthy controls, patients with iRBD, and a positive control group consisting of patients with other, non-PD, neurological disorders. The comparison of these two cohorts is noteworthy as age is the greatest risk factor of developing Parkinson's disease. The centenarians included in our study did not exhibit marked cognitive decline or overt signs of neurodegeneration, therefore, we hypothesised that the resolution and interpretation of the two studies could be improved by comparing the protein expressions from the two plasma studies and thus posed the following questions:

- *Do the Parkinson's disease patients exhibit any signs of accelerated ageing?*
- *Do the centenarians demonstrate a protein expression indicating less susceptibility to neurodegeneration and neuroinflammation?*

The proteins DKK3, PPP3CB, SPP2 and SELE were not detected in the centenarian study and SERPINA1 was not detected in the de novo PD study, therefore no additional information could be extracted for these proteins by the comparative analysis. The reason for the failure to quantify these proteins was due to technical challenges, with levels below the limit of detection, and severe retention time drift for SERPINA1 in the de novo PD study. Out of the 33 proteins identified and quantified in the targeted proteomic studies, 28 proteins were successfully quantified in both.

Reiterating the conclusions from the two targeted plasma studies, in the centenarian study it was proposed that the expression of the proteins could be divided into three groups; one group with protective functions and connections to longevity, one risk group related to ageing and increased risk of mortality, and one pleiotropic group including proteins that related to both risk and protective functions. The proteins and their expression in the centenarians were divided into these groups as follows:

- Protective and assumed linked to longevity: A2M (↑), ADIPOQ (↑), SOD3 (↑), C3 (↓), PKM (↓) and ITIH2 (↓)
- Risk and assumed linked to ageing/risk of mortality: SAA1 (↑), MASP2 (↑), VCAM1 (↑), ICAM1 (↑), SERPINF2 (↓) and BCHE (↓)
- Pleiotropic: PTGDS, CST3, SERPINA3, FABP5 and PRG4

In the de novo PD study, the major conclusions were that the PD patients demonstrated disrupted Wnt signalling as demonstrated by lower levels of DKK3 and PPP3CB, increased complement-mediated inflammation, increased unfolded protein response as indicated by elevated levels of HSPA5 and HSPA1L, and decreased neuroprotection through lower levels of GRN.

As discussed in Chapter 4, the centenarian study is limited by the lack of suitable controls. It was here hypothesised that by comparing the protein expressions from the two studies, it could be possible to tease out which proteins were related to normal ageing (the control samples), pathological ageing (the PD, iRBD and OND samples), and longevity (the centenarian samples). Since the targeted assay was set up as an explorative rather than absolutely quantitative assay, the possibility of direct comparisons between different analytical runs has limitations. Moreover, the assay was not subjected to any validation procedures, it is therefore not known how stable the peptides are over time, or how much different collection centres, sample collection procedures, temperature and time of storage affect the results. With this in mind, some caution needs to be taken when

comparing the different sample cohorts directly as unknown factors may affect the results and thus lead to erroneous interpretations.

7.2.1 *Indirect comparison of the protein expression in the studies of centenarians and newly diagnosed PD patients*

Initially, the results from the two studies were compared indirectly by investigating their individual expression profiles to each other. In both the centenarians and the de novo PD patients, the proteins A2M, ADIPOQ, PTGDS and SERPINA3 were upregulated at a significant level when compared to their respective control groups. None of the proteins were significantly downregulated in both studies. Interestingly, a few proteins demonstrated differences in expression between the two studies. C3 and SERPINF2 were significantly upregulated in the de novo PD patients and significantly downregulated in the centenarians, while MASP2 was significantly upregulated in the centenarians and significantly downregulated in the de novo PD patients. Moreover, several proteins were strongly up- or downregulated in one of the studies while expressing no difference in the other study. The results from the comparison of centenarians to offspring/control, and of de novo PD patients to healthy controls are presented as a network in Figure 7-4, showing fold change direction and p-value significance level of each protein.

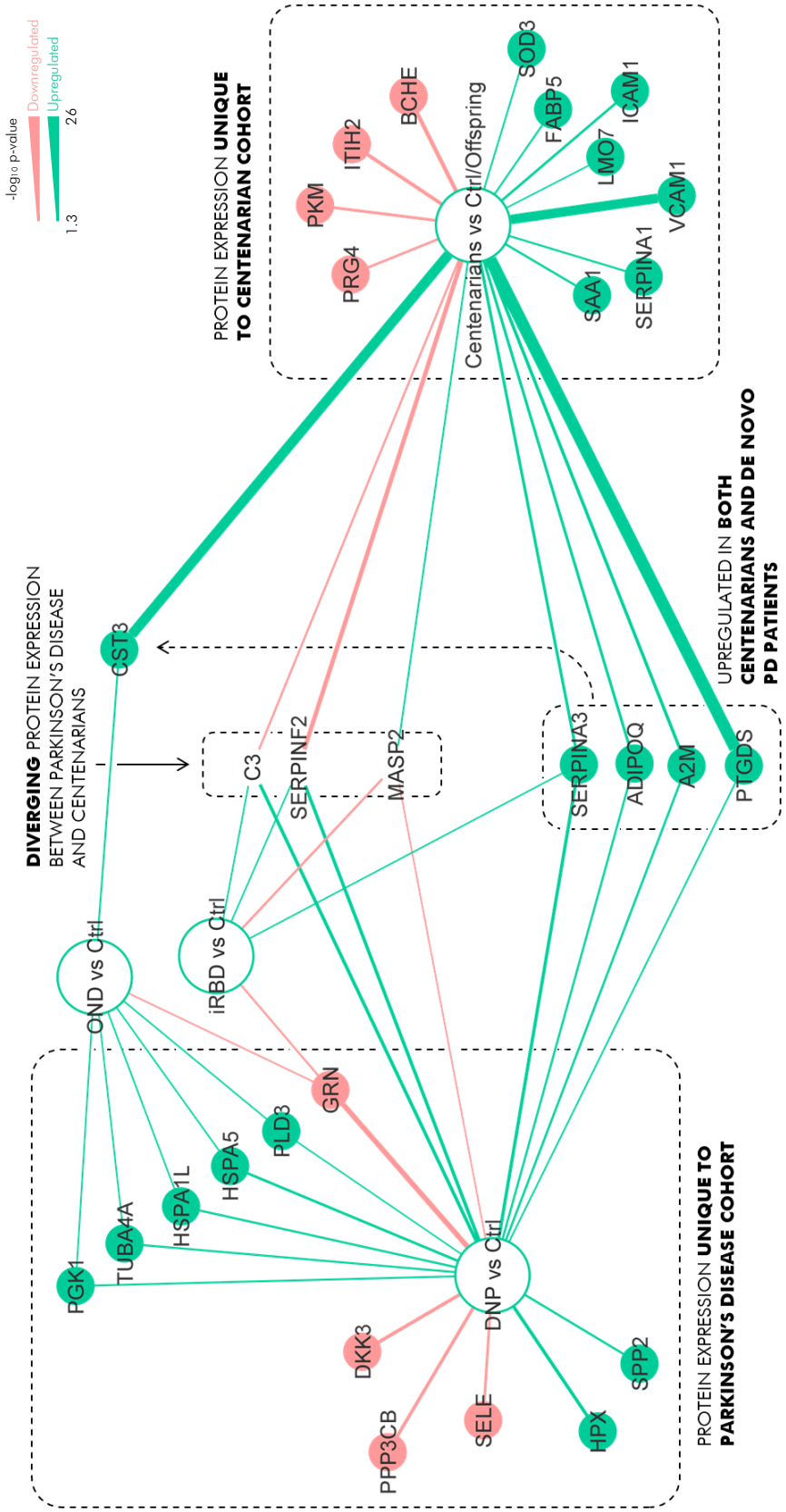


Figure 7-4. Network representation of the protein expressions from the targeted plasma proteomic studies of a cohort including newly diagnosed PD patients, and a cohort including cognitively healthy centenarians. *Proteins upregulated compared to control have green edges, while downregulated proteins have pink edges. The widths of the edges represent p-value significance for each comparison: where wider edges indicate higher significance. The proteins C3, SERPINF2 and MASP2 demonstrated diverging expression when centenarians and newly diagnosed PD patients were compared to their respective controls. SERPINA3, ADIPOQ, A2M and PTGDS were upregulated in both the centenarians and the newly diagnosed PD patients.* ■ *downregulated, and* ■ *upregulated.*

7.2.1.1 *Diverging protein expression between centenarians and newly diagnosed PD patients*

The diverging expression of C3 between the de novo PD patients and centenarians when compared to their respective controls supports the conclusions drawn in Chapters 4 and 5, suggesting inflammation and complement activation in the de novo PD and iRBD patients, while in the centenarians the downregulation of C3 is suggested to constitute part of a protective expression contributing to reducing complement mediated inflammation in the centenarians.

SERPINF2 was found upregulated in the de novo PD and iRBD patients, while downregulated in the centenarians. As mentioned in Chapter 4, SERPINF2 is known to decrease with age, thus it is conceivable that the lower SERPINF2 levels observed in the centenarians are mainly related to their advanced age. However, this explanation does not account for the increased SERPINF2 levels observed in the de novo PD and iRBD patients compared to healthy controls. DNA methylation of SERPINF2 has been linked to Alzheimer's disease [392], but a recent study of SERPINF2 methylation in Parkinson's disease found no differences between PD patients and healthy controls thus suggesting that epigenetic changes in the gene are not related to PD pathology [393]. SERPINF2 is involved in the clotting cascade where it is a major inhibitor of plasmin, which is responsible for breaking down fibrin clots and several other plasma proteins. Thus, upregulation of SERPINF2 could be hypothesised to be correlated with increased blood clotting, and lower levels with decreased ability of clotting. Moreover, it is well-known that interplay exists between the clotting and coagulation cascades, thus suggesting that they may modulate each other's activity [394]. In our data, we found a strong positive correlation between C3 and SERPINF2 with an R2 value of 0.6 (Figure 7-5), thereby suggesting that the two proteins co-vary in our studies and indicate an overall elevated state of inflammation in the de novo PD and iRBD patients, and lower levels of complement-mediated inflammation in the centenarians.

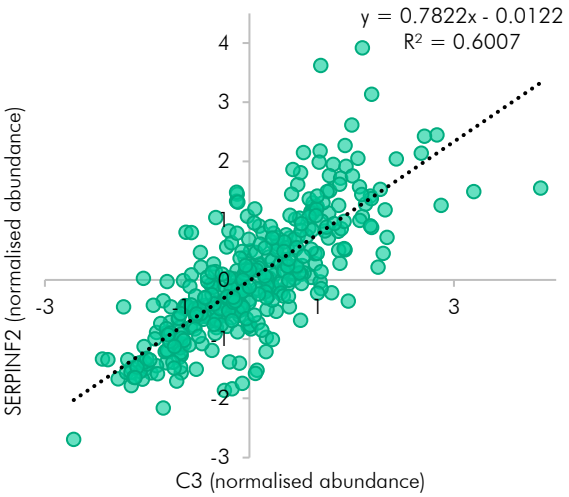


Figure 7-5. Correlation between C3 and SERPINF2 in the targeted plasma proteomics studies of centenarians and newly diagnosed PD patients. *The correlation between the normalised abundance of the proteins is shown.*

The downregulation of MASP2 in the de novo PD patients and upregulation in the centenarians was inverse to the expression of C3 in both studies. In the complement cascade, MASP2 can cleave C2 and C4 and produce C3 [395]. It could be hypothesised that MASP2 is being consumed during activation in the Lectin pathway in the de novo PD patients, but without performing functional studies to explore this inverse relationship, no certain conclusions can be drawn.

7.2.1.2 *Converging protein expression in centenarians and newly diagnosed PD patients*

The proteins ADIPOQ, A2M, SERPINA3 and PTGDS were upregulated in both the de novo PD patients and the centenarians when compared to controls. Upregulation of ADIPOQ and A2M was proposed to contribute to a protective protein expression in the centenarian study, whereas SERPINA3 and PTGDS were categorised as belonging to the group of pleiotropic centenarian proteins.

A2M and ADIPOQ act anti-inflammatory, albeit through different mechanisms. A2M is a protease inhibitor, known to reduce inflammation in cartilages and is therefore used as a treatment against arthritis [200]. ADIPOQ is known to be a systemic anti-inflammatory factor and elevated levels are associated with lower body-mass-index, reduced risk of diabetes, cardiovascular disease, and an overall inverse correlation has been reported between ADIPOQ and inflammatory markers [396]. In the centenarian study, it was stipulated that elevation of A2M and ADIPOQ contributed to the centenarians' longevity. A2M has been identified as a component of both Lewy bodies in PD and senile plaques in AD [397]. This connection has led to numerous studies of A2M's involvement in the two diseases and at least one A2M gene polymorphism has been suggested to be associated with PD [398]. Upregulation of A2M in AD is proposed to aid in the solubility of senile plaques, thus alleviating the effect of the disease. It is possible that a similar mechanism could occur in the Lewy bodies of PD, although no study proving this was encountered. The role of ADIPOQ in neurodegenerative disease has been investigated but remains poorly understood. A review from 2020 concluded that although ADIPOQ may be involved in various protective mechanisms in the brain, its function remains largely unknown. Some studies have found the protein upregulated in neurodegenerative disease while others have found it downregulated. In PD, it has been suggested that ADIPOQ could play a role in lipid rafts and may allow for differentiation between alpha-synucleinopathies and PSP [399].

SERPINA3, upregulated in both centenarians and newly diagnosed PD patients when compared to respective controls, is a protease inhibitor and a positive acute phase protein. SERPINA3 has been implicated in AD and PD, where variants have been identified in patients [400]. SERPINA3 has also been correlated with ageing, where it has been found upregulated with age in the brains of healthy aged individuals compared to younger controls. This study moreover found that SERPINA3 was upregulated in brains from patients with neurodegeneration compared to healthy controls of the same age [401].

PTGDS was upregulated in the centenarian and PD groups when compared to their respective controls. As mentioned in Chapter 4, PTGDS is an inhibitor of platelet aggregation and has been suggested to enhance the anti-inflammatory action of astrocytes in the brain together with the PD-related gene DJ-1. Moreover, a study of PTGDS-immunoreactive isoforms found them to appear in many neurodegenerative disorders, including PD. The same study found no differences in the PTGDS-isoforms over an age-range of 19 to 97 years [402].

7.2.2 *Direct comparison of the centenarian and de novo Parkinson's disease patients*

As stated previously, comparing studies run at different time points from a non-validated assay requires a fair amount of caution to avoid interpreting eventual technical aspects not accounted for as biological variation. Aiming to compare the two studies directly, the normalised data from the two plasma studies were initially z-scored to equalise the variances and centre the averages around zero. The reason for doing this was to remove the effect of small differences in intensity of the variables measured in the two studies, which might otherwise have affected the interpretation. Non-age and sex adjusted data from both studies were utilised.

The data were initially modelled by PCA to obtain an overview of how the groups compared to each other. The PCA score plot is shown in Figure 7-6. As demonstrated by the plot, the control and offspring samples from the different studies distributed close to each other, while the PD and other neurological disorder patient samples distributed closer to the centenarians. The iRBD patient samples were located between these two major clusters.

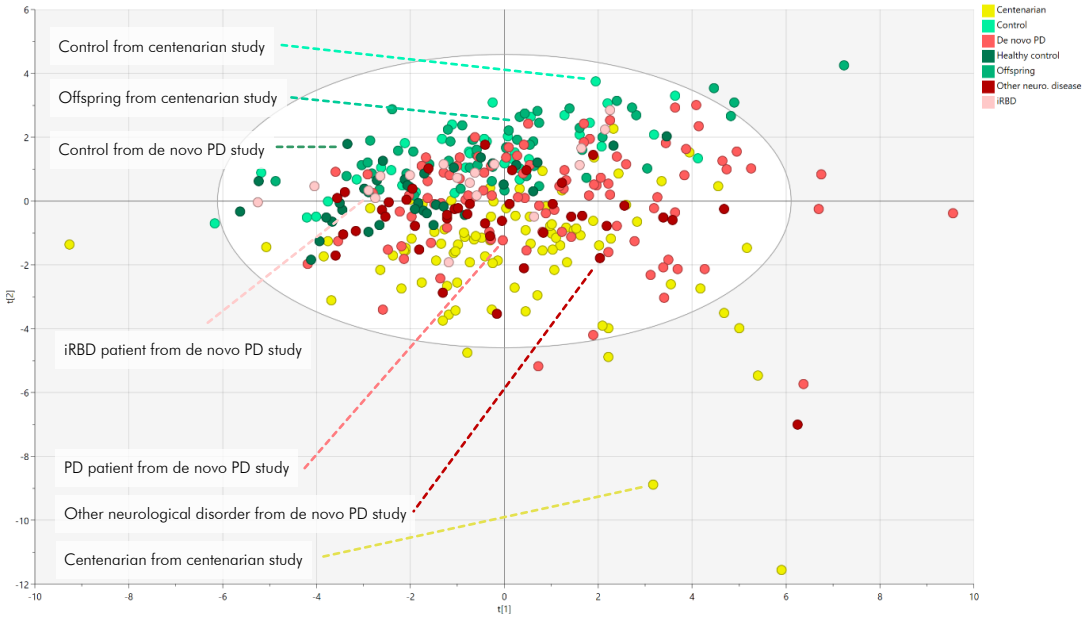


Figure 7-6. PCA scores from the z-scored centenarian and de novo PD studies modelled together. *Principal components 1 and 2 (t[1] and t[2]) are shown. It is demonstrated that the control and offspring samples from the two studies cluster, while the centenarians, PD and other neurological disorder patients distribute close to each other. The iRBD patient samples are found between the two major clusters.*

In conclusion, the initial PCA suggested that there was a shared protein expression between the control and offspring samples from the two studies and that the centenarians, PD patients and other neurological disorder patients had a degree of similarity in protein expression.

Investigating differences between the two centres, the full dataset was modelled by OPLS-DA, with centre (Bologna versus Göttingen) set as the dependent variable. The model was non-significant ($p = 1$), thereby demonstrating that no large, overall centre-to-centre differences existed between the two studies. Next, the offspring samples and controls from both studies were compared in an OPLS-DA. This model proved significant ($p = 5E^{-15}$), thus demonstrating that there was a difference between the control samples from the two studies. The Göttingen controls were attributed with higher levels of GRN and ICAM1, while the Bologna controls had higher levels of SERPINF2, C3 and SERPINA3, among others. No expression differences could be detected for the proteins ADIPOQ, SOD3, PLD3, A2M, FABP5, APOE, FGA, SAA1, PTGDS, VCAM1, CST3, LMO7 and MASP2. The reason for the differences between the controls from the two centres cannot be explained by any known factors in the studies and may therefore be due to technical aspects such as differences in sampling, fasting duration, freeze-thaw cycles or storage duration, or other unknown factors.

The centenarians and the newly diagnosed PD patients were compared in an OPLS-DA model. This model was significant ($p = 2.7\ E^{-12}$) and demonstrated that the centenarians had significantly higher levels of CST3, PTGDS, VCAM1 and ICAM1, while the PD patients had significantly higher levels of HPX, PRG4, PKM, ITIH2, BCHE, C3 and SERPINF2. The remaining proteins did not express any difference between the two groups. The proteins not expressing a difference between the two groups are of special interest as they may pinpoint proteins which could indicate potentially accelerated ageing in the PD patients. To investigate this, the non-differentially expressed proteins from the PD versus centenarian OPLS-DA model were extracted. The proteins differentially expressed in OPLS-DA models between PD and controls, and centenarian versus controls/offspring were moreover extracted. The rationale behind the selection was that proteins demonstrating no difference between centenarians and PD patients, but a difference when comparing PD patients and centenarians to their respective controls, may indicate accelerated aging in the PD patients or other common mechanisms between the two groups. The proteins were illustrated in a Venn diagram (Figure 7-7).

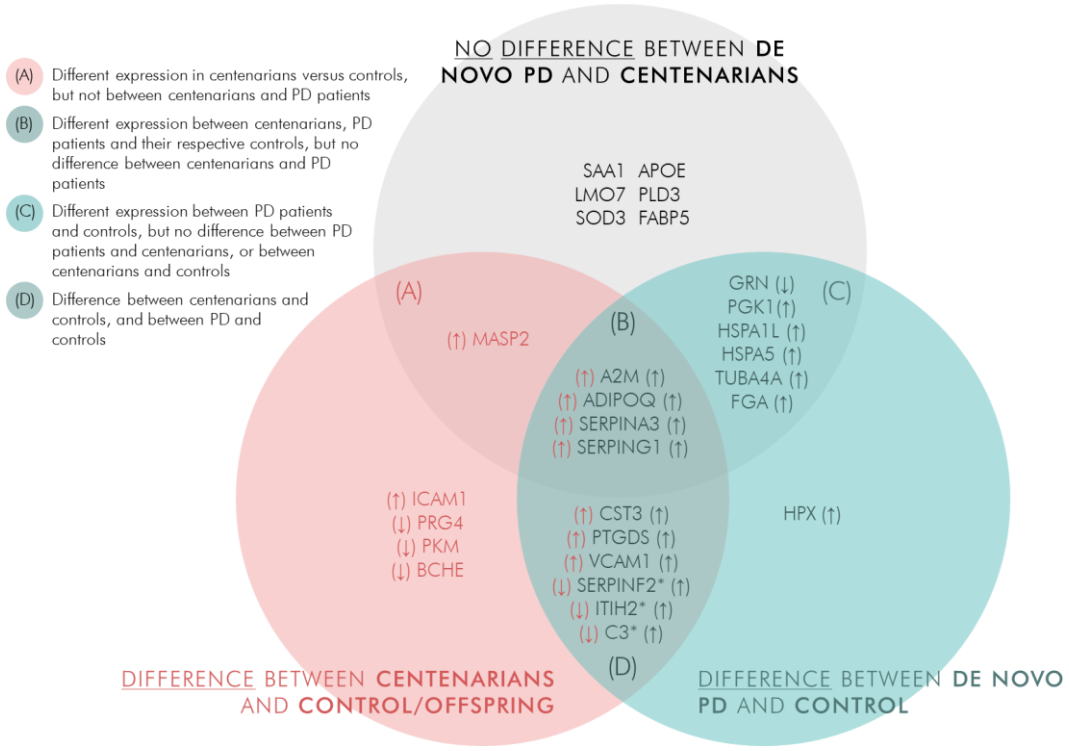


Figure 7-7. Venn diagram of the proteins from discriminant OPLS-DA analyses expressing no difference between centenarians and PD patients, and of the proteins expressing a difference in the comparison of PD versus control, and centenarians versus control. *Difference centenarians/controls (■), difference PD/controls (■), and no difference centenarians/PD (■). Direction of change in the centenarians and PD patients compared to their respective controls is represented by red arrows on the left for the centenarians, and blue arrows on the right for the PD patients. The proteins which did not change in the same direction in the PD patients and the centenarians are denoted by an asterisk after the gene name.*

The Venn diagram demonstrated the following:

- *Venn region A:* MASP2 was upregulated in the centenarians but demonstrated no difference between PD patients and centenarians, and also no difference between PD and healthy controls. This region is proposed as not being related to a protein expression relevant for PD.

- *Venn region B:* Four proteins were common in the three comparisons (A2M, ADIPOQ, SERPINA3 and SERPING1), expressing no difference between the PD patients and the centenarians, but a difference between the groups when compared to their respective controls. These proteins are strongly related to accelerated ageing in the Parkinson's disease patients, or to another common factor with the centenarians which would differentiate the protein expression between PD patients and healthy controls.

- *Venn region C:* The proteins differentially expressed between PD and controls, but not between PD and centenarians, or between centenarians and controls were GRN, PGK1, HSPA1L, HSPA5, TUBA4A and FGA. These proteins are suggested to be related to Parkinson's disease but not longevity in the centenarians

- *Venn region D:* The proteins differentially expressed between PD and controls, and between centenarians and controls, but also between centenarians and PD patients, were CST3, PTGDS and VCAM1 – all upregulated, and SERPINF2, ITIH2 and C3 – downregulated in the centenarians and upregulated in the PD patients. These last three, diverging proteins may therefore differentiate between an expression related to longevity or extremely advanced age in the centenarians compared to the Parkinson's patients, or to mechanisms unique for the PD patients.

The four proteins identified in Venn region B as putatively indicating accelerated ageing in the PD patients, or describing other common mechanisms between the two groups, were input to DAVID Bioinformatics Resources 6.8 and the GO terms were extracted. Although the number of proteins used for enrichment annotations were few and thereby limiting the interpretability of the analysis, the GO annotations still provide insights about the pathway involvement and processes of the proteins. As demonstrated by the GO enrichment p-values, platelet degranulation was the most significant biological process (FDR p = 0.007). Platelet degranulation encompasses the release of compounds from granules within platelets. The granules contain a plethora of different molecules, both metabolites and proteins, and are secreted in response to vascular and endothelial injury and to regulate immune response [403, 404]. Platelet alteration has been studied in

relation to neurodegenerative disease and it has been suggested that alternative platelet activation pathways may be associated with brain function [405]. Platelets have been reported to decrease in count but increase in reactivity after the age of 70. It is suggested that the platelet activation in aged individuals may partially be driven by an upregulation of reactive oxygen species and that that increased platelet activity could be associated with age-related diseases such as cardiovascular disease, neurodegenerative diseases and an overall elevated inflammatory state [406]. The common expression of proteins related to platelet degranulation between PD patients and centenarians thus provides a plausible indication of an elevated state of inflammation in the two groups - in the centenarians caused by ageing, and in the Parkinson's disease patients related to disease.

SERPINF2 and C3 from the three between-group diverging proteins in Venn region D were discussed previously in section 7.2.1.1, where it was concluded that the lower levels of SERPINF2 in the centenarians are likely due to their extremely advanced age, and that downregulation of C3 in the centenarians may be part of a longevity-promoting protein expression, protecting the centenarians from the detrimental effects of excessive complement-mediated inflammation. As mentioned in Chapter 4, ITIH2 has been reported to decrease in expression with older age [178], thus correlating with our observation in the centenarians. The upregulation of C3 observed in the Parkinson's disease patients was stipulated to contribute to increased levels of inflammation in Chapter 5. In section 7.2.1.1, we showed a positive correlation between C3 and SERPINF2 and concluded that this could suggest an overall increased inflammatory state in the PD patients. Unlike the vast majority of the significant proteins, ITIH2 was found upregulated in the PD patients in the OPLS-DA model only, and not in the FDR-adjusted univariate analysis, signifying that although ITIH2 contributed to the OPLS-DA separation between PD and controls, its difference between the two groups may be subtle. ITIH2 was also found upregulated in the pre-PD twins in the discovery study performed in Chapter 5. ITIH2 has been suggested to be involved in inflammation [407], although its potential role in Parkinson's disease pathology remains unexplained.

Returning to the questions posed in the beginning of the section:

- *Do the Parkinson's disease patients exhibit any signs of accelerated ageing?*
- *Do the centenarians demonstrate a protein expression indicating less susceptibility to neurodegeneration and neuroinflammation?*

It cannot be proven beyond doubt that the Parkinson's disease patients exhibit signs of accelerated ageing. Their protein expression does however indicate that they have increased levels of inflammation. Inflammation is also indicated in the centenarians, albeit

not by an expression of identical proteins. The downregulation of the central complement protein C3 implies that complement-mediated inflammation is not present in the centenarians, while upregulation in the PD patients suggests that it is present in this group. Given the potency of the complement system, this downregulation in the centenarians may offer protection from the detrimental effects an uncontrolled complement-response can cause. In order to be able to elucidate any eventual accelerated ageing caused by neurodegeneration in the Parkinson's disease patients an additional, well-defined, group of patients with advanced Parkinson's disease or another neurodegenerative disease, such as Alzheimer's disease, should ideally have been included.

7.3 SUMMARY AND CONCLUSIONS

The comparison of the protein expression in plasma and urine from section 7.1 showed that the correlation between the two matrices was poor for the majority of the proteins. Although scientific literature on the subject was found limited, the lack of co-expression between the matrices in our study correlated with other studies. Thus, the plasma and urine studies were seen as complementary to each other, urine having the advantage of allowing for a greater protein coverage due to a less complex sample matrix.

In the comparison of plasma from centenarians and newly diagnosed PD patients, it was found that the two groups shared upregulation of four proteins when compared to their respective control samples. These proteins were A2M, ADIPOQ, PTGDS and SERPINA3. Moreover, an analysis of the proteins demonstrating no difference between centenarians and de novo PD patients, in conjunction with the proteins differing in each group when compared to their own controls highlighted upregulation of the proteins A2M, ADIPOQ, SERPINA3 and SERPING1. This group of proteins is thus hypothesised to represent a shared expression between centenarians and PD patients, different from controls. Enrichment analysis suggested that platelet degranulation could be an affected pathway, including the proteins A2M, SERPINA3 and SERPING1. Although no certain proof of accelerated ageing could be determined in the de novo PD patients, both the centenarians and the PD patients exhibit a protein expression suggesting inflammation.

Final discussion, conclusions and future work

8

8.1 DISCUSSION

8.1.1 *A brief background and summary of the studies presented in this work*

Parkinson's disease is the second most common neurodegenerative disorder and affects approximately 1% of the population after the age of 65, and 3% after the age of 85 [90]. The disease phenotype is highly heterogenous but commonly manifests as tremor, and stiff and slowing movement, and also apathy, anxiety and depression [95]. The majority of the cases are idiopathic, without any known genetical component and advanced age is recognised as the most important risk factor for non-hereditary Parkinson's disease. Although tremendous efforts have gone into finding PD specific biomarkers they remain elusive and thus, the diagnosis is set clinically with a misdiagnosis rate up to 24% [408]. The most common misdiagnosis is confounding with other parkinsonian disorders such as progressive supranuclear palsy, multiple system atrophy and Lewy-body dementia, to mention a few. The misdiagnosis frequency is generally higher in the early stages of disease [408, 409]. Given the relatively late display of symptoms, diagnosis is often set at a relatively advanced stage of the disease, when a large proportion of the dopaminergic neurons have already been lost. It is therefore imperative to find a way of diagnosing patients earlier and with increased accuracy. It was in this context the experiments presented in this thesis were carried out.

The overall aim was to identify specific proteins which were differentially expressed between PD patients and healthy controls. Utilising untargeted mass spectrometry, we performed three discovery studies in individuals not yet diagnosed with PD, and in patients with PD: two blood-based and one in urine, and additionally one blood-based study of cognitively healthy centenarians. The PD patient studies were aimed at identifying potential biomarker targets, and the centenarian study at discriminating between proteins related to longevity and normal ageing. The discovery studies concluded that a number of proteins were altered in the PD patients and in the centenarians, and also suggested that several pathways – the majority related to inflammation – were affected. Table 8-1 shows the top enriched pathways for each discovery study and the proteins selected for targeted analysis.

Table 8-1. Top pathways, and proteins selected for targeted verification based on the results from the discovery analysis of PD patients and centenarians versus controls. *The pathways in which up- or downregulation were determined are denoted by ↑ for upregulation and ↓ for downregulation. In the cases up- or downregulation could not be determined on a significant level, an asterisk (*) has been added. The pathways which were significantly enriched, but where up- or downregulation could not be established are denoted by §. The proteins are annotated by gene names.*

Discovery study	Top enriched pathways	Proteins selected for targeted analysis
Newly diagnosed, treatment-naïve PD patients and healthy controls (plasma)	Complement system§, acute phase response signalling§, glucocorticoid receptor signalling§, ER stress pathway§, superoxide radicals degradation§	ANXA1, GOLM1, HSPA5, NRP1, UHRF1BP1L, SERPINA3, PRG4, DKK3, C15orf62, SPP2, BCHE, TNNT3, CSF1R, MMP3, PTGDS, ITIH2 and NCAM1
Homozygous twin pairs, discordant for developing PD 4.6 ± 1.7 years from the time of sampling (serum)	Rheumatoid arthritis§, bladder cancer signalling§, Wnt/ß-catenin signalling§, prostanoid biosynthesis§, osteoarthritis pathway§	
Idiopathic PD patients, symptomatic and asymptomatic LRRK2 mutation carriers, and healthy controls (urine)	LXR/RXR activation (↑*), FXR/RXR activation§, 14-3-3-mediated signalling§, apoptosis signalling (↑), axonal guidance signalling§	RAGNAP1, TUBA4A, MAPK12, APOE, FGA, HSPA8, PPP3CB, PLAU, COL4A, THY1, CYCS, CTHRC1, ATIC, CAPN2, DKK3, EFNA5, ENDOU, HBE1, MASP2, MUC5B, NDRG1, SOD3 and TOLLIP
Cognitively healthy centenarians and healthy controls (plasma)	Acute phase response signalling (↑*), LXR/RXR activation (↑), complement system (↑*), FXR/RXR activation§, coagulation system (↓*)	A2M, ADIPOQ, CST3, CTHRC1, FGA, HBE1, PTGDS, SOD3, CSF1R, DKK3, PKM, SAA1and SERPINA1

In the targeted validation phase, the 45 proteins from the discovery studies were developed into a targeted test where several inflammatory proteins from the literature were also included. Here, three studies were performed – one in plasma from newly diagnosed PD patients, one in urine from newly diagnosed PD patients, and one in plasma from centenarians. All targeted studies were matched with controls. In the PD studies, iRBD patients and patients with other (non-PD) neurological disorders were moreover included. In the centenarian study, an additional group of centenarian offspring samples were included. Table 8-2 summarises the top ten significantly altered proteins from each study, and also the major findings.

Table 8-2. Top ten altered proteins, based on p-value significance, and major findings from the targeted analysis of plasma from centenarians, and urine and plasma from newly diagnosed PD patients. ↑ *indicates upregulation, and* ↓ *indicates downregulation in PD/centenarians compared to respective controls. The proteins are denoted by gene names.*

Targeted study	Top ten altered proteins	Major findings
Centenarians, offspring and control (plasma)	PTGDS (↑), CST3 (↑), VCAM1 (↑), SERPINF2 (↓), BCHE (↓), ITIH2 (↓), PKM (↓), A2M (↑), SERPINA3 (↑) and PRG4 (↓)	• Indication of increased inflammation in the centenarians • Protection against oxidative stress? • Reduced complement mediated inflammation?
De novo PD, iRBD, other neurological disorders and control (plasma)	GRN (↓), DKK3 (↓), C3 (↑), PPP3CB (↓), HPX (↑), SERPINF2 (↑), SERPINA3 (↑), SERPING1(↑), SELE (↓) and HSPA5 (↑)	• Predictive SVM/LDA machine learning models classified samples as PD and control with 100% accuracy • Indication of increased inflammation in the PD patients • Suggested involvement of Wnt signalling and unfolded protein response
De novo PD, iRBD, other neurological disorders and control (urine)	ICAM1 (↑), CCL4 (↓), TNNT3 (↑), SPP2 (↑), COL6A3 (↑), NEFM (↑), UBC (↓), PGAM1(↓), SERPINF2 (↓) and CAPN2 (↑)	• Predictive LDA machine learning model classified samples as PD and control with 85.1% accuracy • Indication of increased inflammation in the PD patients • Upregulation of NEFM suggests neuronal damage

In summary, the explorative targeted study verified several proteins identified in the discovery phase. In the centenarian plasma study A2M, ADIPOQ, CST3, PTGDS, PKM, SAA1 and SERPINA1 were confirmed altered with the same expression direction as in the discovery study. In the PD plasma study, HSPA5, SERPINA3 and SPP2 were altered in the same direction as in the discovery study while DKK3 and PTGDS were significantly differentially expressed in the opposite direction compared to the discovery study. In the PD urine study, MAPK12, PPP3CB, CAPN2 and NDRG1 were confirmed, while DKK3 was expressed in the opposite direction compared to the discovery study. Apart from these confirmatory proteins, several others were also found differentially expressed in the targeted study. Below, the findings from our studies will be discussed.

8.1.2 *Inflammation in early Parkinson's disease and ageing*

Signs of elevated inflammation were identified in the plasma from centenarians, with higher levels of several pro-inflammatory proteins compared to controls. In the PD patients' plasma, the central complement cascade protein C3 was strongly upregulated ($p = 1.7\ E^{-9}$), thus indicating complement activation. C3 is a central molecule in the complement cascade; it is formed regardless of the initiating pathway (classical, alternative or lectin). The involvement of complement in neurodegenerative disease is not a new concept but has been identified in several studies. In Parkinson's disease,

complement activation has been associated with the formation of Lewy bodies, where also deposits of iC3b and C9 have been identified [297]. Region-specific neuron loss and cognitive decline were reported decreased in a C3 knockout study of mice [410]. Moreover, C3 has been identified to be involved in the complement-mediated elimination of redundant synapses in the central nervous system, a necessary process, but one that may become detrimental if overactive and may lead to pathological synapse loss and neuronal death [411]. This is indeed interesting and relevant also in the context of the cognitively healthy centenarians, where C3 was found significantly downregulated ($p = 5.2\ E^{-5}$). The complement system makes part of a healthy immune system, where it is activated to mediate a response to pathogens. However, it is also a system prone to "friendly fire", where it can become a self-amplifying loop, fuelling inflammation. This harmful complement activation has been reported to become more pronounced with ageing [207], while downregulation of C3 has been identified in studies of centenarians [209, 412]. The discrepancy in C3 expression between Parkinson's disease patients and cognitively healthy centenarians may therefore suggest that the complement system is involved in both neurodegeneration and longevity, downregulation being protective and promoting healthy ageing, and upregulation potentially contributing to inflammation amplification and neuronal cell death in neurodegenerative disease. Moreover, the expression of the protein SERPINF2 was found to correlate positively with the expression of C3 in our study (section 7.2.1.1). In the centenarians, it was downregulated ($p = 1.1\ E^{-13}$), and in the Parkinson's disease patients upregulated ($p = 1.7\ E^{-9}$). As mentioned previously, SERPINF2 is involved in the clotting cascade where it is a major inhibitor of plasmin, which is responsible for breaking down fibrin clots and several other plasma proteins. It has more recently been implicated in cerebrovascular and cardiovascular disease, where higher levels of SERPINF2 have been found associated with high risk and/or poor outcome. High SERPINF2 levels have also been found to increase the risk of ischemic stroke, and to enhance the expression of the neuroinflammatory protein MMP9, contributing to worse brain injury [413, 414]. It has been strongly suggested that the complement and coagulation cascades may be linked and that they may modulate each other's activities [415]. It could therefore be hypothesised that in Parkinson's disease, these pathways contribute to activating each other and increasing the inflammatory response, while in the centenarians downregulation may act protective.

The expression of proteins related to platelet degranulation in centenarians and Parkinson's disease patients (identified in Chapter 7, section 7.2.2) suggests that there is a common inflammatory component in both groups. Although the centenarians were used as a model of healthy ageing in this work, they are approximately 40 – 50 years older and

known to demonstrate increased levels of low-grade inflammation (inflammageing) [187]. The platelet degranulation-related proteins were upregulated in both groups compared to their respective controls while expressing no difference between the groups. The FDR-adjusted p-values when comparing the groups to controls were - for centenarians and PD patients, respectively: SERPINA3: $p = 1.8\ E^{-7}$, $p = 4.5\ E^{-9}$, SERPING1: $p = 9.8\ E^{-3}$, $p = 7.6\ E^{-8}$, and A2M: $p = 9.0\ E^{-9}$, $p = 5.7\ E^{-5}$. The involvement of platelet alteration in neurodegenerative disease has been investigated in other studies, where it has been suggested that alternative platelet activation pathways may be associated with brain function [405]. In relation to ageing, platelets have been reported to decrease in count, but increase in reactivity after the age of 70. The activation has been suggested to be caused partly by an upregulation of reactive oxygen species and to be associated with an overall elevated inflammatory state [406]. It is conceivable that the proteins SERPINA3 and SERPING1 are not related to a longevity-promoting protein expression in the centenarians, as both proteins have been identified to increase with chronological age in other studies. SERPINA3 has been found upregulated with age in the brains of healthy aged, non-centenarian, individuals compared to younger controls, and also upregulated in brains from patients with neurodegeneration compared to healthy controls of the same age [401]. SERPING1 expression has been demonstrated to correlate positively with chronological age and is therefore suggested to not be related to the longevity of centenarians [416, 417]. As both proteins are known to be upregulated in normal, non-centenarian ageing, their upregulation in the centenarian group is likely a reflection of the extreme age of these individuals and not related to their longevity. The PD patients were compared against controls in the same age range, and thus, the upregulation of SERPINA3 and SERPING1 in the Parkinson's disease group indicates increased inflammation and possibly accelerated ageing or perhaps the concept of "accelerated brain ageing".

The shared upregulation of A2M and ADIPOQ between the PD patients and the centenarians (also identified in Chapter 7, section 7.2.2) is discussed below, in section 8.1.3.

8.1.3 *Adiponectin and A2M – protective or detrimental?*

As per the direct comparison of plasma from PD patients and centenarians in Chapter 7, ADIPOQ and A2M were two out of four proteins significantly upregulated in both centenarians and PD patients and thus hypothesised to indicate inflammation and potentially accelerated ageing in the PD patients, or another common mechanism between the two groups. The FDR-adjusted p-values when comparing the groups to their respective controls were, for ADIPOQ $p = 5.0\ E^{-7}$ in the centenarians and $p = 1.5\ E^{-4}$ in the

PD patients, and for A2M as reported in section 8.1.2. These results were somewhat surprising as the two proteins had been proposed to contribute to a protective and longevity-promoting protein expression in the centenarians in Chapter 4.

ADIPOQ is known to be a systemic anti-inflammatory factor and elevated levels are associated with lower body-mass-index, reduced risk of diabetes, cardiovascular disease, and an overall inverse correlation has been reported between ADIPOQ and inflammatory markers [396]. Several studies have found elevated ADIPOQ expression in centenarians and firmly suggested that it is associated with longevity [418, 419]. Therefore, the upregulation of ADIPOQ observed in the centenarians in our study was ascribed as longevity-promoting. However, the role of ADIPOQ in neurodegenerative disease remains poorly understood. Mice models have suggested that ADIPOQ may be neuroprotective as ADIPOQ deficiency promoted Alzheimer's disease-like symptoms, synapse loss and neuroinflammation [420]. A review from 2020 concluded that although ADIPOQ may be involved in various protective mechanisms in the brain, its function remains largely unknown [399]. A study of adiponectin concentrations in (non-age matched) PD patients, morbidly obese patients and healthy controls found no difference between PD and the control group, but identified a difference between PD and the morbidly obese patients, with higher adiponectin levels in the PD patients [421]. It is well-established that adiponectin levels increase with lower body-mass-index [422]. The body-mass-indices of the individuals included in our studies are not known, and thus we cannot explore if the ADIPOQ expression had a relationship with this parameter. It is therefore concluded that while ADIPOQ likely exerts longevity-promoting effects in the centenarians, its role in Parkinson's disease remains unknown.

A2M is an acute phase protein which can inhibit cytokines and disrupt inflammatory cascades [423]. In normal ageing, it has been found to significantly decrease with age, while longevity studies in the long-lived naked-mole rat found it overexpressed, attributing A2M properties responsible for cancer-protection [168, 199]. A2M has been identified as a component of both Lewy bodies in PD and senile plaques in AD, where upregulation of A2M in AD has been proposed to aid in the solubility of senile plaques, thus alleviating the effect of the disease [397]. Relevant for Parkinson's disease - A2M has moreover been reported to have the ability to bind misfolded proteins, including alpha-synuclein, thereby reducing its neurotoxicity [424]. It is thus plausible that the upregulation of A2M observed in the Parkinson's disease patients, and also in the centenarians, makes part of a protective response – in the PD patients to bind alpha-synuclein oligomers in an attempt to clear the misfolded proteins, and in the centenarians continuously clearing misfolded

proteins and potentially also inhibiting inflammatory processes such as the complement cascade.

8.1.4 *Endoplasmic reticulum stress and the unfolded protein response*

In plasma from the Parkinson's disease patients, the upregulation of HSPA5 ($p = 2.1E^{-6}$) and HSPA1L ($p = 1.2E^{-4}$) when compared to healthy controls, indicated endoplasmic reticulum stress and activation of the unfolded protein response. In the centenarians, HSPA5 and HSPA1L were unchanged compared to the control/offspring group.

The endoplasmic reticulum folds proteins into their correct conformations, adds post-translational modifications and sorts the proteins for their intended destination. Incorrectly folded proteins are detected by the endoplasmic reticulum's quality control system and are sent to the ER-associated degradation pathway for refolding or degradation [298, 299]. If the ER becomes overloaded by excessive amounts of unfolded proteins, the unfolded protein response is activated by HSPA5 (also known as BiP/GRP78), a key regulator of the unfolded protein response, binding to unfolded proteins. This is a response attempting to restore homeostasis by reducing the influx of proteins to the ER and increasing the protein folding capacity [301]. The unfolded protein response must balance the folding capacity and the secretory requirements of the cell. Prolonged endoplasmic reticulum stress may induce neuronal cell death through a number of not fully understood processes, where, among others, calpains appear to act detrimentally and HSPA5 protectively [425]. Given the well-established pathology hallmark of alpha-synuclein oligomerisation in Parkinson's disease, indications of ER stress and unfolded protein response activation are not unexpected and there are several studies confirming a link between these pathways and Parkinson's disease pathology [304]. The unfolded protein response has been targeted in PD studies, investigating its role in reducing ER stress. A rat model of Parkinson's disease found that alpha-synuclein-induced neurotoxicity decreased when overexpressing HSPA5 and attributed this to HSPA5's modulation of the unfolded protein response, acting to downregulate ER stress, thereby reducing apoptosis, and promoting survival of dopaminergic neurons [426]. It is conceivable that the unfolded protein response acts protectively in Parkinson's disease. However, the upregulation of HSPA1L, a component of the ERAD complex, may suggest that ER stress is still present. Moreover, in the targeted urine study, we found the calpain protein CAPN2 significantly upregulated in the PD patients ($p = 2.4\ E^{-3}$). Calpain is implicated in ER stress-induced neuronal cell death and has also been found upregulated in microglia, astrocytes, and neurons [351, 425]. Prolonged ER stress can moreover induce production of pro-inflammatory cytokines [427], and in urine from PD patients, we found

CCL2 upregulated ($p = 2.4E^{-2}$). CCL2 is involved in immune regulation and inflammatory processes which bring upon a strong response [361]. CCL2 has been suggested to contribute to neuroinflammation in PD [362].

In the centenarians, the lack of expression difference for HSPA5 and HSPA1L when compared to controls suggests that the centenarians' endoplasmic reticulum is not affected or protected against mis- or unfolded proteins. There may exist a link to the upregulation of A2M in the centenarians, possibly clearing aggregated proteins more efficiently and thereby protecting them from neurodegeneration.

8.1.5 *Wnt signalling in Parkinson's disease*

In plasma from newly diagnosed PD patients, we found the protein DKK3 strongly downregulated ($p = 5.5\ E^{-10}$) when compared to healthy and age-matched controls. We also found the protein PPP3CB strongly downregulated in the same comparison ($p = 1.7\ E^{-9}$). In the centenarians, DKK3 and PPP3CB could not be quantitated as they were below the detection limit or potentially due to other reasons, such as the proteins being labile.

DKK3 is a modulator of the canonical Wnt/ß-catenin signalling pathway, and PPP3CB (also known as calcineurin A2) is a component of the Wnt/Ca^{2+} signalling pathway. Wnt signalling encompasses a set of complex pathways involved in several aspects of cell development [307, 308]. In the adult brain, Wnt signalling governs a range of crucial functions, including neuronal survival, synapse formation, neurogenesis and regeneration [312]. Wnt signalling is vital for the development and maintenance of dopaminergic neurons and it has been suggested that activation of the pathway could provide an important protective role, possibly preventing the loss of dopaminergic neurons [293]. Indeed, this makes the pathway highly relevant for Parkinson's disease. Given the putative potential of restoring dopaminergic neurons' functions, Wnt signalling has been the attention of a number of studies in recent years [311]. Downregulation of Wnt signalling has been proposed to foster dysfunction and/or death of dopaminergic neurons, however, a mice study demonstrated that the dopaminergic neurons could be restored by activating ß-catenin in situ [313]. In another mice study, neural stem cells were transplanted to the substantia nigra of medically PD-induced mice, re-expression of Wnt signalling promoters and repair of dopaminergic neurons could be noticed [314].

Upregulation of DKK3 activates the Wnt/ß-catenin signalling pathway [311], which in theory acts protectively on the dopaminergic neurons. It is therefore interesting that the

newly diagnosed Parkinson's patients express significantly lower levels of this potentially PD-protective protein, and of the Wnt/Ca^{2+} component PPP3CB, compared to the healthy controls. Our results thus suggest that Wnt signalling may be disrupted in the early stages of Parkinson's disease, potentially acting detrimental on the survival of the dopaminergic neurons. As previously stated, to our knowledge, this is the first time a targeted proteomic study of Parkinson's disease finds DKK3 expression significantly downregulated in patients.

Apart from its role in neurogenesis and maintenance of neurons in the brain, Wnt signalling is also proposed to modulate neuroinflammation. Studies have suggested that Wnt/ß-catenin signalling is activated after injury to the brain, acting neuroprotective [428, 429]. It has been suggested that components of the Wnt signalling pathway (Wnt1) can reduce oxidative stress and repress pro-inflammatory microglial activation, possibly through crosstalk between Wnt and inflammatory signalling pathways [430]. We found the pro-inflammatory matrix metalloprotease MMP3 and the pro-inflammatory cytokine CCL2 significantly upregulated ($p = 5.3\ E^{-3}$ and $p = 2.4\ E^{-2}$, respectively) in urine from PD patients when compared to healthy controls. These proteins are known to be involved in neuroinflammation, with MMP3 inducing cytokine release from microglia, and CCL2 being a potent chemoattractant, recruiting white blood cells [431, 432]. Moreover, we also found the protein MAPK12 significantly upregulated in the PD patients' urine compared to controls ($p = 9.0\ E^{-3}$). MAPK12 is one of four proteins in the p38 MAP kinase family. The p38 MAPK proteins are activated by a range of stress stimuli, including cytokines, such as CCL2, and oxidative stress; the p38 pathway increases production of cytokines and is also involved in apoptosis [381, 383]. Increased p38 signalling may furthermore inhibit Wnt signalling by promoting the expression of one of its inhibitors, as have been reported in cancer studies [433, 434]. This would indicate a self-amplifying loop of neuroinflammation – with increased production of pro-inflammatory species by microglia in conjunction with downregulation of the neuroprotective Wnt pathway.

It would have been beneficial to the study to compare the expression of the Wnt proteins, CCL, MMP3 and MAPK12 between the Parkinson's patients and the centenarians but as the Wnt proteins could not be detected in the centenarian study potentially due to technical restraints, and CCL2, MMP3 and MAPK12 were only detected in urine, this was not possible.

In summary, the protein expression observed in our studies demonstrates that although there is a degree of similarity between the newly diagnosed Parkinson's patients and the

centenarians, a number of proteins and pathways differ, and may thus better the understanding of why some individuals develop neurodegenerative disease as they age, and some do not. Our results suggest that the Parkinson's disease patients may be negatively affected by complement activation and may suffer from prolonged endoplasmic reticulum stress and neuroinflammation. They may moreover be subjected to disrupted or downregulated Wnt signalling, possibly aggravating neuroinflammation and reducing their capacity to protect dopaminergic neurons from cell death. In the centenarians, we suggest that a reduced or lower level of complement activation may promote longevity. The upregulation of A2M may protect the centenarians from accumulation of mis- and unfolded proteins, as suggested by the lack of ER stress and unfolded protein response indications. Upregulation of ADIPOQ may further exert an overall anti-inflammatory and longevity-promoting influence on the centenarians. Figure 8-1 provides an overview of the observed protein expression and possible implications for the two groups.

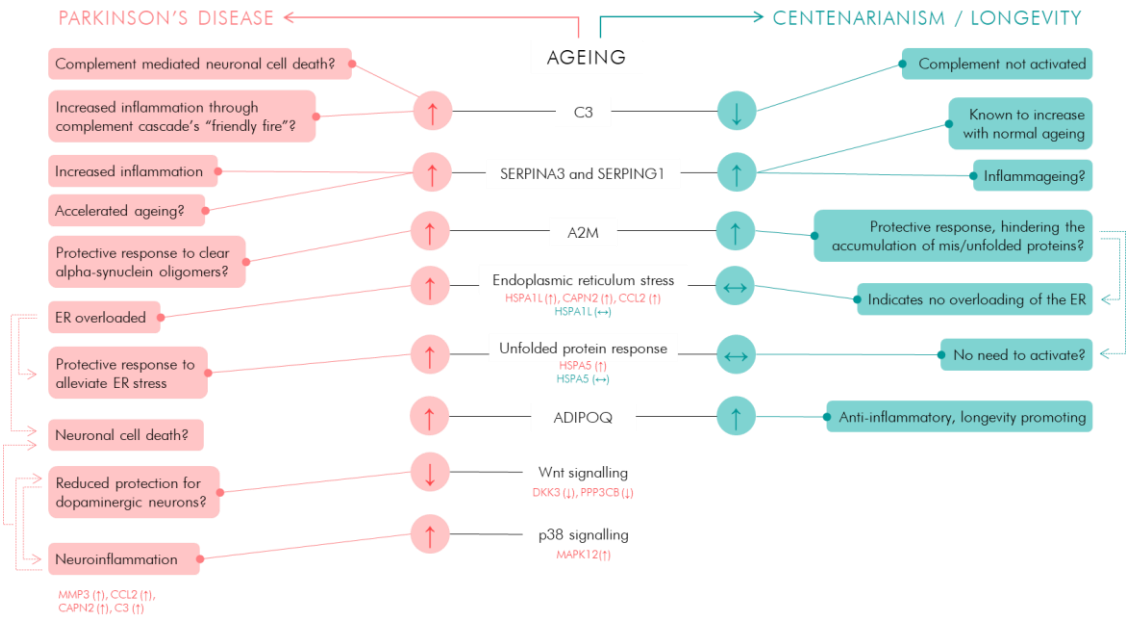


Figure 8-1. Observed proteins, involvement in pathways and possible implications of the protein expression from the studies of centenarians and newly diagnosed Parkinson's disease patients. *In the middle of the plot, proteins and pathways are shown, and their expression in the Parkinson's patients (■) and in the centenarians (■) when compared to their respective control group are represented by arrows, where ↑ upregulated, ↓ downregulated, and ↔ no difference. C3 was downregulated in the centenarians, indicating no complement activation, and upregulated in the PD patients thus indicating activation and possible involvement in complement mediated neuronal cell death. SERPINA3 and SERPING1 were upregulated in both groups, putatively indicating inflammageing in the centenarians and increased inflammation/accelerated ageing in the PD patients. A2M was upregulated in both groups and may act to clear mis- or unfolded proteins. Signs of endoplasmic reticulum stress were observed in the PD patients but not in the centenarians. The centenarians did not demonstrate indications of the unfolded protein response being activated, while the PD patients did. The longevity promoting ADIPOQ was upregulated in the centenarians and in the PD patients. The role of ADIPOQ in Parkinson's disease remains unclear. The PD patients showed signs of decreased Wnt signalling and increased p38 MAPK signalling, the pathways possibly modulating each other and aggravating neuroinflammation and promoting neuronal cell death.*

8.1.6 *General considerations and study limitations*

There are a few considerations and study limitations to recognise in this work. Firstly, as discussed in Chapter 4, the lack of suitable controls in the centenarian study does contain a caveat as there is no entirely certain approach for determining if the centenarians' protein expression is related to their extremely advanced age, their longevity, or the fact that they have lived through different times than their younger controls. Our solution for identifying proteins putatively linked to longevity despite this constraint was to study the centenarians' protein expression in conjunction with the available literature from previous proteomic studies on normal ageing and longevity. Another possible solution for differentiating between protein expression related to normal ageing and longevity would have been to generate models of protein expression related to age from the control group, and then predict the centenarians to determine if their protein expression indicated them as younger or at their biological age. This was attempted in Chapter 4, section 4.3.2.3, but as the age-range of the control/offspring group was too narrow (mean age = 70.6 ± 6.8 years, range: 54 – 89 years), the strategy was not feasible.

Another consideration is the need for further validation of the proteins identified in the targeted phase. The targeted methodology in this work consisted of a fusion between a strictly validatory and an explorative targeted assay. We included targets from several studies of neurodegeneration, and also pro- and anti-inflammatory proteins from literature, and applied this to all the targeted cohorts. The rationale behind setting the targeted assay up in this fashion, rather than as separate validation studies only monitoring the putative biomarkers form each separate study, was to allow for measurement of low-abundant inflammatory proteins from literature, which might not have been detected in the discovery phase. It was further to allow for comparison of the protein expressions between the studies, something which would not have been possible in a strict validation methodology as the same proteins would not have been measured across the different cohorts. Thanks to this strategy, we were able to identify new putative targets also in the targeted phase which would not have been discovered if we had performed a strict validation study. However, this consequently infers that another validation study will need to be performed to confirm the newly identified targets.

The machine learning models developed in Chapters 5 and 6 could classify samples as PD or control with exceptional accuracy in plasma and with good accuracy in urine. The models have the distinct advantage of being trained on newly diagnosed and treatment-naïve Parkinson's disease patients, thereby closely resembling the conditions for the

individuals they would be used to classify. Models trained on patients with advanced PD, on treatment, may have limited utility for screening patients for PD as they are based on a state which may not be representative for individuals in very early stages of the disease and not on treatment. However, as discussed previously, the selectivity of the models was deemed lacking as they could not adequately separate the group of other neurological disorders from Parkinson's disease. This was largely seen as a result of the protein expression between the other neurological disorders and the PD groups being very similar for a number of the proteins included in the models. Further work would need to be undertaken to improve the selectivity and specificity, but we argue that the models are nonetheless highly useful as they could be used to screen individuals for Parkinson's disease as a starting point and be followed up by clinical evaluation to verify the diagnosis.

8.2 CONCLUSIONS

The overarching aim of the three studies herein presented, was to find novel biomarkers for Parkinson's disease and to identify proteins and pathways related to healthy ageing in centenarians. In the two Parkinson's disease studies of urine and plasma, a number of proteins was found which, when modelled as a panel, separated Parkinson's patients from control with very good accuracy. The models could successfully predict 100% of the samples as Parkinson's disease or control in plasma and 85.1% in urine. In the centenarian study, it was hypothesised that the centenarians displayed a protein expression related to both risk and protective mechanisms. Comparing the protein expression between Parkinson's disease patients and centenarians, proteins and pathways putatively involved in the divergence between neurodegeneration and healthy ageing were identified. Without performing functional studies to validate the presented theories, we can only hypothesise about the roles of the differentially expressed proteins, however, it is clear that several of the findings correlate with the pathology observed in Parkinson's disease. In conclusion, the protein expression from our experiments suggests the following:

- ***Complement C3 may be involved in Parkinson's disease, and in longevity.*** The upregulation in PD patients indicates complement activation and may also contribute to synapse elimination. The downregulation in the centenarians suggests that reduced complement activation may promote longevity.

- ***Upregulation of A2M is a protective response in the early stages of Parkinson's disease, and promotes longevity in centenarians.*** A2M can bind misfolded proteins and is hypothesised to, in the PD patients, attempt to

bind and clear alpha-synuclein oligomers. In the centenarians, it is believed to continuously be clearing misfolded proteins and inhibit inflammatory processes such as the complement cascade.

- ***Endoplasmic reticulum stress and activation of the unfolded protein response are present in Parkinson's disease but not in centenarians.*** The unfolded protein response is likely a protective response in the early stages of Parkinson's disease, however the upregulation of HSPA1L, CAPN2 and CCL2 indicate that ER stress is still present. Possibly due to efficient clearance of unfolded proteins through the upregulation of A2M, the centenarians do not exhibit signs of ER stress or activation of the unfolded protein response.

- ***Wnt signalling may be downregulated in Parkinson's disease.*** The strong downregulation of the Wnt signalling activating protein DKK3 in PD patients suggests that the Wnt pathway may be downregulated or disrupted. Wnt signalling is theorised to act protectively on dopaminergic neurons, and therefore, dysregulation of the pathway would be detrimental to Parkinson's disease. Moreover, Wnt signalling may protect against neuroinflammation and thus, downregulation of the pathway would act to aggravate the PD pathology. In our study, the upregulation of the neuroinflammatory proteins MMP3, CCL2 and CAPN2 supports this. p38 signalling may further suppress Wnt signalling, as indicated by the upregulation of the p38 protein MAPK12 in our study.

8.3 FUTURE WORK, STUDIES AND PERSPECTIVES

Although the presented studies have identified promising mechanisms and biomarker targets for Parkinson's disease, there are further experiments necessary to verify the validity of the proteomic biomarkers. Therefore, the proposed future work and studies are:

Validate the protein panels in larger sample cohorts. Biomarker studies, especially for non-hereditary neurodegenerative diseases, have often been described as "one hit wonders" where findings from the discovery phase do not validate in targeted analyses of larger sample cohorts. Our approach to the targeted study following the discovery phase was a hybrid between a validatory and an explorative assay, as we included targets identified in all the different discovery studies in one targeted assay, and also included inflammatory markers from literature. This augmented assay was applied to all validation sample cohorts, and we identified new targets in the validation phase. For this reason, the new findings which emerged from the targeted phase would need to be verified in a new study.

Analyse urine samples from LRRK2 mutation carriers. At the time of analysis of the targeted urine validation study, collection of urine samples from LRRK2 mutation carriers was still ongoing and therefore this group was not included in the validation phase. A large number of LRRK2 mutation carrier samples are currently being collected by Dr Mie Rizig (Institute of Neurology, UCL) as part of a grant resulting from the work presented in this thesis and will be analysed in late 2022.

Translate the validated proteins into a clinical assay. The proteins which are confirmed in the extended validation phase need to be translated into a clinical assay. For the clinical assay to be applicable to large-scale patient screening, it should ideally be rapid to prepare, run and analyse the samples. The sample preparation procedure may thus need to be refined to allow for speedy sample processing and to ensure that low-abundant compounds, such as DKK3, are easily detectable. One option for sample preparation may be utilising a polyclonal antibody pull out, specific for the monitored peptides. This would leave the samples containing only the pulled-out peptides, thus highly enriched and without possible interfering species. Moreover, the required time of digestion may be evaluated and possibly shortened, thereby reducing the sample preparation time. Stable isotope labelled peptides should be used as internal standards for each peptide, ensuring accurate quantitation. Method development will further need to be undertaken to create an analytical LCMS method which is as rapid as possible, while not compromising sensitivity and specificity. This work will moreover include validating the assay thoroughly to set up suitable calibration curve ranges for all peptides and determine limit of detection and limit of quantitation. A suitable quality control system must also be set up to ensure that robustness between runs is maintained. Stability studies need to be undertaken to determine the effects of different anti-coagulants used for plasma samples, the difference between plasma and serum, the influence of sample storage and freeze-thaw cycles, and the peptides' stability over time. A suitable machine learning model may be developed from the proteins in the assay and this model needs to be trained and assessed to set limits for sample classification.

Investigate the role of Wnt signalling and neuroinflammation in Parkinson's disease. The possible link to reduced Wnt signalling and neuroinflammation is exceptionally interesting and relevant to Parkinson's disease and would thus benefit from further exploration. This could be examined in cell or animal models, where neurons might be co-cultured with microglia and the effect of the levels of DKK3 on the neuronal and microglial expressions evaluated. The identification of proteins that are elevated or reduced could be studies further by using chemical inhibitors, gene editing or by silencing techniques.

The ability to enhance or inhibit proteomic pathways observed in this thesis and examine if some phenotypes are observed in the models would go a long way to confirming the hypotheses why some people develop Parkinson's disease and some live extraordinarily long lives.

References

1. Strimbu, K. and J.A. Tavel, *What are biomarkers?* Curr Opin HIV AIDS, 2010. **5**(6): p. 463-6.
2. Olivier, M., et al., *The Need for Multi-Omics Biomarker Signatures in Precision Medicine.* Int J Mol Sci, 2019. **20**(19).
3. American Diabetes, A., *Diagnosis and classification of diabetes mellitus.* Diabetes Care, 2010. **33 Suppl 1**: p. S62-9.
4. Goldman, R., *Creatinine excretion in renal failure.* Proc Soc Exp Biol Med, 1954. **85**(3): p. 446-8.
5. Pepys, M.B., *C-reactive protein fifty years on.* Lancet, 1981. **1**(8221): p. 653-7.
6. Khan, M.H., et al., *Serum C-reactive protein levels correlate with clinical response in patients treated with antibiotics for wound infections after spinal surgery.* Spine J, 2006. **6**(3): p. 311-5.
7. Chace, D.H. and T.A. Kalas, *A biochemical perspective on the use of tandem mass spectrometry for newborn screening and clinical testing.* Clin Biochem, 2005. **38**(4): p. 296-309.
8. Rashed, M.S., *Clinical applications of tandem mass spectrometry: ten years of diagnosis and screening for inherited metabolic diseases.* J Chromatogr B Biomed Sci Appl, 2001. **758**(1): p. 27-48.
9. Conway, S.R. and H.R. Wong, *Biomarker Panels in Critical Care.* Crit Care Clin, 2020. **36**(1): p. 89-104.
10. Yamada, R., et al., *Interpretation of omics data analyses.* J Hum Genet, 2021. **66**(1): p. 93-102.
11. Rifai, N., M.A. Gillette, and S.A. Carr, *Protein biomarker discovery and validation: the long and uncertain path to clinical utility.* Nat Biotechnol, 2006. **24**(8): p. 971-83.
12. Spagou, K., et al., *HILIC-UPLC-MS for exploratory urinary metabolic profiling in toxicological studies.* Anal Chem, 2011. **83**(1): p. 382-90.
13. Lindon, J.C. and J.K. Nicholson, *Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery.* Trac-Trends in Analytical Chemistry, 2008. **27**(3): p. 194-204.
14. Xu, P., D.M. Duong, and J.M. Peng, *Systematical Optimization of Reverse-Phase Chromatography for Shotgun Proteomics.* Journal of Proteome Research, 2009. **8**(8): p. 3944-3950.
15. Dorsey, J.G. and K.A. Dill, *The molecular mechanism of retention in reversed-phase liquid chromatography.* Chemical Reviews, 1989. **89**(2): p. 331-346.
16. Zhang, T., et al., *Evaluation of coupling reversed phase, aqueous normal phase, and hydrophilic interaction liquid chromatography with Orbitrap mass spectrometry for metabolomic studies of human urine.* Anal Chem, 2012. **84**(4): p. 1994-2001.
17. Harris Daniel, C., *Quantitative chemical analysis.* 2010, New York: W H Freeman and Company. Various paginations ill., photos, charts 29 cm.
18. Swartz, M.E., *UPLC (TM): An introduction and review.* Journal of Liquid Chromatography & Related Technologies, 2005. **28**(7-8): p. 1253-1263.
19. Novakova, L., L. Matysova, and P. Solich, *Advantages of application of UPLC in pharmaceutical analysis.* Talanta, 2006. **68**(3): p. 908-18.
20. Wilson, S.R., et al., *Nano-LC in proteomics: recent advances and approaches.* Bioanalysis, 2015. **7**(14): p. 1799-815.
21. Gilar, M., et al., *Orthogonality of separation in two-dimensional liquid chromatography.* Anal Chem, 2005. **77**(19): p. 6426-34.
22. Wang, H., et al., *An off-line high pH reversed-phase fractionation and nano-liquid chromatography-mass spectrometry method for global proteomic profiling of cell lines.* J Chromatogr B Analyt Technol Biomed Life Sci, 2015. **974**: p. 90-5.
23. SOCIETY, A.P., *April 1946: First Concept of Time-of-Flight Mass Spectrometer*, in *This Month in Physics History*, A. Chodos, Editor. 2001.
24. Jones, M., *Organic chemistry.* 2000, New York: W.W. Norton.
25. Gross, J.H. and SpringerLink, *Mass spectrometry : a textbook.* 2004, Berlin : London: Springer. 510 : ill.; 24 cm.

26. Cohen, A., *Mass Spectrometry, Review of the Basics: Electrospray, MALDI and Commonly Used Mass Analyzers (vol 44, pg 210, 2009)*. Applied Spectroscopy Reviews, 2009. **44**(4): p. 362-362.
27. Awad, H., M.M. Khamis, and A. El-Aneed, *Mass Spectrometry, Review of the Basics: Ionization*. Applied Spectroscopy Reviews, 2015. **50**(2): p. 158-175.
28. Fenn, J.B., *Electrospray wings for molecular elephants (Nobel lecture)*. Angew Chem Int Ed Engl, 2003. **42**(33): p. 3871-94.
29. Kromidas, S., *More practical problem solving in HPLC*. 2005, Weinheim ; Great Britain: Wiley-VCH. xv, 294 p.
30. Iribarne, J.V. and B.A. Thomson, *On the evaporation of small ions from charged droplets*. Journal of Chemical Physics, 1976. **64**(6): p. 2287-2294.
31. Banerjee, S. and S. Mazumdar, *Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte*. Int J Anal Chem, 2012. **2012**: p. 282574.
32. Kebarle, P., *A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry*. Journal of Mass Spectrometry, 2000. **35**(7): p. 804-817.
33. Batey, J.H., *The physics and technology of quadrupole mass spectrometers*. Vacuum, 2014. **101**: p. 410-415.
34. Ho, C.S., et al., *Electrospray ionisation mass spectrometry: principles and clinical applications*. Clin Biochem Rev, 2003. **24**(1): p. 3-12.
35. Hart-Smith, G. and S.J. Blanksby, *Mass Analysis*, in *Mass Spectrometry in Polymer Chemistry*. 2012. p. 5-32.
36. Price, D., *Time-of-flight mass spectrometry: the early years as chronicled by the European time-of-flight symposia*. 1994, ACS Publications.
37. Cotter, R.J., *Time-of-Flight Mass-Spectrometry - an Increasing Role in the Life Sciences*. Biomedical and Environmental Mass Spectrometry, 1989. **18**(8): p. 513-532.
38. Shalliker, R.A., *Hyphenated and alternative methods of detection in chromatography*. 2011: CRC Press.
39. Cotter, R.J., *Peer Reviewed: The New Time-of-Flight Mass Spectrometry*. Anal Chem, 1999. **71**(13): p. 445A-51A.
40. Mamyrin, B., et al., *The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution*. Zh. Eksp. Teor. Fiz, 1973. **64**(1): p. 82-89.
41. Kanu, A.B., et al., *Ion mobility-mass spectrometry*. Journal of Mass Spectrometry, 2008. **43**(1): p. 1-22.
42. Revercomb, H. and E.A. Mason, *Theory of plasma chromatography/gaseous electrophoresis. Review*. Analytical chemistry, 1975. **47**(7): p. 970-983.
43. Medhe, S., *Mass spectrometry: detectors review*. Chem Biomol Eng, 2018. **3**: p. 51-58.
44. Van Bramer, S.E., *An introduction to mass spectrometry*. Widener University, Department of Chemistry, One University Place, Chester, PA, 1998. **19013**.
45. Arnaud, C.H., *As the triple quadrupole turns 40, mass spec gurus look back on what it's meant to chemistry*, in *C&EN*. 2018.
46. Sleno, L. and D.A. Volmer, *Ion activation methods for tandem mass spectrometry*. Journal of Mass Spectrometry, 2004. **39**(10): p. 1091-1112.
47. Yang, L.Y., et al., *Investigation of an enhanced resolution triple quadrupole mass spectrometer for high-throughput liquid chromatography/tandem mass spectrometry assays*. Rapid Communications in Mass Spectrometry, 2002. **16**(21): p. 2060-2066.
48. Yates, J.R., *A century of mass spectrometry: from atoms to proteomes*. Nature Methods, 2011. **8**(8): p. 633-637.
49. Chait, B.T., *Chemistry. Mass spectrometry: bottom-up or top-down?* Science, 2006. **314**(5796): p. 65-6.
50. Graves, P.R. and T.A. Haystead, *Molecular biologist's guide to proteomics*. Microbiol Mol Biol Rev, 2002. **66**(1): p. 39-63; table of contents.

51. Rodriguez, J., et al., *Does trypsin cut before proline?* J Proteome Res, 2008. **7**(1): p. 300-5.
52. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.
53. Waters, *An Overview of the Principles of MSE, The Engine that Drives MS Performance [White paper]*. 2011.
54. Marx, V., *Targeted proteomics.* Nature methods, 2013. **10**(1): p. 19.
55. Goldman, M., *Education in Medicine: Moving the Boundaries to Foster Interdisciplinarity.* Frontiers in Medicine, 2016. **3**: p. 1-3.
56. Wiederin, J. and P. Ciborowski, *6 - Immunoaffinity Depletion of Highly Abundant Proteins for Proteomic Sample Preparation*, in *Proteomic Profiling and Analytical Chemistry (Second Edition)*, P. Ciborowski and J. Silberring, Editors. 2016, Elsevier: Boston. p. 101-114.
57. Fang, X. and W.W. Zhang, *Affinity separation and enrichment methods in proteomic analysis.* J Proteomics, 2008. **71**(3): p. 284-303.
58. Zelena, E., et al., *Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum.* Anal Chem, 2009. **81**(4): p. 1357-64.
59. Dunn, W.B., et al., *Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry.* Nat Protoc, 2011. **6**(7): p. 1060-83.
60. Jonsson, P., et al., *Constrained randomization and multivariate effect projections improve information extraction and biomarker pattern discovery in metabolomics studies involving dependent samples.* Metabolomics, 2015. **11**(6): p. 1667-1678.
61. Collins, G.S. and K.G.M. Moons, *Reporting of artificial intelligence prediction models.* Lancet, 2019. **393**(10181): p. 1577-1579.
62. Chen, J.H. and S.M. Asch, *Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations.* New England Journal of Medicine, 2017. **376**(26): p. 2507-2509.
63. Magnello, M.E., *Karl Pearson and the Establishment of Mathematical Statistics.* International Statistical Review / Revue Internationale de Statistique, 2009. **77**(1): p. 3-29.
64. Pearson, K., *On lines and planes of closest fit to systems of points in space.* Philosophical Magazine, 1901. **2**(7-12): p. 559-572.
65. Trygg, J., E. Holmes, and T. Lundstedt, *Chemometrics in metabonomics.* J Proteome Res, 2007. **6**(2): p. 469-79.
66. Eriksson, L., et al., *Multi- and Megavariate Data Analysis (part I)*. 2 ed. 2006, Umeå: Umetrics Academy.
67. Wold, S., M. Sjostrom, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems, 2001. **58**(2): p. 109-130.
68. Trygg, J. and S. Wold, *Orthogonal projections to latent structures (O-PLS).* Journal of Chemometrics, 2002. **16**(3): p. 119-128.
69. Eriksson, L., et al., *Multi- and Megavariate Data Analysis (part II)*. 2 ed. 2006, Umeå: Umetrics Academy.
70. Hoerl, R.W., *Ridge Regression: A Historical Context.* Technometrics, 2020. **62**(4): p. 420-425.
71. Jain, R.K., *Ridge-Regression and Its Application to Medical Data.* Computers and Biomedical Research, 1985. **18**(4): p. 363-368.
72. Tibshirani, R., *Regression shrinkage and selection via the Lasso.* Journal of the Royal Statistical Society Series B-Methodological, 1996. **58**(1): p. 267-288.
73. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005)*. Journal of the Royal Statistical Society Series B-Statistical Methodology, 2005. **67**: p. 768-768.
74. Plant, R.E., *Linear Discriminant Analysis.* 2019.

75. Balakrishnama, S. and A. Ganapathiraju, *Linear discriminant analysis-a brief tutorial.* Institute for Signal and information Processing, 1998. **18**(1998): p. 1-8.

76. Tharwat, A., et al., *Linear discriminant analysis: A detailed tutorial.* Ai Communications, 2017. **30**(2): p. 169-190.

77. Wang, W. *Sustainable Development Evaluation Based on Support Vector Machine and Principle Component Analysis.* in *2015 International Conference on Education, Management, Information and Medicine.* 2015. Atlantis Press.

78. Noble, W.S., *What is a support vector machine?* Nature Biotechnology, 2006. **24**(12): p. 1565-1567.

79. Pisner, D.A. and D.M. Schnyer, *Chapter 6 - Support vector machine*, in *Machine Learning*, A. Mechelli and S. Vieira, Editors. 2020, Academic Press. p. 101-121.

80. Cai, J., et al., *Feature selection in machine learning: A new perspective.* Neurocomputing, 2018. **300**: p. 70-79.

81. Xue, B., et al., *A Survey on Evolutionary Computation Approaches to Feature Selection.* Ieee Transactions on Evolutionary Computation, 2016. **20**(4): p. 606-626.

82. Stone, M., *Cross-Validatory Choice and Assessment of Statistical Predictions.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1974. **36**(2): p. 111-147.

83. Wong, T.T., *Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation.* Pattern Recognition, 2015. **48**(9): p. 2839-2846.

84. Celisse, A. and S. Robin, *Nonparametric density estimation by exact leave-p-out cross-validation.* Computational Statistics & Data Analysis, 2008. **52**(5): p. 2350-2368.

85. Jung, Y., *Multiple predicting K-fold cross-validation for model selection.* Journal of Nonparametric Statistics, 2018. **30**(1): p. 197-215.

86. Berrar, D., *Cross-Validation.* 2019.

87. Schapira, A.H.V., K.R. Chaudhuri, and P. Jenner, *Non-motor features of Parkinson disease.* Nat Rev Neurosci, 2017. **18**(8): p. 509.

88. Shulman, L.M., et al., *Non-recognition of depression and other non-motor symptoms in Parkinson's disease.* Parkinsonism Relat Disord, 2002. **8**(3): p. 193-7.

89. Ferreira, M. and J. Massano, *An updated review of Parkinson's disease genetics and clinicopathological correlations.* Acta Neurol Scand, 2017. **135**(3): p. 273-284.

90. Poewe, W., et al., *Parkinson disease.* Nat Rev Dis Primers, 2017. **3**: p. 17013.

91. Ascherio, A. and M.A. Schwarzschild, *The epidemiology of Parkinson's disease: risk factors and prevention.* Lancet Neurol, 2016. **15**(12): p. 1257-1272.

92. Elbaz, A., et al., *Epidemiology of Parkinson's disease.* Rev Neurol (Paris), 2016. **172**(1): p. 14-26.

93. Postuma, R.B., et al., *Parkinson risk in idiopathic REM sleep behavior disorder: preparing for neuroprotective trials.* Neurology, 2015. **84**(11): p. 1104-13.

94. Parkinson, J., *An essay on the shaking palsy. 1817.* J Neuropsychiatry Clin Neurosci, 2002. **14**(2): p. 223-36; discussion 222.

95. Rodriguez-Oroz, M.C., et al., *Initial clinical manifestations of Parkinson's disease: features and pathophysiological mechanisms.* Lancet Neurol, 2009. **8**(12): p. 1128-39.

96. Dickson, D.W., *Neuropathology of Parkinson disease.* Parkinsonism Relat Disord, 2018. **46 Suppl 1**: p. S30-S33.

97. Marques, O. and T.F. Outeiro, *Alpha-synuclein: from secretion to dysfunction and death.* Cell Death Dis, 2012. **3**: p. e350.

98. UniProt. *UniProtKB - P37840 (SYUA_HUMAN).* [cited 2019 08-07]; Available from: https://www.uniprot.org/uniprot/P37840.

99. Blesa, J., et al., *Oxidative stress and Parkinson's disease.* Front Neuroanat, 2015. **9**: p. 91.

100. Rocha, E.M., B. De Miranda, and L.H. Sanders, *Alpha-synuclein: Pathology, mitochondrial dysfunction and neuroinflammation in Parkinson's disease.* Neurobiol Dis, 2018. **109**(Pt B): p. 249-257.

101. Ferreira, S.A. and M. Romero-Ramos, *Microglia Response During Parkinson's Disease: Alpha-Synuclein Intervention.* Front Cell Neurosci, 2018. **12**: p. 247.

102. Stirnemann, J., et al., *A Review of Gaucher Disease Pathophysiology, Clinical Presentation and Treatments.* Int J Mol Sci, 2017. **18**(2).

103. Sidransky, E., et al., *Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease.* N Engl J Med, 2009. **361**(17): p. 1651-61.

104. Postuma, R.B., et al., *MDS clinical diagnostic criteria for Parkinson's disease.* Mov Disord, 2015. **30**(12): p. 1591-601.

105. Marsili, L., G. Rizzo, and C. Colosimo, *Diagnostic Criteria for Parkinson's Disease: From James Parkinson to the Concept of Prodromal Disease.* Front Neurol, 2018. **9**: p. 156.

106. Lewitt, P.A., *Levodopa for the treatment of Parkinson's disease.* N Engl J Med, 2008. **359**(23): p. 2468-76.

107. Katzenschlager, R., et al., *Apomorphine subcutaneous infusion in patients with Parkinson's disease with persistent motor fluctuations (TOLEDO): a multicentre, double-blind, randomised, placebo-controlled trial.* Lancet Neurology, 2018. **17**(9): p. 749-759.

108. Okun, M.S., *Deep-brain stimulation for Parkinson's disease.* N Engl J Med, 2012. **367**(16): p. 1529-38.

109. Li, T. and W. Le, *Biomarkers for Parkinson's Disease: How Good Are They?* Neurosci Bull, 2020. **36**(2): p. 183-194.

110. Cova, I. and A. Priori, *Diagnostic biomarkers for Parkinson's disease at a glance: where are we?* J Neural Transm (Vienna), 2018. **125**(10): p. 1417-1432.

111. WHO. *Ageing and health.* 2020 2020-06-11]; Available from: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

112. Hoffman, J.M., et al., *Proteomics and metabolomics in ageing research: from biomarkers to systems biology.* Essays Biochem, 2017. **61**(3): p. 379-388.

113. Nkuipou-Kenfack, E., et al., *Proteome analysis in the assessment of ageing.* Ageing Res Rev, 2014. **18**: p. 74-85.

114. Rowe, J.W. and R.L. Kahn, *Successful aging.* Gerontologist, 1997. **37**(4): p. 433-440.

115. Rowe, J.W. and R.L. Kahn, *Human Aging - Usual and Successful.* Science, 1987. **237**(4811): p. 143-149.

116. Franceschi, C. and M. Bonafe, *Centenarians as a model for healthy aging.* Biochemical Society Transactions, 2003. **31**: p. 457-461.

117. Passarino, G., et al., *Male/female ratio in centenarians: a possible role played by population genetic structure.* Experimental gerontology, 2002. **37**(10-11): p. 1283-1289.

118. Samaras, T.T., *Longevity of Specific Populations*, in *International Encyclopedia of Public Health (Second Edition)*, S.R. Quah, Editor. 2017, Academic Press: Oxford. p. 464-468.

119. Hitt, R., et al., *Centenarians: the older you get, the healthier you have been.* The Lancet, 1999. **354**(9179): p. 652.

120. Motta, M., et al., *Successful aging in centenarians: myths and reality.* Archives of Gerontology and Geriatrics, 2005. **40**(3): p. 241-251.

121. Jin, K., *Modern Biological Theories of Aging.* Aging Dis, 2010. **1**(2): p. 72-74.

122. Weinert, B.T. and P.S. Timiras, *Invited review: Theories of aging.* J Appl Physiol (1985), 2003. **95**(4): p. 1706-16.

123. Trindade, L.S., et al., *A novel classification system for evolutionary aging theories.* Front Genet, 2013. **4**: p. 25.

124. Baraibar, M.A., et al., *Expression and modification proteomics during skeletal muscle ageing.* Biogerontology, 2013. **14**(3): p. 339-52.

125. Woods, J.A., et al., *Exercise, inflammation and aging.* Aging Dis, 2012. **3**(1): p. 130-40.

126. Franceschi, C. and J. Campisi, *Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases.* J Gerontol A Biol Sci Med Sci, 2014. **69 Suppl 1**: p. S4-9.

127. Ferrucci, L. and E. Fabbri, *Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty.* Nature Reviews Cardiology, 2018. **15**(9): p. 505-522.

128. Finkel, T. and N.J. Holbrook, *Oxidants, oxidative stress and the biology of ageing.* Nature, 2000. **408**(6809): p. 239-47.

129. Harper, M.E., et al., *Ageing, oxidative stress, and mitochondrial uncoupling.* Acta Physiologica Scandinavica, 2004. **182**(4): p. 321-331.

130. Vitale, G., S. Salvioli, and C. Franceschi, *Oxidative stress and the ageing endocrine system.* Nature Reviews Endocrinology, 2013. **9**(4): p. 228.

131. Aubert, G. and P.M. Lansdorp, *Telomeres and aging.* Physiol Rev, 2008. **88**(2): p. 557-79.

132. Kelly, D.P., *Cell biology: Ageing theories unified.* Nature, 2011. **470**(7334): p. 342-3.

133. Perls, T., *Centenarians who avoid dementia.* Trends Neurosci, 2004. **27**(10): p. 633-6.

134. Perls, T.T., *Cognitive Trajectories and Resilience in Centenarians-Findings From the 100-Plus Study.* JAMA Netw Open, 2021. **4**(1): p. e2032538.

135. Mollenhauer, B., et al., *Nonmotor and diagnostic findings in subjects with de novo Parkinson disease of the DeNoPa cohort.* Neurology, 2013. **81**(14): p. 1226-1234.

136. Karolinska Institutet. *The Swedish Twin Registry*. 2021; Available from: https://ki.se/en/research/the-swedish-twin-registry.

137. Osborne, J., *Improving your data transformations: Applying the Box-Cox transformation.* Practical Assessment, Research, and Evaluation, 2010. **15**(1): p. 12.

138. Millioni, R., et al., *High Abundance Proteins Depletion vs Low Abundance Proteins Enrichment: Comparison of Methods to Reduce the Plasma Proteome Complexity.* Plos One, 2011. **6**(5).

139. Lee, P.Y., et al., *Plasma/serum proteomics: depletion strategies for reducing high-abundance proteins for biomarker discovery.* Bioanalysis, 2019. **11**(19): p. 1799-1812.

140. Kovacs, A. and A. Guttman, *Medicinal Chemistry Meets Proteomics: Fractionation of the Human Plasma Proteome.* Current Medicinal Chemistry, 2013. **20**(4): p. 483-490.

141. Catherman, A.D., O.S. Skinner, and N.L. Kelleher, *Top Down proteomics: facts and perspectives.* Biochem Biophys Res Commun, 2014. **445**(4): p. 683-93.

142. Yu, Y.B., et al., *Urine Sample Preparation in 96-Well Filter Plates for Quantitative Clinical Proteomics.* Analytical Chemistry, 2014. **86**(11): p. 5470-5477.

143. Barratt, J. and P. Topham, *Urine proteomics: the present and future of measuring urinary protein components in disease.* Canadian Medical Association Journal, 2007. **177**(4): p. 361-368.

144. MacLean, B., et al., *Skyline: an open source document editor for creating and analyzing targeted proteomics experiments.* Bioinformatics, 2010. **26**(7): p. 966-8.

145. UniProt. *BLAST*. 2019-11-03]; Available from: https://www.uniprot.org/blast/.

146. Kessner, D., et al., *ProteoWizard: open source software for rapid proteomics tools development.* Bioinformatics, 2008. **24**(21): p. 2534-6.

147. Seabold, S. and J. Perktold. *Statsmodels: Econometric and statistical modeling with python.* in *Proceedings of the 9th Python in Science Conference*. 2010. Austin, TX.

148. Cleveland, W.S., *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association, 1979. **74**(368): p. 829-836.

149. Wada, K., *Outliers in official statistics.* Japanese Journal of Statistics and Data Science, 2020. **3**(2): p. 669-691.

150. Leys, C., et al., *Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median.* Journal of Experimental Social Psychology, 2013. **49**(4): p. 764-766.

151. Choudhary, S. and A. Kothari, *Green data center using spearmans ranking algorithm.* International Journal of Computer Science and Information Technologies, 2015. **6**(2): p. 1672-1676.

152. Pham-Gia, T. and T.L. Hung, *The mean and median absolute deviations.* Mathematical and Computer Modelling, 2001. **34**(7-8): p. 921-936.

153. Hagg, S. and J. Jylhava, *Sex differences in biological aging with a focus on human studies.* Elife, 2021. **10**.

154. Zhu, W., N. Zeng, and N. Wang, *Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations.* NESUG proceedings: health care and life sciences, Baltimore, Maryland, 2010. **19**: p. 67.

155. Van Stralen, K.J., et al., *Diagnostic methods I: sensitivity, specificity, and other measures of accuracy.* Kidney international, 2009. **75**(12): p. 1257-1263.

156. Poljsak, B., et al., *Nature Versus Nurture: What Can be Learned from the Oldest-Old's Claims About Longevity?* Rejuvenation Res, 2021.

157. *IPA QIAGEN Inc.,*. IPA®; Available from: https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis.

158. Gabay, C. and I. Kushner, *Acute-phase proteins and other systemic responses to inflammation.* N Engl J Med, 1999. **340**(6): p. 448-54.

159. Zelcer, N. and P. Tontonoz, *Liver X receptors as integrators of metabolic and inflammatory signaling.* Journal of Clinical Investigation, 2006. **116**(3): p. 607-614.

160. Kalaany, N.Y. and D.J. Mangelsdorf, *LXRs AND FXR: The Yin and Yang of cholesterol and fat metabolism.* Annual Review of Physiology, 2006. **68**: p. 159-191.

161. Mangelsdorf, D.J. and R.M. Evans, *The Rxr Heterodimers and Orphan Receptors.* Cell, 1995. **83**(6): p. 841-850.

162. Shi, Q., et al., *Complement C3 deficiency protects against neurodegeneration in aged plaque-rich APP/PS1 mice.* Sci Transl Med, 2017. **9**(392).

163. Furie, B. and B.C. Furie, *Molecular and cellular biology of blood coagulation.* N Engl J Med, 1992. **326**(12): p. 800-6.

164. Mackman, N., R.E. Tilley, and N.S. Key, *Role of the extrinsic pathway of blood coagulation in hemostasis and thrombosis.* Arterioscler Thromb Vasc Biol, 2007. **27**(8): p. 1687-93.

165. Gorbet, M.B. and M.V. Sefton, *Biomaterial-associated thrombosis: roles of coagulation factors, complement, platelets and leukocytes.* Biomaterials, 2004. **25**(26): p. 5681-703.

166. Johnson, A.A., et al., *Systematic review and analysis of human proteomics aging studies unveils a novel proteomic aging clock and identifies key processes that change with age.* Ageing Research Reviews, 2020. **60**.

167. Lehallier, B., et al., *Undulating changes in human plasma proteome profiles across the lifespan.* Nat Med, 2019. **25**(12): p. 1843-1850.

168. Birkenmeier, G., et al., *Human alpha2-macroglobulin: genotype-phenotype relation.* Exp Neurol, 2003. **184**(1): p. 153-61.

169. Huffman, D.M., et al., *Distinguishing between longevity and buffered-deleterious genotypes for exceptional human longevity: the case of the MTP gene.* J Gerontol A Biol Sci Med Sci, 2012. **67**(11): p. 1153-60.

170. Atzmon, G., et al., *Adiponectin levels and genotype: a potential regulator of life span in humans.* J Gerontol A Biol Sci Med Sci, 2008. **63**(5): p. 447-53.

171. Brimijoin, S., et al., *Physiological roles for butyrylcholinesterase: A BChE-ghrelin axis.* Chem Biol Interact, 2016. **259**(Pt B): p. 271-275.

172. Ho, J.E., et al., *Protein Biomarkers of Cardiovascular Disease and Mortality in the Community.* J Am Heart Assoc, 2018. **7**(14).

173. Cribbs, D.H., et al., *Extensive innate immune gene activation accompanies brain aging, increasing vulnerability to cognitive decline and neurodegeneration: a microarray study.* J Neuroinflammation, 2012. **9**: p. 179.

174. Blumenau, S., et al., *Investigating APOE, APP-Abeta metabolism genes and Alzheimer's disease GWAS hits in brain small vessel ischemic disease.* Sci Rep, 2020. **10**(1): p. 7103.

175. Orwoll, E.S., et al., *Proteomic assessment of serum biomarkers of longevity in older men.* Aging Cell, 2020: p. e13253.

176. Shavlakadze, T., et al., *Age-Related Gene Expression Signature in Rats Demonstrate Early, Late, and Linear Transcriptional Changes from Multiple Tissues.* Cell Reports, 2019. **28**(12): p. 3263–+.

177. Gorgoulis, V.G., et al., *p53-dependent ICAM-1 overexpression in senescent human cells identified in atherosclerotic lesions.* Lab Invest, 2005. **85**(4): p. 502-11.

178. Wang, H., et al., *Quantitative iTRAQ-based proteomic analysis of differentially expressed proteins in aging in human and monkey.* BMC Genomics, 2019. **20**(1): p. 725.

179. Triplett, J.C., et al., *Metabolic clues to salubrious longevity in the brain of the longest-lived rodent: the naked mole-rat.* J Neurochem, 2015. **134**(3): p. 538-50.

180. Kostrominova, T.Y. and S.V. Brooks, *Age-related changes in structure and extracellular matrix protein expression levels in rat tendons.* Age (Dordr), 2013. **35**(6): p. 2203-14.

181. Chen, P.J., et al., *Age-related changes in the cartilage of the temporomandibular joint.* Geroscience, 2020. **42**(3): p. 995-1004.

182. Vanni, S., et al., *Differential overexpression of SERPINA3 in human prion diseases.* Sci Rep, 2017. **7**(1): p. 15637.

183. Portis, S.M., et al., *Effects of nutraceutical intervention on serum proteins in aged rats.* Geroscience, 2020. **42**(2): p. 703-713.

184. Yousef, H., et al., *Aged blood impairs hippocampal neural precursor activity and activates microglia via brain endothelial cell VCAM1.* Nat Med, 2019. **25**(6): p. 988-1000.

185. Abondio, P., et al., *The Genetic Variability of APOE in Different Human Populations and Its Implications for Longevity.* Genes (Basel), 2019. **10**(3).

186. Lu, J., et al., *Profiling plasma peptides for the identification of potential ageing biomarkers in Chinese Han adults.* PLoS One, 2012. **7**(7): p. e39726.

187. Franceschi, C., et al., *Inflamm-aging - An evolutionary perspective on immunosenescence.* Molecular and Cellular Gerontology, 2000. **908**: p. 244-254.

188. Jumeau, C., et al., *Expression of SAA1, SAA2 and SAA4 genes in human primary monocytes and monocyte-derived macrophages.* PLoS One, 2019. **14**(5): p. e0217005.

189. Yu, M.H., et al., *SAA1 increases NOX4/ROS production to promote LPS-induced inflammation in vascular smooth muscle cells through activating p38MAPK/NF-kappaB pathway.* BMC Mol Cell Biol, 2019. **20**(1): p. 15.

190. Abouelasrar Salama, S., et al., *Serum Amyloid A1 (SAA1) Revisited: Restricted Leukocyte-Activating Properties of Homogeneous SAA1.* Front Immunol, 2020. **11**: p. 843.

191. Mercurio, D., et al., *Targeted deletions of complement lectin pathway genes improve outcome in traumatic brain injury, with MASP-2 playing a major role.* Acta Neuropathol Commun, 2020. **8**(1): p. 174.

192. Videm, V. and M. Albrigtsen, *Soluble ICAM-1 and VCAM-1 as markers of endothelial activation.* Scandinavian Journal of Immunology, 2008. **67**(5): p. 523-531.

193. GeneCards. *SERPINF2 Gene*. [cited 2021 2021-02-21]; Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=SERPINF2.

194. GeneCards, *BCHE Gene.*

195. Das, U.N., *Acetylcholinesterase and butyrylcholinesterase as markers of low-grade systemic inflammation.* Annals of Hepatology, 2012. **11**(3): p. 409-411.

196. Chen, V.P., et al., *Butyrylcholinesterase gene transfer in obese mice prevents postdieting body weight rebound by suppressing ghrelin signaling.* Proc Natl Acad Sci U S A, 2017. **114**(41): p. 10960-10965.

197. Monti, D., et al., *Inflammaging and human longevity in the omics era.* Mech Ageing Dev, 2017. **165**(Pt B): p. 129-138.

198. GeneCards. *A2M Gene.* Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=A2M&keywords=a2m.

199. Kurz, S., et al., *The anti-tumorigenic activity of A2M-A lesson from the naked mole-rat.* PLoS One, 2017. **12**(12): p. e0189514.

200. Zhu, M., et al., *alpha-2-Macroglobulin, a Native and Powerful Proteinase Inhibitor, Prevents Cartilage Degeneration Disease by Inhibiting Majority of Catabolic Enzymes and Cytokines.* Current Molecular Biology Reports, 2021. **7**(1): p. 1-7.

201. Cuellar, J.M., V.G. Cuellar, and G.J. Scuderi, *alpha2-Macroglobulin: Autologous Protease Inhibition Technology.* Phys Med Rehabil Clin N Am, 2016. **27**(4): p. 909-918.

202. Lagunas-Rangel, F.A. and V. Chavez-Valencia, *Learning of nature: The curious case of the naked mole rat.* Mech Ageing Dev, 2017. **164**: p. 76-81.

203. Straub, L.G. and P.E. Scherer, *Metabolic Messengers: Adiponectin.* Nat Metab, 2019. **1**(3): p. 334-339.

204. Achari, A.E. and S.K. Jain, *Adiponectin, a Therapeutic Target for Obesity, Diabetes, and Endothelial Dysfunction.* Int J Mol Sci, 2017. **18**(6).

205. Nguyen, N.H., G.B. Tran, and C.T. Nguyen, *Anti-oxidative effects of superoxide dismutase 3 on inflammatory diseases.* J Mol Med (Berl), 2020. **98**(1): p. 59-69.

206. Laurila, J.P., et al., *SOD3 reduces inflammatory cell migration by regulating adhesion molecule and cytokine expression.* PLoS One, 2009. **4**(6): p. e5786.

207. Ricklin, D., E.S. Reis, and J.D. Lambris, *Complement in disease: a defence system turning offensive.* Nat Rev Nephrol, 2016. **12**(7): p. 383-401.

208. Schartz, N.D. and A.J. Tenner, *The good, the bad, and the opportunities of the complement system in neurodegenerative disease.* Journal of Neuroinflammation, 2020. **17**(1).

209. Fu, S.H., et al., *Centenarian longevity is positively correlated with IgE levels but negatively correlated with C3/C4 levels, abdominal obesity and metabolic syndrome.* Cellular & Molecular Immunology, 2020. **17**(11): p. 1196-1197.

210. Xiong, Y., et al., *Regulation of glycolysis and gluconeogenesis by acetylation of PKM and PEPCK.* Cold Spring Harb Symp Quant Biol, 2011. **76**: p. 285-9.

211. Lewandowski, S.L., et al., *Pyruvate Kinase Controls Signal Strength in the Insulin Secretory Pathway.* Cell Metabolism, 2020. **32**(5): p. 736-+.

212. Haspula, D., et al., *Influence of a Hyperglycemic Microenvironment on a Diabetic Versus Healthy Rat Vascular Endothelium Reveals Distinguishable Mechanistic and Phenotypic Responses.* Frontiers in Physiology, 2019. **10**.

213. Salto, R., et al., *Dietary Complex and Slow Digestive Carbohydrates Prevent Fat Deposits During Catch-Up Growth in Rats.* Nutrients, 2020. **12**(9).

214. Shlipak, M.G., et al., *Cystatin C and mortality risk in the elderly: the health, aging, and body composition study.* J Am Soc Nephrol, 2006. **17**(1): p. 254-61.

215. Zou, J., et al., *Cystatin C as a potential therapeutic mediator against Parkinson's disease via VEGF-induced angiogenesis and enhanced neuronal autophagy in neurovascular units.* Cell Death & Disease, 2017. **8**(6): p. e2854-e2854.

216. White, C.A., S. Ghazan-Shahi, and M.A. Adams, *β-Trace Protein: A Marker of GFR and Other Biological Pathways.* American Journal of Kidney Diseases, 2015. **65**(1): p. 131-146.

217. Choi, D.J., et al., *A Parkinson's disease gene, DJ-1, regulates anti-inflammatory roles of astrocytes through prostaglandin D2 synthase expression.* Neurobiol Dis, 2019. **127**: p. 482-491.

218. Das, N., et al., *Proteoglycan 4: From Mere Lubricant to Regulator of Tissue Homeostasis and Inflammation: Does proteoglycan 4 have the ability to buffer the inflammatory response?* Bioessays, 2019. **41**(1): p. e1800166.

219. Human Protein Atlas. *Cell Atlas - FABP5*. 2019-11-01]; Available from: https://www.proteinatlas.org/ENSG00000164687-FABP5/cell.

220. Gene Cards. *FABP5 Gene - GeneCards*. [cited 2019 1 November]; Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=FABP5.

221. Hoekstra, M., et al., *Microarray analysis indicates an important role for FABP5 and putative novel FABPs on a Western-type diet.* J Lipid Res, 2006. **47**(10): p. 2198-207.

222. Xie, X., et al., *Proteomics analyses of subcutaneous adipocytes reveal novel abnormalities in human insulin resistance.* Obesity (Silver Spring), 2016. **24**(7): p. 1506-14.

223. Montecinos, L., J.D. Eskew, and A. Smith, *What Is Next in This "Age" of Heme-Driven Pathology and Protection by Hemopexin? An Update and Links with Iron.* Pharmaceuticals (Basel), 2019. **12**(4).

224. Garatachea, N., et al., *The ApoE Gene Is Related with Exceptional Longevity: A Systematic Review and Meta-Analysis.* Rejuvenation Research, 2015. **18**(1): p. 3-13.

225. Kannel, W.B., et al., *Fibrinogen and risk of cardiovascular disease. The Framingham Study.* JAMA, 1987. **258**(9): p. 1183-6.

226. Lowe, G.D., A. Rumley, and I.J. Mackie, *Plasma fibrinogen.* Ann Clin Biochem, 2004. **41**(Pt 6): p. 430-40.

227. Jian, J.L., J. Konopka, and C.J. Liu, *Insights into the role of progranulin in immunity, infection, and inflammation.* Journal of Leukocyte Biology, 2013. **93**(2): p. 199-208.

228. Wang, J., et al., *HSPA5 Gene encoding Hsp70 chaperone BiP in the endoplasmic reticulum.* Gene, 2017. **618**: p. 14-23.

229. GeneCards. *PGK1 Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=PGK1&keywords=pgk1.

230. Eisenstein, M., *Centenarians: Great expectations.* Nature, 2012. **492**(7427): p. S6-8.

231. Stebbins, G.T., et al., *How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: comparison with the unified Parkinson's disease rating scale.* Mov Disord, 2013. **28**(5): p. 668-70.

232. Sauerbier, A., et al., *Non motor subtypes and Parkinson's disease.* Parkinsonism Relat Disord, 2016. **22 Suppl 1**: p. S41-6.

233. Rizzo, G., et al., *Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis.* Neurology, 2016. **86**(6): p. 566-76.

234. Parnetti, L., et al., *CSF and blood biomarkers for Parkinson's disease.* Lancet Neurol, 2019. **18**(6): p. 573-586.

235. He, R., et al., *Recent Advances in Biomarkers for Parkinson's Disease.* Front Aging Neurosci, 2018. **10**: p. 305.

236. Shtilbans, A. and C. Henchcliffe, *Biomarkers in Parkinson's disease: an update.* Curr Opin Neurol, 2012. **25**(4): p. 460-5.

237. Romeo, M.J., et al., *CSF proteome: a protein repository for potential biomarker identification.* Expert Rev Proteomics, 2005. **2**(1): p. 57-70.

238. Maurer, M.H., *Proteomics of brain extracellular fluid (ECF) and cerebrospinal fluid (CSF).* Mass Spectrom Rev, 2010. **29**(1): p. 17-28.

239. Dayon, L., et al., *Proteomes of Paired Human Cerebrospinal Fluid and Plasma: Relation to Blood-Brain Barrier Permeability in Older Adults.* J Proteome Res, 2019. **18**(3): p. 1162-1174.

240. Whelan, C.D., et al., *Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer's disease.* Acta Neuropathol Commun, 2019. **7**(1): p. 169.

241. Wu, S., et al., *Il34-Csf1r Pathway Regulates the Migration and Colonization of Microglial Precursors.* Dev Cell, 2018. **46**(5): p. 552-563 e4.

242. Xicola, R.M., et al., *Clinical features and cancer risk in families with pathogenic CDH1 variants irrespective of clinical criteria.* J Med Genet, 2019. **56**(12): p. 838-843.

243. Tian, X., et al., *E-cadherin/beta-catenin complex and the epithelial barrier.* J Biomed Biotechnol, 2011. **2011**: p. 567305.

244. Suhaimi, S.A., S.C. Chan, and R. Rosli, *Matrix Metallopeptidase 3 Polymorphisms: Emerging genetic Markers in Human Breast Cancer Metastasis.* J Breast Cancer, 2020. **23**(1): p. 1-9.

245. Kim, E.M. and O. Hwang, *Role of matrix metalloproteinase-3 in neurodegeneration.* J Neurochem, 2011. **116**(1): p. 22-32.

246. Choi, D.H., et al., *Matrix Metalloproteinase-3 Causes Dopaminergic Neuronal Death through Nox1-Regenerated Oxidative Stress.* Plos One, 2014. **9**(12).

247. Weekman, E.M. and D.M. Wilcock, *Matrix Metalloproteinase in Blood-Brain Barrier Breakdown in Dementia.* Journal of Alzheimers Disease, 2016. **49**(4): p. 893-903.

248. Busceti, C.L., et al., *Dickkopf-3 Causes Neuroprotection by Inducing Vascular Endothelial Growth Factor.* Front Cell Neurosci, 2018. **12**: p. 292.

249. Zenzmaier, C., L. Sklepos, and P. Berger, *Increase of Dkk-3 blood plasma levels in the elderly.* Exp Gerontol, 2008. **43**(9): p. 867-70.

250. Bras, J., R. Guerreiro, and J. Hardy, *SnapShot: Genetics of Parkinson's disease.* Cell, 2015. **160**(3): p. 570-570 e1.

251. Yang, X., et al., *Depletion of microglia augments the dopaminergic neurotoxicity of MPTP.* FASEB J, 2018. **32**(6): p. 3336-3345.

252. Ruseva, M.M., et al., *An anticomplement agent that homes to the damaged brain and promotes recovery after traumatic brain injury in mice.* Proc Natl Acad Sci U S A, 2015. **112**(46): p. 14319-24.

253. Fluiter, K., et al., *Inhibition of the membrane attack complex of the complement system reduces secondary neuroaxonal loss and promotes neurologic recovery after traumatic brain injury in mice.* J Immunol, 2014. **192**(5): p. 2339-48.

254. Orsini, F., et al., *Versatility of the complement system in neuroinflammation, neurodegeneration and brain homeostasis.* Front Cell Neurosci, 2014. **8**: p. 380.

255. Alawieh, A., et al., *Identifying the Role of Complement in Triggering Neuroinflammation after Traumatic Brain Injury.* J Neurosci, 2018. **38**(10): p. 2519-2532.

256. Baker, M., et al., *Mutations in progranulin cause tau-negative frontotemporal dementia linked to chromosome 17.* Nature, 2006. **442**(7105): p. 916-919.

257. Kao, A.W., et al., *Progranulin, lysosomal regulation and neurodegenerative disease.* Nature Reviews Neuroscience, 2017. **18**(6): p. 325-333.

258. Mateo, I., et al., *Reduced serum progranulin level might be associated with Parkinson's disease risk.* European Journal of Neurology, 2013. **20**(12): p. 1571-1573.

259. Zhang, J.Z., et al., *A WNT1-regulated developmental gene cascade prevents dopaminergic neurodegeneration in adult En1(+/-) mice.* Neurobiology of Disease, 2015. **82**: p. 32-45.

260. Fukusumi, Y., et al., *Dickkopf 3 Promotes the Differentiation of a Rostrolateral Midbrain Dopaminergic Neuronal Subset In Vivo and from Pluripotent Stem Cells In Vitro in the Mouse.* Journal of Neuroscience, 2015. **35**(39): p. 13385-13401.

261. Rusnak, F. and P. Mertz, *Calcineurin: Form and function.* Physiological Reviews, 2000. **80**(4): p. 1483-1521.

262. De, A., *Wnt/Ca2+ signaling pathway: a brief overview.* Acta Biochimica Et Biophysica Sinica, 2011. **43**(10): p. 745-756.

263. Lam, S., et al., *A systems biology approach for studying neurodegenerative diseases.* Drug Discovery Today, 2020. **25**(7): p. 1146-1159.

264. Caraveo, G., et al., *Calcineurin determines toxic versus beneficial responses to alpha-synuclein.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(34): p. E3544-E3552.

265. Sanchez-Navarro, A., et al., *An integrative view of serpins in health and disease: the contribution of SerpinA3.* American Journal of Physiology-Cell Physiology, 2020. **320**(2): p. C106-C118.

266. Vanni, S., et al., *Differential overexpression of SERPINA3 in human prion diseases.* Scientific Reports, 2017. **7**.

267. Fissolo, N., et al., *CSF SERPINA3 Levels Are Elevated in Patients With Progressive MS.* Neurol Neuroimmunol Neuroinflamm, 2021. **8**(2).

268. Yu, H., et al., *Platelet biomarkers for a descending cognitive function: A proteomic approach.* Aging Cell, 2021: p. e13358.

269. Kim, K.S., et al., *Proteolytic Cleavage of Extracellular alpha-Synuclein by Plasmin IMPLICATIONS FOR PARKINSON DISEASE.* Journal of Biological Chemistry, 2012. **287**(30): p. 24862-24872.

270. Seo, M.H. and S. Yeo, *Association of increase in Serping1 level with dopaminergic cell reduction in an MPTP-induced Parkinson's disease mouse model.* Brain Research Bulletin, 2020. **162**: p. 67-72.

271. Gene Cards. *HPX Gene - GeneCards.* Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=HPX.

272. Hahl, P., et al., *Identification of oxidative modifications of hemopexin and their predicted physiological relevance.* J Biol Chem, 2017. **292**(33): p. 13658-13671.

273. Chen-Roetling, J., et al., *Hemopexin increases the neurotoxicity of hemoglobin when haptoglobin is absent.* Journal of Neurochemistry, 2018. **145**(6): p. 464-473.

274. Kurvits, L., et al., *Transcriptomic profiles in Parkinson's disease.* Exp Biol Med (Maywood), 2021. **246**(5): p. 584-595.

275. Silva, M., P.A. Videira, and R. Sackstein, *E-Selectin Ligands in the Human Mononuclear Phagocyte System: Implications for Infection, Inflammation, and Immunotherapy.* Frontiers in Immunology, 2018. **8**.

276. Silva, M., P.A. Videira, and R. Sackstein, *E-Selectin Ligands in the Human Mononuclear Phagocyte System: Implications for Infection, Inflammation, and Immunotherapy.* Front Immunol, 2017. **8**: p. 1878.

277. Gene Cards. *HSPA5 Gene - GeneCards.* [cited 2019 2019-11-01]; Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=HSPA5.

278. Hetz, C. and S. Saxena, *ER stress and the unfolded protein response in neurodegeneration.* Nature Reviews Neurology, 2017. **13**(8): p. 477-491.

279. Gene Cards, *HSPA6 Gene - GeneCards.*

280. Gene Cards, *HSPA1L Gene - GeneCards.*

281. Mayer, M.P. and B. Bukau, *Hsp70 chaperones: Cellular functions and molecular mechanism.* Cellular and Molecular Life Sciences, 2005. **62**(6): p. 670-684.

282. MalaCards. *Gerstmann-Straussler Disease (GSD).* Available from: https://www.malacards.org/card/gerstmann_straussler_disease.

283. Turner, M.E., et al., *Secreted Phosphoprotein 24 is a Biomarker of Mineral Metabolism.* Calcified Tissue International, 2021. **108**(3): p. 354-363.

284. Cai, R., et al., *Enhancing glycolysis attenuates Parkinson's disease progression in models and clinical databases.* Journal of Clinical Investigation, 2019. **129**(10): p. 4539-4549.

285. Janke, C. and M.M. Magiera, *The tubulin code and its role in controlling microtubule properties and functions.* Nature Reviews Molecular Cell Biology, 2020. **21**(6): p. 307-326.

286. Okada, K., et al., *An autopsy case of pure nigropathy with TUBA4A nonsense mutation.* Neuropathol Appl Neurobiol, 2021.

287. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.

288. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

289. Szklarczyk, D., et al., *STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.* Nucleic Acids Res, 2019. **47**(D1): p. D607-D613.

290. Goksuluk, D., et al., *easyROC: An Interactive Web-tool for ROC Curve Analysis Using R Language Environment.* R Journal, 2016. **8**(2): p. 213-230.

291. Fawcett, T., *An introduction to ROC analysis.* Pattern Recognition Letters, 2006. **27**(8): p. 861-874.

292. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.

293. Arenas, E., *Wnt signaling in midbrain dopaminergic neuron development and regenerative medicine for Parkinson's disease.* J Mol Cell Biol, 2014. **6**(1): p. 42-53.

294. Marchetti, B., *Wnt/beta-Catenin Signaling Pathway Governs a Full Program for Dopaminergic Neuron Survival, Neurorescue and Regeneration in the MPTP Mouse Model of Parkinson's Disease.* Int J Mol Sci, 2018. **19**(12).

295. Hajishengallis, G., et al., *Novel mechanisms and functions of complement.* Nat Immunol, 2017. **18**(12): p. 1288-1298.

296. Carpanini, S.M., M. Torvell, and B.P. Morgan, *Therapeutic Inhibition of the Complement System in Diseases of the Central Nervous System.* Frontiers in Immunology, 2019. **10**.

297. Loeffler, D.A., D.M. Camp, and S.B. Conant, *Complement activation in the Parkinson's disease substantia nigra: an immunocytochemical study.* Journal of Neuroinflammation, 2006. **3**.

298. Gorlach, A., P. Klappa, and T. Kietzmann, *The endoplasmic reticulum: Folding, calcium homeostasis, signaling, and redox control.* Antioxidants & Redox Signaling, 2006. **8**(9-10): p. 1391-1418.

299. Smith, M.H., H.L. Ploegh, and J.S. Weissman, *Road to Ruin: Targeting Proteins for Degradation in the Endoplasmic Reticulum.* Science, 2011. **334**(6059): p. 1086-1090.

300. Colla, E., *Linking the Endoplasmic Reticulum to Parkinson's Disease and Alpha-Synucleinopathy.* Frontiers in Neuroscience, 2019. **13**.

301. Walter, P. and D. Ron, *The Unfolded Protein Response: From Stress Pathway to Homeostatic Regulation.* Science, 2011. **334**(6059): p. 1081-1086.

302. Martinez, A., et al., *Targeting of the unfolded protein response (UPR) as therapy for Parkinson's disease.* Biol Cell, 2019. **111**(6): p. 161-168.

303. Bertolotti, A., et al., *Dynamic interaction of BiP and ER stress transducers in the unfolded-protein response.* Nature Cell Biology, 2000. **2**(6): p. 326-332.

304. da Costa, C.A., et al., *The Endoplasmic Reticulum Stress/Unfolded Protein Response and Their Contributions to Parkinson's Disease Physiopathology.* Cells, 2020. **9**(11).

305. Choi, M.L. and S. Gandhi, *Crucial role of protein oligomerization in the pathogenesis of Alzheimer's and Parkinson's diseases.* Febs Journal, 2018. **285**(19): p. 3631-3644.

306. Mercado, G., et al., *ER stress and Parkinson's disease: Pathological inputs that converge into the secretory pathway.* Brain Research, 2016. **1648**: p. 626-632.

307. Komiya, Y. and R. Habas, *Wnt signal transduction pathways.* Organogenesis, 2008. **4**(2): p. 68-75.

308. Barker, N., *The Canonical Wnt/β-Catenin Signalling Pathway*, in *Wnt Signaling: Pathway Methods and Mammalian Models*, E. Vincan, Editor. 2008, Humana Press: Totowa, NJ. p. 5-15.

309. van Amerongen, R. and A. Berns, *Knockout mouse models to study Wnt signal transduction.* Trends in Genetics, 2006. **22**(12): p. 678-689.

310. Huelsken, J. and J. Behrens, *The Wnt signalling pathway.* Journal of Cell Science, 2002. **115**(21): p. 3977-3978.

311. Marchetti, B., *Wnt/beta-Catenin Signaling Pathway Governs a Full Program for Dopaminergic Neuron Survival, Neurorescue and Regeneration in the MPTP Mouse Model of Parkinson's Disease.* International Journal of Molecular Sciences, 2018. **19**(12).

312. Marchetti, B., et al., *Parkinson's disease, aging and adult neurogenesis: Wnt/beta-catenin signalling as the key to unlock the mystery of endogenous brain repair.* Aging Cell, 2020. **19**(3).

313. L'Episcopo, F., et al., *Wnt/beta-Catenin Signaling Is Required to Rescue Midbrain Dopaminergic Progenitors and Promote Neurorepair in Ageing Mouse Model of Parkinson's Disease.* Stem Cells, 2014. **32**(8): p. 2147-2163.

314. L'Episcopo, F., et al., *Neural Stem Cell Grafts Promote Astroglia-Driven Neurorestoration in the Aged Parkinsonian Brain via Wnt/beta-Catenin Signaling.* Stem Cells, 2018. **36**(8): p. 1179-1197.

315. An, M. and Y. Gao, *Urinary Biomarkers of Brain Diseases.* Genomics Proteomics Bioinformatics, 2015. **13**(6): p. 345-54.

316. Lerma, E.V., *Approach to the patient with renal disease.* Prim Care, 2008. **35**(2): p. 183-94, v.

317. Thongboonkerd, V. and P. Malasit, *Renal and urinary proteomics: current applications and challenges.* Proteomics, 2005. **5**(4): p. 1033-42.

318. Beasley-Green, A., *Urine Proteomics in the Era of Mass Spectrometry.* Int Neurourol J, 2016. **20**(Suppl 2): p. S70-75.

319. Nguyen, A.P.T., et al., *Dopaminergic neurodegeneration induced by Parkinson's disease-linked G2019S LRRK2 is dependent on kinase and GTPase activity.* Proc Natl Acad Sci U S A, 2020. **117**(29): p. 17296-17307.

320. Foote, M. and Y. Zhou, *14-3-3 proteins in neurological disorders.* Int J Biochem Mol Biol, 2012. **3**(2): p. 152-64.

321. Stoeckli, E.T., *Understanding axon guidance: are we nearly there yet?* Development, 2018. **145**(10).

322. Lee, S.H., et al., *Sirtuin signaling in cellular senescence and aging.* BMB Rep, 2019. **52**(1): p. 24-34.

323. Jiang, J., et al., *Activation of neuronal nitric oxide synthase (nNOS) signaling pathway in 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD)-induced neurotoxicity.* Environ Toxicol Pharmacol, 2014. **38**(1): p. 119-30.

324. Wang, J. and H. Knaut, *Chemokine signaling in development and disease.* Development, 2014. **141**(22): p. 4199-205.

325. Suffredini, A.F., et al., *New insights into the biology of the acute phase response.* J Clin Immunol, 1999. **19**(4): p. 203-14.

326. Bamberger, M.E. and G.E. Landreth, *Inflammation, apoptosis, and Alzheimer's disease.* Neuroscientist, 2002. **8**(3): p. 276-83.

327. Hong, C. and P. Tontonoz, *Coordination of inflammation and metabolism by PPAR and LXR nuclear receptors.* Curr Opin Genet Dev, 2008. **18**(5): p. 461-7.

328. Mendes, K.L., D.F. Lelis, and S.H.S. Santos, *Nuclear sirtuins and inflammatory signaling pathways.* Cytokine Growth Factor Rev, 2017. **38**: p. 98-105.

329. Dieterle, F., et al., *Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics.* Anal Chem, 2006. **78**(13): p. 4281-90.

330. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

331. GeneCards. *ICAM1 Gene.* Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=ICAM1&keywords=icam.

332. Li, W., et al., *Interaction between ICAM1 in endothelial cells and LFA1 in T cells during the pathogenesis of experimental Parkinson's disease.* Exp Ther Med, 2020. **20**(2): p. 1021-1029.

333. McGeer, P.L. and E.G. McGeer, *Glial reactions in Parkinson's disease.* Mov Disord, 2008. **23**(4): p. 474-83.

334. GeneCards. *CCL4 Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=CCL4&keywords=ccl4.

335. Zhu, M., et al., *Age-related brain expression and regulation of the chemokine CCL4/MIP-1beta in APP/PS1 double-transgenic mice.* J Neuropathol Exp Neurol, 2014. **73**(4): p. 362-74.

336. GeneCards. *TNNT3 Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=TNNT3&keywords=tnnt3#summaries.

337. Cards, G. *SPP2 Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=SPP2&keywords=spp2.

338. Ochieng, J. and G. Chaudhuri, *Cystatin superfamily.* J Health Care Poor Underserved, 2010. **21**(1 Suppl): p. 51-70.

339. GeneCards. *COL6A3 Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=COL6A3&keywords=COL6A3.

340. Demir, E., et al., *Mutations in COL6A3 cause severe and mild phenotypes of Ullrich congenital muscular dystrophy.* Am J Hum Genet, 2002. **70**(6): p. 1446-58.

341. Jin, C.Y., et al., *Study of the collagen type VI alpha 3 (COL6A3) gene in Parkinson's disease.* BMC Neurol, 2021. **21**(1): p. 187.

342. GeneCards. *NEFM Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=NEFM&keywords=nefm#summaries.

343. Liu, Q., et al., *Neurofilament proteins in neurodegenerative diseases.* Cellular and Molecular Life Sciences, 2004. **61**(24): p. 3057-3075.

344. Khalil, M., et al., *Neurofilaments as biomarkers in neurological disorders.* Nature Reviews Neurology, 2018. **14**(10): p. 577-589.

345. GeneCards. *UBC Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=UBC&keywords=ubc.

346. Chung, K.K.K., V.L. Dawson, and T.M. Dawson, *The role of the ubiquitin-proteasomal pathway in Parkinson's disease and other neurodegenerative disorders.* Trends in Neurosciences, 2001. **24**(11): p. S7-S14.

347. GeneCards. *PGAM1 Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=PGAM1&keywords=pgam1.

348. Jung, H.Y., et al., *Phosphoglycerate Mutase 1 Prevents Neuronal Death from Ischemic Damage by Reducing Neuroinflammation in the Rabbit Spinal Cord.* International Journal of Molecular Sciences, 2020. **21**(19).

349. GeneCards. *CAPN2 Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=CAPN2&keywords=capn2.

350. Mahaman, Y.A.R., et al., *Involvement of calpain in the neuropathogenesis of Alzheimer's disease.* Medicinal Research Reviews, 2019. **39**(2): p. 608-630.

351. Chera, H., et al., *Immunofluorescent labeling of increased calpain expression and neuronal death in the spinal cord of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-treated mice.* Brain Research, 2004. **1006**(2): p. 150-156.

352. GeneCards. *FGF21 Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=FGF21&keywords=fgf21.

353. Kang, K., et al., *FGF21 attenuates neurodegeneration through modulating neuroinflammation and oxidant-stress.* Biomedicine & Pharmacotherapy, 2020. **129**.

354. GeneCards. *NCAM1 Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=NCAM1&keywords=ncam1.

355. GeneCards. *NDRG1 Gene*. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=NDRG1&keywords=NDRG1.

356. Berger, P., A. Niemann, and U. Suter, *Schwann cells and the pathogenesis of inherited motor and sensory neuropathies (Charcot-Marie-Tooth disease).* Glia, 2006. **54**(4): p. 243-257.

357. Wang, H., et al., *Identification of Regulatory Relationships in Parkinson's Disease.* Journal of Molecular Neuroscience, 2013. **51**(1): p. 9-12.

358. GeneCards. *MMP3 Gene.* Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=MMP3&keywords=mmp3.

359. GeneCards. *MAPK12 Gene.* Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=MAPK12&keywords=mapk12.

360. Bohush, A., G. Niewiadomska, and A. Filipek, *Role of Mitogen Activated Protein Kinase Signaling in Parkinson's Disease.* International Journal of Molecular Sciences, 2018. **19**(10).

361. GeneCards. *CCL2 Gene.* Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=CCL2&keywords=ccl2.

362. Kempuraj, D., et al., *Neuroinflammation Induces Neurodegeneration.* J Neurol Neurosurg Spine, 2016. **1**(1).

363. Shen, R.N., et al., *Association of Two Polymorphisms in CCL2 With Parkinson's Disease: A Case-Control Study.* Frontiers in Neurology, 2019. **10**.

364. GeneCards. *HSP90AB1 Gene.* Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=HSP90AB1&keywords=HSP90AB1.

365. Xie, H., et al., *Identification of chaperones in a MPP(+)-induced and ATRA/TPA-differentiated SH-SY5Y cell PD model.* Am J Transl Res, 2016. **8**(12): p. 5659-5671.

366. Aridon, P., et al., *Protective role of heat shock proteins in Parkinson's disease.* Neurodegener Dis, 2011. **8**(4): p. 155-68.

367. Falsone, S.F., et al., *The molecular chaperone Hsp90 modulates intermediate steps of amyloid assembly of the Parkinson-related protein alpha-synuclein.* J Biol Chem, 2009. **284**(45): p. 31190-9.

368. Dong, M.X., et al., *Serum Butyrylcholinesterase Activity: A Biomarker for Parkinson's Disease and Related Dementia.* Biomed Res Int, 2017. **2017**: p. 1524107.

369. GeneCards. *IL5 Gene.* Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=IL5&keywords=il5.

370. Pal, R., et al., *Role of neuroinflammation and latent transcription factors in pathogenesis of Parkinson's disease.* Neurol Res, 2016. **38**(12): p. 1111-1122.

371. Hasson, S.A., et al., *High-content genome-wide RNAi screens identify regulators of parkin upstream of mitophagy.* Nature, 2013. **504**(7479): p. 291-5.

372. Frantz, C., K.M. Stewart, and V.M. Weaver, *The extracellular matrix at a glance.* Journal of Cell Science, 2010. **123**(24): p. 4195-4200.

373. Bonneh-Barkay, D. and C.A. Wiley, *Brain Extracellular Matrix in Neurodegeneration.* Brain Pathology, 2009. **19**(4): p. 573-585.

374. Ghosh, D., et al., *alpha-synuclein aggregation and its modulation.* International Journal of Biological Macromolecules, 2017. **100**: p. 37-54.

375. Lehri-Boufala, S., et al., *New roles of glycosaminoglycans in alpha-synuclein aggregation in a cellular model of Parkinson disease.* PLoS One, 2015. **10**(1): p. e0116641.

376. Raghunathan, R., et al., *A glycomics and proteomics study of aging and Parkinson's disease in human brain.* Scientific Reports, 2020. **10**(1).

377. Zhang, W. and H.T. Liu, *MAPK signal pathways in the regulation of cell proliferation in mammalian cells.* Cell Research, 2002. **12**(1): p. 9-18.

378. Kim, E.K. and E.J. Choi, *Compromised MAPK signaling in human diseases: an update.* Archives of Toxicology, 2015. **89**(6): p. 867-882.

379. Cebrian, C., J.D. Loike, and D. Sulzer, *Neuroinflammation in Parkinson's Disease Animal Models: A Cell Stress Response or a Step in Neurodegeneration?* Behavioral Neurobiology of Huntington's Disease and Parkinson's Disease, 2015. **22**: p. 237-270.

380. Hunter, R.L., et al., *Inflammation induces mitochondrial dysfunction and dopaminergic neurodegeneration in the nigrostriatal system.* Journal of Neurochemistry, 2007. **100**(5): p. 1375-1386.

381. Zarubin, T. and J.H. Han, *Activation and signaling of the p38 MAP kinase pathway.* Cell Research, 2005. **15**(1): p. 11-18.

382. Banisadr, G., et al., *Highly regionalized neuronal expression of monocyte chemoattractant protein-1 (MCP-1/CCL2) in rat brain: Evidence for its colocalization with neurotransmitters and neuropeptides.* Journal of Comparative Neurology, 2005. **489**(3): p. 275-292.

383. Cho, J. and D.L. Gruol, *The chemokine CCL2 activates p38 mitogen-activated protein kinase pathway in cultured rat hippocampal cells.* Journal of Neuroimmunology, 2008. **199**(1-2): p. 94-103.

384. Block, M.L., L. Zecca, and J.S. Hong, *Microglia-mediated neurotoxicity: uncovering the molecular mechanisms.* Nature Reviews Neuroscience, 2007. **8**(1): p. 57-69.

385. Pal, R., et al., *Role of neuroinflammation and latent transcription factors in pathogenesis of Parkinson's disease.* Neurological Research, 2016. **38**(12): p. 1111-1122.

386. Lee, Y.C., et al., *Neurofilament Proteins as Prognostic Biomarkers in Neurological Disorders.* Current Pharmaceutical Design, 2019. **25**(43): p. 4560-4569.

387. UniProt, C., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D480-D489.

388. Jia, L., et al., *An attempt to understand kidney's protein handling function by comparing plasma and urine proteomes.* PLoS One, 2009. **4**(4): p. e5146.

389. Nobles, C., et al., *Correlation of urine and plasma cytokine levels among reproductive-aged women.* Eur J Clin Invest, 2015. **45**(5): p. 460-5.

390. Farrah, T., et al., *State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project.* J Proteome Res, 2014. **13**(1): p. 60-75.

391. Togashi, Y., et al., *Urinary cystatin C as a biomarker for acute kidney injury and its immunohistochemical localization in kidney in the CDDP-treated rats.* Exp Toxicol Pathol, 2012. **64**(7-8): p. 797-805.

392. Delgado-Morales, R. and M. Esteller, *Opening up the DNA methylome of dementia.* Molecular Psychiatry, 2017. **22**(4): p. 485-496.

393. Sharma, A., et al., *Comprehensive Profiling of Blood Coagulation and Fibrinolysis Marker Reveals Elevated Plasmin-Antiplasmin Complexes in Parkinson's Disease.* Biology (Basel), 2021. **10**(8).

394. Oikonomopoulou, K., et al., *Interactions between coagulation and complement--their role in inflammation.* Semin Immunopathol, 2012. **34**(1): p. 151-65.

395. Noris, M. and G. Remuzzi, *Overview of complement activation and regulation.* Semin Nephrol, 2013. **33**(6): p. 479-92.

396. Ouchi, N. and K. Walsh, *Adiponectin as an anti-inflammatory factor.* Clin Chim Acta, 2007. **380**(1-2): p. 24-30.

397. Kruger, R., et al., *Genetic analysis of the alpha2-macroglobulin gene in early- and late-onset Parkinson's disease.* Neuroreport, 2000. **11**(11): p. 2439-42.

398. Guo, X., et al., *Association between two alpha-2-macroglobulin gene polymorphisms and Parkinson's disease: a meta-analysis.* Int J Neurosci, 2016. **126**(3): p. 193-8.

399. Polito, R., et al., *Adiponectin Role in Neurodegenerative Diseases: Focus on Nutrition Review.* Int J Mol Sci, 2020. **21**(23).

400. Zhang, Y., et al., *Effect of ApoA4 on SERPINA3 mediated by nuclear receptors NR4A1 and NR1D1 in hepatocytes.* Biochemical and biophysical research communications, 2017. **487**(2): p. 327-332.

401. Vanni, S., et al., *Brain aging: A Janus-faced player between health and neurodegeneration.* Journal of Neuroscience Research, 2020. **98**(2): p. 299-311.

402. Harrington, M.G., et al., *Prostaglandin D synthase isoforms from cerebrospinal fluid vary with brain pathology.* Dis Markers, 2006. **22**(1-2): p. 73-81.

403. Whiteheart, S.W., *Platelet granules: surprise packages.* Blood, 2011. **118**(5): p. 1190-1.

404. Leiter, O. and T.L. Walker, *Platelets in Neurodegenerative Conditions-Friend or Foe?* Front Immunol, 2020. **11**: p. 747.

405. Ferrer-Raventos, P. and K. Beyer, *Alternative platelet activation pathways and their role in neurodegenerative diseases.* Neurobiol Dis, 2021. **159**: p. 105512.

406. Montenont, E., M.T. Rondina, and R.A. Campbell, *Altered functions of platelets during aging.* Curr Opin Hematol, 2019. **26**(5): p. 336-342.

407. Scavenius, C., et al., *Human inter-alpha-inhibitor is a substrate for factor XIIIa and tissue transglutaminase.* Biochimica Et Biophysica Acta-Proteins and Proteomics, 2011. **1814**(12): p. 1624-1630.

408. Tolosa, E., G. Wenning, and W. Poewe, *The diagnosis of Parkinson's disease.* Lancet Neurol, 2006. **5**(1): p. 75-86.

409. Beach, T.G. and C.H. Adler, *Importance of low diagnostic Accuracy for early Parkinson's disease.* Mov Disord, 2018. **33**(10): p. 1551-1554.

410. Shi, Q.Q., et al., *Complement C3-Deficient Mice Fail to Display Age-Related Hippocampal Decline.* Journal of Neuroscience, 2015. **35**(38): p. 13029-13042.

411. Stevens, B., et al., *The classical complement cascade mediates CNS synapse elimination.* Cell, 2007. **131**(6): p. 1164-1178.

412. Fu, S.H., et al., *Inverse association between periumbilical fat and longevity mediated by complement C3 and cardiac structure.* Aging-Us, 2020. **12**(22): p. 23296-23305.

413. Singh, S., S. Saleem, and G.L. Reed, *Alpha2-Antiplasmin: The Devil You Don't Know in Cerebrovascular and Cardiovascular Disease.* Frontiers in Cardiovascular Medicine, 2020. **7**.

414. Dwir, D., et al., *MMP9/RAGE pathway overactivation mediates redox dysregulation and neuroinflammation, leading to inhibitory/excitatory imbalance: a reverse translation study in schizophrenia patients.* Molecular Psychiatry, 2020. **25**(11): p. 2889-2904.

415. Amara, U., et al., *Molecular Intercommunication between the Complement and Coagulation Systems.* Journal of Immunology, 2010. **185**(9): p. 5628-5636.

416. Tanaka, T., et al., *Plasma proteomic signature of age in healthy humans.* Aging Cell, 2018. **17**(5).

417. Santos-Lozano, A., et al., *Successful aging: insights from proteome analyses of healthy centenarians.* Aging-Us, 2020. **12**(4): p. 3502-3515.

418. Roszkowska-Gancarz, M., et al., *Total and high molecular weight adiponectin and level-modifying polymorphisms of ADIPOQ in centenarians.* Endokrynologia Polska, 2012. **63**(6): p. 439-446.

419. Arai, Y., K. Kamide, and N. Hirose, *Adipokines and Aging: Findings From Centenarians and the Very Old.* Frontiers in Endocrinology, 2019. **10**.

420. Ng, R.C.L. and K.H. Chan, *Potential Neuroprotective Effects of Adiponectin in Alzheimer's Disease.* International Journal of Molecular Sciences, 2017. **18**(3).

421. Cassani, E., et al., *Serum Adiponectin Levels in Advanced-Stage Parkinson's Disease Patients.* Parkinsons Disease, 2011. **2011**.

422. Qi, Y., et al., *Adiponectin acts in the brain to decrease body weight.* Nature Medicine, 2004. **10**(5): p. 524-529.

423. Ehlting, C., S.D. Wolf, and J.G. Bode, *Acute-phase protein synthesis: a key feature of innate immune functions of the liver.* Biological Chemistry, 2021. **402**(9): p. 1129-1145.

424. Whiten, D.R., et al., *Single-Molecule Characterization of the Interactions between Extracellular Chaperones and Toxic alpha-Synuclein Oligomers.* Cell Reports, 2018. **23**(12): p. 3492-3500.

425. Lindholm, D., H. Wootz, and L. Korhonen, *ER stress and neurodegenerative diseases.* Cell Death and Differentiation, 2006. **13**(3): p. 385-392.

426. Gorbatyuk, M.S., et al., *Glucose Regulated Protein 78 Diminishes alpha-Synuclein Neurotoxicity in a Rat Model of Parkinson Disease.* Molecular Therapy, 2012. **20**(7): p. 1327-1337.

427. Fernandez, D., et al., *The Unfolded Protein Response in Immune Cells as an Emerging Regulator of Neuroinflammation.* Frontiers in Aging Neuroscience, 2021. **13**.

428. Mastroiacovo, F., et al., *Induction of the Wnt antagonist, Dickkopf-1, contributes to the development of neuronal death in models of brain focal ischemia.* Journal of Cerebral Blood Flow and Metabolism, 2009. **29**(2): p. 264-276.

429. Toledo, E.M., M. Colombres, and N.C. Inestrosa, *Wnt signaling in neuroprotection and stem cell differentiation.* Progress in Neurobiology, 2008. **86**(3): p. 281-296.

430. Marchetti, B. and S. Pluchino, *Wnt your brain be inflamed? Yes, it Wnt!* Trends in Molecular Medicine, 2013. **19**(3): p. 144-156.

431. Connolly, C., et al., *Enhanced Immune Response to Mmp3 Stimulation in Microglia Expressing Mutant Huntingtin.* Neuroscience, 2016. **325**: p. 74-88.

432. Bose, S. and J. Cho, *Role of chemokine CCL2 and its receptor CCR2 in neurodegenerative diseases.* Archives of Pharmacal Research, 2013. **36**(9): p. 1039-1050.

433. Browne, A.J., et al., *p38 MAPK regulates the Wnt inhibitor Dickkopf-1 in osteotropic prostate cancer cells.* Cell Death & Disease, 2016. **7**.

434. Rachner, T.D., et al., *P38 regulates the Wnt inhibitor Dickkopf-1 in breast cancer.* Biochemical and Biophysical Research Communications, 2015. **466**(4): p. 728-732.

Supplementary material

Supplementary table 1. Significantly different proteins from discovery proteomics study of centenarians which was used for the pathway analysis in IPA. *The table shows the p-values and the fold changes for the proteins (denoted by gene names)*

Gene	p-value	Fold change
PZP	1.94E-05	1.684701
SERPINA1	6.17E-05	1.73181
ITIH3	7.62E-05	1.708078
CLEC3B	0.000298	-1.65156
LRG1	0.000467	1.527141
CTBS	0.000988	1.768556
CD5L	0.001642	1.760714
SERPINA4	0.001675	1.555738
OAF	0.002554	1.586098
CHIT1	0.002974	1.601173
CD14	0.004617	1.542748
DKK3	0.004811	1.558184
UGT2B7	0.004896	1.422519
C1RL	0.005041	1.489035
C9	0.005087	1.5195
CACNB3	0.005574	-1.6446
A1BG	0.006678	1.468933
GFAP	0.006889	1.517302
HYI	0.007417	1.544166
PI16	0.007772	1.450476
C22orf15	0.007945	1.477272
PKM	0.008547	-1.37904
SAA1	0.009027	1.448169
AMBP	0.009468	1.421611
CP	0.009814	1.502154
STX11	0.010424	-1.45947
LMNA	0.011123	-1.4286
CFD	0.011455	1.448416
SERPINA3	0.011601	1.478006
A2M	0.011729	1.431837
ORM1	0.012515	1.390967
NoGene_04	0.013317	-1.59076
METTL4	0.014453	1.470671
TGFBI	0.015632	1.439106
AMPD3	0.015927	1.423705
ADIPOQ	0.016633	1.511633
B2M	0.017582	1.441804
SPARCL1	0.018008	1.419579
AZGP1	0.018094	1.410957
SOD3	0.018419	1.494747
FAM184A	0.018426	1.35335
ITIH4	0.018635	1.368747
APMAP	0.01868	1.407449
HBE1	0.018863	1.429083
C1QB	0.019699	1.43234

Gene	p-value	Fold change
TPM2	0.020764	-1.44907
DYTN	0.021606	1.355401
KIF7	0.023071	1.376944
PTGDS	0.024648	1.428652
KLKB1	0.026836	1.329518
PPBP	0.027282	1.436044
OLFM1	0.027748	1.364223
DOCK10	0.027983	1.398321
CST3	0.028926	1.423519
GSN	0.031905	1.386684
KNG1	0.032963	1.363328
ORM2	0.033104	1.367543
C7	0.033771	1.409285
FHL5	0.034531	1.360578
C1QC	0.03576	1.379908
SRFBP1	0.035972	-1.54587
ERICH6B	0.037066	1.409262
RAB40AL	0.037917	1.374689
S100A9	0.038328	-1.49063
C4BPA	0.040151	1.339755
FBLN1	0.040687	1.373487
C6	0.042612	1.320121
FGA	0.043262	1.398148
PROCR	0.045712	1.298981
CD163	0.046332	1.343045
BTD	0.048208	1.339022
VCL	0.048623	1.298372
CFI	0.049081	-1.61582
ALB	0.049706	-1.47564
NRP1	0.049734	-1.52566

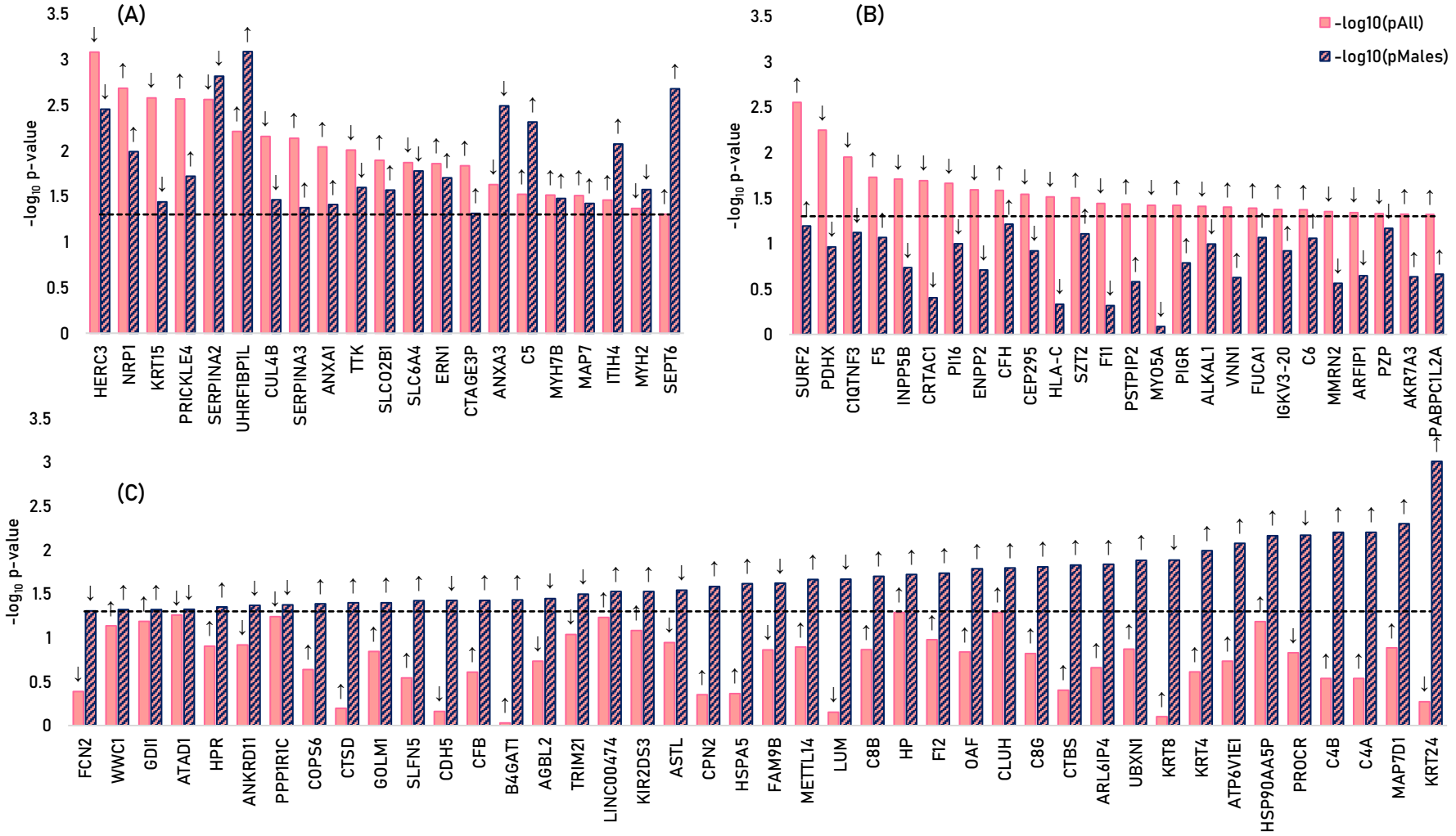
Supplementary table 2. Beta coefficients resulting from the regression analysis of the discovery proteomics study of centenarians and controls, relating protein expression and age. *The table shows the beta coefficients and the p-value significance of the regression for the groups (i) all, (ii) centenarians, and (iii) controls. NS p > 0.05, *p < 0.05, **p < 0.01, ***p < 0.001 and ****p < 0.0001*

	(i) All		(ii)Centenarians		(iii) Controls	
	Beta coefficient	p-value	Beta coefficient	p-value	Beta coefficient	p-value
PZP	76.9	****	12.6	NS	18.4	NS
SERPINA1	42.3	***	6.6	NS	-8.1	NS
ITIH3	37.7	***	8.0	*	-4.2	NS
CLEC3B	-26.1	***	-1.2	NS	2.8	NS
LRG1	28.1	***	6.3	**	0.1	NS
CTBS	38.7	**	6.2	*	-9.6	NS
CD5L	29.4	***	2.9	NS	9.3	NS
SERPINA4	60.1	**	11.7	NS	1.9	NS
OAF	30.5	**	0.0	NS	6.4	NS
CHIT1	26.1	**	-2.0	NS	6.0	NS
CD14	39.0	**	7.6	NS	-6.4	NS
DKK3	15.3	**	-0.3	NS	2.3	NS
UGT2B7	47.8	**	2.3	NS	18.6	NS
C1RL	26.8	**	1.5	NS	0.0	NS
C9	48.4	**	9.6	NS	0.8	NS
CACNB3	-8.4	**	-1.2	NS	0.7	NS
A1BG	43.2	**	11.5	**	-4.4	NS
GFAP	38.5	*	0.1	NS	-6.0	NS
HYI	31.8	*	0.1	NS	-17.9	NS
PI16	47.7	**	2.0	NS	8.1	NS
C22orf15	87.6	***	25.3	****	29.6	NS
PKM	-33.2	**	-6.5	*	-4.9	NS
SAA1	38.9	**	7.4	NS	-3.6	NS
AMBP	29.5	**	6.6	*	-0.4	NS
CP	38.5	*	7.5	NS	-3.6	NS
STX11	-9.5	*	-1.6	NS	1.3	NS
LMNA	-42.7	**	-2.5	NS	-13.3	NS
CFD	27.6	*	0.5	NS	2.6	NS
SERPINA3	28.4	**	5.2	*	-1.6	NS
A2M	27.6	*	1.7	NS	-1.5	NS
ORM1	22.6	**	3.4	NS	3.5	NS
A6NIZ1	-10.7	*	-0.4	NS	1.9	NS
METTL4	34.4	**	8.6	*	1.3	NS
TGFBI	43.3	*	2.5	NS	2.2	NS
AMPD3	42.1	*	6.2	NS	2.0	NS
ADIPOQ	16.9	**	2.6	*	2.8	NS
B2M	17.5	*	2.9	NS	-5.0	NS
SPARCL1	28.0	*	-2.0	NS	-2.0	NS
AZGP1	31.7	**	8.4	**	2.2	NS
SOD3	21.3	**	4.5	***	0.8	NS
FAM184A	21.4	*	5.8	**	-0.4	NS
ITIH4	43.5	*	13.5	**	-6.2	NS

	(i) All		(ii)Centenarians		(iii) Controls	
	Beta coefficient	p-value	Beta coefficient	p-value	Beta coefficient	p-value
APMAP	44.7	*	-1.5	NS	2.9	NS
HBE1	20.2	*	3.4	NS	-2.8	NS
C1QB	33.8	*	1.7	NS	-1.4	NS
TPM2	-10.9	*	-2.5	NS	0.4	NS
DYTN	15.1	**	2.9	*	3.7	NS
KIF7	30.1	*	8.0	**	-0.6	NS
PTGDS	24.7	*	6.1	NS	1.0	NS
KLKB1	35.0	*	8.8	**	3.8	NS
PPBP	11.7	*	-0.8	NS	-1.5	NS
OLFM1	19.7	*	1.2	NS	-0.8	NS
DOCK10	45.8	**	9.3	*	13.1	NS
CST3	37.6	*	7.2	*	6.7	NS
GSN	83.1	*	6.4	NS	9.0	NS
KNG1	44.1	*	13.3	**	2.0	NS
ORM2	22.0	*	5.3	**	2.3	NS
C7	36.5	NS	-3.0	NS	-2.9	NS
FHL5	22.0	*	1.4	NS	1.7	NS
C1QC	42.7	*	5.0	NS	3.3	NS
SRFBP1	-8.2	NS	-1.6	NS	1.5	NS
ERICH6B	45.3	NS	-6.5	NS	-8.8	NS
RAB40AL	12.3	*	1.9	NS	-0.7	NS
S100A9	-21.1	*	-1.9	NS	-0.2	NS
C4BPA	26.4	*	2.5	NS	0.4	NS
FBLN1	30.5	*	5.5	NS	2.3	NS
C6	47.2	*	11.6	*	6.3	NS
FGA	32.4	NS	3.4	NS	-4.4	NS
PROCR	30.5	*	1.2	NS	10.3	NS
CD163	22.3	*	3.5	NS	3.5	NS
BTD	36.6	NS	8.8	NS	-28.0	*
VCL	30.6	NS	10.4	**	-7.5	NS
CFI	-22.3	NS	-5.5	NS	1.5	NS
ALB	-8.0	NS	-0.2	NS	0.3	NS
NRP1	-12.0	*	-1.7	NS	-0.6	NS

Supplementary table 3. Outlier values detected in the targeted centenarian study at ten median absolute deviations. *The highest outlier values are observed for SAA1 in the female centenarians, apart from this, the outliers were observed to be randomly distributed among the samples. M = male, and F = female*

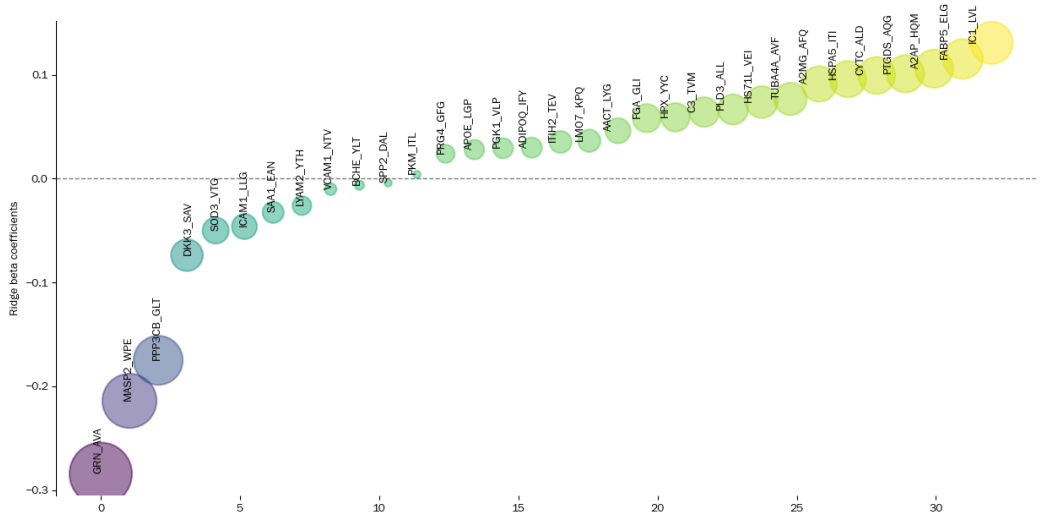
Sample group	Sex	Age	SERPINA3	FABP5	HSPA1L	PTGDS	SAA1	SOD3	TUBA4A	VCAM1
Control	F	82						10.6		
Control	M	71		32.6						
Offspring	F	56			4.0					
Offspring	F	67						11.8		
Offspring	F	68						10.4		
Centenarian	F	101					145.3			
Centenarian	F	102					130.1			
Centenarian	F	103		11.3						
Centenarian	F	105	4.4				675.1			
Centenarian	F	107							12.0	
Centenarian	F	107						12.9		
Centenarian	F	109					142.9			
Centenarian	M	100				7.5				6.6
Centenarian	M	100		10.5						



Supplementary figure 1. Significance of differentially expressed proteins between de novo PD and controls ($-\log_{10}$ p-value). *(A) Significantly different in the comparison between DNP and controls in all samples and males only. (B) Significantly different in the comparison between DNP and control in all samples. (C) Significantly different in males only. The horizontal dashed line represents the nominal p-value cut-off limit of 0.05. The arrow above each bar shows the expression of the proteins, ↑ upregulated in de novo PD, ↓ downregulated in de novo PD*

Supplementary table 4. Outlier values from the targeted plasma proteomics study of healthy controls, de novo PD, iRBD patients and patients with other neurological disorders. *Outliers were mainly detected in SAA1, possibly indicating a severe inflammatory response, and in SOD3. M = male, and F = female*

Sample group	Sex	Age	FABP5	PLD3	PPP3CB	SAA1	SOD3
Control	F	60				33.0	
De novo PD	F	77		7.7			21.2
De novo PD	F	61					
De novo PD	M	81				42.1	
De novo PD	F	72					11.4
De novo PD	F	53	9.5				
De novo PD	F	75					13.1
De novo PD	F	66					10.2
De novo PD	M	64				28.1	
OND	M	80				34.5	



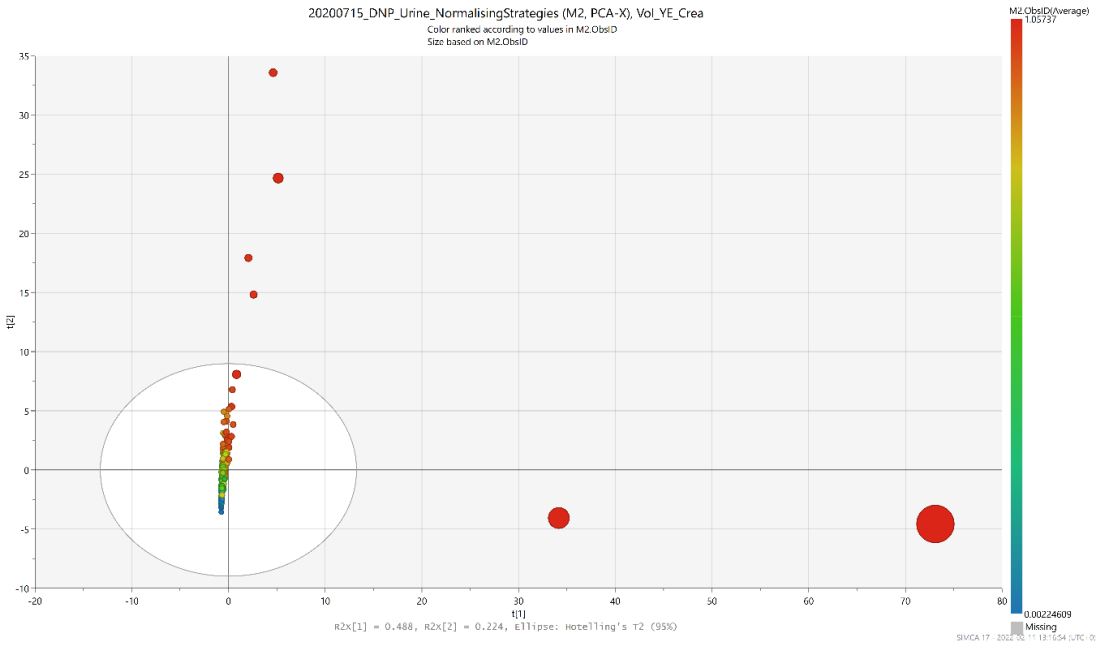
Supplementary figure 2. Coefficients in the predictive plasma PD Ridge classifier model. *The size of the coefficients are represented by circle radii. The coefficient with the greatest absolute values have the largest influence on the classifier model*

Supplementary table 5. Pathway analysis results from the urine discovery analysis of idiopathic Parkinson's disease patients versus healthy controls. *The table shows the significantly enriched pathways, the -log$_{10}$ p-value of the enrichment, the z-score denoting activation (>2) or deactivation (<2), and the proteins in the pathway*

Ingenuity Canonical Pathways	-log$_{10}$ p-value	z-score	Proteins
LXR/RXR Activation	4.76	1.63	AGT, VTN, TNFRSF1A, FGA, APOE, HPX
FXR/RXR Activation	4.68		AGT, VTN, MAPK12, FGA, APOE, HPX
Glycerol Degradation I	3.71		GK2, GPD1
Gap Junction Signalling	3.66		EGF, PPP3CB, TUBA4A, PLCG1, ACTA1, TUBA3C/TUBA3D
14-3-3-mediated Signalling	3.53		TNFRSF1A, MAPK12, TUBA4A, PLCG1, TUBA3C/TUBA3D

Ingenuity Canonical Pathways	-log$_{10}$ p-value	z-score	Proteins
Apoptosis Signalling	3.16	2.00	TNFRSF1A, PLCG1, CYCS, CAPN2
Axonal Guidance Signalling	3.06		EGF, PPP3CB, EFNA5, TUBA4A, PLCG1, ACE, EPHA7, TUBA3C/TUBA3D
FAK Signalling	3.01		EGF, PLCG1, ACTA1, CAPN2
Acute Phase Response Signalling	3.01	2.24	AGT, TNFRSF1A, MAPK12, FGA, HPX
Germ Cell-Sertoli Cell Junction Signalling	3.01		TNFRSF1A, MAPK12, TUBA4A, ACTA1, TUBA3C/TUBA3D
Sertoli Cell-Sertoli Cell Junction Signalling	2.97		TNFRSF1A, MAPK12, TUBA4A, ACTA1, TUBA3C/TUBA3D
Sirtuin Signalling Pathway	2.77	-1.63	SOD3, MAPK12, NDRG1, TUBA4A, NDUFB11, TUBA3C/TUBA3D
Renin-Angiotensin Signalling	2.67		AGT, MAPK12, PLCG1, ACE
Parkinson's Signalling	2.64		MAPK12, CYCS
Clathrin-mediated Endocytosis Signalling	2.63		EGF, HSPA8, PPP3CB, ACTA1, APOE
Calcium-induced T Lymphocyte Apoptosis	2.61		PPP3CB, PLCG1, CAPN2
Induction of Apoptosis by HIV1	2.61		TNFRSF1A, MAPK12, CYCS
Remodelling of Epithelial Adherens Junctions	2.49		TUBA4A, ACTA1, TUBA3C/TUBA3D
Phagosome Maturation	2.46		CTSH, TUBA4A, ACE, TUBA3C/TUBA3D
EGF Signalling	2.45		EGF, MAPK12, PLCG1
Epithelial Adherens Junction Signalling	2.41		EGF, TUBA4A, ACTA1, TUBA3C/TUBA3D
UDP-N-acetyl-D-galactosamine Biosynthesis I	2.35		GALE,
Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes	2.27		MAPK12, PPP3CB, PLCG1
Death Receptor Signalling	2.09		TNFRSF1A, ACTA1, CYCS
Hepatic Fibrosis / Hepatic Stellate Cell Activation	2.03		AGT, EGF, TNFRSF1A, COL4A2
Dendritic Cell Maturation	2.02	1.00	TNFRSF1A, MAPK12, IGHG1, PLCG1
ErbB Signalling	2.02		EGF, MAPK12, PLCG1
Coagulation System	1.97		PLAU, FGA
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	1.96	1.00	TNFRSF1A, MAPK12, PLCG1, APOE
Complement System	1.95		MASP2, CFH
UVA-Induced MAPK Signalling	1.93		MAPK12, PLCG1, CYCS
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	1.93		DKK3, TNFRSF1A, PPP3CB, IGHG1, PLCG1
Type I Diabetes Mellitus Signalling	1.91		TNFRSF1A, MAPK12, CYCS
Glycerol-3-phosphate Shuttle	1.87		GPD1,
Inosine-5'-phosphate Biosynthesis II	1.87		ATIC,
Neuroprotective Role of THOP1 in Alzheimer's Disease	1.86		AGT, ENDOU, ACE
Leukocyte Extravasation Signalling	1.86	1.00	MAPK12, PLCG1, ACTA1, THY1
Role of PKR in Interferon Induction and Antiviral Response	1.84		TNFRSF1A, CYCS
Glucocorticoid Receptor Signalling	1.76		AGT, PLAU, HSPA8, MAPK12, PPP3CB
Phagosome Formation	1.75		VTN, IGHG1, PLCG1

Ingenuity Canonical Pathways	-log$_{10}$ p-value	z-score	Proteins
nNOS Signalling in Neurons	1.74		PPP3CB, CAPN2
TNFR1 Signalling	1.71		TNFRSF1A, CYCS
CD28 Signalling in T Helper Cells	1.71		MAPK12, PPP3CB, PLCG1
Iron homeostasis signalling pathway	1.71		EGF, HPX, HBE1
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	1.71		DKK3, TNFRSF1A, MAPK12, PPP3CB
Amyloid Processing	1.68		MAPK12, CAPN2
Galactose Degradation I (Leloir Pathway)	1.65		GALE,
CD27 Signalling in Lymphocytes	1.64		MAPK12, CYCS
Nur77 Signalling in T Lymphocytes	1.63		PPP3CB, CYCS
autophagy	1.61		CTSH, ACE
Regulation of Cellular Mechanics by Calpain Protease	1.6		EGF, CAPN2
Huntington's Disease Signalling	1.6		EGF, HSPA8, CYCS, CAPN2
Ephrin A Signalling	1.53		EFNA5, EPHA7
PCP pathway	1.53		CTHRC1, MAPK12
Superoxide Radicals Degradation	1.45		SOD3,
Agrin Interactions at Neuromuscular Junction	1.44		MAPK12, ACTA1
Chemokine Signalling	1.43		MAPK12, PLCG1
Mitochondrial Dysfunction	1.42		MAPK12, CYCS, NDUFB11
Glioma Invasiveness Signalling	1.41		VTN, PLAU
Myc Mediated Apoptosis Signalling	1.41		MAPK12, CYCS
Tec Kinase Signalling	1.4		MAPK12, PLCG1, ACTA1
Caveolar-mediated Endocytosis Signalling	1.39		EGF, ACTA1
Ephrin Receptor Signalling	1.37		EGF, EFNA5, EPHA7
Toll-like Receptor Signalling	1.35		MAPK12, TOLLIP
GDNF Family Ligand-Receptor Interactions	1.34		MAPK12, PLCG1
IL-15 Signalling	1.34		MAPK12, PLCG1
Non-Small Cell Lung Cancer Signalling	1.33		EGF, PLCG1
Neuroinflammation Signalling Pathway	1.33		TNFRSF1A, MAPK12, PPP3CB, PLCG1
Purine Nucleotides De Novo Biosynthesis II	1.32		ATIC
UDP-N-acetyl-D-galactosamine Biosynthesis II	1.32		GALE
IL-17A Signalling in Airway Cells	1.32		MAPK12, MUC5B



Supplementary figure 3. PCA of the creatinine normalised targeted urine proteomics analysis of de novo PD patients. The PCA shows that six samples are extreme outliers.

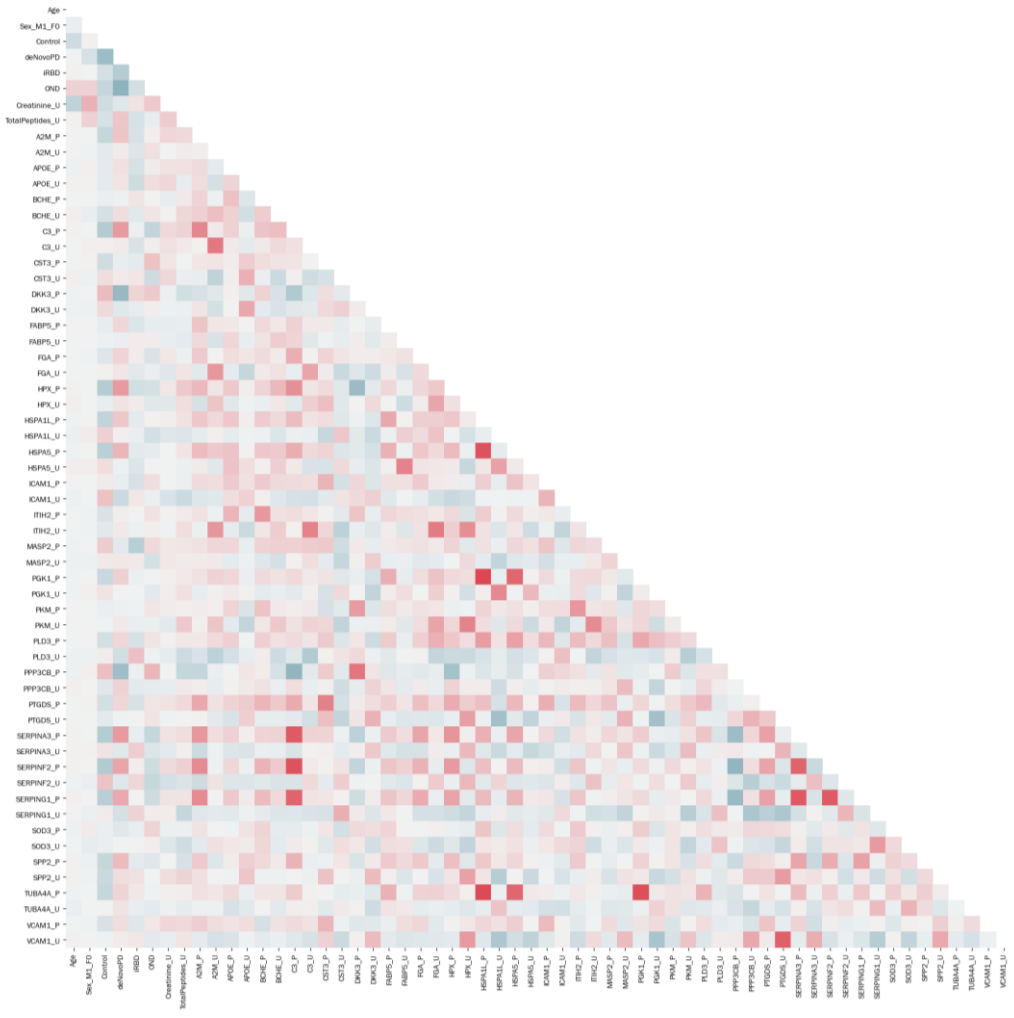
Supplementary table 6. Outliers detected at ten median absolute deviations in the targeted proteomics analysis of the de novo PD urine cohort. *In total, 37 outlier values were detected in 27 different proteins. DKK = healthy controls, DNP = de novo PD, and OND = other neurological disorders. M = male, F = female, and NA = not available*

Sample	Sex	Age	SERPINA1	A2M	APOE	ATIC	BCHE	C3	CAPN2	CYCS	CST3	DKK3	FABP5	FGA	FGF21	HPX	HSPA8	ITIH2	MAPK12	MMP3	MSN	NEFM	PKM	PPP3CB	TEK	TXN	TOLLIP	TUBA4A	UBC
DKK	F	56																								x			
	F	73																										x	
	F	70																											x
DKS	M	65															x												
	F	62																							x				
	M	72				x					x	x																	
DNP	F	77		x										x				x											
	M	81	x		x		x		x					x				x		x			x						
	M	87																				x							
	M	76																							x				
	F	71						x																					
	F	74	x																		x								
	F	54																								x			
	F	50											x																
	F	73																									x		
	F	80			x																								
	F	73																						x					
	M	80																	x										
	M	72					x								x														
	NA	NA																						x					
	F	80																			x								
OND	M	74												x															
	M	80														x													
	F	71								x																			

Supplementary table 7. Benjamini-Hochberg FDR-adjusted p-values from the comparison of control to de novo PD, iRBD and other, non-PD, neurological disorders in the targeted urine proteomics study

Protein (gene)	Control versus vs de novo PD	Control versus iRBD	Control versus other neurological disorders
Intercellular adhesion molecule 1 (ICAM1)	9.1E-04	5.8E-01	5.6E-02
C-C motif chemokine 4 (CCL4)	9.2E-04	7.9E-01	7.0E-01
Troponin T3, fast skeletal type (TNNT3)	1.1E-03	6.5E-01	3.4E-03
Secreted Phosphoprotein 2 (SPP2)	1.1E-03	4.3E-02	6.1E-03
Collagen alpha-3(VI) chain (COL6A3)	1.1E-03	2.6E-01	6.0E-02
Neurofilament medium polypeptide (NEFM)	1.1E-03	5.8E-01	1.9E-02
Polyubiquitin-C (UBC)	1.1E-03	1.7E-03	1.9E-02
Phosphoglycerate mutase 1 (PGAM1)	1.2E-03	7.4E-01	1.9E-01
Alpha-2-antiplasmin (SERPINF2)	1.3E-03	7.9E-01	1.7E-04
Calpain-2 catalytic subunit (CAPN2)	2.4E-03	1.0E+00	8.8E-03
Fibroblast growth factor 21 (FGF21)	2.6E-03	6.5E-01	6.8E-02
Neural cell adhesion molecule 1 (NCAM1)	3.5E-03	3.7E-01	5.9E-01
N-myc downstream regulated 1 (NDRG1)	5.3E-03	5.0E-01	2.0E-02
Matrix metallopeptidase 3 (MMP3)	5.3E-03	1.0E+00	6.7E-02
Mitogen-activated protein kinase 12 (MAPK12)	9.0E-03	1.0E+00	8.4E-03
C-C motif chemokine 2 (CCL2)	2.4E-02	1.7E-03	1.9E-02
Heat shock protein HSP 90-beta (HSP90AB1)	2.4E-02	2.7E-01	3.1E-01
Serine/threonine-protein phosphatase 2B catalytic subunit (PPP3CB)	3.0E-02	8.8E-01	5.9E-01
Cholinesterase (BCHE)	3.6E-02	7.6E-01	5.6E-02
Interleukin-5 (IL5)	4.3E-02	2.6E-01	5.4E-01
Alpha-1-antichymotrypsin (SERPINA3)	4.4E-02	7.9E-01	6.7E-02
Dickkopf WNT signaling pathway inhibitor 3 (DKK3)	4.4E-02	3.9E-01	5.7E-01
Heat shock 70 kDa protein 1-like (HSPA1L)	5.0E-02	3.5E-01	1.7E-02
Plasma protease C1 inhibitor (SERPING1)	9.5E-02	6.5E-01	6.8E-01
Cystatin-C (CST3)	9.5E-02	7.3E-01	3.4E-03
C-C motif chemokine 17 (CCL17)	1.0E-01	1.0E+00	5.9E-01
Heat shock cognate 71 kDa protein (HSPA8)	1.0E-01	6.0E-01	1.8E-01
Interleukin-1 beta (IL1B)	1.1E-01	6.5E-01	8.5E-01
Angiopoietin-1 receptor (TEK)	1.3E-01	7.1E-01	6.2E-01
NACHT, LRR and PYD domains-containing protein 3 (NLRP3)	1.4E-01	6.0E-01	5.9E-01
Amyloid-beta precursor protein (APP)	1.4E-01	3.1E-03	4.1E-01
L-xylulose reductase (DCXR)	3.1E-01	1.0E+00	9.4E-01
Erythropoietin (EPO)	3.6E-01	9.9E-01	5.9E-01
Thioredoxin (TXN)	3.8E-01	6.0E-01	8.5E-01
Inter-alpha-trypsin inhibitor heavy chain 2 (ITIH2)	4.0E-01	8.8E-01	5.9E-01
Annexin A2 (ANXA2)	4.0E-01	5.8E-01	7.6E-01
Tubulin alpha-4A chain (TUBA4A)	4.0E-01	9.4E-01	9.0E-01
Hemopexin (HPX)	4.1E-01	6.5E-01	8.9E-01
Alpha-2-macroglobulin (A2M)	4.4E-01	6.5E-01	2.4E-01
Apolipoprotein E (APOE)	5.5E-01	1.1E-01	4.9E-01
60 kDa heat shock protein, mitochondrial (HSPD1)	5.6E-01	6.5E-01	8.9E-01
Collagen alpha-2(IV) chain (COL4A2)	5.7E-01	1.0E+00	7.6E-01
Prostaglandin-H2 D-isomerase (PTGDS)	5.7E-01	8.5E-01	5.7E-01
Pyruvate kinase M (PKM)	5.7E-01	1.0E+00	8.5E-01

Protein (gene)	Control versus vs de novo PD	Control versus iRBD	Control versus other neurological disorders
Phospholipase D Family Member 3 (PLD3)	5.7E-01	5.5E-02	9.2E-01
Alpha-1-antitrypsin (SERPINA1)	5.7E-01	6.8E-01	8.5E-01
Extracellular superoxide dismutase [Cu-Zn] (SOD3)	5.7E-01	6.5E-01	6.2E-01
Fatty acid binding protein 5 (FABP5)	6.2E-01	6.5E-01	5.7E-01
Bifunctional purine biosynthesis protein PURH (ATIC)	6.2E-01	3.7E-01	8.5E-01
Phosphoglycerate kinase 1 (PGK1)	6.9E-01	6.8E-01	3.4E-01
Toll-interacting protein (TOLLIP)	7.0E-01	7.1E-01	5.9E-01
Endoplasmic reticulum chaperone BiP (HSPA5)	7.0E-01	5.0E-01	7.2E-01
Cytochrome C (CYCS)	7.0E-01	7.9E-01	8.9E-01
Urokinase-type plasminogen activator (PLAU)	7.5E-01	1.1E-01	4.7E-01
Mannan binding lectin serine peptidase 2 (MASP2)	8.6E-01	1.0E+00	8.8E-02
Fibrinogen alpha chain (FGA)	8.6E-01	5.0E-01	9.0E-01
Complement C3 (C3)	8.6E-01	1.9E-01	8.9E-01
Vascular cell adhesion protein 1 (VCAM1)	9.0E-01	8.7E-01	6.8E-01
Moesin (MSN)	9.7E-01	6.5E-01	9.0E-01



Supplementary figure 4. Pearson correlation matrix of the targeted proteomic studies of de novo PD patients, healthy controls, iRBD patients and patients with other neurological disorders.