FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

Jan Stanstrup

DEPARTMENT OF NUTRITION, EXERCISE AND SPORTS

Metabolomics Investigation of Whey Intake:
Discovery of Markers and Biological Effects Supported by a Computer-Assisted Compound Identification Pipeline

PhD thesis

Jan Stanstrup

Metabolomics investigation of whey intake:
Discovery of markers and biological effects supported by a computer-assisted compound identification pipeline

Principal Supervisor
Professor Lars Ove Dragsted

Co-supervisors
Thaer Barri, Ph.D.

Metabolomics Investigation of Whey Intake:
Discovery of Markers and Biological Effects
Supported by a Computer-Assisted Compound Identification Pipeline

ISBN 978-87-7611--704-7

Printed by SL grafik, Frederiksberg C, Denmark (www.slgrafik.dk)

“*Measure what can be measured, and make measurable what cannot be measured."*

Galileo Galilei

PREFACE

This thesis presents the results of my work as a Ph.D. student and it is submitted to meet the requirements for attaining the Ph.D. degree at the Faculty of Science, University of Copenhagen.

The grant was provided by the Danish Obesity Research Centre (DanORC) and Nutritional Biomics and Innovation (NuBI). DanORC was supported by the Danish Council for Strategic Research and NuBI was supported by The Danish Ministry of Food, Agriculture and Fisheries.

The reported work has been carried out in the Bioactive Foods and Health group at the Department of Nutrition, Exercise and Sports, Faculty of Sciences, University of Copenhagen under the principal supervision of Professor Lars Ove Dragsted and at Dr. Steffen Neumann's Bioinformatics & Mass Spectrometry group at IPB, Halle, Germany.

This Ph.D. thesis describes an investigation of the effects of whey protein intake on the human metabolome. Two of the studies are described in the following papers:

- **Stanstrup J**, Schou SS, Holmer-Jensen J, Hermansen K, Dragsted LO. Whey protein delays gastric emptying and suppresses plasma fatty acids and their metabolites compared to casein, gluten and fish protein. 2013. (*submitted*)
- **Stanstrup J**, Rasmussen JE, Ritz C, Holmer-Jensen J, Hermansen K, Dragsted LO. Intakes of whey protein hydrolysate and whole whey proteins are discriminated by LC–MS. Metabolomics. 2013.

The results of an additional study are summarized in section 5.3 of this thesis.

To assist the main investigation related to whey protein a pipeline for semi-automated compound identification for use in metabolomics studies was developed. This work was published in:

- **Stanstrup J**, Gerlich M, Dragsted LO, Neumann S. Metabolite Profiling and beyond: Approaches for the rapid processing and annotation of human blood serum mass spectrometry data. Anal. Bioanal. Chem. 2013. 405(15):5037–48.

The mentioned publications are included in Appendix II, III and I, respectively.

During the Ph.D. period contributions were also made to the following additional papers, which are, however, not included in this thesis:

- Jiang P, Dragsted LO, Barri T, **Stanstrup J**, Jensen ML, Wan JM, Sangild PT. Antibiotics markedly alter the urinary and plasma metabolome of preterm pigs susceptible to necrotizing enterocolitis. (*in preparation*)
- Jiang P, Dragsted LO, **Stanstrup J**, Thymann P, Sangild PT. Severe malnutrition alters the urinary and plasma metabolome of neonatal pigs. (*in preparation*)
- Johansson A, Barri T, Ulmius M, **Stanstrup J**, Önning G, Dragsted LO. LC-QTOF/MS metabolomic profiles in human urine after a 5-week high dietary fiber intake. (*in preparation*)

ACKNOWLEDGEMENTS

Many persons have contributed to making my Ph.D. period unforgettable through scientific guidance and discussions as well as friendship, encouragements and support and I would like to express my gratitude to them here.

First of all I would like to thanks my supervisors Lars Ove Dragsted, Rasmus Bro and Thaer Barri that made all this possible by selecting me for the position in the first place.

I would especially like to thank Lars for his always overwhelmingly enthusiastic support and willingness to let me go where my curiosity led me and for his manner in which to gently pull me back to the middle of the road when needed.

During my Ph.D. I had the great fortune to spend three months and a half in the Bioinformatics & Mass Spectrometry group of Steffen Neumann. The time I spend there was one of the most rewarding experiences of my Ph.D. I learned an incredible amount in this short period of time and gained a number of skills that I am sure will serve me well going forward. I would also like to thank Steffen, Michael Gerlich and Carsten Kuhl not only for their continuous willingness to help with the project, or any other problem I might have, but also for making me feel welcome in their group and creating an always jovial atmosphere. Many of the tools described in this thesis was invented in their group and all credit belongs to the respective developers.

I am also grateful to Jakob Ewald Rasmussen, Christian Ritz and Simon Stubbe Schou for unhesitatingly providing assistance indispensable to the success of the project.

I would moreover like to thank all the technical and administrative staff for always kind and helpful assistance. A special thanks goes to Ümmühan Celik for assistance in the lab and to Suzanne Møller for tirelessly taking care of any issue that might arise with a smile (including but not limited to tracking down my luggage in Oslo when I had instead intended it to accompany me to Glasgow).

I have been very lucky to serve with wonderful fellow convicts: Daniela, Gözde, Maj-Britt and Mette. Thanks to all of you for fruitful discussions, for creating a team atmosphere in an otherwise competitive world, for indispensable help and for insisting to sometimes pull me out of my Ph.D. bubble – and occasionally succeeding.

I have learned a lot during my Ph.D. but I am keenly aware that all of this was only possible because of the people that helped prepared me for this task.

At this time I would like to recall and acknowledge Bent Kirkegaard, my high school chemistry teacher, who sparked my interest in spectrometry and spectroscopy. I have found the foundation he laid out both solid and comforting in my studies at FARMA and during this Ph.D.

The supervision of Professor Dan Stærk during my master thesis was also a key formative experience. He set me on a path to research and I am grateful to him for believing that I belonged in the world of academia when it took a while for others to agree.

Above all I would like to thank Sara for sharing this journey with me, for listening to "discoveries" that were probably *marginally* more interesting to me than to her but most importantly for being a spirit of fire when lights seemed otherwise to be dim around me.

Lastly I want to thanks my parents. Without their example of hard work, curiosity and dedication this would not have been possible.

Jan Stanstrup

Frederiksberg, December 2013

ABSTRACT

Classical human intervention studies are typically carried out to prove a hypothesis or answer a specific question. Therefore the design of the experiment and the performed analyses are based only on a few parameters related to the specific effect to be investigated while a wealth of information contained in the collected samples is not even recorded.

In recent years another approach, *metabolomics*, has proved to be a very valuable technique in nutrition research. In the metabolomics approach an intervention is undertaken without pre-defining which variables will be monitored – on the contrary the aim is to acquire as much information as possible in an unbiased way. In the present Ph.D. thesis the LC-MS metabolomics approach was applied to human nutrition intervention studies.

The plasma and urine samples collected in connection with three human nutrition intervention studies were analyzed using a metabolomics approach for the purpose of finding biomarkers of milk-derived whey protein intake and investigating the effects of whey intake on the human metabolome. It was an additional aim to devise computer-assisted methods to rationalize the process of compound identification.

In the first study the metabolomics profiles were compared following high-fat meals containing either cod, gluten, casein or whey as the protein source, while in the second study whole whey, subfractions of whey (α-lactalbumin or caseinoglycomacropeptide) and whey hydrolysate were compared. Both studies were performed with obese non-diabetics and the last study was repeated with diabetics as well.

We demonstrated that intake of whey causes a decreased rate of gastric emptying compared to other protein sources. This is in contrast to previous findings suggesting that whey is cleared faster than other proteins.

Paradoxically, we also find disproportionately elevated levels and shorter T_{max} of some aromatic and branched-chain amino acids following the whey meal. This suggests that whey affects absorption of amino acids in a way independent from, or at least not wholly controlled by, gastric emptying.

In addition, we find that whey caused decreased levels of a number of fatty acids due to increased insulin levels, which in turn is likely induced by the exaggerated amino acid levels.

We found no differences between the subfractions however, except for those explained by their different amino acid compositions. However, the hydrolysate contained a number of cyclic dipeptides that may be causing the hypoglycaemic effects observed for the hydrolysate. In addition we found that the manufacturing process for the hydrolysate caused methionine oxidation products, which were metabolized endogenously to metabolites not observed previously in humans.

We did not succeed in finding highly specific exposure markers of whey as the effects were confined to modifying levels of endogenous metabolites. Whey hydrolysate, on the other hand, contained unusual cyclic dipeptides; they are however, unlikely to be whey specific but rather a result of the hydrolysis process. These results could not have been recognized using traditional hypothesis testing approaches.

We also found a number of markers of the cod meal. While most are likely also markers of meat intake, arsenobetaine may be a specific marker of recent salt water fish intake.

We did not find any difference between obese non-diabetics and diabetics in their responses to whey. This, however, might be due to the semi-quantitative and separate nature of the analyses of the two sample sets, not allowing direct comparison of the "true" plasma levels.

At present, the major bottleneck in metabolomics studies of this kind is compound identification. Therefore this thesis will also present and discuss state-of-the-art tools for computer-assisted compound identification, including: annotation of adducts and fragments, determination of the molecular ion, *in silico* fragmentation, retention time mapping between analytical systems and *de novo* retention time prediction. A pipeline combining these tools in a single workflow is described, and the potential impact in the field of metabolomics highlighted. These tools were applied in the reported metabolomics studies.

RESUMÉ (ABSTRACT IN DANISH)

Klassiske humane interventionsstudier bliver sædvanligvis udført for at bevise en hypotese eller for at svare på et specifikt spørgsmål. Derfor er studiedesignet og de udførte analyser baseret på nogle få parametre relateret til de specifikke effekter, der ønskes undersøgt, mens en stor mængde information indeholdt i de indsamlede prøver ikke engang bliver målt.

I de seneste år har en anden metode, kaldet *metabolomics*, vist sig at være en særdeles fordelagtig teknik i ernæringsforskning. I metabolomics-tilgangen bliver interventionsstudiet udført uden at predefinere, hvilke variable der skal måles – derimod er målet at opsamle så meget information som muligt på en ikke selektiv måde. I denne ph.d.-afhandling blev væskekromatografi-massespektrometri (LC-MS) metabolomics-tilgangen anvendt i forbindelse med humane ernærings-interventionsstudier.

Plasma- og urinprøverne, opsamlet i forbindelse med tre humane ernærings-interventionsstudier, blev analyseret med metabolomics-tilgangen med det formål at finde biomarkører for mælkeproteinet valle og for at undersøge påvirkningen af valle på det menneskelige metabolom. Det var yderligere et mål at udarbejde computerassisterede metoder til at rationalisere processen ledende til identifikation af stoffer.

I det første studie blev metabolomics-profiler efter indtag af et måltid med højt fedtindhold, der indeholdt torsk, gluten, kasein eller valle som proteinkilde, sammenlignet, mens valle i det andet studie blev sammenlignet med underinddelinger af valle (α-laktalbumin eller caseinoglycomacropeptid) og hydrolyseret valle. Begge studier blev udført med overvægtige ikke-diabetikere og det sidste studie blev derudover gentaget med diabetikere.

Vi viser, at indtag af valle forårsager en lavere ventrikeltømningshastighed sammenlignet med andre proteinkilder. Dette står i kontrast til tidligere undersøgelser, som viser, at valle tømmes hurtigere fra ventriklen end andre proteiner.

Paradoksalt i forhold til ovenstående så finder vi disproportionalt høje niveauer og kortere T_{max} af nogle aromatiske og forgrenede aminosyrer efter indtag af valle-måltidet. Dette indikerer, at absorptionen af aminosyrer påvirkes uafhængigt af, eller i det mindste ikke fuldstændigt kontrolleres af ventrikeltømningshastigheden.

Derudover finder vi, at valle nedsætter niveauet af et antal fedtsyrer på grund af øget insulinniveau, hvilket sandsynligvis er induceret af de forhøjede aminosyreniveauer.

Vi fandt ingen forskel mellem underinddelingerne af valle, men hydrolysatet indeholdt en række cykliske dipeptider, der måske er årsagen til den hypoglykæmiske effekt observeret for hydrolysatet. Tillige fandt vi, at produktionsprocessen for hydrolysatet førte til dannelse af oxidationsprodukter af methionin, som metaboliseres endogent til tidligere ukendte metabolitter.

Det lykkedes ikke at finde eksponeringsmarkører med høj specificitet for valle, da effekterne var begrænset til at ændre niveauerne af endogene metabolitter. Valle-hydrolysatet derimod indeholdt usædvanlige cykliske dipeptider; det er dog ikke sandsynligt, at de er valle-specifikke, men snarere et udtryk for hydrolyseringsprocessen. Disse resultater kunne ikke være opnået ved de traditionelle hypotese-testende tilgange.

Vi fandt desuden en række markører for torskemåltidet. De fleste er sandsynligvis også markører for kødindtag, men arsenobetain kan være en specifik markør for fiskeindtag.

Vi fandt ingen forskel mellem overvægtige ikke-diabetikere og diabetikere. Det kan dog skyldes, at analysen ikke blev udført på en måde, der tillod direkte sammenligning af plasmaniveauer.

På nuværende tidspunkt er identifikation af stofferne den største flaskehals i metabolomics studier af denne type. Derfor vil denne afhandling præsentere og diskutere nogle af de nyeste værktøjer til computerassisteret identifikation af stoffer, herunder: annotering af fragmenter, bestemmelse af molekylær-ionen, *in silico* fragmentering, retentionstids-overførsel mellem analytiske systemer og *de novo* retentionstidsforudsigelse. Et sammenhængende system, der kombinerer disse værktøjer, bliver beskrevet, og den potentielle indflydelse på metabolomics-feltet understreges. Dette system blev brugt i de rapporterede studier.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AA	Amino acid
ALPH	α-lactalbumin test meal
APAP	Acetyl-*p*-aminophenol, also known as paracetamol or acetaminophen
AUC	Area under the curve
BCAA	Branched-chain amino acid
BMI	Body mass index
CAS	Casein test meal
CGMP	Caseinoglycomacropeptide test meal
CoA	Coenzyme A
COD	Cod test meal
DKP	2,5-diketopiperazines
EDTA	Ethylenediaminetetraacetic acid
ESI	Electrospray ionization
GC	Gas chromatography
GIP	Gastric inhibitory polypeptide
GLU	Gluten test meal
GPAT	Glutamine *N*-phenylacetyltransferase
HDL	High-density lipoprotein
iAUC	Incremental area under the curve
LC	Liquid chromatography
LC-MS	Liquid chromatography coupled to mass spectrometry
LDL	Low-density lipoprotein
LOESS	Locally weighted scatterplot smoothing
LogD	Logarithm to the distribution coefficient
LogP	Logarithm to the partition coefficient
m/z	Mass to charge ratio
MCS	Maximum common substructure
MetSO	Methionine sulfoxide
NIPALS	Non-linear iterative partial least squares
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MSR	(Peptide) methionine sulfoxide reductase
NMR	Nuclear magnetic resonance
PAG	*N*-phenylacetyl-glutamine
PAM	*N*-phenylacetyl-methionine
PAMSO	*N*-phenylacetyl-methionine sulfoxide
Q-TOF	Quadrupole-time-of-flight
RT	Retention time
SEM	Standard error of the mean
TIC	Total ion chromatogram
WH	Whey hydrolysate test meal
WHO	World Health Organization
WI	Whey isolate test meal

C h a p t e r 1

INTRODUCTION

In this chapter the context in which this thesis is placed will be presented together with its aim and outline.

1.1 BACKGROUND

During the last one hundred years the populations in most regions of the world have undergone dramatic changes in terms of wealth, health [1] and lifestyle. In the Western World this has led to societies of affluence. While this is arguably a positive development a strong increase in the prevalence of obesity [2], metabolic syndrome and diabetes [3] have become a consequence of availability of an unrestricted amount of food and a more sedentary lifestyle.

Obesity and related risk factors are today some of the leading causes of preventable death [4] and thus of immense humanitarian and economic importance. The development of obesity is complex with serious social and psychological dimensions affecting virtually all ages and socioeconomic groups [5] and is not caused by a high-fat diet and low exercise level alone. In fact it has not been proven in any convincing way that there is a strict dose-response relationship between intake of calories and body fat; the quality of the fat plays an important role as well [6]. Genetic disposition, pre- and postnatal factors as well as intake of nutrients, independent of their caloric value, may also play a role. The question is which.

In the quest to answer this question the qualitative and quantitative intake of protein is of special interest since it has been shown that it is possible to achieve greater weight loss on a low fat/high protein diet compared to a low fat/high carbohydrate diet and that protein has a higher calorie per calorie satiety power [7]. The mechanism is so far unknown but possible causes include inhibition of energy intake due to release of gut peptides, liver metabolism and/or direct effects of certain amino acids [7].

Furthermore, it has been demonstrated that specifically *whey* proteins have certain biological properties that might be beneficial in the treatment and prevention of the metabolic syndrome related to obesity and diabetes. Whey proteins are derived from milk where it constitutes approximately 20 % of total proteins in bovine milk. The remaining milk proteins are caseins. Casein is the protein precipitated when milk is curdled by rennet (or by acidification) and is the main constituent of cheese. Conversely, whey is the proteins that remain in solution and is considered a waste product of cheese production. Whey proteins are therefore inexpensive, but also of high nutritional value and serves as an excellent additive to improve food formulations. These consideration, together with the possibly health promoting effects of whey protein, have sparked intense research into the effects and biochemical actions of whey protein. In fact several hundred journal papers describing research related to whey have been published during the timespan of the Ph.D. studies described in this thesis.

Whey proteins improve fasting lipids and insulin levels in overweight and obese individuals following a period of whey supplementation [8] and reduces short-term food intake compared to other protein sources [9] and thus might be valuable in the prevention and treatment of the metabolic syndrome.

While fasting glucose and lipid levels are important indicators of homeostasis, the deleterious effects of the metabolic syndrome are perhaps even more strongly associated with postprandial hyperglycemia and hyperlipaemia.

Holmer-Jensen *et al.* recently demonstrated that whey also has beneficial effects immediately following intake of a high fat meal. Whey caused lower postprandial lipemia (plasma triglycerides and free fatty acids), lower blood glucose and higher insulin levels compared to supplementation with cod and gluten [10].

While whole whey has been extensively studied, the role of individual whey proteins and peptides has not been thoroughly explored. Bovine-derived whey protein consists of approximately 50-55 % β-lactoglobulin, 20-25 % α-lactalbumin, 10-15 % caseinoglycomacropeptide, 10-15 % immunoglobulins, 5-10 % albumin in addition to small amounts of lactoferrin and lactoperoxidase [11]. Caseinoglycomacropeptide is a hydrophilic glycopeptide released from κ-casein during cheese production (using rennin) [12].

To investigate whether the activity of whey protein could be attributed to one of these subfractions Holmer-Jensen *et al.* conducted a new meal study similar to the one above but comparing whey isolate to products with enhanced proportions of α-lactalbumin and caseinoglycomacropeptide, respectively. In addition, a whey hydrolysate product was included in the study. This study did not show any difference between whey isolate or any of the two subfractions, however, the whey hydrolysate induced a larger incremental area under the plasma insulin curve (iAUC) at 30 min.

In this thesis I report on the results from the analysis of plasma samples collected in conjunction with the two studies by Holmer-Jensen *et al.* described above using a *metabolomics* approach.

1.2 NUTRITIONAL METABOLOMICS

The traditional approach to investigating the effect of different nutrients, i.e. intervention studies, is both time-consuming and expensive. The conclusions you can draw from such studies are also inherently obscured by methodological limitations. It is for example well-known that weight conscious individuals systematically underreport energy intake [13] and thus reliable information is difficult to obtain through such studies. Alternatively the study design could include stricter control of food intake, however, this is uncomfortable to both subject and researcher in all but the shortest studies and accurate assessment of intake of specific nutrients is still difficult to ascertain.

For these reasons it could be advantageous if surrogate measures, so called biomarkers, for the exposure under investigation could substitute self-reporting of intake.

World Health Organization (WHO) has defined a biomarker [14] as

> *"any measurement reflecting an interaction between a biological system and an environmental agent, which may be chemical, physical or biological".*

In addition WHO distinguish between two classes of biomarkers namely "*biomarkers of exposure*" defined as

> *"an exogenous substance or its metabolite or the product of an interaction between a xenobiotic agent and some target molecule or cell that is measured in a compartment within an organism"*

and *"biomarkers of effect"* defined as

> *"a measurable biochemical, physiological, behavioral or other alteration within an organism that, depending upon the magnitude, can be recognized as associated with an established or possible health impairment or disease".*

In the studies described in this thesis the purpose was two-fold. We sought both to establish biomarkers of intake of different protein sources, especially whey, but also to examine the effect of whey intake on the *metabolome.*

The term *metabolome* was first introduced by Oliver *et al.* in 1998 [15] and is generally taken to mean the collection of all low molecular weight molecules (metabolites) present in biological system in a particular physiological or developmental state [16].

Traditionally the investigation of perturbations to the metabolome of biological systems has been hypothesis-driven. In a hypothesis-driven approach the metabolites (or macromolecules or micronutrients) are predefined. This allows highly accurate targeted analytical methods to be employed. However, the degree to which the results of such studies reflect the biologically most important alterations are constrained by the strength and appropriateness of the *a priori* hypothesis.

In contrast, new developments in analytical instrumentation and computer power have opened the door to investigate changes to the metabolome in a holistic data-driven fashion.

Describing the holistic analysis of changes to the metabolome Nicholson *et al.* introduced the term "*metabonomics*" in 1999 [17] as

> *"the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification"*

while Oliver Fiehn in 2001 [18] defined the term "*metabolomics*" as

> *"A comprehensive and quantitative analysis of all metabolites"*.

The difference between the terms is mostly philosophical and the terms are used interchangeably. In this thesis I will use the more general term *metabolomics,* even though the studies described fall under the sub-set of metabolomics studies designated by the *metabonomics* definition.

Metabolomics is ideally suited to study the biochemical response to a stimulus when no clear hypothesis of the expected response can be formed beforehand, as is the case with the studies described here. The major challenge of metabolomics is the massive amount of data gathered (Chapter 2) when attempting to characterize and quantify the whole metabolome. First the data gathered need to pre-processed into a form appropriate for statistical treatment (Chapter 3). However, the main bottleneck in metabolomics is that initially the molecular structures of the "metabolites" found to characterize a certain stimulus (intervention) response are

unknown. Therefore the structure of the compounds (metabolites) of interest need to be elucidated (Chapter 4) which is a time consuming process.

1.3 THE AIM OF THE THESIS

The aim of the projects described in this thesis was to establish biomarkers of whey intake and to investigate the effects of whey intake on the human metabolome. Because compound identification constitutes the major bottleneck in any metabolomics study it was also the aim to develop a pipeline for computer-assisted compound identification to more rationally achieve the above goals.

1.4 OUTLINE OF THE THESIS

In this thesis I will first introduce the analytical platform and the associated terms and characteristics of the data associated with a liquid-chromatography-mass-spectrometry metabolomics analysis (Chapter 2).

Next, in Chapter 3, data pre-processing and statistical analysis will be described briefly.

In Chapter 4 I will describe state-of-the-art methods for annotation of liquid-chromatography -mass-spectrometry (LC-MS) data and computer-assisted compound identification. Because the description of these methods are scattered between technical papers and software documentations I have compiled a comprehensive summary of the individual steps.

In Chapter 5 my three metabolomics studies on the investigation of the effect of whey protein on the metabolome will be summarized.

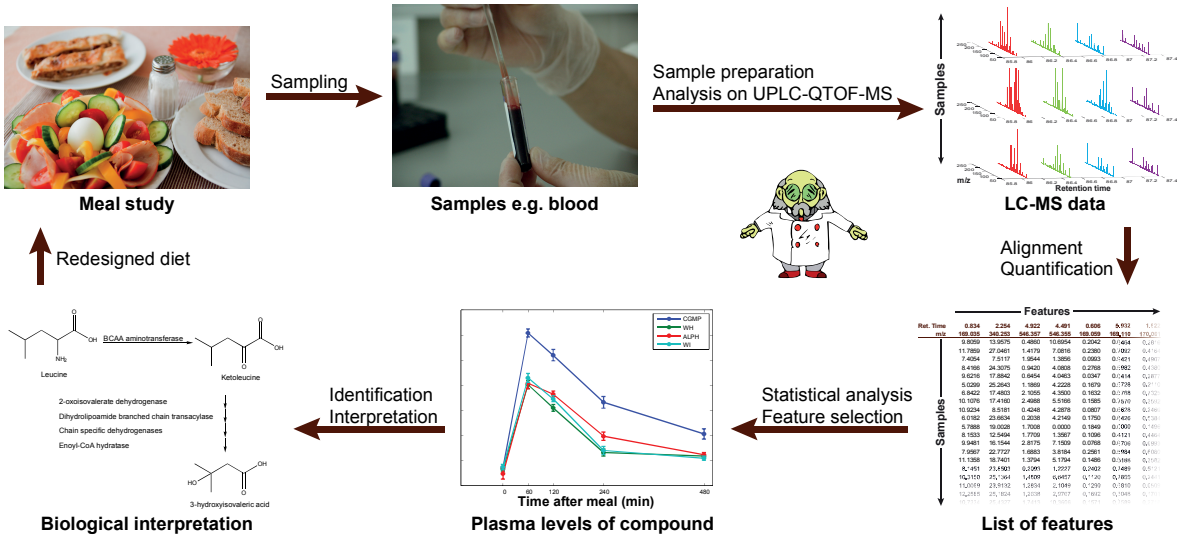
Finally concluding remarks and perspectives are offered in Chapter 6 and 7, respectively.

Chapter 2

ANALYTICAL PLATFORM

In this chapter the metabolomics platform will be introduced. Basic aspects of liquid chromatography and mass spectrometry will be introduced and characteristics important in relation to metabolomics will be highlighted. This chapter serves to introduce and define concepts that will be used in later chapters.

Nutritional metabolomics pipeline



When a metabolomics study is planned the analytical method to be used should be considered carefully. In metabolomics the most used analytical methods are nuclear magnetic resonance (NMR) and mass spectrometry (MS) coupled to separation techniques such as liquid chromatography (LC) or gas chromatography (GC).

In the studies described in this thesis metabolomics using LC-MS methods will be described since LC-MS offers superior sensitivity and selectivity compared to most other methods while also having a relatively high throughput and broad coverage of compounds.

Metabolomics studies can be divided in two main approaches:

metabolomics *fingerprinting* can be defined as an

> *"unbiased, global screening approach to classify samples based on metabolite patterns or 'fingerprints' that change in response to disease, environmental or genetic perturbations with the ultimate goal to identify discriminating metabolites*" [19]

and thus also includes unknowns compounds while metabolomics *profiling* can be defined as the

> *"quantitative analysis of set of metabolites in a selected biochemical pathway or a specific class of compounds. This includes target analysis, the analysis of a very limited number of metabolites, e.g. single analytes as precursors or products of biochemical reactions*" [19].

Metabolomics fingerprinting is used for exploratory studies aiming to investigate the molecular source of a known difference between two study groups or to find new biomarkers of a certain metabolic state (e.g. disease) or a certain exposure (e.g. food or environmental). In metabolomics fingerprinting the analysis should be as unbiased as possible in terms of coverage of compounds.

Metabolomics profiling is used to investigate effects of known biochemical origin or to validate results from fingerprinting studies. In metabolomics profiling the analytical platform can be designed to have optimum sensitivity and selectivity and accurate quantification for the preselected compounds. This is done by optimizing sample preparation such that unwanted compounds are removed so that they do not interfere with the quantification of the compounds of interest. In addition, the analytical parameters can be selected to target only the selected compounds which increase sensitivity and selectivity.

The terms fingerprinting and profiling (and sometimes even semi-targeted) are often used interchangeably but in the following, approaches for and studies utilizing metabolomics fingerprinting as defined above will be described.

2.1 SAMPLE PREPARATION

The first step in analyzing the samples is sample preparation. Most combinations of sample type and analytical platform require some form of sample preparation and sample preparation is a crucial step since the downstream analysis cannot be more accurate than the sample preparation.

The appropriate sample preparation depends on the nature of the sample type (the matrix). Urine samples, for example, often require no other treatment than a simple dilution; though some prefer to remove urea as the large amounts interferes with many analytical platforms.

Plasma and serum samples on the other hand need more careful treatment. Plasma samples can be advantageous compared to serum since samples do not need to wait to be frozen and yield is usually higher and the risk of hemolysis and thrombocytolysis is lower [20].

On the other hand for plasma samples an anti-coagulant need to be added. These can interfere with the analysis of analytes through ion-suppression (see section 2.2.4). A recent analysis by Barri and Dragsted [21] showed that heparin plasma is the most appropriate anti-coagulant for LC-MS metabolomics analysis since citrate and EDTA interfere with the quantification of other compounds.

Typically sample preparation of plasma samples involve removal of proteins, extraction of protein bound molecules and subsequent re-dissolution of the analytes. A reproducible method for the extraction of analytes is necessary to avoid obscuring the results. Two different methods of protein precipitation were recently compared by Barri *et al.* [22], which showed that a filtration method was better in terms of recovery of compounds and reproducibility compared to a centrifugation method.

It is not possible to re-dissolve all compounds equally well following protein removal and hence a completely unbiased sample preparation method does not exist; possibly with the exception of methods that sequentially extract different compound classes using different solvents. However, the resulting samples would then have to be analyzed separately.

In the studies described herein, we have used a method that primarily extracts polar and medium-polar compounds to allow a high sample throughput.

2.2 LC-Q-TOF-MS

There exist a variety of mass instrument all with their inherent pros and cons. A detailed discussion is outside the scope of this thesis and the reader is referred to a recent review of current state-of-the-art spectrometers by Holčapek *et al.* [23]. Quadrupole-time-of-flight (Q-TOF) instruments are commonly used in metabolomics since they offer high mass accuracy and fast sampling (many sample points over each chromatographic peak such that the area under the peak can be better quantified) and therefore this is the type of instrument that will be discussed in the following and the instrument used the studies described.

2.2.1 LIQUID CHROMATOGRAPHY

LC is a method of separating compounds according to their polarity (lipophilic/hydrophilic characteristics). The analytes are injected into a flow of solvents and carried to a chromatographic column; in the studies described in this thesis a so called reversed-phase columns is used. A reversed-phase chromatographic column consists of an apolar stationary phase typically made of straight chain alkyl groups bound to silica particles. How long it takes for the analytes to *elute* from the column is determined by hydrophobic interactions with the column material. This has the effect of separating the analytes such that polar (hydrophilic) compounds elute faster compared to highly apolar (lipophilic) analytes that are more *retained* on the column. See Figure 2.1 for a schematic representation of a chromatographic separation.

In our setup we have used a relatively short run-time for the analysis. Many compounds are therefore not well separated which puts high demands on the mass analyzer as well as on the subsequent analysis of the data. Some methods for dealing with this problem are described in Chapter 4.

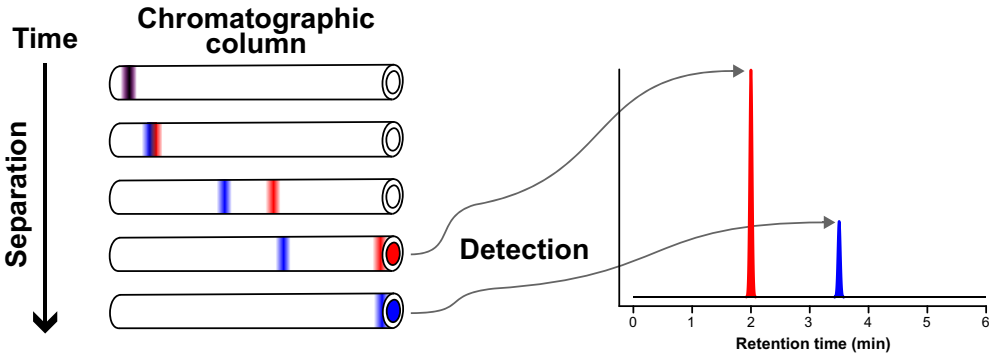


Figure 2.1. Schematic representation of the principle of liquid chromatography. The two analytes (red and blue) are separated as they pass through a chromatographic column because they bind with different strength to the stationary chromatographic material. The compounds are then passed to a detector resulting in a chromatogram where the area under the peaks is proportional to the concentration of the analyte.

2.2.2 MASS SPECTROMETRY

Following elution from the chromatographic column the flow of solvent is dispersed into an aerosol of charged (ionized) droplets (in electrospray, ESI). The analytes are ionized by the addition or removal of a proton (H^+) from the analyte. The instrument can be operated in both positive and negative ionization mode. This is advantageous since not all compounds can carry a positive or negative charge. Compounds that cannot carry a charge (are not ionizable) cannot be analyzed by this method.

The droplets become increasingly smaller as the solvent evaporates. The small droplets of charged analytes are then sucked into the first stage of the mass spectrometer under high vacuum (see Figure 2.2).

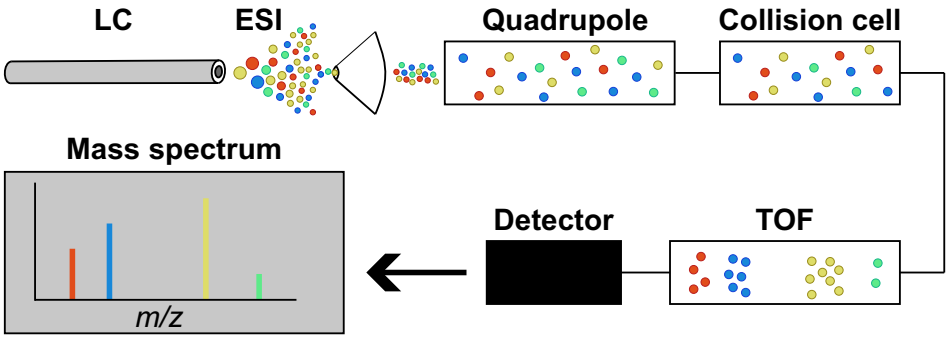


Figure 2.2. Simplified schematic representation of the stages comprised in a mass spectrometer. LC, liquid chromatography, ESI, electrospray ionization, TOF, time-of-flight, *m/z*, mass to charge ratio. Figure after The University of Leeds (http://www.astbury.leeds.ac.uk/facil/MStut/mstutorial.htm) and The University of Bristol (http://www.chm.bris.ac.uk/ms/theory/esi-ionisation.html**).**

The quadrupole of the spectrometer is not used in the standard analysis of samples and functions only to focus the ion beam into the next stages of the instrument. The collision cell can be used to fragment the molecules into smaller pieces, which is advantageous for structure elucidation (discussed below); however in the initial profiling experiments it is advantageous to avoid fragmentation as it simplifies data analysis. Therefore the *collision energy* in the collision cell is set low to limit fragmentation. Next, the molecules reach the time-of-flight (TOF) compartment of the instrument. In the TOF compartment the molecules are passed through a magnetic field. The ions with the lowest mass to charge ratio (*m/z*) will pass through the TOF the fastest and it is thus possible to determine the *m/z* of the analytes.

Typically small molecules only carry a single charge (for positive mode ionization written as $[M+H]^+$), such that the mass of the analyte can be calculated directly from the observed *m/z*.

A mass spectrum is constructed by scanning a selected mass interval (typically 50-1000 Da in metabolomics) several times per second (Figure 2.3C). A total ion chromatogram (TIC) is constructed by summing the intensities of all the *m/z* measurements in each scan and plotting the sum *versus* the retention time (Figure 2.3A). An extracted ion chromatogram (EIC) is constructed by plotting the intensity of a specific *m/z versus* retention time.

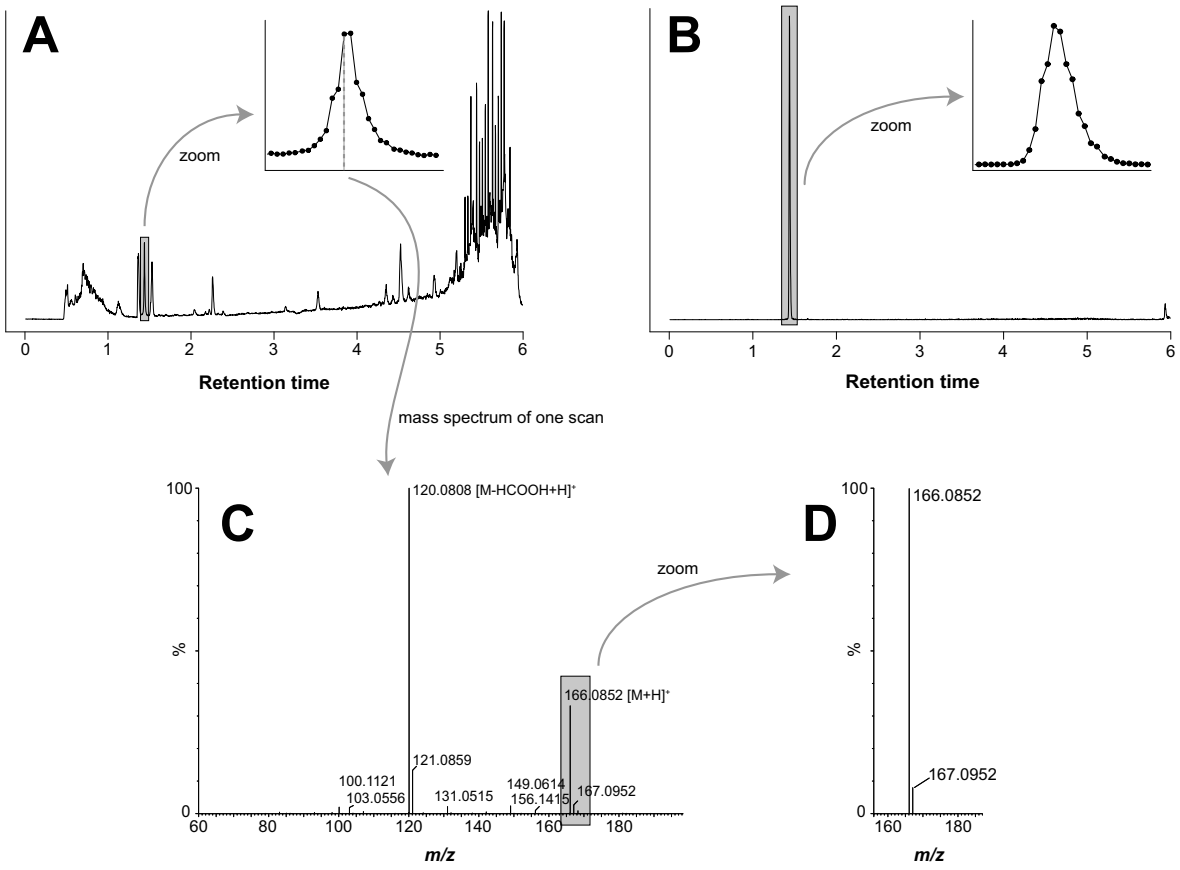


Figure 2.3. A) Total ion chromatogram (TIC), B) extracted ion chromatogram of $[M+H]^+$ species of phenylalanine, C) Mass spectrum of phenylalanine, D) zoom of the $[M+H]^+$ isotope peaks.

Several features of the mass spectrum will be used in the following discussion and are therefore explain briefly.

In the mass spectrum of phenylalanine (Figure 2.3C) we can see that in addition to the charged phenylalanine species $[M+H]^+$ a number of lower mass peaks appear. These are caused by phenylalanine breaking up into *fragments*. Simple fragmentation involves breaking of covalent bonds in the molecule to release *neutral losses* of a piece of the molecule. The remaining part of the compound (the fragment) is still charged and therefore detected by the spectrometer.

Another feature is that all peaks have an associated peak with *m/z* one mass unit higher (when the sensitivity of the instrument allows observing this) as seen in Figure 2.3D. These peaks are due to *isotopes*. Most elements can exists with a different number of neutrons in the nucleus; that is several isotopes exist. Taking carbon as an example the most common carbon isotope is ^{12}C. However, about 1 % of naturally occurring carbon exists as ^{13}C which has an extra neutron in the nucleus. A smaller fraction of molecules will also exists where there are two ^{13}C atoms present and an *m/z* two units higher will also be present in the mass spectrum. Theoretically there will exist molecules with all combinations of stables isotopes. The relative intensities of all the resulting mass peaks are called the *isotopic pattern* of the molecule (or rather of the molecule's molecular sum formula). However, due to the low abundance of most heavier isotopes, often only the first one or two higher mass isotope peaks are observed; though exceptions exists where more prevalent heavier isotopes exist. If we only consider carbon then we can appreciate that the more carbons are present in a molecule the higher the chance that one or more of the carbons will be a ^{13}C atom. Therefore the larger the compound the more intense the higher mass peaks will be. However, in metabolomics we analyze only small compounds and the second mass peak will typically have an intensity about 10 % of the peak from the lowest and most abundant isotope, ^{12}C.

In the following I will refer to all but the lowest mass peak in the isotopic pattern of a molecule as the *isotope peaks*.

2.2.3 TANDEM MASS SPECTROMETRY

As mentioned above, molecules tend to form fragments in the spectrometer. In the profiling of the samples this is an undesirable effect as it complicates the data analysis and obscures which peaks correspond to the complete molecule, referred to as the *pseudo-molecular peak*.

However, for the purpose of elucidating the structure of the compounds, fragmentation is a useful phenomenon. Therefore compounds of interest, where the structure is unknown, are selectively fragmented in so called MS/MS experiments.

MS/MS experiments function by using the quadrupole to filter out all compounds with the exception of the compound of interest. Then the analyte passes to the collision cell where the analyte is allowed to collide with neutral molecules (such as argon) at a higher collision energy than is used for the profiling experiments. The collision causes some of the kinetic energy to be converted to internal energy which causes covalent bonds to break and fragments are created by releasing neutral losses. The mass of the neutral loss and the fragment gives information about the structure of part of the molecule and the combination of all fragments and neutral losses can thus ideally be used to deduce the structure of the compound. Usually the fragmentation will not provide sufficient information to assign a single unique structure to the unknown molecule. A standard compound of known identity will then have to be analyzed under identical conditions to compare the mass spectra and confirm the structure.

2.2.4 LIMITATIONS

Though Q-TOF-based metabolomics is a very versatile method for untargeted metabolomics it has some notable shortcomings.

First of all, the sample preparation determines which compounds are injected into the LC-MS system and as mentioned a truly unbiased extraction is not possible.

Second the chromatographic method is not efficient at separating compounds of any polarity. In reversed-phase chromatography no separation of very polar compounds are achieved and their quantification if therefore not possible. To analyze very polar compounds a normal-phase or HILIC type system can be used instead.

Following separation the compounds need to be ionized to enter the mass spectrometer and hence only ionizable compounds can be detected. In addition compounds compete for the charge in the ESI source and compounds more able to sustain a charge are therefore favored. Co-eluting compounds with a high ability to sustain a charge can therefore lead to decreased ionization of other compounds and therefore suppress their signal. This phenomenon is known as *ion-suppression*, and it is one of the main problems in using an ESI source for quantification. Ion-suppression can lead to a lower signal of an analyte based on the matrix in which it is analyzed. The result is that the quantification has an inherent analytic uncertainty and Q-TOF-based metabolomics is therefore described as semi-quantitative.

In targeted analyses the problems caused by differential ionization are overcome by adding a known amount of an isotopically labeled internal standard for each compound included in the analysis. The analytes can then be quantified relative to the internal standards that experience the same ionization conditions.

Chapter 3

DATA ANALYSIS

In this chapter data pre-processing and data treatment are introduced. Focus will be on describing the data processing needed to make the data suitable for statistical analysis. Next, different approaches to the statistical analysis is introduced with focus on issues specific to metabolomics-type data.

3.1 PRE-PROCESSING

Data pre-processing encompasses the transformations performed to the raw data prior to statistical analysis. Due to the three-dimensional (*m/z*, retention time and intensity dimensions) nature of LC-MS data, statistical analysis cannot be performed directly and pre-processing is therefore an integral step of analyzing LC-MS data.

The first step in pre-processing is so called *peak-picking*. Peak-picking procedures attempt to locate all chromatographic peaks for each *m/z* value. The areas under these peaks are then integrated and are proportional to the concentration of the analyte. Each of these integrated peaks are thus defined by their *m/z* value and their retention time and termed *features* or *mass features*. The output of peak-picking is therefore a samples × features table appropriate for statistical analysis (see Figure 3.1). Typically the peak-picking results in several thousand features. These features can be considered variables that describe each sample.

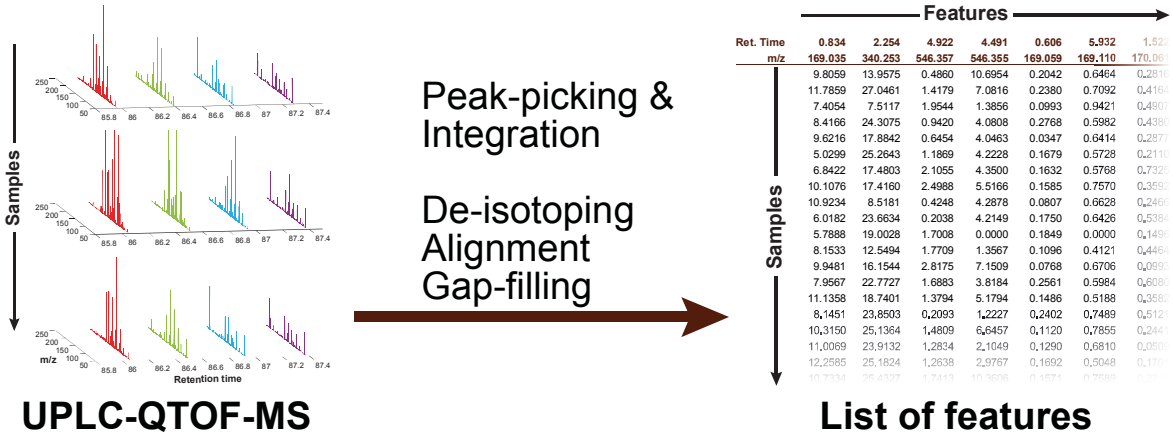


Figure 3.1. Section of the input (left) and output (right) data for pre-processing.

In the simplest form peak-picking could be achieved by dividing the mass range into intervals (bins), constructing the extracted ion chromatogram (EIC) for each bin and integrating the EIC (though the chromatogram needs to be split in the case of several chromatographic peaks in the same EIC). It is necessary to bin the *m/z* values since the experimental determination of *m/z* inherently are subject to small experimental variation. This could also be seen as the spectrometers having finite resolution in the *m/z dimension.* The experimental variation is usually below 10 ppm for Q-TOF equipment or even lower for current state-of-the-art equipment.

A simple approach, as the one sketched above, would, however, be prone to detecting noise and would not deal well with co-eluting peaks. Several different programs exists that offer more appropriate pre-processing of LC-MS data. The most used are open source programs such as XCMS [24], MZmine [25] or proprietary vendor specific software such as MarkerLynx by Waters [26]. Available software has been reviewed by Castillo *et al.* [27] and the result of peak-picking have been compared by Gürdeniz *et al*. [28,29]. The peak-picking algorithms employed by these programs are different and upwards of 40 % of detected features are algorithm-specific depending on the optimization of the user-definable parameters [29,30].

For the analysis of the studies I have performed pre-processing in both XCMS and MarkerLynx. The quantification from MarkerLynx was used since it appeared to better detect small peaks. XCMS was used to take advantage of the annotation features discussed in Chapter 4.

The sensitivity of MarkerLynx, however, comes at a price. When the parameters are set to increase sensitivity for small peaks also a large number of features are created that are clearly not part of any well-defined chromatographic peak. Many of these “noise” peaks are caused by contaminants that do not exhibit normal chromatographic behavior. I have developed a procedure for filtering out these contaminant features so that the problem of these “false-positives” can be eliminated (but only those caused by contaminants) while retaining the additional “true-positives” produced by MarkerLynx (see Section 4.7).

Some programs (MarkerLynx, MZmine etc.) also remove features associated with isotope peaks while other programs (XCMS, see Section 4.2) simply annotate the features as isotopes.

Peak-picking is performed separately for each sample. The next important pre-processing step is therefore matching features originating from the same ion between samples. This is referred to as alignment. Some programs simply match features with similar *m/z* and retention time. Other programs, like XCMS, use the matched features to create a retention time correction model to correct for retention time shift between samples; and then use the "corrected retention" times to match features using more strict parameters.

Typically, the last step in pre-processing is *gap-filling*. When the peaks were initially picked for each sample it was necessary to set an intensity threshold to avoid picking up noise. However, that means that the final data matrix contains "zero intensity" values for the samples in which a certain feature was not detected. These zero values are problematic for the statistical treatment and therefore not desirable. Therefore most programs re-examine the samples with zero intensity values and integrate the EIC in the area where the features should occur.

3.2 DATA TREATMENT

Prior to statistical analysis it is often advantageous to correct for systematic variation between samples that are not a consequence of the intervention you wish to study.

3.2.1 SAMPLE NORMALIZATION

This variation could for example arise from a systematic overall difference in concentration between samples. For plasma samples this is rarely an issue since homeostasis controls plasma analyte concentrations tightly.

For urine samples, however, the situation is very different. The volume of urine, and thereby the concentration of analytes, can vary considerably depending on water consumption and other physiological factors [20]. Commonly it is attempted to correct for this dilution effect by normalization using the urine volume, creatinine concentration, osmolality, protein/creatinine or simply the sum of the total ion chromatogram (TIC) as a measure of overall concentration [20]. The suitability of these methods have, however, not been thoroughly validated. All of these methods attempt to establish a dilution factor for each sample but it is unclear if it is appropriate to use an overall dilution factor for all analytes. It might not be appropriate due to concentration-dependent ion-suppression and the fact that some of the methods are dominated by high abundance species that might not be a good measure of overall concentration.

We have therefore opted not to use any sample-wise normalization apart from the inter-person variation correction implicit in the univariate analysis described below.

3.2.2 CORRECTION FOR SENSITIVITY CHANGES

Systematic variation can also arise from the analytical procedure. In an LC-MS platform the major source of analytical variation is the ion source. Ionization efficiency is modified primarily by buildup of analyte and analyte matrix, in addition to contaminants from other parts of the platform (for example the solvents used for liquid chromatography (LC)), causing ion-suppression. This causes differences in instrument sensitivity between analytical batches, but can also cause a sensitivity drift during a batch.

One approach to correcting for this could be to use the sum of the intensities (of all features) for each sample as a measure of general instrument sensitivity; this would be called normalization to the sum and should be equivalent of using the sum of the TIC, though less influenced by noise than using the TIC. This approach, however, is dominated by the major peaks. And in addition any real effect involving a large number of analytes would be attenuated and obscured by this procedure. As above, this procedure is also not able to account for matrix/analyte-specific ion-suppression.

Another approach would be to correct batch-wise under the assumptions that, since the samples have been randomized, there should be no global difference between batches. In Figure 3.2 I have shown the effect of using the batch-wise mean of the sum of the intensities. This approach, however, does not account for intra-batch drift.

An approach adding an intra-batch regression has therefore been proposed among others by Wang *et al.* [31], who also reviewed other normalization methods.

I implemented a similar correction using a standard common sample analyzed several times per batch to assess sensitivity. In this way the matrix would also be the same. The effect on the sum of intensities can be seen in Figure 3.2.

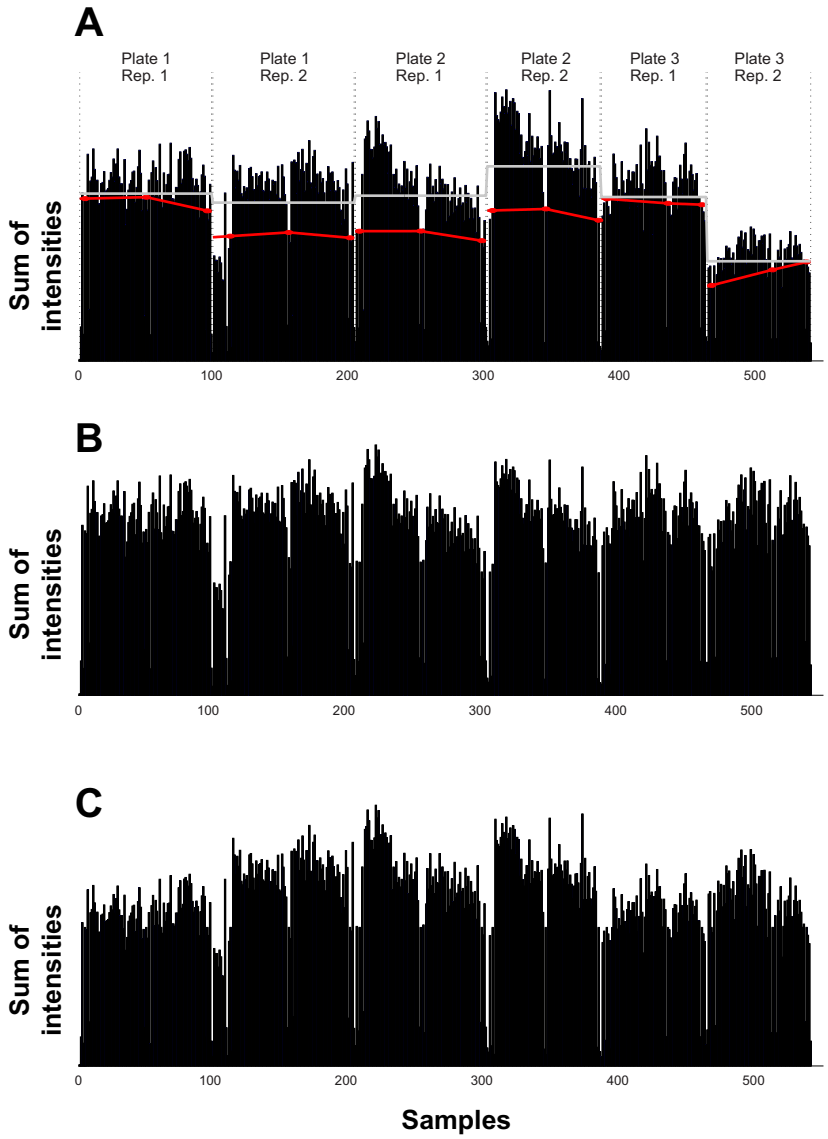


Figure 3.2. Sum of intensities before any correction (A), after correction to the sum of intensities using the average for each batch (B) and after correction using regression across each batch based on the sum of intensities from standard mixtures (C). In gray are the batch average and in red a standard mixture of compounds.

The sum of intensities is of course hardly a good measure of the appropriateness of the corrections. I therefore performed a principal component analysis (PCA, explained briefly in section 3.3) to assess the effect. As can be seen in Figure 3.3A-C none of the normalization strategies are able to correct for the large difference between batches.

As hinted to before this is because ion-suppression (and ion-enhancing) is not uniform between analytes. This is also the reason why even using a selection of internal standards is

insufficient to correct for batch effects across all analytes assessed in a metabolomics analysis as have been noted by others [32].

Therefore, instead of assuming that the sum of all intensities is constant between batches, I use a batch correction that assumes that the sum of intensities for each *individual* feature is constant between batches. In other words the intensities are corrected *feature-wise* to have the same sum for each batch. The effect of this correction is seen in Figure 3.3D; the batch-effect has been removed and the relevant effects can now be studied.

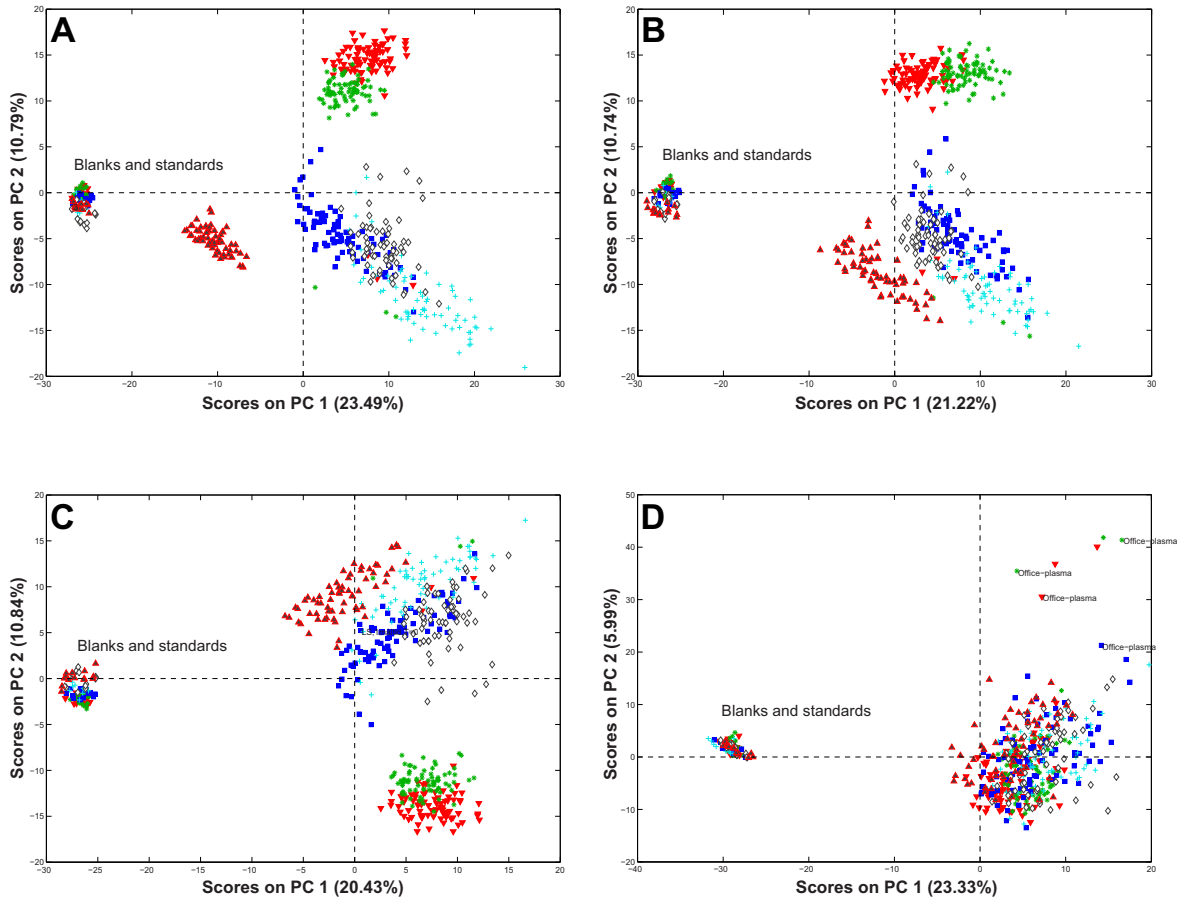


Figure 3.3. Score plot from the PCA of an example study (study discussed in section 5.1.1). Uncorrected data (A), after correction to the sum of intensities using the average for each batch (B), after correction using regression across each batch based on the sum of intensities from standard mixtures (C) and after features-wise batch normalization (D). All datasets have been auto-scaled and the 80 % rule applied (see below). Colored by batch.

3.3 MULTIVARIATE ANALYSIS

Due to the multivariate nature of metabolomics data (in the case of LC-MS data each feature represents a variable describing each sample) the data are usually analyzed using multivariate statistics. The description of these methods is outside the scope of this thesis, however, the basic principles will be briefly presented to justify the choice of an alternative analysis strategy.

The most simple multivariate method is Principal Component Analysis (PCA). PCA aims to extract the major patterns in a dataset. This greatly reduces the complexity of the data analysis when the dataset contains a large number of correlated features. The linear combination of correlated features that account for most of the variation in the dataset is extracted into a number of so-called principal components. The result of a PCA analysis can be seen in Figure 3.4. In the scores plot it can be seen that the samples taken at different time points following the test meals can be somewhat discriminated and in the loadings plot it can be seen that mainly a number of amino acids account for the differences between the T = 60 and T = 120 samples and the remaining samples.

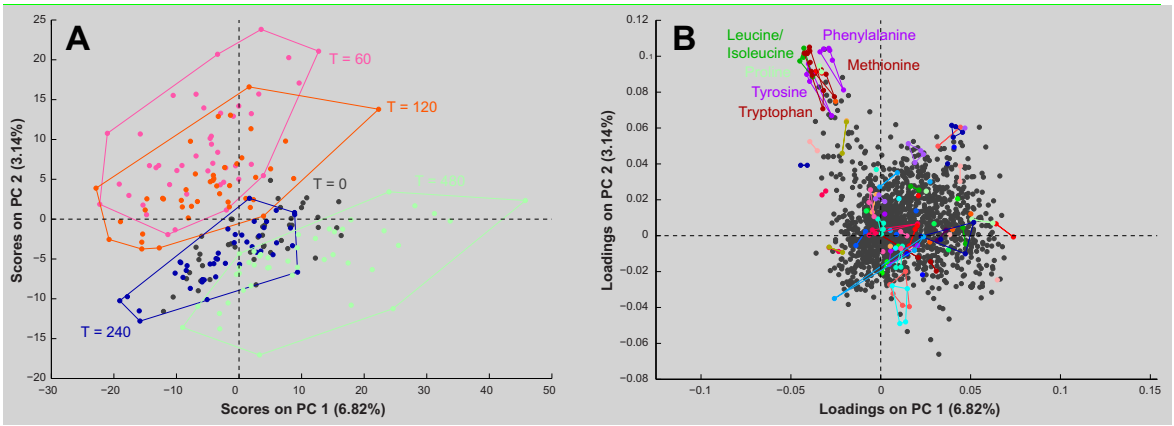


Figure 3.4. Scores (A) and loadings (B) plots from PCA analysis. Study described in section 5.1.1. The number of variables have been reduced using the 80 % rule [33] stating that a feature must be non-zero in at least 80 % of samples within a sample group (time, meal combination). Missing data imputed using the NIPALS algorithm. The data has been autoscaled. Samples are colored by time point and groups outlined while variables are colored such that features origination from the same mother ion have the same color (this grouping is explained in section 4.1).

For this study, however, we were mostly interested in the differences between the test meals and less so in the difference over time. It turned out that PCA was not able to capture the differences between the meals. This is because the differences are too subtle and not described by a few major homogeneous trends [34].

A method is therefore needed that focusses on finding features that differ between meals. Examples of such supervised methods include partial least squares discriminant analysis (PLS-DA) and sparse principal component analysis (SPCA) [35], both of which have recently been utilized for metabolomics studies in our group [36,37]. However, these are not appropriate for a study with as many factor combinations as is present in the design of the studies discussed here. For this purpose analysis of variance – simultaneous component analysis (ASCA) [38] have been successfully applied [39] and would also be appropriate for the studies described in this thesis. Multivariate methods designed specifically for studies with a "replicated time course experiment" [40] that increase the power of analysis of such data have also recently been proposed.

3.4 UNIVARIATE ANALYSIS

Instead of using multivariate statistics I decided to use a univariate approach, linear mixed models, since implementations are currently more readily available and it requires less model optimization and validation compared to multivariate approaches. This approach was able to reveal features that differ between the test meals, however, patterns of multiple metabolites that might have been revealed by multivariate analysis are not recovered by this approach.

3.4.1 LINEAR MIXED MODELS

Linear mixed models (LMM) are the natural univariate choice for analyzing metabolomics data. In contrast to other familiar methods such as Student's t-test and ANOVA they are better at handling missing data, which is common in metabolomics, and at handling correlated observations (such as a repeated measures design where observations are not independent between time points for the same subject). A detailed description of LMMs are beyond the scope of this thesis, however, the model used will be presented briefly, followed by a discussion of issues specific to the analysis of metabolomics data.

Since it is the effect of the test meals we are interested in examining, two models are built; one model that takes into account the different meals (A) while the other does not (B):

$$Model_A: \quad response = M \cdot T + D + G + Z + p + b + \varepsilon$$
$$Model_B: \quad response = \quad T + D + G + Z + p + b + \varepsilon$$

M denotes the test meal group, T denotes the time point, D denotes the experiment day, G denotes gender, p denotes the specific study participant and b denotes the analysis batch. Z is a quantitative explanatory variable with the baseline (T=0) value for the respective meal/time/person combination; baseline values was thus not considered as time points in these models. Upper case letters specify fixed effects and lower case letters specify random effects. P-values for the comparison of these models were then calculated using a likelihood ratio test for each feature. Comparison of the models allows the selection of features where the meal affects the metabolite concentration. Note that we do not include a meal effect but only a meal × time interaction since it is not expected that there is a meal effect not modified by time.

3.4.2 THE FALSE DISCOVERY RATE

One of the drawbacks of using univariate statistics instead of multivariate methods is the issue of multiple testing. If you apply the common criterion of accepting all tests where $p < 0.05$ as significant, a number of false positives equal to 5 % of the total number of features tested is expected. Given the large number of features tested in a metabolomics study this could amount to several hundred features erroneously being accepted as significantly different between treatments (meals in this case). Clearly this is not an acceptable situation when there are typically a hundred or less true positives.

The classical method of compensating for this problem is to control the family-wise error rate (FWER, methods for example by Bonferroni [41], Holm [42], Hochberg [43] and Šidák [44]). Controlling the FWER means ensuring that the chance of making one false discovery (a false positive, a type I error) is less than the chosen level of significance. This in turn is a very conservative approach when testing thousands of features and expecting true positives in the order of a hundred – since the chance of rejecting true positives (false negative) is heavily increased by this approach.

Especially for exploratory studies it is preferable to allow *some* false positives to avoid rejecting many true positives. For this reason the concept of controlling the false discovery *rate* (FDR) instead was formally introduced by Benjamini and Hochberg [45] and later developed further [46–48]. Controlling the FDR means allowing a known fraction (typically 0.05) of the rejected null hypothesis (e.g. the features accepted as different) to be false positives.

The formal methods developed by Benjamini and Hochberg was inspired by the suggestion from Schweder and Spjøtvoll [49] to use the distribution of p-values to estimate the FDR (see example in Figure 3.5). For the true negatives part of the distribution a uniform distribution is expected. We can see that there is an overrepresentation of features in the 0 - 0.05 bracket. This overrepresentation is expected to corresponds to the true positives. In the Schweder and Spjøtvoll paper this was used to calculate a less conservative cut-off to control the FWER such that the new cut-off would be $\frac{\alpha}{false\ positives}$; in this case the new cut off would be ~0.05/600 = 0.000083. This would result in deeming 59 features significantly different, whereas traditional FWER methods would consider only 48 as different.

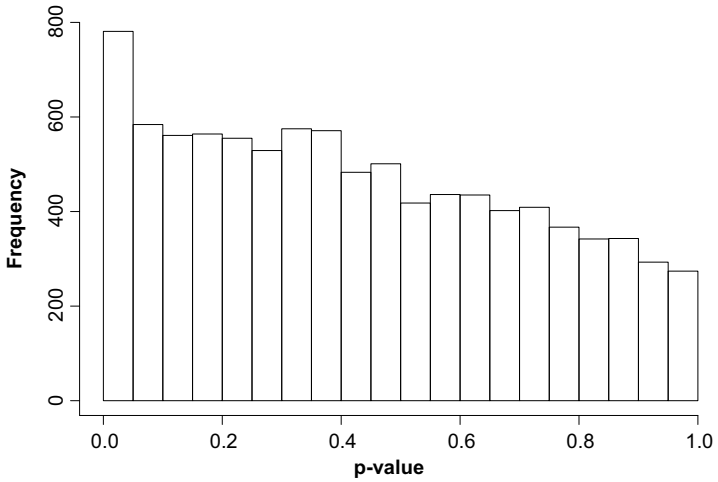


Figure 3.5. Histogram of p-values (N=9523) for the study described in section 5.1.1 (positive mode).

In the original formulation by Benjamini and Hochberg the FDR is controlled by calculating a "corrected" p-value, q, as $q = p \cdot \frac{m}{i}$, where i is the rank of the p-value among p-values from all tests, m. This q-value is then used to set an FDR-based criterion, for example q<0.05; resulting in this case in 85 significant features.

This formulation assumes independence between tests, however, this is not the case for the later developed two-stage Benjamini and Hochberg step-up FDR-controlling procedure [47] used to access the data described in this thesis. Practically most FDR models give identical results with the data used in this thesis. The exception is the Benjamini and Yekutieli [48] method that gives more conservative results, between the FWER methods and the other FDR methods.

IDENTIFICATION

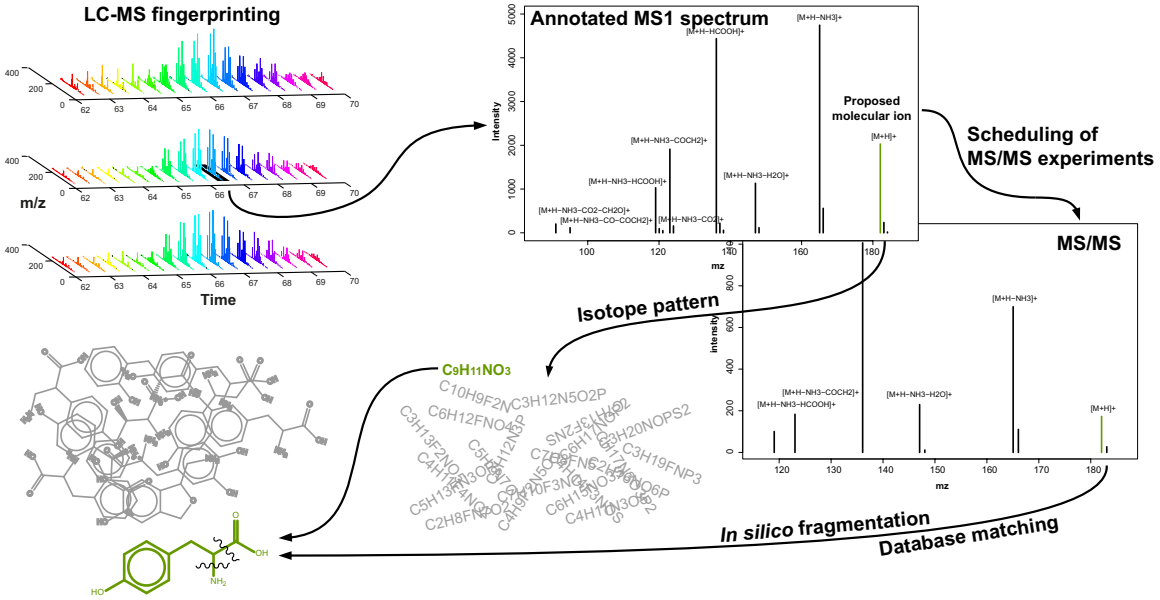


Figure reproduced after Stanstrup *et al.* **[50]** with kind permission from Springer Science and Business Media.

In this chapter a number of computerized methods for compound identification through liquid-chromatography-mass-spectrometry (LC-MS) analysis are presented. Most of the methods discussed are implemented in software packages for the R programming language [51] and build on data processed using the XCMS package [24].

Specific functions, their usage and optimization, are not described in detail but the chapter is designed to give the reader an overview of the available functions and how they can be applied to assist compound identification. For details on implementation the reader is referred to the individual package documentation, paper I and the code examples supplied at msbi.ipb-halle.de/msbi/BeyondProfiling. Further discussions and code examples can be found at the metabolomics forum (metabolomics-forum.com).

In section 4.1 several methods for grouping ion species originating from the same compound are presented and several pitfalls discussed. In section 4.2, this is followed by a presentation of the methods available in the R package CAMERA [52] for annotating fragments and adducts, I show the results of such an annotation and the information that can be retrieved.

In section 4.3 I briefly describe methods for determining the isotopic pattern including an improved method for determining the isotopic ratio using multiple samples. In section 4.4 current state-of-the-art methods for computer-assisted identification based on MS/MS fragmentation is presented. In section 4.5 I describe a method for mapping ("translating") the retention time from one chromatographic system to another, while in section 4.6 a method for predicting retention time based on a compound's structure is presented and the applications to metabolomics studies are discussed. Finally in section 4.7 I describe how we implemented the methods summaries in the preceding sections into a complete semi-automatic pipeline for compound identification.

Unless otherwise stated data from the study described in section 5.1.1 was used to illustrate the methods.

4.1 GROUPING OF RELATED IONS

When attempting to identify features the first step is to understand which ones originate from the same molecule and understanding the relationship between the features, *i.e.* which feature is the pseudo-molecular ion, which are adducts and which are fragments. Features originating from the same molecule will have the same retention time (disregarding small variations caused by the peak picking). Therefore you could initially simply group features that are close in retention time.

However, there are many unrelated co-eluting compounds. To separate those feature that are co-eluting but not related the correlation between features could be investigated. Related features will be highly correlated whereas unrelated features are not correlated.

There are two conceptually different types of correlation that can be assessed:

1) Correlation across samples: If there is a strong signal from of a compound in a sample all features originating from that compound must have proportionally high intensities and conversely for a sample with a low amount of the same molecule. See Figure 4.1A.
2) Correlation across peaks (or inside sample or peak shape correlation): The intensities in the extracted ion chromatograms (EICs) of two related features must be correlated due to co-elution. This correlation analysis is thus a further refinement of the naïve retention time grouping described above. See Figure 4.1B and C.

To perform a correlation analysis in the complete dataset automatically a MATLAB function was developed that utilizes correlation across samples. The method implemented works in the in following way: a matrix of size $N_{feature} \times N_{feature}$ is created where each element indicates if two features are considered related. For two features to be considered related the following requirements need to be met:

1) The features need to be closer in retention time than a user-defined threshold
2) 80 % of samples with non-zero intensities in one of the features must have non-zero intensities for the other feature
3) At least 10 samples in the dataset need to have non-zero values for both features

4) The Pearson correlation coefficient needs to be above a user-defined threshold

Because all features belonging to a group do not necessarily fulfill the requirements in relation to *all* other members of the group the algorithm in the next step fills the "missing" correlations as seen in Figure 4.2. After this fill-in step each *unique* row defines the members of each compound group.

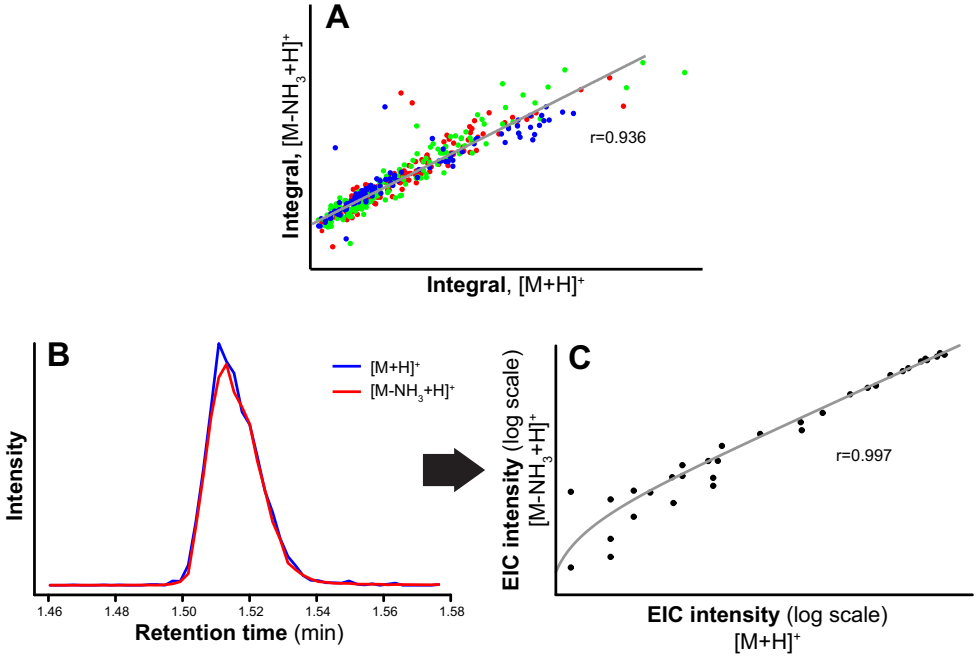


Figure 4.1. A) Peak integral for all samples for the $[M+H]^+$ and $[M-NH_3+H]^+$ feature of tryptophan. In different colors are the different analytical batches.
B) Extracted ion chromatogram of the $[M+H]^+$ (blue) and $[M-NH_3+H]^+$ (red, down-scaled by a factor of 12) feature of tryptophan.
C) Intensities for each scan of the extracted ion chromatogram of a single sample for the $[M+H]^+$ and $[M-NH_3+H]^+$ feature of tryptophan.

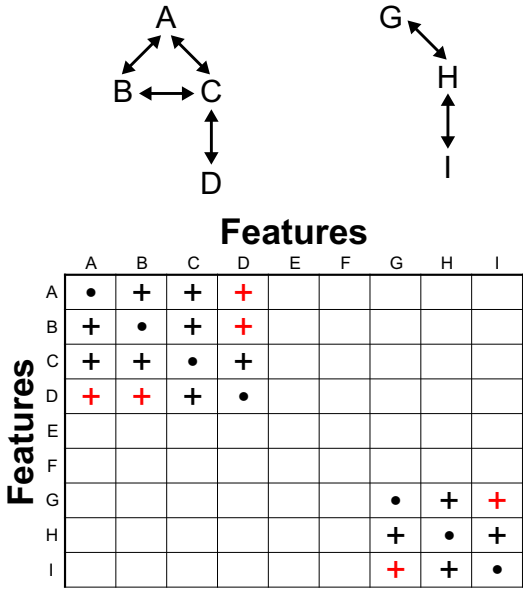


Figure 4.2. Correlation matrix with two features groups. The features A, B, C and D are related and G, H and I are related. Feature pairs marked with "+" fulfill the requirement to be considered related. In red are filled-in "missing" correlations.

Table 4.1. Grouping of features related to tryptophan based on correlation.

		Correlation threshold						
		Custom MATLAB method					CAMERA	
m/z	**Suggested ion**	0.9	0.8	0.7	0.6	0.5	0.7-0.9	0.7, across only
100.1126[1]	*Contaminant*		X	X	X	X		
118.0657	$[M\text{-}NH_3\text{-}CO\text{-}COCH_2+H]^+$		X	X	X	X	X	X
130.0655	$[M\text{-}NH_3\text{-}CO_2\text{-}CH_2+H]^+$					X	X	X
132.0808	$[M\text{-}NH_3\text{-}CO\text{-}CO+H]^+$	X	X	X	X	X	X	X
142.0647	$[M\text{-}NH3\text{-}HCOOH+H]^+$					X	X	X
144.0416[2]	*Unknown*						X	X
144.0814	$[M\text{-}NH_3\text{-}CO_2+H]^+$	X	X	X	X	X	X	X
146.0599	$[M\text{-}NH_3\text{-}COCH_2+H]^+$	X	X	X	X	X	X	X
159.0922	$[M\text{-}HCOOH+H]^+$	X	X	X	X	X	X	X
170.0608	$[M\text{-}NH_3\text{-}H_2O+H]^+$		X	X	X	X	X	X
188.0711	$[M\text{-}NH_3+H]^+$	X	X	X	X	X	X	X
205.0979	$[M+H]^+$	X	X	X	X	X	X	X
245.1301	$[M+(CH_3)_2CO\text{-}H_2O+H]^+$			X	X	X	X	X
276.1823	*Unrelated*						X	
374.1463[1]	*Unrelated*				X	X		
409.1873	$[2M+H]^+$		X	X	X	X	X	X
447.1336	$[2M+K]^+$						X	X
817.3606[1]	$[4M+H]^+$			X	X	X		

Isotope features have been removed for simplicity.
[1]Peak only found by MassLynx and not XCMS.
[2]Peak only found by XCMS and not MassLynx. Correlation across samples to the $[M+H]^+$ is 0.84. The *m/z* values cannot be explained by a simple neutral loss but is possible assuming complex rearrangements.

The method has been applied and the resulting group of features related to tryptophan is listed in Table 4.1 and it is indicated which features are included at which correlation threshold level. A correlation coefficient of 0.7 was chosen as a good compromise between false positives and false negatives. At this level, however, the algorithm did include one feature (*m/z* = 100.1126) that appears to be a false positive (two apparent false positives (*m/z* = 100.1126 and 374.1463) are included if the threshold is lowered to 0.6).

Since any cut-off leads to either false negatives or false positives it can sometimes be more instructive to plot a heat map of correlations of the features with similar retention time as the feature of interest. As can be seen in Figure 4.3 the two false positives are still included in the cluster, but the correlation is evidently based on relatively few features and there are many missing correlations to other features. The heat map can also be used to visualize ion-suppression which will be evident as negative correlations. In the heat map in Figure 4.3A, however, the negative correlations have a poor correlation (see figure Figure 4.3B) and the apparently clusters are a product of the cluster-based ordering of the features.

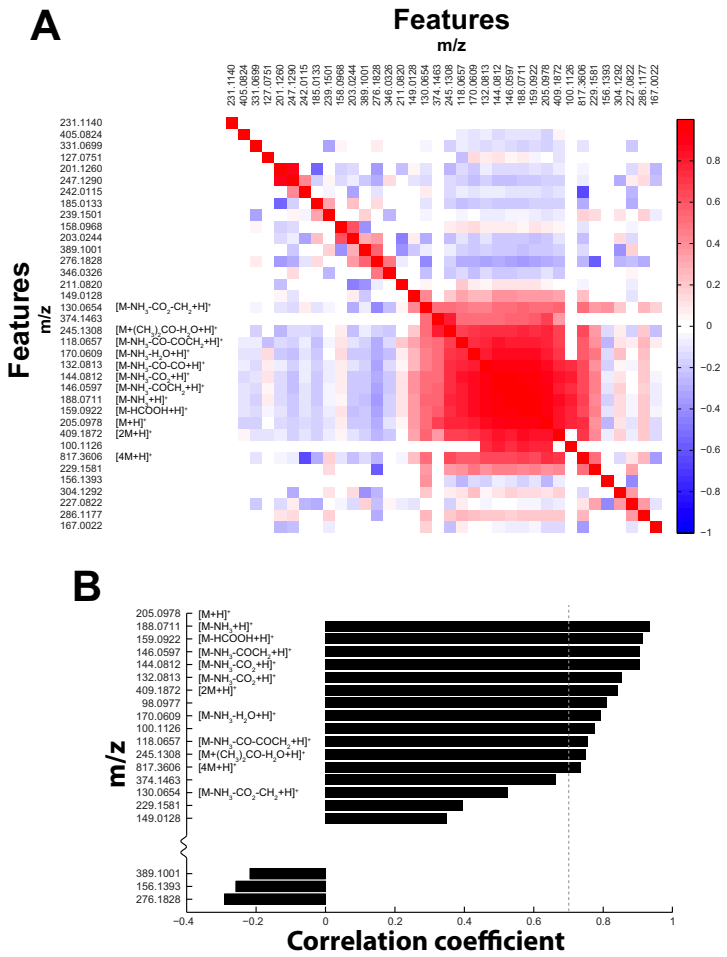


Figure 4.3. A) Heat map of correlations 0.01 min around the retention time of tryptophan. Features reordered by k-nearest neighbor (KNN) clustering using a function from PLS Toolbox **[53]** modified to handle missing values. In white are feature pairs with correlation coefficient of zero or pairs where not enough values were present in both features.
B) Correlations of close elution features to the $[M+H]^+$ feature.

The two false positives are a consequence of MarkerLynx not having gap-filling (see section 3.1). To avoid a correlation calculation based only on very few values the algorithm implemented in MATLAB requires ten samples to have non-zero values for both features that are compared. Taking the *m/z* = 100.1126 feature as an example it does have ten samples with non-zero values. When using these ten samples the correlation coefficient to the $[M+H]^+$ ion is 0.75, probably by chance as shown in Figure 4.4. Using this few samples to calculate the correlation is clearly not optimal and can lead to false positives as was the case for these two features within the tryptophan group.

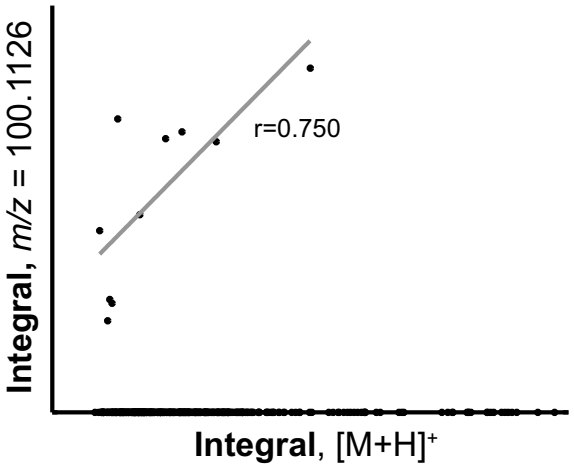


Figure 4.4. Peak integral for all samples for the $[M+H]^+$ feature of tryptophan and the *m/z* = 100.1126 feature.

While the MATLAB function was developed to work on the final peak table generated by any pre-processing software, Kuhl *et al.* have released the CAMERA [52] package for R which among other functions provide correlation analysis. I compared the performance of the MATLAB function to the equivalent CAMERA functions.

CAMERA also offers grouping of features based on correlation across samples. The implementation is, however, rather different. The CAMERA approach in short works in the following way:

1) It starts from the most intense feature and assigns all features within a given retention time interval to a group and continues until all features have been assigned to a group.
2) The Pearson correlation is then calculated between all features in each group.
3) The Pearson correlation is used as the edge weight in a graph based clustering. The feature groups are then refined such that each sub-graph represents the final groups.

In addition to the correlation across samples CAMERA can also include the correlation across peaks and isotope information (see section 4.2) in the edge weight. The correlation across peaks is calculated from the EICs for each feature and a Pearson correlation between the intensities at each scan point. The edge score thus becomes:

$$score(x,y) = CAS_{xy} + \frac{\sum_{i=1}^{N}(CPS_{ixy})}{N} + ISO_{xy}$$

where CAS is the correlation across samples, CPS_i is the correlation across peaks for sample *i* and ISO the binary encoded presence or absence of an isotope relationship.

The option to calculate correlations "inside samples" i.e. across the peaks is very useful especially when dealing with few samples where a correlation across samples cannot be reliably established. However, correlation across peaks can also lead to false positive correlations if unrelated peaks are closely eluting. Take the *m/z* = 276.1823 features as an example. When across peak correlations are included in the score the score is driven up despite the very low correlation across samples and the feature is erroneously included in the group. In Figure 4.5B and C we can observe that the peaks are co-eluting yet Figure 4.5A shows that there is no correlation across samples.

Using only correlations across samples in CAMERA did not lead to any truly related features not being included in the tryptophan group and I therefore suggest using only correlations across samples when sufficient samples are available to reliably calculate correlations. This is especially crucial for early eluting peaks where many compounds are co-eluting and when you use a chromatographic system with a short total run time where peaks are less well separated as was the case in the studies described in this thesis.

Though both the MATLAB and CAMERA functions group most ions correctly in the example above it is clear that the CAMERA function is more robust to noisy, and therefore not perfectly correlated, peaks than the MATLAB function and is not as sensitive to the selected correlation threshold.

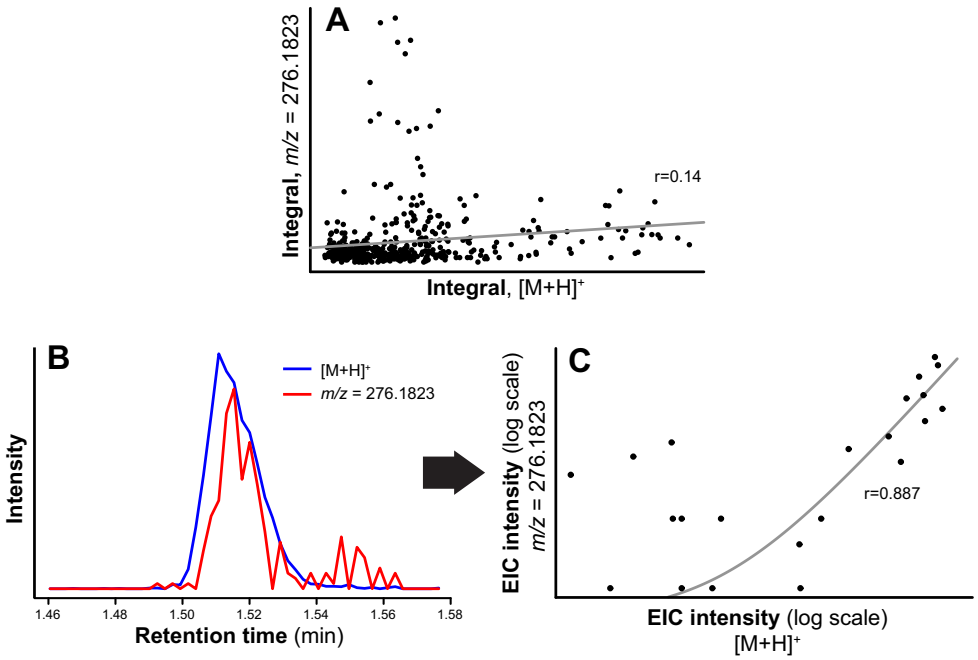


Figure 4.5. A) Peak integral for all samples for the $[M+H]^+$ and *m/z* = 276.1823 feature of tryptophan. In different colors are the different analytical batches.
B) Extracted ion chromatogram of the $[M+H]^+$ (blue, down-scaled by a factor of 15) and *m/z* = 276.1823 (red) feature of tryptophan.
C) Intensities for each scan of the extracted ion chromatogram of the $[M+H]^+$ and *m/z* = 276.1823 feature of tryptophan.

4.2 ASSIGNING THE PSEUDO-MOLECULAR ION, FRAGMENTS AND ADDUCTS

Before attempting to identify a feature it should be established if the feature is the pseudo-molecular ion, an ion from a related isotope, an adduct or a fragment. Annotation of the ion species reduces the number of features to be considered for identification and even provides significant hints towards the structure of the compound. This is most easily and convincingly achieved when the feature is grouped with other features described as above. Once the features have been grouped into feature groups (also referred to as pseudo-spectra) the relationship between the features can be established.

Several commercial programs incorporate automatic assignment of feature relationships (pseudo-molecular/fragment/adduct relationship). In this section the functions and application of the methods available in the free and open source CAMERA package will be described.

First isotope relationships are annotated. CAMERA first checks which feature pairs have *m/z* ratio differences compatible with the difference between isotopes. Since the intensity of the first isotope is primarily determined by the ^{13}C isotope, the minimum and maximum number of carbons in the molecular formula is calculated. Based on this estimation the possible interval of the ratio between the intensity of the monoisotopic peak and the first isotope is compared to the observed ratio. Only isotope clusters matching this ratio interval are annotated.

Next, the adducts and fragments are annotated based on a list of known adducts and neutral losses. The adduct or fragment annotations are assigned to feature pairs when the difference in *m/z* values between the features match the mass of the adduct or neutral loss. With each annotation a hypothesis for the mass of the underlying compound is created. Annotations that presume the same hypothesis constitute an annotation group. For example three features annotated as $[M+H]^+$, $[M+Na]^+$ and $[M-H_2O+H]^+$ belong to the same annotation group. Another annotation group could exist that assumes that the feature annotated as $[M-H_2O+H]^+$ above is actually the $[M+H]^+$ of another compound with other features annotated as fragments of this compound. That means that each feature group can contain several annotation groups. Therefore it is critical to have an extensive list of possible neutral losses such that fragments originating from the same compound are not split between annotation

groups simply because the list of neutral losses does not contain the neutral loss needed to make the connection between the features.

CAMERA provides a default list of common adducts such as Na^+, K^+ and Cl^- adducts and common neutral loses such as H_2O, CO_2 and NH_3 for a total of 50 and 33 possible annotations for positive and negative mode, respectively. Based on my manual interpretation of spectra I have extended the list to 148 and 137 adducts and fragments in positive and negative mode, respectively.

Statistics of the annotation can be found in Table 4.2. It can be seen that more features are annotated in positive mode than in negative mode. This is likely because more fragments are formed in positive mode as indicated by the higher number of feature groups containing more than a single feature. Only in negative mode does my extended list of annotations manage to annotate significantly more features than the short default list. But more importantly the extended list lowers the mean number of annotation groups per feature group such that fever hypotheses for the pseudo-molecular ion exist and the fragments originating from one compound do not split between several annotation groups.

The distribution of annotations found for the dataset discussed in section 5.1.1 is reported in Table 4.3 and Table 4.4 and selected annotations are highlighted in Figure 4.6.

Table 4.2. Statistics of the annotation of the dataset described in section 5.1.1.

Mode	Positive	Negative
Grouping statistics		
# Feature	4232	5096
# Feature groups	721	1429
# Features groups with several features	301 (42 %)	494 (35 %)
# Features in feature groups with several features	3812 (90 %)	4161 (82 %)
Annotation statistics		
# Features annotated as isotopes	813 (21 %[1])	878 (21 %)
Default list		
# Feature with adducts/fragments annotation	1436 (42 %[2])	993 (24 %)
# Feature groups with adduct/fragment annotation	135 (45 %[3])	147 (30 %)
# Annotation groups	801	594
Mean number of annotation groups per feature group with annotations	5.9	4.0
Median number of annotation groups per feature group with annotations	3	2
Extended list		
# Feature with adducts/fragments annotation	1651 (48 %[2])	1681 (40 %)
# Feature groups with adduct/fragment annotation	197 (65 %[3])	257 (52 %)
# Annotation groups	571	713
Mean number of annotation groups per feature group with annotations	2.9	2.8
Median number of annotation groups per feature group with annotations	2	1

[1]Of features in feature groups with several features
[2]Of non-isotope features
[3]Of features groups with several features

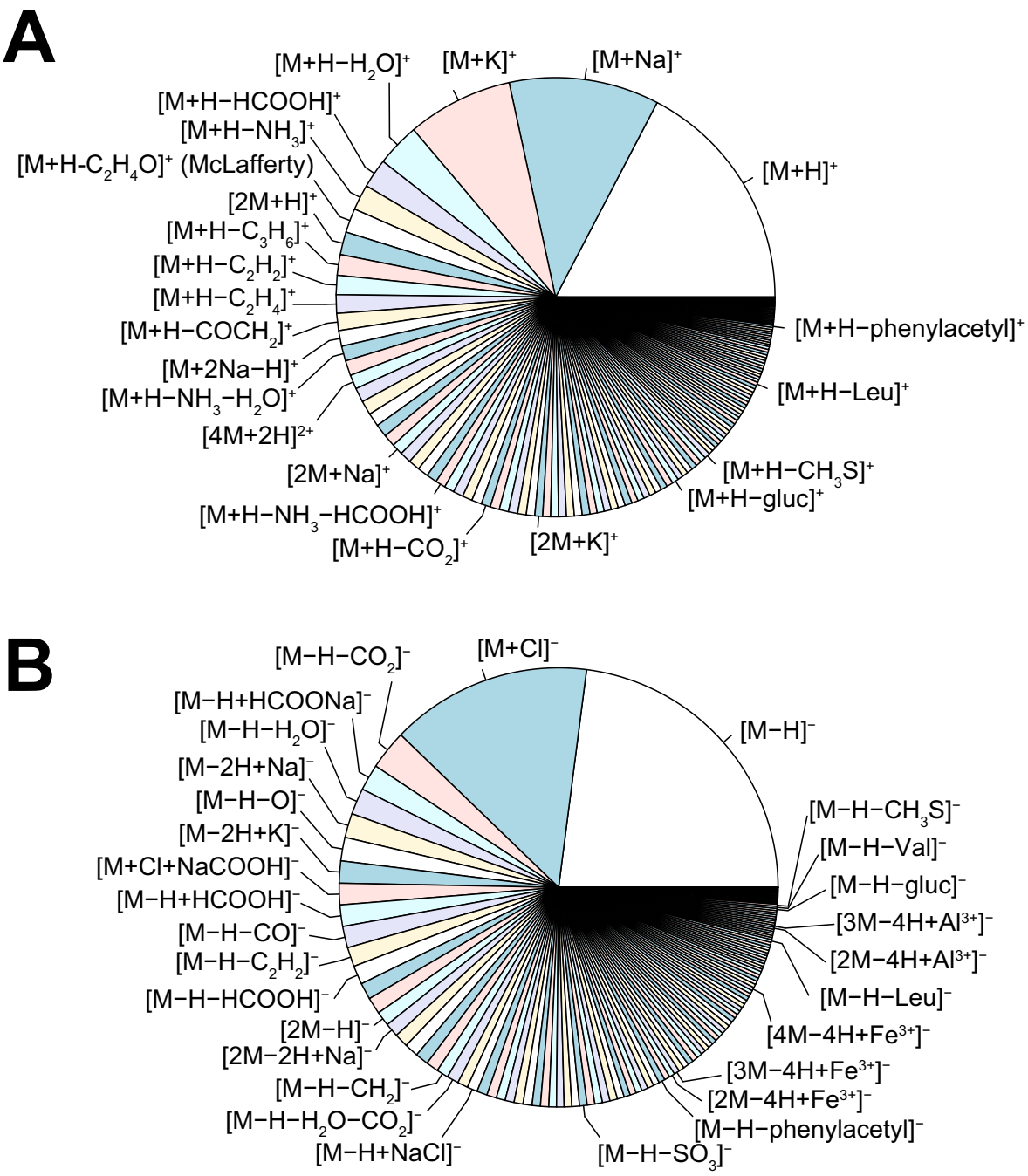


Figure 4.6. Distribution of annotations in positive (A) and negative (B) ionization mode.

Table 4.3. Annotations assigned in positive ionization mode.

Annotation	Count	N mol	Mass difference
$[M+H]^+$	331	1	1.0073
$[M+Na]^+$	211	1	22.9892
$[M+K]^+$	149	1	38.9632
$[M+H-H_2O]^+$	61	1	-17.0033
$[M+H-HCOOH]^+$	41	1	-44.9982
$[M+H-NH_3]^+$	36	1	-16.0193
$[M+H-C_2H_4O]^+$ (McLafferty)	34	1	-43.0189
$[2M+H]^+$	32	2	1.0073
$[M+H-C_3H_6]^+$	29	1	-41.0397
$[M+H-C_2H_2]^+$	27	1	-25.0084
$[M+H-C_2H_4]^+$	26	1	-27.0240
$[M+H-COCH_2]^+$	24	1	-41.0033
$[M+2Na-H]^+$	22	1	44.9712
$[M+H-NH_3-H_2O]^+$	21	1	-34.0298
$[M+H-CH_2]^+$	21	1	-13.0084
$[4M+2H]^{2+}$	20	4	2.0146
$[M+H-CO]^+$	20	1	-26.9876
$[M+H-C_3H_6O]^+$	20	1	-57.0346
$[M+H-C_6H_{12}]^+$	19	1	-83.0866
$[M+H-C_4H_6O]^+$	18	1	-69.0346
$[M+H-C_4H_8O_2]^+$	17	1	-87.0452
$[2M+Na]^+$	16	2	22.9892
$[M+H-(H_2O)_2]^+$	16	1	-35.0139
$[M+H-C_4H_6]^+$	16	1	-53.0397
$[M+H-C_4H_8]^+$	15	1	-55.0553
$[M+H-C_2H_4O_2]^+$	15	1	-59.0139
$[M+H-NH_3-HCOOH]^+$	14	1	-62.0248
$[M+H-C_5H_{10}]^+$	14	1	-69.0710
$[M+H-C_3H_4]^+$	14	1	-39.0240
$[M+H-COCH_2-C_4H_8]^+$	14	1	-97.0659
$[M+H-C_4H_6-H_2O]^+$	14	1	-71.0502
$[M+H-CO_2]^+$	13	1	-42.9826
$[M+H-O]^+$	13	1	-14.9876
$[M+H-CH_2O]^+$	13	1	-29.0033
$[M+H-C_5H_8O]^+$	13	1	-83.0502
$[M+H-C_3H_4O]^+$	12	1	-55.0189
$[M+H-CH_4]^+$	12	1	-15.0240
$[2M+K]^+$	11	2	38.9632
$[4M+2Na]^{2+}$	11	4	45.9784
$[M+H-CH_3OH]^+$	11	1	-31.0189
$[M+H-C_2H_4-CO_2]^+$	11	1	-71.0139
$[M+H-NH_3-H_2O-H_2O]^+$	11	1	-52.0404
$[M+H-C_4H_8-C_4H_6]^+$	11	1	-109.1020
$[M+H-(H_2O)_3]^+$	11	1	-53.0244
$[M+Na+K-H]^+$	10	1	60.9451
$[M+H+CHOONa]^+$	10	1	68.9947
$[M+H-C_2H_6]^+$	10	1	-29.0397
$[M+H-NH_3-CO_2-NH_3-H_2O]^+$	10	1	-95.0462
$[M+H-H_2O-C_2H_2O_2]^+$	10	1	-75.0088
$[M+H-C_4H_6-COCH_2]^+$	10	1	-95.0502
$[4M+2K]^{2+}$	9	4	77.9263
$[M+H-NH_3-COCH_2]^+$	9	1	-58.0298
$[M+H-S]^+$	9	1	-30.9648

$[M+H-C_4H_6-C_2H_4]^+$	9	1	-81.0710
$[M+H-C_3H_4O-C_4H_6]^+$	9	1	-109.066
$[M+H-NH_3-C_3H_4-COCH_2]^+$	9	1	-98.0611
$[M+H-C_3H_9N]^+$	9	1	-58.0662
$[M+H-gluc]^+$	9	1	-175.0250
$[3M+H]^+$	8	3	1.0073
$[M+H-C_5H_8]^+$	8	1	-67.0553
$[M+H-H_2O-H_2O-C_2H_4O$ (McLafferty)$]^+$	8	1	-79.0401
$[M+H-C_2H_2O_2]^+$	8	1	-56.9982
$[M+H-HCOOH-HCOOH]^+$	8	1	-91.0037
$[M+H-C_4H_6-C_4H_6O]^+$	8	1	-123.0820
$[M+H-CH_3S]^+$	8	1	-45.9883
$[3M+2H]^{2+}$	6	3	2.0146
$[2M-2H+Al^{3+}]^+$	6	2	24.9670
$[M+H-SO_3-H_2O]^+$	6	1	-96.9601
$[M+H-NH_3-CO-COCH_2]^+$	6	1	-86.0248
$[M+H-C_8H_6O-NH_3]^+$	6	1	-134.0610
$[M+H-C_2H_4-HCOOH]^+$	6	1	-73.0295
$[M+H-NH_3-NH_3-C_3H_4]^+$	6	1	-73.0771
$[M+H-C_3H_6O-CH_3OH]^+$	6	1	-89.0608
$[M+H-C_4H_6-NH_3-H_2O]^+$	6	1	-88.0768
$[M+H-NH_3-HCOOH-CH_3OH]^+$	6	1	-94.0510
$[2M+3H]^{3+}$	5	2	3.0218
$[4M+3H]^{3+}$	5	4	3.0218
$[3M+Na]^+$	5	3	22.9892
$[3M+K]^+$	5	3	38.9632
$[4M-2H+Al^{3+}]^+$	5	4	24.9670
$[M+H+(CH_3)_2CO-H_2O]^+$ (acetone cond.)	5	1	41.0386
$[M+H-H_2O-HCOOH]^+$	5	1	-63.0088
$[M+H-NH_3-C_3H_4]^+$	5	1	-56.0506
$[M+H-NH_3-CO_2-C_5H_8]^+$	5	1	-128.0720
$[M+H-C_2H_4O_2-CH_3OH]^+$	5	1	-91.0401
$[M+H-NH_3-C_2H_6]^+$	5	1	-46.0662
$[M+H-Leu]^+$	5	1	-130.0870
$[4M+H]^+$	4	4	1.0073
$[2M+2Na-H]^+$	4	2	44.9712
$[M+H+K]^{2+}$	4	1	39.9704
$[3M+H+K]^{2+}$	4	3	39.9704
$[3M+2K]^{2+}$	4	3	77.9263
$[M+H+KCl]^+$	4	1	74.9398
$[M+H-NH_3-CO_2]^+$	4	1	-60.0091
$[M+H-CO_2-C_3H_6]^+$	4	1	-85.0295
$[M+H-H_2O-CO_2]^+$	4	1	-60.9931
$[M+H-NH_3-CO_2-C_3H_4O]^+$	4	1	-116.0350
$[M+H-C_3H_9N-C_2H_4O_2]^+$	4	1	-118.0870
$[M+H-gluc-(H_2O)_2]^+$	4	1	-211.0460
$[M+H-(H_2O)_3-CO]^+$	4	1	-81.0193
$[2M+2K-H]^+$	3	2	76.9190
$[2M+H+K]^{2+}$	3	2	39.9704
$[4M+Na+K]^{2+}$	3	4	61.9524
$[2M-2H+Fe^{3+}]^+$	3	2	53.9204
$[M+H+NaCl]^+$	3	1	58.9659
$[M+H-NH_3-CO-CO]^+$	3	1	-72.0091
$[M+H-C_8H_6O]^+$	3	1	-117.0350
$[M+H-NH_3-CO_2-CH_2O]^+$	3	1	-90.0197
$[M+H-gluc-(H_2O)_3]^+$	3	1	-229.0570
$[M+H-phenylacetyl]^+$	3	1	-117.0350
$[M+2H]^{2+}$	2	1	2.0146
$[M+2K-H]^+$	2	1	76.9190

$[3M+2K-H]^+$	2	3	76.9190
$[M+H+Na]^{2+}$	2	1	23.9965
$[3M+2Na]^{2+}$	2	3	45.9784
$[M+Na+K]^{2+}$	2	1	61.9524
$[2M+Na+K]^{2+}$	2	2	61.9524
$[M+H+(NaCl)_2]^+$	2	1	116.9245
$[M+H-SO_3]^+$	2	1	-78.9495
$[M+H-C_8H_6O-H_2O]^+$	2	1	-135.0450
$[M+H-NH_3-CO-COCH_2-C_4H_6O]^+$	2	1	-156.0670
$[M+H-C_3H_4O-C_4H_8O_2]^+$	2	1	-143.0710
$[M+H-gluc+H_2O]^+$	2	1	-193.0350
$[M+3H]^{3+}$	1	1	3.0218
$[3M+2Na-H]^+$	1	3	44.9712
$[4M+Na]^+$	1	4	22.9892
$[4M+K]^+$	1	4	38.9632
$[4M+2K-H]^+$	1	4	76.9190
$[M+2Na]^{2+}$	1	1	45.9784
$[M+2K]^{2+}$	1	1	77.9263
$[4M+H+K]^{2+}$	1	4	39.9704
$[3M+Na+K]^{2+}$	1	3	61.9524
$[3M-2H+Fe^{3+}]^+$	1	3	53.9204
$[M+H-S-NH_3-HCOOH]^+$	1	1	-93.9968
$[M+H-SO_3-H_2O-NH_3]^+$	1	1	-113.9870
$[M+H-Val]^+$	1	1	-116.0720
$[2M+2H]^{2+}$	0	2	2.0146
$[3M+3H]^{3+}$	0	3	3.0218
$[4M+2Na-H]^+$	0	4	44.9712
$[2M+H+Na]^{2+}$	0	2	23.9965
$[2M+2Na]^{2+}$	0	2	45.9784
$[3M+H+Na]^{2+}$	0	3	23.9965
$[4M+H+Na]^{2+}$	0	4	23.9965
$[2M+2K]^{2+}$	0	2	77.9263
$[4M-2H+Fe^{3+}]^+$	0	4	53.9204
$[3M-2H+Al^{3+}]^+$	0	3	24.9670
$[M+H-NH_3-C_8H_6O-CH_2]^+$	0	1	-148.0770
$[M+H-gluc-(H_2O)_3-CO]^+$	0	1	-257.051

Table 4.4. Annotations assigned in negative ionization mode.

Annotation	Count	N mol	Mass difference
$[M-H]^-$	451	1	-1.0073
$[M+Cl]^-$	291	1	34.9694
$[M-H-CO_2]^-$	57	1	-44.9971
$[M-H+HCOONa]^-$	38	1	66.9802
$[M-H-H_2O]^-$	38	1	-19.0178
$[M-2H+Na]^-$	35	1	20.9747
$[M-H-O]^-$	34	1	-17.0022
$[M-2H+K]^-$	32	1	36.9486
$[M+Cl+NaCOOH]^-$	31	1	102.9568
$[M-H+HCOOH]^-$	31	1	44.9982
$[M-H-CO]^-$	31	1	-29.0022
$[M-H-C_2H_2]^-$	28	1	-27.0229
$[M-H-HCOOH]^-$	27	1	-47.0128
$[M-H-C_3H_6]^-$	24	1	-43.0542
$[M-H-CH_4]^-$	24	1	-17.0386
$[2M-H]^-$	21	2	-1.0073
$[2M-2H+Na]^-$	20	2	20.9747
$[M-H-C_2H_2O_2]^-$	20	1	-59.0128
$[M-H-C_2H_4O_2]^-$	20	1	-61.0284
$[M-H-CH_3OH]^-$	19	1	-33.0335
$[M-H-C_4H_8O_2]^-$	18	1	-89.0597
$[M-H-CH_2]^-$	17	1	-15.0229
$[M-H-H_2O-CO_2]^-$	17	1	-63.0077
$[4M-2H]^{2-}$	15	4	-2.0146
$[M-H+NaCl]^-$	15	1	56.9513
$[M-H-COCH_2]^-$	15	1	-43.0178
$[M-H-CH_2O]^-$	14	1	-31.0178
$[M-H-NH_3]^-$	13	1	-18.0338
$[M-H-NH_3-CO_2]^-$	13	1	-62.0237
$[M-H-C_5H_8O]^-$	13	1	-85.0648
$[M-H-H_2O-C_2H_2O_2]^-$	13	1	-77.0233
$[M-H-C_4H_6O]^-$	12	1	-71.0491
$[M-H-C_3H_4]^-$	12	1	-41.0386
$[2M-2H+K]^-$	11	2	36.9486
$[M-H-C_2H_4]^-$	11	1	-29.0386
$[M-H-C_4H_8]^-$	11	1	-57.0699
$[M-H-C_3H_6O]^-$	11	1	-59.0491
$[M-H-SO_3]^-$	11	1	-80.9641
$[M-H-NH_3-CO-COCH_2]^-$	11	1	-88.0393
$[M-H-C_2H_4-CO_2]^-$	11	1	-73.0284
$[M-H-C_4H_6-H_2O]^-$	11	1	-73.0648
$[M-H-(H_2O)_3]^-$	11	1	-55.039
$[M-H-C_2H_6]^-$	10	1	-31.0542
$[M+Cl+HCOOH]^-$	9	1	80.9749
$[M-H-(H_2O)_2]^-$	9	1	-37.0284
$[M-H-C_8H_6O]^-$	9	1	-119.0491
$[M-H-NH_3-CO_2-CH_2O]^-$	9	1	-92.0342
$[M-H-C_2H_4-HCOOH]^-$	9	1	-75.0441
$[M-H-(H_2O)_3-CO]^-$	9	1	-83.0339
[M-H-phenylacetyl]$^-$	9	1	-119.0492
$[2M+Cl]^-$	8	2	34.9694
$[2M-4H+Fe^{3+}]^-$	8	2	51.9042
$[3M-4H+Fe^{3+}]^-$	8	3	51.9042

$[M-H-C_2H_4O$ (McLafferty)$]^-$	8	1	-45.0335
$[M-H-C_3H_4O]^-$	8	1	-57.0335
$[M-H-C_5H_8]^-$	8	1	-69.0699
$[M-H-COCH_2-C_4H_8]^-$	8	1	-99.0804
$[M-H-C_3H_6O-CH_3OH]^-$	8	1	-91.0754
$[3M-2H]^{2-}$	7	3	-2.0146
$[4M+2Cl]^{2-}$	7	4	69.9388
$[M-H+(NaCl)_2]^-$	7	1	114.91
$[M-H+KCl]^-$	7	1	72.9253
$[M-H-NH_3-H_2O]^-$	7	1	-36.0444
$[M-H-C_8H_6O-H_2O]^-$	7	1	-137.0597
$[M-H-NH_3-CO_2-NH_3-H_2O]^-$	7	1	-97.0608
$[M-H-C_4H_6-C_2H_4]^-$	7	1	-83.0855
$[M-H-C_4H_6-COCH_2]^-$	7	1	-97.0648
$[M-H-C_3H_4O-C_4H_8O_2]^-$	7	1	-145.0859
$[4M-2H+Na]^-$	6	4	20.9747
$[4M-2H+K]^-$	6	4	36.9486
$[2M-3H+Fe^{2+}]^-$	6	2	52.9115
$[3M-3H+Fe^{2+}]^-$	6	3	52.9115
$[M-H-C_2H_4O_2-CH_3OH]^-$	6	1	-93.0546
$[M-H-NH_3-HCOOH-CH_3OH]^-$	6	1	-96.0655
$[3M-2H+K]^-$	5	3	36.9486
$[4M-H+Cl]^{2-}$	5	4	33.9621
$[4M-4H+Fe^{3+}]^-$	5	4	51.9042
$[M-H-C_5H_{10}]^-$	5	1	-71.0855
$[M-H-CO_2-C_3H_6]^-$	5	1	-87.0441
$[M-H-H_2O-H_2O-C_2H_4O$ (McLafferty)$]^-$	5	1	-81.0546
$[M-H-NH_3-H_2O-H_2O]^-$	5	1	-54.055
$[M-H-C_4H_8-C_4H_6]^-$	5	1	-111.1168
$[3M-H]^-$	4	3	-1.0073
$[4M-H]^-$	4	4	-1.0073
$[3M+Cl]^-$	4	3	34.9694
$[M-H+(CH_3)_2CO-H_2O]^-$ (acetone cond.)	4	1	39.024
$[M-H-NH_3-HCOOH]^-$	4	1	-64.0393
$[M-H-S]^-$	4	1	-32.9793
$[M-H-S-NH_3-HCOOH]^-$	4	1	-96.0114
$[M-H-C_4H_6]^-$	4	1	-55.0542
$[M-H-C_6H_{12}]^-$	4	1	-85.1012
$[M-H-SO_3-H_2O]^-$	4	1	-98.9747
$[M-H-C_3H_4O-C_4H_6]^-$	4	1	-111.0804
$[M-H-Leu]^-$	4	1	-132.102
$[M-2H]^{2-}$	3	1	-2.0146
$[4M-3H]^{3-}$	3	4	-3.0218
$[3M-2H+Na]^-$	3	3	20.9747
$[4M+Cl]^-$	3	4	34.9694
$[3M+2Cl]^{2-}$	3	3	69.9388
$[2M-4H+Al^{3+}]^-$	3	2	22.9524
$[3M-4H+Al^{3+}]^-$	3	3	22.9524
$[M-H-NH_3-COCH_2]^-$	3	1	-60.0444
$[M-H-H_2O-HCOOH]^-$	3	1	-65.0233
$[M-H-SO_3-H_2O-NH_3]^-$	3	1	-116.0012
$[M-H-C_8H_6O-NH_3]^-$	3	1	-136.0757
$[M-H-NH_3-NH_3-C_3H_4]^-$	3	1	-75.0917
$[M-H-NH_3-C_3H_4-COCH_2]^-$	3	1	-100.0757
$[M-H-C_3H_9N]^-$	3	1	-60.0808
$[M-H-gluc]^-$	3	1	-177.0394
$[M-H-Val]^-$	3	1	-118.0863
$[M-H-CH_3S]^-$	3	1	-48.0028
$[M-H+Cl]^{2-}$	2	1	33.9621

$[M-H-NH_3-C_3H_4]^-$	2	1	-58.0651
$[M-H-NH_3-CO_2-C_3H_4O]^-$	2	1	-118.0499
$[M-H-HCOOH-HCOOH]^-$	2	1	-93.0182
$[M-H-NH_3-CO-COCH_2-C_4H_6O]^-$	2	1	-158.0812
$[M-H-C_4H_6-C_4H_6O]^-$	2	1	-125.0961
$[M-H-NH_3-C_2H_6]^-$	2	1	-48.0808
$[M-H-C_3H_9N-C_2H_4O_2]^-$	2	1	-120.1019
$[M-3H]^{3-}$	1	1	-3.0218
$[M+2Cl]^{2-}$	1	1	69.9388
$[4M-3H+Fe^{2+}]^-$	1	4	52.9115
$[M-H-NH_3-CO-CO]^-$	1	1	-74.0237
$[M-H-C_4H_6-NH_3-H_2O]^-$	1	1	-90.0913
$[M-H-NH_3-C_8H_6O-CH2]^-$	1	1	-150.0913
$[M-H-gluc-H_2O]^-$	1	1	-195.05
$[M-H-gluc-(H_2O)_3-CO]^-$	1	1	-259.066
$[2M-2H]^{2-}$	0	2	-2.0146
$[2M-3H]^{3-}$	0	2	-3.0218
$[3M-3H]^{3-}$	0	3	-3.0218
$[2M-H+Cl]^{2-}$	0	2	33.9621
$[2M+2Cl]^{2-}$	0	2	69.9388
$[3M-H+Cl]^{2-}$	0	3	33.9621
$[4M-4H+Al^{3+}]^-$	0	4	22.9524
$[M-H-NH_3-CO_2-C_5H_8]^-$	0	1	-130.0863
$[M-H-gluc-(H_2O)_2]^-$	0	1	-213.0605
$[M-H-gluc-(H_2O)_3]^-$	0	1	-231.0711

Not surprisingly the most prevalent adducts are the well-known Na^+, K^+ and Cl^- adducts and the most prevalent fragments are the common neutral losses like H_2O, CO_2, NH_3 and HCOOH. Polymers such as $[2M+H]^+$ are also prevalent.

An interesting result of the analysis is that the negative mode neutral adduct of sodium formate (HCOONa) is even more prevalent than formic acid (HCOOH) adducts. Compound databases like HMDB [54] and METLIN [55] normally check if a query mass could be a formic acid adduct while it is not checked if the mass could represent a sodium formate adduct. Given these results sodium formate should be checked too.

Unusual iron adducts such as $[2M-4H+Fe^{3+}]^-$ species were also found. It has previously been noted that likely the most important characteristic determining if a compound generates sodium adducts is the ability of the compound to generate sodium chelates [56]. The same appears to be true for iron, and likely any metal adduct, as the iron adduct was found to be present for di-ionic compounds such as 2-hydroxyisovaleric acid that likely form a strong chelate with the ferric ion as shown in Figure 4.7. However, also mono-ionic (when carrying a net negative charge) compounds such as tryptophan was found to form ferric adducts. In these cases four molecules are needed to form the chelate ($[4M+Fe^{3+}-4H]^-$).

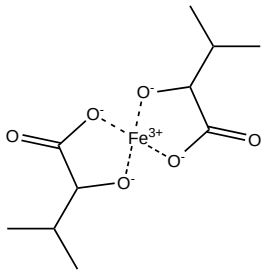


Figure 4.7. In source generated 2-hydroxy-isovaleric acid ferric chelate ([2M-4H+Fe^{3+}]$^-$).

As can be seen from the lists in Table 4.3 and Table 4.4 I have included many sequential fragmentations (such as [M+H-NH_3-H_2O-H_2O]$^+$). This is necessary to correctly annotate the relationship when the feature group contains more than one fragment. But it also means that creating the list of possible fragments becomes very cumbersome. To reduce the amount of manual curation of the list CAMERA recently added functions to generate the sequential fragmentation rules by automatically combining the "basic" neutral losses; however these functions currently do not appear to be working optimally.

It should be emphasized, that while a larger list of possible annotations increase the number of features that can be annotated, the larger the list of possible annotations the higher the chance of incorrect annotations that by chance fit a mass difference between unrelated features. Despite the fact that I have confirmed a [2M-4H+Al^{3+}]$^-$ species for 2-hydroxyisovaleric acid (in analysis of a pure standard) the features annotated as such in the dataset from biological samples appear to be chance findings.

Reducing the number of features to consider for identification is one purpose of annotation. But the annotation can also give very direct evidence of the structure of the compound. This is the case when the fragments are highly specific such as loss of glucuronide, sulfate, CH_3S (methionine indicator), leucine, valine and phenylacetyl.

With the features grouped and the pseudo-molecular ion identified it is possible to use the feature groups, or pseudo-spectra, as pseudo-MS/MS spectra. These pseudo-MS/MS spectra can under optimal conditions be interpreted as regular MS/MS spectra for the purpose of identification, both manual and computer-assisted as discussed in section 4.4.

4.3 MOLECULAR FORMULA DETERMINATION

When the pseudo-molecular ion has been determined the typical next step is determining the molecular sum formula of the compound. This is done by calculation all combinations of atoms that would sum up to the mass of the unknown compound under investigation. With the exception of very small parent ion masses this leads to a high number of candidate formulas; the number increasing dramatically with increasing mass.

A number of methods have therefore been developed to eliminate candidate formulas that are not chemically possible. Classical rules include the nitrogen rule stating that [57]:

> *organic compounds containing exclusively hydrogen, carbon, nitrogen, oxygen, silicon, phosphorus, sulfur, and the halogens either have*
> *1) an odd nominal (unionized) mass that indicates an odd number of nitrogen atoms are present or*
> *2) an even nominal mass that indicates an even number of nitrogen atoms are present in the molecular ion*

The nitrogen rule, while useful for small molecules and fragments, should not be applied to higher mass compounds since the non-nominal mass contributions from each element add up to the "next" integer for high mass compounds. Kind and Fiehn showed that this leads to failure of the nitrogen rule for a large portion of compounds [58].

Another classical approach is calculation of the Ring Double Bond Equivalents (RDBE), also called the degree of saturation defined as [58]:

$$RDBE = C + Si - \frac{H + F + Cl + Br + I}{2} + \frac{N + P}{2} + 1$$

where each element symbol represents the count of the element in the structure. This formula, however, only takes into account compounds in their lowest valence state [58]. But since nitrogen, phosphorous and sulfur can exist in several valence states this formula is not always accurate. This is exemplified by the compounds methionine ($C_5H_{11}NO_2S$), methionine sulfoxide ($C_5H_{11}NO_3S$) and methionine sulfone ($C_5H_{11}NO_4S$) that all have the same RDBE. The RDBE might, however, be used to exclude compounds with an unrealistically high or low RDBE.

Kind and Fiehn have proposed 7 “golden” alternative rules rigorously validated against compound databases [58]. These rules cause fewer correct formulas to be disregarded.

Another, more complex, approach is to use the observed fragments to rule out formulas where the neutral losses are not a subset of the sum formula. Such an approach has been implemented in the program Sirius [59]. Since I wanted a pipeline that could be fully automated I have not used this approach. Future releases that enable automation could be implemented in the identification pipeline. The annotation from CAMERA could even be used directly instead of actual MS/MS spectra.

In the following I have used the R package Rdisop [59] to generate possible molecular formulas.

Rdisop only disregards

1) compounds that violate Senior’s third theorem [60] stating that the sum of valences has to be greater than or equal to twice the number of atoms minus one
2) radicals with odd sum of valences

though RDBE and adherence to the nitrogen rule is computed for manual inspection.

After all possible formulas have been computed the formulas should be ranked according to how likely each formula is compared to the experimental data. The simplest way to rank the formulas would be to rank them according to closeness to the experimental mass of the compound. However, it has been shown that the experimental isotopic pattern has equal or even higher discrimination potential [61]. We therefore compare the theoretical ratio between the monoisotopic mass and the first isotope to the observed isotopic pattern [50]. Rdisop provides a score for the fit of the isotopic distribution (all isotopes) and mass error. Unfortunately, I have found that the score occasionally gives curious results where a compound is scored lower even when it is more close to the experimental mass values and isotope distribution than another compound. For the purpose of presenting a more intuitive ranking list of candidate formulas I have therefore devised a score that gives more intuitive results. The score gives equal importance to the mass error and difference between experimental and theoretical isotopic ratios:

$$score_i = \frac{|E(IR)_i|^{-1} \cdot |E(mz)_i|^{-1}}{\sum_{i=1}^{N} \left(|E(IR)_i|^{-1} \cdot |E(mz)_i|^{-1}\right)} \cdot 100$$

where $E(IR)_i$ is the relative error in the ratio between the monoisotopic peak and the first isotope and $E(mz)_i$ is the relative mass error for formula *i*. I have not attempted to validate the appropriateness of the score. A better score should take into account all isotope peaks giving less importance to low intensity peaks since the accuracy of the mass and the intensity is lower for low intensity peaks. Also it would be appropriate to punish very aberrant values more than the above score does. A sigmoid function could be used for this purpose as was used for MetFusion scoring [62] (see section 4.4).

The formulas resulting from decomposing the observed mass and isotopic pattern of tryptophan is reported in Table 4.5. The correct formula is ranked first with the scoring suggested above while the Rdisop score would rank it second (and assign a rather low absolute value despite the excellent fit). Only the formulas $C_{11}H_{12}N_2O_2$, $C_4H_{12}N_8S$, $C_9H_{17}O_3P$ and $C_5H_{13}N_6OP$ have non-zero positive integer RDBEs and adheres to the nitrogen rule. By applying the rules suggest by Kind and Fiehn regarding element ratios $C_4H_{12}N_8S$ can be regarded as highly unlikely due to the unusual N to C ratio. $C_9H_{17}O_3P$ and $C_5H_{13}N_6OP$ do not appear to violate any of the rules proposed by Kind and Fiehn. However, no $C_5H_{13}N_6OP$ compounds appear to be known and such a structure would require some very unusual connectivity in the structure and fit the isotopic pattern very poorly. $C_9H_{17}O_3P$ cannot easily be excluded in this case and indeed $C_9H_{17}O_3P$ compounds do exists. This illustrates clearly that even for small molecules it can be impossible to assign a unique molecular formula based only on pseudo-molecular species. Fortunately, the fact that neutral losses of $[M\text{-}NH_3+H]^+$ and $[M\text{-}HCOOH+H]^+$ were found exclude all other formulas than the correct one. Regretfully only the stand-alone program Sirius is currently able to use this information as explained above and it is therefore not yet used in the identification pipeline.

Table 4.5. Decomposition of the recorded *m/z* value attributable to tryptophan with a 10 ppm tolerance. Ranked according to the score suggested above.

Rank	Formula	Rdisop score	Nitrogen rule	RDBE	Isotope ratio error (%)	ppm	New score
1	$C_{11}H_{12}N_2O_2$	0.0141	Valid	7	-0.79	3.45	43.5
2	$C_4H_{12}N_8S$	4.8E-75	Valid	3	57.0	0.09	23.8
3	$C_5H_{18}NO_5S$	4.6E-24	Invalid	-2.5	82.8	0.06	23.5
4	$C_3H_{18}N_4O_2P_2$	4.3E-73	Valid	-2	155	0.40	1.92
5	$C_7H_{15}N_3O_2P$	6.3E-16	Invalid	2.5	42.8	1.92	1.44
6	$C_9H_{10}N_5O$	3.1E-34	Invalid	7.5	9.03	10.00	1.30
7	$C_9H_{17}O_3P$	0.986	Valid	2	26.4	-4.66	0.96
8	$C_6H_{22}NS_3$	2.9E-27	Invalid	-3.5	34.7	-4.19	0.81
9	$C_6H_{14}N_5OS$	7.6E-43	Invalid	2.5	37.4	-6.49	0.49
10	$H_{15}N_9PS$	8.4E-264	Invalid	-1.5	204	-1.44	0.40
11	$C_3H_{26}OP_2S_2$	8.1E-53	Valid	-8	146	2.69	0.30
12	$CH_{21}N_2O_5PS$	4.8E-140	Valid	-7	318	-1.46	0.25
13	$C_5H_{13}N_6OP$	3.8E-61	Valid	3	64.1	8.50	0.22
14	$C_5H_{20}NO_3P_2$	4.6E-30	Invalid	-2.5	107	-6.18	0.18
15	$C_2H_{25}N_2PS_3$	5.3E-93	Valid	-8	130	-5.72	0.16
16	$C_3H_{16}N_4O_4S$	1.6E-80	Valid	-2	119	6.64	0.15
17	$C_3H_{24}O_3S_3$	1.2E-66	Valid	-8	112	8.94	0.12
18	$C_2H_{17}N_6OPS$	6.3E-128	Valid	-2	138	-8.02	0.11
19	$H_{141}NO_3$	2.8E-93	Valid	-69	405	-2.78	0.11
20	$CH_{16}N_7OP_2$	1.3E-198	Invalid	-1.5	231	6.98	0.07
21	$H_{28}O_3S_4$	1.1E-166	Valid	-13	255	-7.58	0.06
22	$CH_{24}N_3P_2S_2$	5.3E-156	Invalid	-7.5	216	9.27	0.06
23	$C_2H_{22}O_6P_2$	1.5E-102	Valid	-7	371	6.95	0.05
24	$H_{20}N_4O_4S_2$	1.6E-232	Valid	-7	275	-9.88	0.04
25	$CH_{23}N_2O_3P_3$	5.9E-174	Valid	-7	469	-7.71	0.03

Normally the isotopic pattern would be observed in the spectrum from a single sample, usually even in a single scan. However, for Q-TOF instruments the quantification and therefore the isotopic pattern is not nearly accurate enough to give a precise estimate of the true isotopic distribution of the compound of interest.

Therefore we have proposed a method where all samples from an experiment is used instead of a single sample. In this way an averaged spectrum is obtained where the random quantification error is averaged out.

In short we integrate the peak of interest in all samples. In this case a dataset with 220 samples analyzed in duplicate. We then investigated several approached to using the combined data to get an estimate of the true isotopic ratio:

1) The arithmetic mean of the isotope ratios for all samples
2) The median of the isotope ratios for all samples
3) A standard linear regression between the intensities of the monoisotopic peak against the isotope peaks

4) A robust linear regression method from the R package MASS [63]
5) A weighted mean such that the contribution to the overall mean of the isotope ratios, $\bar{R}_{I,p}$, for isotope peak p, for each sample s, was weighted by the intensity of the monoisotopic peak $I_{0,s}$, such that:

$$\overline{R}_{I,p} = \frac{\sum_{s=1}^{n}\left(I_{0,s} \cdot \frac{I_{p,s}}{I_{0,s}}\right)}{\sum_{s=1}^{n}(I_{0,s})} = \frac{\sum_{s=1}^{n}(I_{p,s})}{\sum_{s=1}^{n}(I_{0,s})}$$

The weighted mean approach performed better than the other approaches though the median was not dramatically worse, see Table 4.6. Using the full dataset also improved the accuracy substantially compared to using a single sample or just a few samples.

Table 4.6. Analysis of five approaches to calculating the isotopic ratio utilizing all available samples.

	Absolute error			
Ratio calculation method	**Mean**	**Median**	**95 percentile**	**Max**
Mean	60.58 %	4.81 %	28.87 %	1735.01 %
Median	6.37 %	3.83 %	18.90 %	23.98 %
Normal regression	6.84 %	3.65 %	18.69 %	35.25 %
Robust regression	7.00 %	4.23 %	19.87 %	37.06 %
Weighted mean	5.34 %	2.94 %	17.84 %	21.72 %

The XCMS method "integrated peak intensity" (into) was used for peak quantification.

As will be explained in the next section the identification pipeline leads to a large number of candidate molecules for each feature that is to be identified. Giving the accuracy of the experimental isotope ratio as determined by our method 43 % of molecular formulas and 25 % of compound candidates could on average been disregarded among the compounds identified using the pipeline if the maximum observed error in the isotope ratio was used as a strict filter.

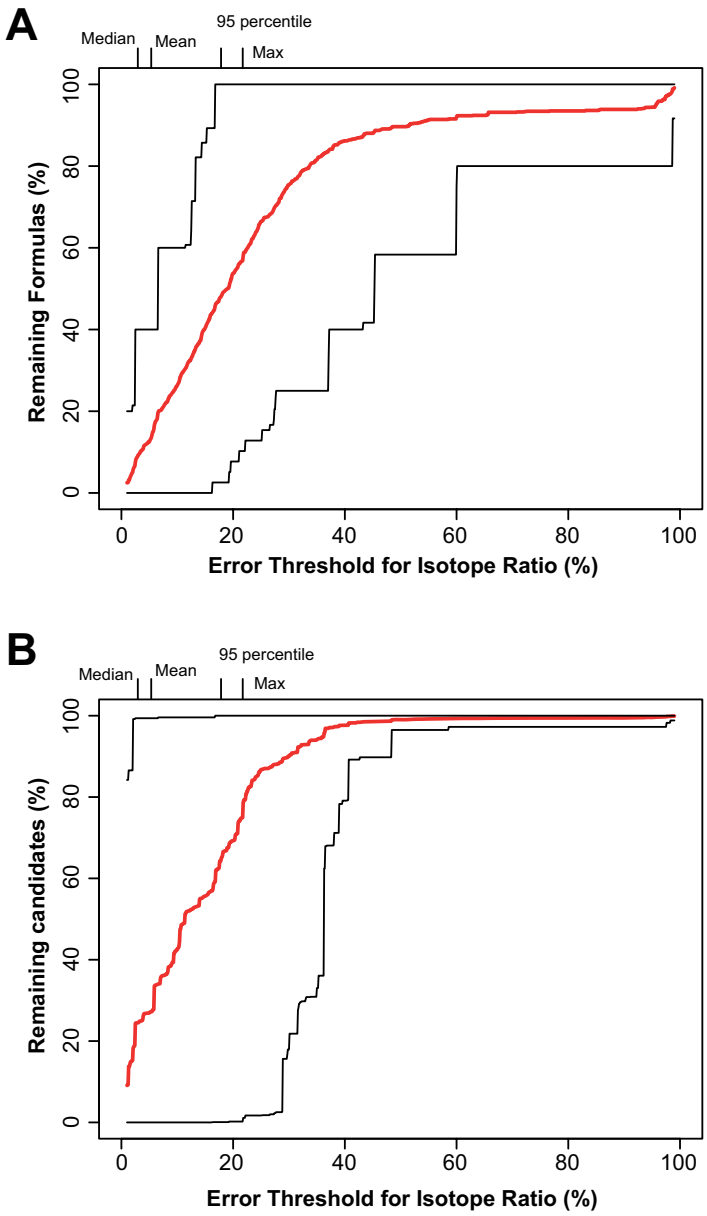


Figure 4.8. A) Remaining molecular formulas as a function of the allowed relative error of the ratio of the first isotope relative to the pseudo-molecular ion. The mean (red) between candidate lists as well as the minimum and maximum are shown. The median, mean, 95 percentile and the maximum error observed for the identified compounds have been marked at the top of the plot.
B) Remaining candidates as a function of the allowed relative error of the ratio of the first isotope relative to the pseudo-molecular ion. The mean (red) between candidate lists as well as the minimum and maximum are shown. The median, mean, 95 percentile and the maximum error observed for the identified compounds have been marked at the top of the plot.
Figure reproduced after Stanstrup et al. **[50]** with kind permission from Springer Science and Business Media.

In Figure 4.8, the effect of lowering the tolerance can be observed. Relatively modest improvements in the determination of the isotope pattern could enable a much more effective filtering. If for example a maximum error of 10 % could be guaranteed, on average 75 % of formulas and 60 % of candidates could be disregarded solely based on the isotopic ratio.

For more details on the evaluation of different calculation approaches the reader is referred to Paper I and supplemental file 3 of the same paper.

4.4 MASS FRAGMENTATION AND *IN SILICO* TOOLS

While the molecular formula of a compound is significant information important for final identification it is rarely sufficient on its own.

Traditionally the next step is acquisition of MS/MS data of the unknown and subsequent manual interpretation of the spectra. A number of candidate structures (candidate compounds below) would then be proposed. These candidate compounds then had to be bought or synthesized to compare the MS/MS spectra against the spectra of the unknown. This method has the disadvantage of being heavily reliant on the expertise of the expert spectrometrist and tremendously time-consuming due to the high number of compounds that are in all principal unknowns in a metabolomics study.

To save time researchers therefore query the spectrum of the unknown against comprehensive spectral databases of known compounds such as the Wiley [64], NIST [65], MassBank [66], METLIN [55], or HMDB [54] databases. While this is a sensible first step it can only identify a compound when the spectrum is in the spectral database. Unfortunately the spectral databases, while being comprehensive in their own right, only cover a fraction of known compounds and obviously do not cover novel compounds. In addition, ESI spectra are subject to substantial variability between instruments and it can therefore be difficult to definitively associate the spectrum of the unknown to a single spectrum in the database.

To expand the amount of available reference data it has been attempted to predict the MS/MS spectra of compounds based on their structure. However, an approach powerful enough to reliably predict the spectrum of any compound have not been developed and hence *in silico* generated spectral libraries are limited to certain compound classes such as lipids [67] and peptides [68] having more uniform and predictable fragmentation.

Instead of building *in silico* MS/MS spectral libraries an alternative approach has been developed called MetFrag [69]. MetFrag retrieves compound structures from compound libraries such as Chemspider [70], KEGG [71] or PubChem [72] based on the recorded mass or determined molecular formula of the unknown. This typically results in thousands of candidates. These structures are then *in silico* fragmented, which means that MetFrag breaks the molecule in every position and calculates the *m/z* ratio of resulting fragments. It should be noted that MetFrag cannot consider possible rearrangements and therefore an

experimentally observed fragment that MetFrag cannot explain does not in itself rule out a certain structure. The expected fragments are then compared to the observed fragments in the MS/MS spectrum of the unknown. Finally the candidate structures are ranked according to a score based on similarity of *m/z* values and their intensities. See Figure 4.9 for the MetFrag scheme.

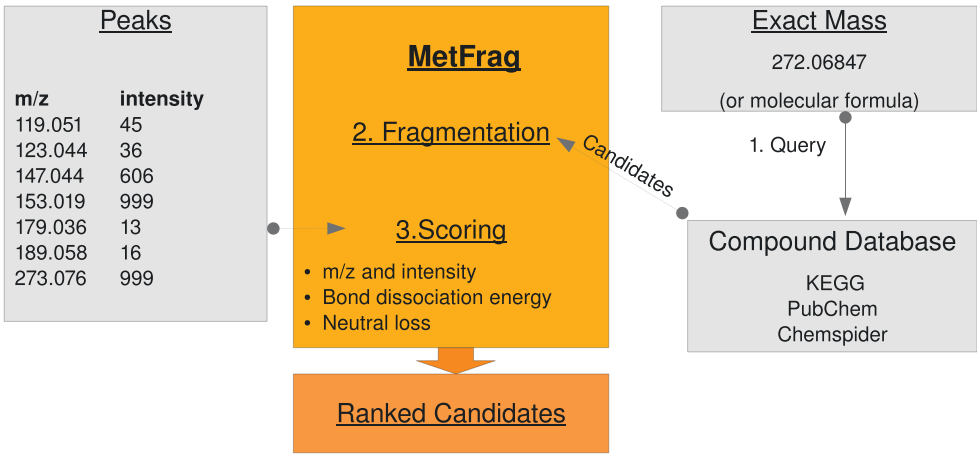


Figure 4.9. MetFrag workflow. Figure by Wolf *et al.* **[69]**.

The MetFrag approach is still limited to the compounds found in the compound libraries and hence novel compounds will be missed. Recently attempts have been made to combine a compound generator that generates all possible structures from a molecular formula and then applies MetFrag to the generated structures [73]. With the exception of very small molecules, this approach, unfortunately leads to an enormous number of theoretical compounds. To limit the number of candidate compounds, similar MS spectra from spectral libraries (more specifically mass spectral trees, i.e. extensive MS^n data) were used to attempt to determine the "maximum common substructure" (MCS) of the compounds in the spectral library with similar spectra [73]. It was then assumed that the unknown shared this MCS and the structure generator could consequently be restricted to only generate compounds with this MCS. However, very often such an MCS cannot be established and the number of theoretical candidates is therefore too large to limit the number of candidates to a manageable number.

Most recently, *in silico* fragmentation (i.e. MetFrag) has been combined with spectral databases to get the best of both worlds: the known validity of experimental spectra in spectral databases with the broader coverage achieved by *in silico* fragmentation.

This approach was implemented in MetFusion [62,74]. MetFusion combines the MetFrag score with the spectral similarity to spectra in the spectral library (usually MassBank). The structures from the compound library are used and the exact compounds do not need to be in the spectral library. Instead the spectral similarity, for the most similar spectra in the spectral library, is weighted by their structural similarity to the compound candidates from the compound database (retrieved by MetFrag). See Figure 4.10 and refer to the MetFusion publication for further details.

The performance of the MetFusion approach will be discussed in section 4.7.

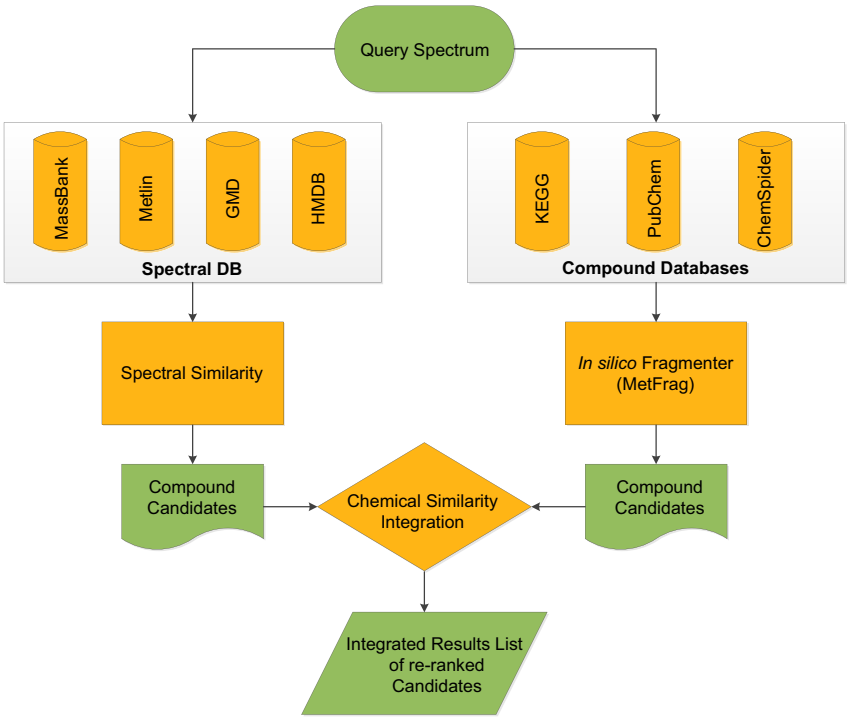


Figure 4.10. MetFusion workflow. Figure by Michael Gerlich **[74]**.

4.5 RETENTION TIME MAPPING

In LC-MS the mass spectrum of a compound is the most discriminating characteristic of a compound and hence the tools to effectively use the spectral information are well-developed as illustrated in the previous chapter. But solely inspecting the fragmentation is only utilizing half of the available information and often it does not permit assigning a feature to a single unique structure.

Another key characteristic of compounds that can be used to discriminate between isomers is the retention time (RT). RT information has unfortunately largely been ignored when building databases. The reason RT information has been neglected is that RT is specific to a specific chromatographic setup and it has therefore been considered of little value to other researchers to share RT information obtained in a specific chromatographic setup.

Currently researchers consequently purchase or synthesize all the compounds that their feature of interest could represent based on the mass spectral data so that the retention time of these known compounds can be compared to the unknown in their specific setup. This is both time-consuming and expensive.

We therefore wanted to devise a system to predict the retention time of compounds. One way to predict the retention time in our own system would be to use RT information obtained in another system. This would require a model able to translate the retention time in one system to the retention time in another.

In GC/MS setups retention indices have long been commonplace as system independent measures of retention, but for LC-MS systems there is no consensus yet.

Vaughan *et al.* recently described a RT mapping method able to translate the retention time from one system to another [75]. In this work similarity in *m/z* values and intensity correlation across a set of samples was used to match features between chromatographic setups and thus build a model able to translate retention times between the two systems. This system could be used to predict the retention time of a compound in one of the systems if it is known in the other.

In paper I [50], we instead developed a method able to create the RT mapping based on the manually determined RT of 39 well-characterized metabolites in positive mode in two

chromatographic systems (A and B). A monotonic increasing locally weighted scatterplot smoothing (LOESS) function was fitted between the RTs in these two systems to obtain a calibration curve between them.

Once this "rough" model was used to change all the RTs of the features in system A to the corresponding RT in system B we could then use the normal RT correction and feature grouping implemented in XCMS to map a large number of features (of unknown identity) and use them to refine the model. Our method does not rely on intensity correlations across a range of samples and can thus be used already if only a single sample has been analyzed in different systems.

The analysis of the serum sample in positive mode resulted in 1,632 and 6,247 features, respectively, in systems A and B. After the initial mapping, based only on the 39 known metabolites, 344 features could be matched. Using this set of matched features for a new calibration curve, 361 features could now be matched between systems. Figure 4.11 shows the RT mapping.

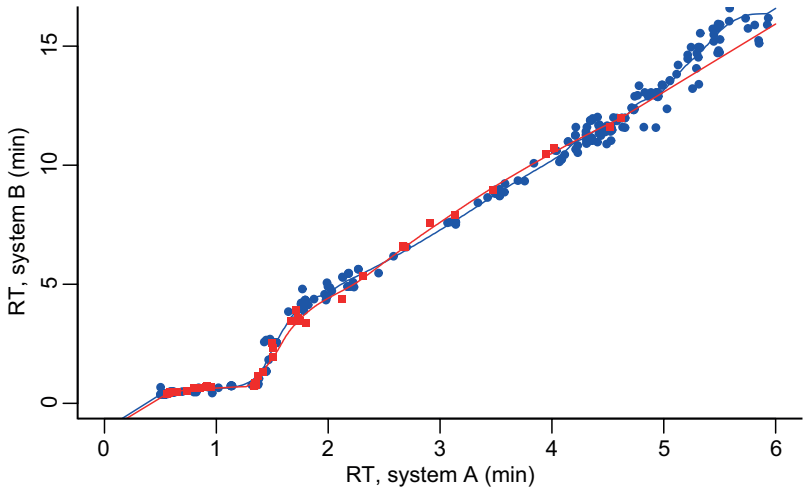


Figure 4.11. RT mapping between LC-MS systems A and B. The red squares and line show the mapping according to 39 well characterized metabolites. The blue circles and line show the mapping based on the complete set of mapped features.
Figure reproduced after Stanstrup *et al.* **[50]** with kind permission from Springer Science and Business Media.

When mapping RTs from systems A to B (that had a total runtime of 20 min), the predicted RTs had mean and median absolute errors of 0.18 min and 0.08 min, respectively, and 90 % of features had a deviation of less than 0.45 min. With this system, it is thus possible to re-

find features of interest in different chromatographic systems with an accuracy close to or a few times the instrumental “between batches” deviation; it should be noted, though that a larger error was observed for several features.

We developed this retention time mapping as a proof of concept. As no databases currently exist that systematically collect RT information the system cannot be used to identify compounds on a large scale. We did, however, use it to translate retention times from a previous version of our current chromatographic method system such that the RT information of standards recorded in our previous method could still be used to identify compounds in our new setup.

The potential usefulness of RT databases coupled with our mapping method was exemplified by the study discussed in section 5.1.1. In this study, it was necessary to identify six dipeptides. Because amino acids can be combined in many ways to give dipeptides with the same mass, we had to synthesize 25 dipeptides to cover all combinations. Had retention time information been available for these 25 dipeptides, nine could have been excluded safely using a model with similar error range as the one described above. This was even in a worst-case scenario where the possible compounds were chemically very similar and therefore had similar retention times. In other cases, the possible compounds might not be commercially available and synthesis not feasible.

In the future it should be sought to build the required framework for a compound database that include RT information and use it to build a system such that the RT of a compound can be predicted in any system as soon as it is known in another system.

4.6 RETENTION TIME PREDICTION

Since an extensive database of compounds that include retention times does not exist we investigated the possibility of predicting the retention time of a compound solely based on its structure.

Several different strategies for retention time (RT) prediction have been developed. One approach has been to project RTs from a system using isocratic conditions to a gradient system [76,77]. This approach is extremely accurate, but can only predict the RT of compounds previously characterized in the isocratic system. Solvents, column and column temperature must also be the same as originally used which severely limit general applicability.

Some systems rely on complex models based on physicochemical descriptors of compounds. In the MolFind system [78] the retention index was predicted from 33 Molconn topological molecular descriptors [79] and Creek *et al.* published a multiple linear regression (MLR) model for ZIC-HILIC column systems using six physiochemical descriptors (log*D*, negative and positive, number of rotatable bonds, number of phosphate groups and number of hydrogen bond donors divided by molecular weight) [80]. These systems share the characteristic that they require a large number of training compounds and the more complex models risk causing severe overfitting of models making them less generally applicable.

We, instead, developed a much simpler model [50]. The RT of a compound in a reversed phase elution system is primarily determined by the lipophilic/hydrophilic characteristics of the compound. The lipophilicity of a compound is traditionally quantified by the familiar log*P-value*. The log*P-value* is defined as the ratio of concentrations of un-ionized compound between a lipophilic solvent (traditionally octanol) and water:

$$logP_{octanol/water} = log\left(\frac{[solute_{un-ionized}]_{octanol}}{[solute_{un-ionized}]_{water}}\right)$$

The log*P-value* describes only the behavior of the un-ionized compound and is therefore determined at a pH where the compound is primarily un-ionized. This is not the condition present in the solvent systems used for LC, however, where many compounds are in the ionized form. Therefore the log*D* value is a more appropriate measure of the behavior in an

LC system. log*D* is defined as the pH dependent ratio of the sum of the concentrations of all forms of the compound in each of the two phases:

$$logD_{octanol/water,pH} = log\left(\frac{[solute_{un-ionized}]_{octanol}}{[solute_{un-ionized}]_{water} + [solute_{ionized}]_{water}}\right)$$

We decided to build a model solely based on predicted log*D* values to see if such a simple model would be efficient at all. We used ChemAxon's Jchem libraries (Marvin v5.9.0, 2012) to predict the log*D* value of 43 standard compounds and measured their retention time. For the calculation, the pH was set to 2.7 to match the 0.1 % solution of formic acid used in the chromatographic method. A calibration curve between observed RT and predicted log*D* was fitted using a monotonic increasing LOESS function [81]. The resulting calibration curve can be found in Figure 4.12.

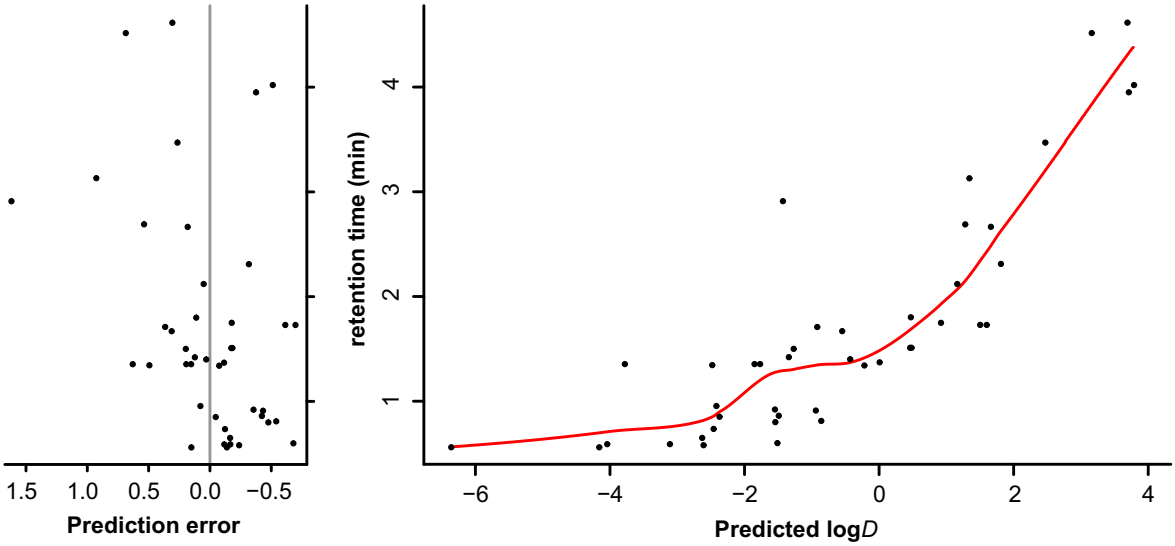


Figure 4.12. Retention times vs. predicted log*D* values.

This retention time prediction is clearly not nearly as accurate as the retention time mapping described in the previous section. Of the compounds identified using the complete pipeline, the mean and median absolute errors of the predicted RT were 0.35 min and 0.23 min, respectively in the 6 min gradient run. Ninety percent of the compounds had absolute errors of less than 0.7 min. The mean and median *relative* errors of the predicted RT were 15 % and 14 %, respectively. Ninety percent of the compounds had relative errors of less than 32 %.

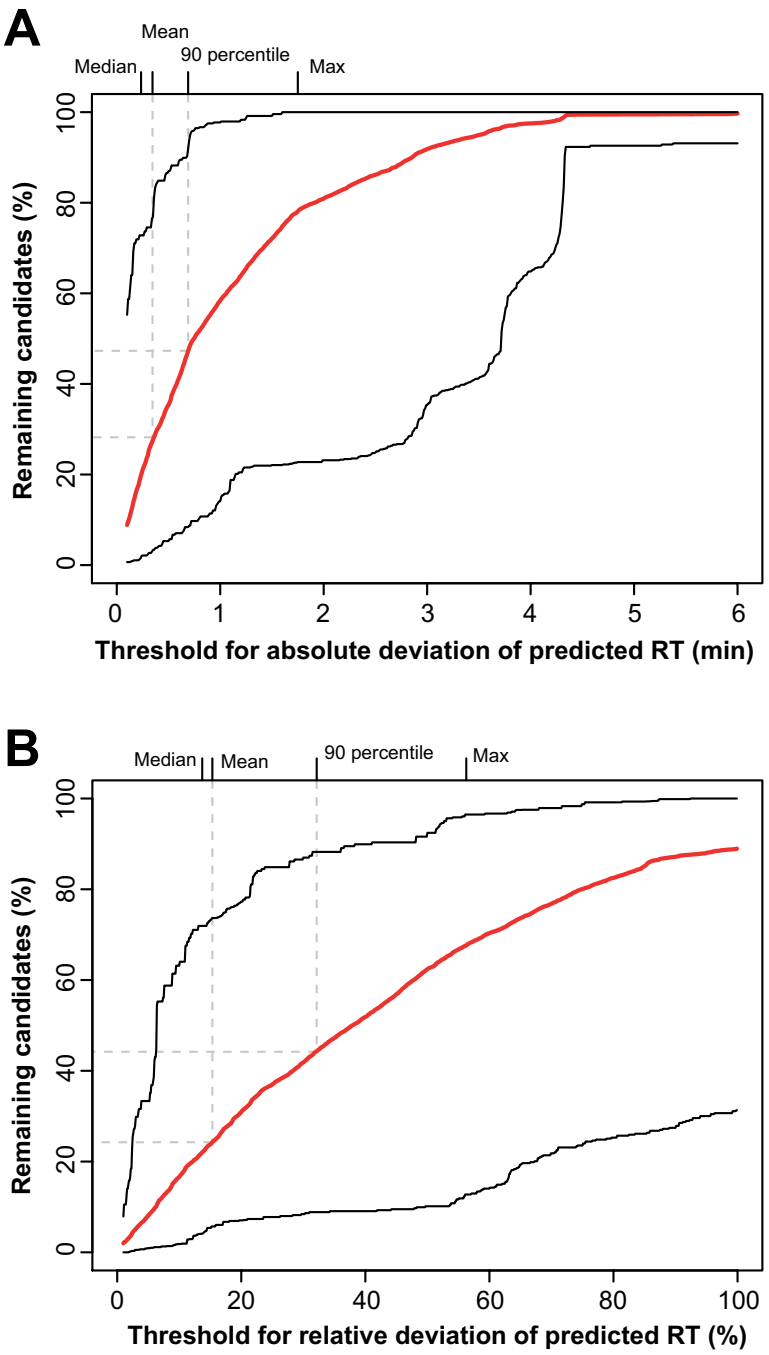


Figure 4.13.
A) Remaining molecular formulas as a function of the absolute error between observed and predicted retention time. The mean (red) between candidate lists as well as the minimum and maximum are shown. The median, mean, 90 percentile and the maximum error observed for the identified compounds have been marked at the top of the plot.
B) Remaining candidates as a function of the allowed relative difference between predicted and observed retention time. The mean (red) between candidate lists as well as the minimum and maximum are shown. The median, mean, 90 percentile and the maximum error observed for the identified compounds have been marked at the top of the plot.
Figure reproduced after Stanstrup *et al*. **[50]** with kind permission from Springer Science and Business Media.

While the accuracy is certainly not high enough for use as reliable evidence for the identity of a compound it can be used to filter out unrealistic compound candidates with predicted retention times very deviant from the experimental retention time. With our prediction model 35 % of candidate structures could on average be disregarded if a filter was applied such that all candidates with a higher deviation than the maximum observed were disregarded. The number of candidates that can be removed based on RT is quite variable; in the current study, between 4 and 87 % of the candidates could be disregarded. Since the true maximum error of the prediction is unknown, we do not suggest using this approach for filtering but rather as an additional hint for selecting the most plausible candidates.

Frequently, high scoring candidate (based on *in silico* fragmentation) structures are chemically similar and as a consequence they will have similar predicted (and actual) RTs. There are cases though where compounds expected to fragment similarly will have highly dissimilar retention times. An example of this is the dipeptide Phe–Phe. Most of the high scoring candidates based on *in silico* fragmentation contain a carbamide group instead of the primary amine in Phe-Phe and the predicted RTs for those candidates are thus much longer than the observed RT, thereby allowing such candidates to be disregarded.

The error range we found with our model is similar to the one achieved by Creek *et al.* [80] using the MLR model described above who also found log*D* to be the most dominant predictor. This suggests that the inclusion of other features improve the accuracy of the RT prediction only slightly, if at all.

Since the model relies on an *in silico* prediction of log*D*, the accuracy of this prediction is imperative. We found that the studies investigating state-of-the-art prediction of *logP-values* show that for most compounds log*P* can be predicted to about 0.5 log units [82]. In our setup, this corresponds to an error of the predicted RT up to 0.5 min depending on the RT. The errors are thus in the range where the observed errors could be attributed to error in the log*D* prediction itself. No larger studies on the accuracy of log*D* prediction could be found in the literature. However, the log*D* prediction is commonly achieved by combining the predicted log*P*-value and the logarithm to the acid dissociation constant (pK_a) value. Inaccuracy of pK_a prediction influences log*D* prediction most when the pK_a of the compound is close to the pH of the solvent. However, we did not observe such a trend which suggests that the log*P*-value

is the critical step in building a reliable RT predictor. It therefore does not appear that there is much room for further development until such a time that the underlying physicochemical descriptors can be more accurately predicted and prediction of RT from a compound structure is therefore a challenge not currently solved to a high degree of accuracy.

4.7 SEMI-AUTOMATED IDENTIFICATION PIPELINE

In the previous sections in this chapter I have described the tools available for compound identification. In this section I describe the overall workflow for the semi-automated identification pipeline we developed and described in paper I. The pipeline is outlined in Figure 4.14 and this figure will serve as the point of reference through this section.

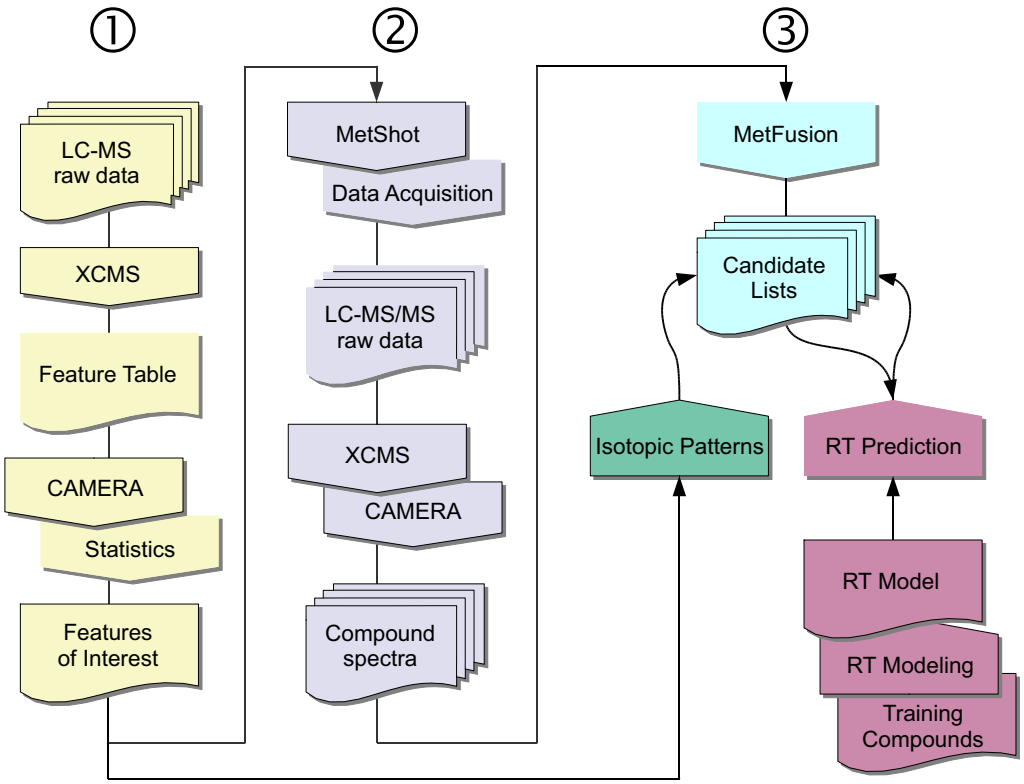


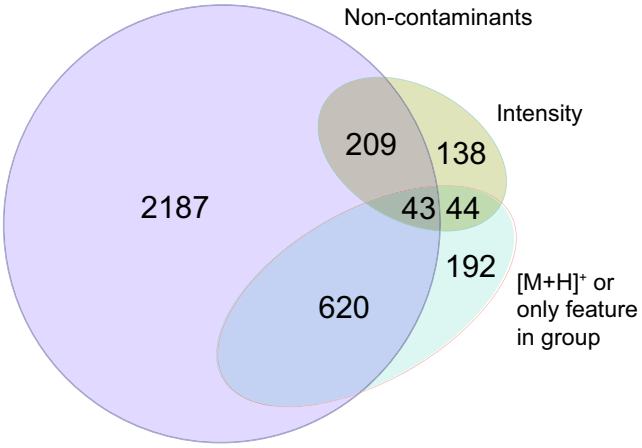
Figure 4.14. Identification pipeline. 1, The raw data is processed with XCMS, annotated with CAMERA, and the features of interest determined. 2, MetShot is used to generate MS/MS experiment files for use with the spectrometer that subsequently acquires the spectra. The spectra are processed, annotated, and combined across ionization energies. 3, MetFusion retrieves and ranks candidate compounds compatible with the acquired data. The observed isotopic pattern is compared with the theoretical pattern of all retrieved candidate compounds. Finally, the RTs of all candidates are predicted and compared with the observed RT.
Figure reproduced after Stanstrup *et al.* **[50]** with kind permission from Springer Science and Business Media.

For the first part of the pipeline (in yellow) the full dataset is pre-processed using XCMS (section 3.1), the features are grouped (section 4.1) and annotated with regards to fragments and adducts (section 4.2). In the final step features of interest are selected. In our case features were selected that:

1) were not detected as probable contaminants
2) had an intensity above 200 counts (arbitrary system dependent scale)
3) were annotated as the pseudo-molecular ion or is present in a feature group with only one feature

In a study setting where there is only interest in features of relevance to the exposure, a p-value limit or similar should be added to the selection of features.

Forty-three and 55 features in positive and negative ionization mode, respectively, fulfilled these requirements (Figure 4.15).



Contaminant features: 1173 (28 %)

Figure 4.15. Venn diagram of selection of features of interest in positive ionization mode. Ellipses are to scale according to the number of members in each group.

We selected our pre-processing settings to be able to detect very low intensity peaks, many of which are caused by contaminants. Unfortunately this inevitably leads to a large number of "false peaks" being detected. I find that two types of contaminant behaviors can be found. The first kind of contaminants show up as regular peaks in the chromatogram and are thus compounds retained by the LC column. However, they are present in the blank samples and are therefore rarely relevant to the matrix under investigation. I detected this kind of contaminants by selecting features that have more than 3 scans above 50 counts in a 0.2 min window around the retention time of the feature. The second kind of contaminants is contaminants not retained by the column to a high degree and reaching the detector continuously; either because they originate from the solvent or are released from the LC system. Such a contaminant causes a constant presence in the EIC. Therefore the chance that

the peak picking algorithm detects a "peak" at several points in the EIC is high and a relatively small number of contaminants cause a large number of contaminant features. I detected this kind of contaminants by selecting features where the EIC of the feature's *m/z* value is above 50 counts for 1 consecutive minute at any point in the chromatogram of the blank samples.

After features are selected for identification the appropriate MS/MS experiments need to be planned. Instead of painstakingly setting up the experiments manually in the spectrometer software we used the R package MetShot [83]. MetShot automatically generates the experiment files needed for the spectrometer and it optimizes the experiments such that the same injection can be used to monitor several ions at different retention times. This greatly reduces the number of injections needed. Instead of optimizing collision energies we opted to conduct all MS/MS experiments at 10, 20 and 30 eV to cover the optimum collision energy for most compounds' fragmentation. A screenshot of an optimized sequence of MS/MS experiments can be seen in Figure 4.16.

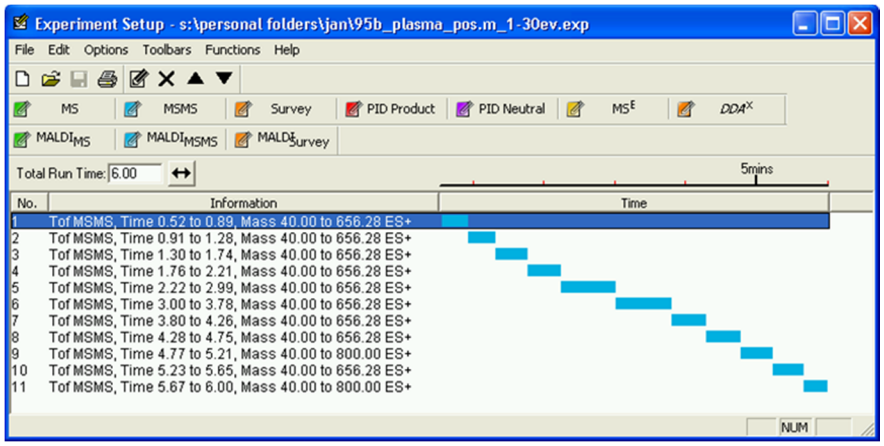


Figure 4.16. Summary of an optimized sequence of MS/MS experiments.

After the MS/MS data have been acquired the data is processed using XCMS and features again grouped using CAMERA to extract the spectrum of each compound.

These spectra are then queried using MetFusion (section 4.4). The resulting − typically thousands − candidate structures (compounds with similar mass in a compound library) are then ranked by the integrated (spectral similarity and spectra library match) MetFusion score.

MetFusion can be used manually using a web-interface available at msbi.ipb-halle.de/MetFusion. However, for the pipeline to be as automatic as possible we call the MetFusion Java libraries directly from R.

Next we import the MetFusion candidate lists into R so that we can add additional information to the results summary. We calculate scores for the fit of the experimental isotopic pattern to the theoretical pattern (section 4.3) and we compare the predicted retention time with the experimental retention time (section 4.6) for each compound candidate. An example of the output for the first four candidates from the analysis of the feature identified as tryptophan can be seen in Figure 4.17.

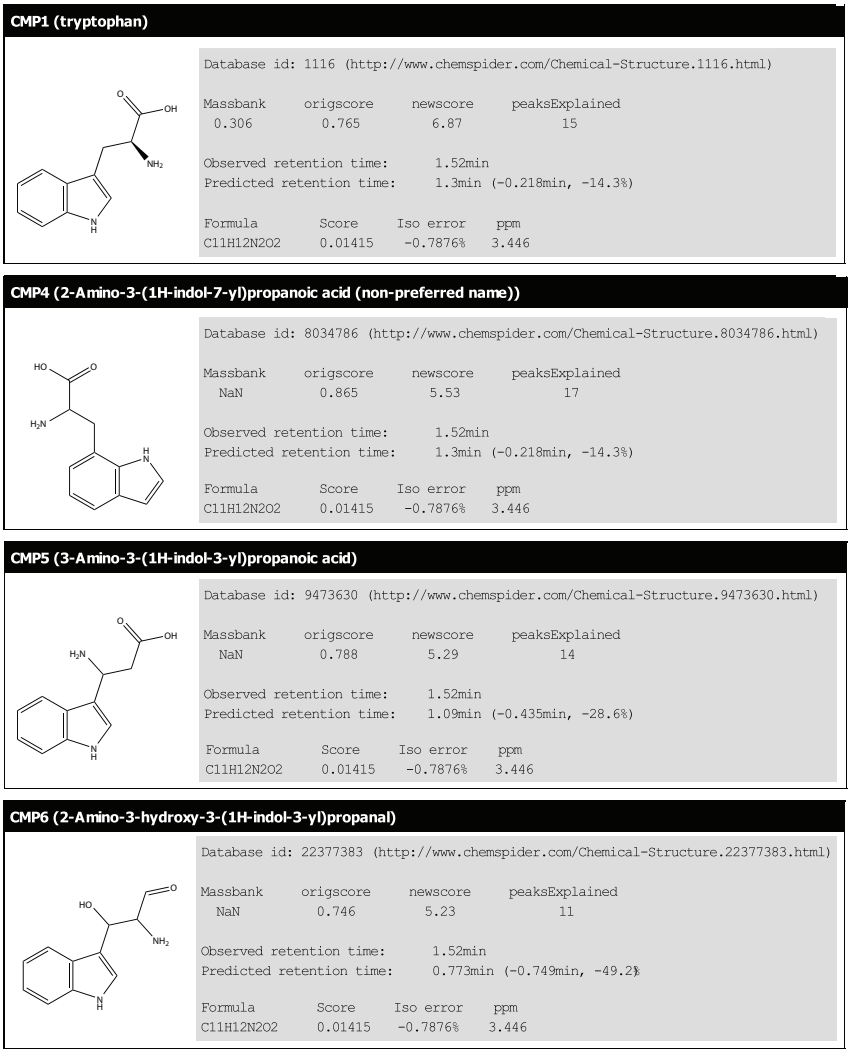


Figure 4.17. First four candidates associated with the feature identified as tryptophan from the output of the pipeline for semi-automatic identification. Origscore is the score attributed when only *in silico* fragmentation (MetFrag) is considered. Newscore is the combined MetFusion score. peaksExplained indicate how many peaks can be explained by *in silico* bond dissociation.

While the correct compound is ranked first of the candidates shown in Figure 4.17, the MetFusion score is not distinctly higher than for the other candidates. The fourth candidate, though, can be excluded based on the predicted retention time. Only the nature of the analyzed samples implies that probably tryptophan is the most likely compound. The next step is thus a comparison with the spectra of an authentic standard analyzed under the same conditions such that intensities of each fragment peak can be compared.

It is unfortunately not always the case that the nature of the sample indicates which compounds are plausible. For example, several biologically distinct isomers of hydroxybutyric acid exist. These isomers will score similarly with MetFusion and even comparison with authentic standards is not guaranteed to allow a definite identification since the spectra of the isomers are close to identical.

The metabolomics standards initiative (MSI) has defined four levels of identification for non-novel compounds, ranging from the highest level 1 if at least two properties (such as mass and retention time) are compared relative to an authentic compound analyzed under identical experimental conditions, or level 2 for *putatively* annotated compounds if only database information or literature values reported for authentic samples are available. Identifications at level 3 are only accurate down to the compound class, while level 4 covers unknown, but quantifiable, compounds [84].

These definitions have been adapted by the metabolomics community. From the example of hydroxybutyric acid discussed above it is, however, clear that even MSI level identification 1 should *not* be taken to mean "identified beyond reasonable doubt" in LC-MS studies. For identification beyond a reasonable doubt it is necessary to consider all plausible isomers even if the spectrum of a standard has a very high degree of similarity to the spectrum of the unknown. The list of MetFusion candidates can be helpful in this task as isomers which are expected to have similar spectra are ranked similarly. It is then up to the researcher to decide which compound candidates are relevant in the context of the study. It is fair to mention that this problem mostly applies to very small molecules where few or no distinct fragments are formed. As has been pointed out by others there is a need to update current reporting standards and to make the certainty of an identification more clear in published works [85].

With the semi-automatic identification pipeline we were able to identify 17 and 27 compounds, respectively. The major reason for failure to identify a compound was insufficient intensity of the peaks from the MS/MS analysis (Figure 4.18).

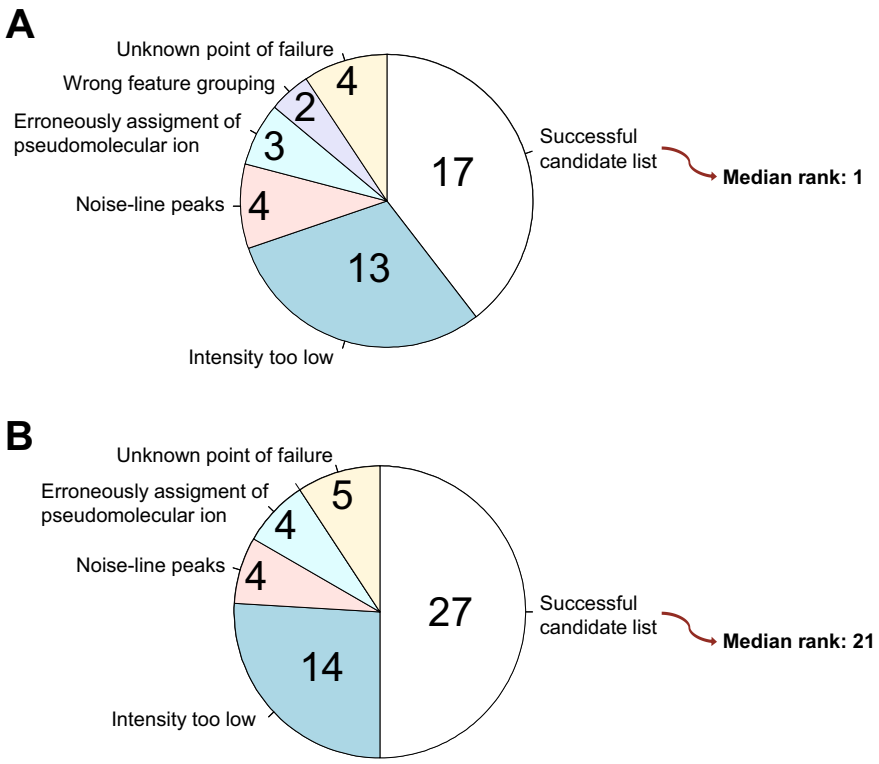


Figure 4.18. Recognized points of failure in the identification pipeline.

The complete lists of automatically identified compounds can be found in Table 4.7 and Table 4.8. The number of identified compounds is admittedly quite low. In Figure 4.15 it can be appreciated that the main cause of the low number of compounds that can be identified automatically using our pipeline is lack of sensitivity. Six hundred twenty additional features were annotated as the pseudo-molecular ion (positive ionization mode) but the intensities of the signals were lower than what would be required for MS/MS experiments. Many of these compound can be identified based on retention time – as indeed was done for the studies discussed in the next chapters – but not automatically until retention time mapping databases (section 4.5) become prevalent and robust. It should be noted that the newest generation of mass spectrometers offer dramatically improved sensitivity and I would expect that it is possible to automatically identify upwards of ten times as many compounds automatically with the currently most sensitive spectrometers.

Table 4.7. Identified or putatively identified compounds in positive mode.

Compound	m/z	RT (min)	MetFusion	Rank	RT pred (min)
Acetyl carnitine[1]	204.1225	0.92	0.235	22	0.68 (-0.24, -26.1 %)
Caffeine[1]	195.0877	1.68	1.35	1	1.36 (-0.33, -19.4 %)
Decanoyl carnitine	316.2489	3.11	0.284	2	1.36 (-1.76, -56.4 %)
Isoleucine[1,3]	132.1019	1.35	14.4	1	1.20 (-0.15, -11.1 %)
Leucine[1,3]	132.1019	1.35	14.4	2	1.16 (-1.19, -14.1 %)
LysoPC(16:0/0:0)[2]	496.3408	4.52	1.39	2	3.83 (-0.69, -15.3 %)
LysoPC(16:1/0:0)	494.3246	4.20	0.357	1	3.51 (-0.69, -16.5 %)
LysoPC(18:0/0:0)[2]	524.3724	4.92	0.519	3	4.48 (-0.44, -8.93 %)
LysoPC(18:1/0:0)[2]	522.3558	4.61	0.575	1	4.31 (-0.31, -6-61 %)
LysoPC(20:5/0:0)[1]	542.3252	4.10	0.524	1	3.80 (-0.30, -7.35 %)
LysoPC(22:6/0:0)	568.3410	4.30	0.643	2	4.27 (-0.02, -0.55 %)
Phe-Phe	313.1539	1.99	1.30	2	1.35 (-0.63, -31.9 %)
Phenylalanine[1]	166.0861	1.43	8.39	1	1.30 (-0.13, -9.33 %)
Proline[1]	116.0704	0.67	12.4	1	0.81 (+0.15, +22.4 %)
Tryptophan[1]	205.0979	1.52	6.87	1	1.30 (-0.22, -14.3 %)
Tyrosine[1]	182.0814	1.11	5.84	1	1.27 (+0.16, +14.4 %)
Uric acid[1]	169.0353	0.83	1.69	1	1.28 (+0.45, +53.8 %)
Valine[1]	118.0862	0.83	2.41	10	0.98 (+0.15, +17.8 %)

[1]Identified at MSI level 1 by comparison of retention time, accurate mass and MS/MS spectra.
[2]Identified at MSI level 1 by comparison of retention time and accurate mass (for these compounds MS/MS data of authentic standards was not acquired).
[3]Coeluting. Indistinguishable.
An extended summary giving compound accession numbers, the isotope analysis and complete results of MetFusion is supplied in table 1 in the supplementary material 2.

Table 4.8. Identified or putatively identified compounds in negative mode.

Compound	m/z	RT (min)	MetFusion	Rank	RT pred (min)
2-Hydroxyhexadecanoic acid	271.2275	5.07	0.0119	124	4.98 (-0.09, -1.82 %)
2-Hydroxyisovaleric acid[1]	117.0548	1.66	0.365	5	1.65 (-0.01, -0.57 %)
5α-Dihydrotestosterone sulfate[7]	369.1732	3.11	1.21	54	2.31 (-0.81, -25.9 %)
Chenodeoxycholic acid glycine conjugate[4]	448.3062	3.53	N/A[8]	21	3.30 (-0.23, -6.49 %)
Chenodeoxycholic acid glycine conjugate[5]	448.3062	3.61	N/A[8]	19	3.30 (-0.32, -8.79 %)
Chenodeoxyglycocholic acid[4]	448.3062	3.53	N/A[8]	6	3.30 (-0.23, -6.49 %)
Cresol sulfate[2]	187.0059	1.99	0.307	21	1.35 (-0.64, -31.9 %)
Deoxycholic acid glycine conjugate[4]	448.3062	3.53	N/A[8]	23	3.37 (-0.16, -4.41 %)
Deoxycholic acid glycine conjugate[5]	448.3062	3.61	N/A[8]	20	3.37 (-0.24, -6.76 %)
DHEA sulfate[6]	367.1575	2.95	0.241	14	2.00 (-0.95, -32.1 %)
eicosapentaenoic acid (20:5(n-3))[2]	301.2172	4.98	1.14	89	5.30 (+0.32, +6.49 %)
Epitestosterone sulfate or Testosterone sulfate (stereo isomers)[6]	367.1575	2.95	0.240	22	2.01 (-0.94, -31.9 %)
Etiocholanolone sulfate or Androsterone sulfate (stereo isomers)[7]	369.1732	3.11	1.21	54	2.31 (-0.81, -25.9 %)
Glycoursodeoxycholic acid[4]	448.3062	3.53	N/A[8]	28	3.30 (-0.23, -6.49 %)
Glycoursodeoxycholic acid[5]	448.3062	3.61	N/A[8]	24	3.30 (-0.32, -8.79 %)
Hippuric acid[2]	178.0497	1.81	1.12	1	1.69 (-0.13, -6.91 %)
Hydroxyphenyl sulfate	188.9853	1.58	0.307	6	1.35 (-0.23, -14.5 %)
Isoleucine[1,3]	130.0863	1.35	5.32	25	1.20 (-0.15, -10.8 %)
Leucine[1,3]	130.0863	1.35	5.59	23	1.16 (-0.18, -13.7 %)
LysoPC(0:0/18:1)[2]	566.3470	4.53	0.0828	142	4.31 (-0.23, -4.98 %)
LysoPC(14:0/0:0)	512.2992	4.08	0.0390	402	3.03 (-0.96, -23.5 %)
LysoPC(0:0/16:0)	540.3307	4.42	N/A[8]	1	3.83 (-0.60, -13.5 %)
LysoPC(18:1/0:0)[2]	566.3464	4.61	0.0678	135	4.31 (-0.30, -6.59 %)
LysoPC(20:5/0:0)	586.3150	4.10	0.367	17	3.80 (-0.62, -14.1 %)
LysoPC(22:6/0:0)	612.3305	4.29	N/A[8]	1	4.27 (-0.02, -0.55 %)
LysoPC(0:0/18:0)	568.3619	4.84	0.237	35	4.64 (-0.16, -3.49 %)
LysoPE(18:0)	480.3095	4.93	0.0287	19	4.60 (-0.34, -6.81 %)
Phenylalanine[2]	164.0704	1.43	1.65	1	1.30 (-0.13, -9.1 %)
Phenylsulfate	172.9903	1.65	0.316	6	1.32 (-0.34, -20.4 %)
Salicylic acid[1]	137.0235	2.47	8.61	2	2.54 (+0.07, +2.95 %)
Sphingosine 1-phosphate	378.2414	4.07	0.306	10	4.06 (-0.01, -0.25 %)
Tryptophan[1]	203.0813	1.52	2.64	1	1.30 (-0.22, -14.2 %)
Tyrosine[1]	180.0653	1.11	2.35	2	1.27 (+0.16, +14.5 %)
Uric acid[2]	167.0198	0.83	0.277	23	1.28 (+0.45, +54.3 %)
α-hydroxybutyric acid[2,3]	103.0395	1.39	1.02	20	1.49 (0.10, +7.51 %)
α-hydroxyisobutyric acid[2,3]	103.0395	1.39	0.786	32	1.46 (0.08, +5.47 %)
β-hydroxyisobutyric acid[2,3]	103.0395	1.39	1.03	23	1.40 (0.01, +1.06 %)

[1]Identified at MSI level 1 by comparison of retention time, accurate mass and MS/MS spectra.
[2]Identified at MSI level 1 by comparison of retention time and accurate mass (for these compounds MS/MS data of authentic standards was not acquired).
[3]Coeluting. Indistinguishable.
[4,5,6,7]Standards were not acquired and available data does not allow for distinguishing these candidates.
[8]No fragments could be matched in Massbank and the rank is thus based solely on Metfrag *in silico* fragmentation.
An extended summary giving compound accession numbers, the isotope analysis and complete results of MetFusion is supplied in table 2 in the supplementary material 2.

In positive ionization mode the median and mean ranks of the correct compound in the candidate list was 1 and 3, respectively. In negative ionization mode the median and mean ranks were 21 and 39, respectively. We note that the chemical diversity of the identified compounds is dominated by amino acids and phospholipids. We believe that this is due to their high abundance in serum and good ionizability rather than bias in our selection of features. However, our selection is dependent on the CAMERA annotation and the list of possible fragments was generated studying similar compounds.

With our data the approach performed impressively with positive ionization mode data, while for negative ionization mode the correct compound often had a very high rank. We believe that the lower performance for negative ionization mode is caused by lower coverage of negative ionization mode spectra in spectral databases. We therefore believe that further augmentations of spectral databases will profoundly increase the performance of our pipeline. It should also be noted that since the compounds selected were those with high intensities, and likely high concentrations in human plasma, their structure, or very similar structures, were often present in the spectral databases. The MetFusion approach was recently tested in the "Critical Assessment of Small Molecule Identification" (CASMI) challenge where it performed acceptably but the mean rank of the correct compound was much higher. This was due to the more exotic structure of the compounds studied, and MetFusion therefore cannot be expected to perform similarly in less well-defined matrixes such as plant extracts.

Chapter 5

WHEY PROTEIN

In this chapter three examples of nutritional metabolomics studies will be described. These studies seek to shed additional light on the molecular mechanism behind the effects of whey protein described in Chapter 1. Paper II and III describe the first two studies in full and the results will be summarized in the following. The results of the last study are briefly reported in section 5.3.

5.1 WHEY PROTEIN VS. OTHER PROTEIN SOURCES

As described in Chapter 1 it has been demonstrated that whey has a positive effect on postprandial insulin and hyperlipemia compared to other protein sources. So far the underlying biochemical processes have not been elucidated. Therefore, understanding the underlying mechanisms causing this effect might prove to be extremely valuable in the understanding, prevention and treatment of the metabolic syndrome that is closely interlinked with blood levels of both insulin and lipids.

We therefore analyzed the data collected by Holmer-Jensen *et al.* comparing whey protein to cod, gluten and casein in a meal study [10]. The results are summarized below, while further details can be found in paper II.

5.1.1 STUDY DESIGN

The study was conducted as a randomized, single-blinded, crossover meal study. All volunteers ingested an isocaloric test meals consisting of an energy-free soup added 100 g of butter and 45 g of one of four protein products. Forty five grams of carbohydrate was given in the form of white bread. The protein products compared were whey isolate (WI), casein (CAS) gluten (GLU) or cod (COD) protein. Refer to paper II for further details. The participants arrived fasted and blood samples were drawn immediately before intake of the meal (T = 0) and at 1, 2, 4 and 8 hours after T = 0 as depicted in Figure 5.1.

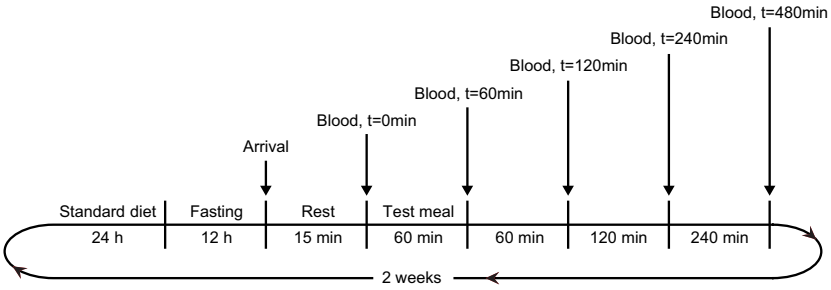


Figure 5.1. Study design. The participants arrived fasted and after a period of rest blood samples were drawn before intake of the meal (T = 0) and at 1, 2, 4 and 8 hours after T = 0. After 2 weeks the participants returned to test another meal until all participants had tested all four test meals.

5.1.2 RESULTS

The most unexpected result from the study was that gastric emptying was delayed after the WI meal. This is highly surprising since whey is usually considered a "fast" protein source since it is highly soluble in the gastric juices and is therefore rapidly emptied from the stomach when compared to other protein sources such as casein. Gastric emptying of the liquid phase was assessed using paracetamol (also called acetaminophen or acetyl-*p*-aminophenol, APAP) since it is passively and quickly absorbed only from the upper small intestine and gastric emptying thus determine the time taken to reach the absorption site [86].

The unexpected finding would in itself lead to questions regarding the applicability of APAP as a marker of gastric emptying; one might imagine that whey had some idiosyncratic interaction with APAP that lowered the uptake of APAP into the plasma. However, as was noted in the original study [10], the results are supported by the observation that the gastric inhibitory polypeptide (GIP) response also was lower for the first hours after the WI meal. GIP is released following direct interaction between meal components and the intestinal mucosa [87] and the lower response should therefore indicate that the gastric emptying indeed was delayed. See Figure 5.2 for the kinetic profiles of APAP and GIP.

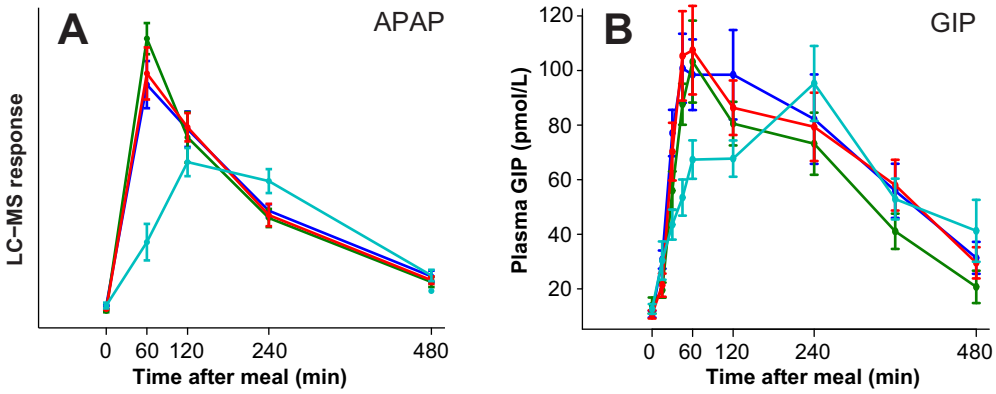


Figure 5.2. Kinetic plasma profiles of plasma APAP (A) and GIP (B) after intake of casein (●), cod (●), gluten (●) and whey (●). Error bars represent the standard errors of the mean (± SEM).

Given the slower gastric emptying of WI it would be expected that the amino acids (AAs) and other components would also be absorbed slower into the blood.

Paradoxically, we instead found that the AAs in the present study were disproportionately increased postprandially, most notable for tryptophan as can be appreciated in Figure 5.3.

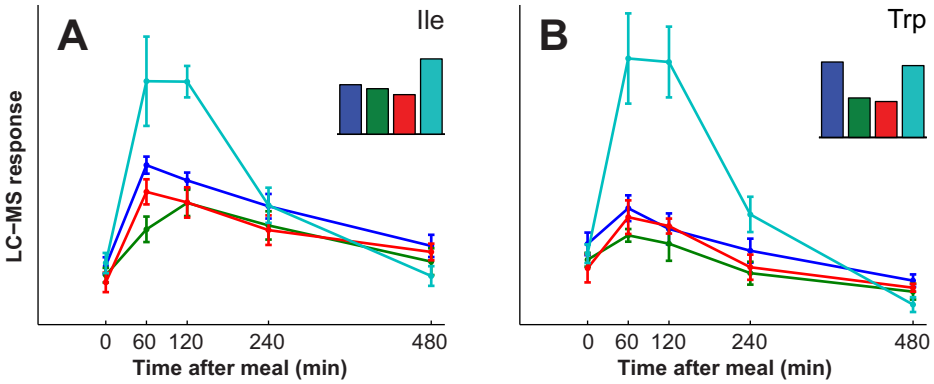


Figure 5.3. Kinetic plasma profiles of plasma isoleucine (A) and tryptophan (B) after intake of casein (●), cod (●), gluten (●) and whey (●). Error bars represent the standard errors of the mean (± SEM). The bar plots represent the relative content of the amino acids in each test meal.

It has previously been demonstrated that leucine plasma levels increase faster after intake of whey compared to after intake of casein [88]. However, in that study casein led instead to prolonged elevated levels of leucine as would be expected if casein had slower gastric emptying. It therefore appears that the "meal matrix" affects the relative gastric emptying of the different protein products.

From the perspective of obese and diabetics the most important result is the insulin secreting (see Figure 5.4) and consequent glucose lowering properties of the WI meal.

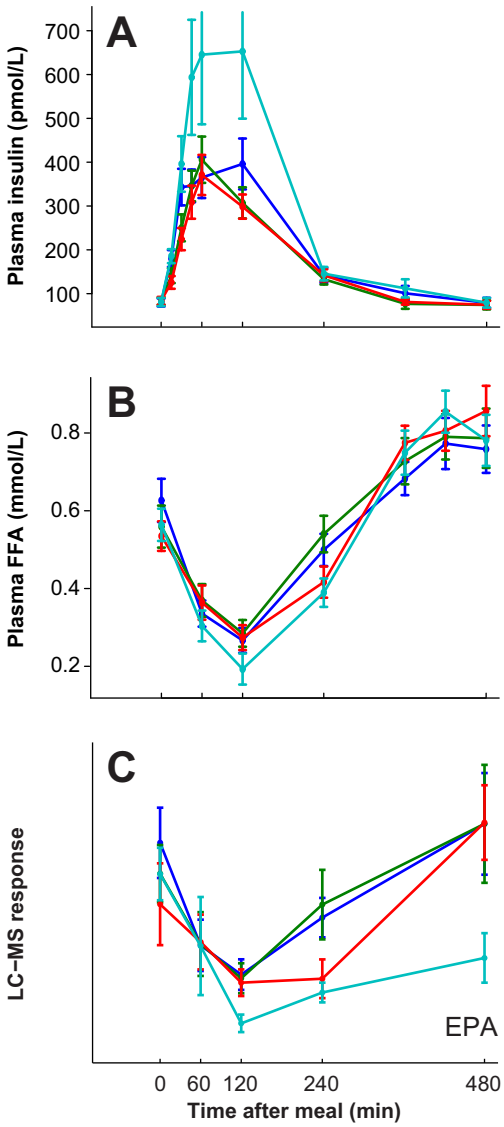


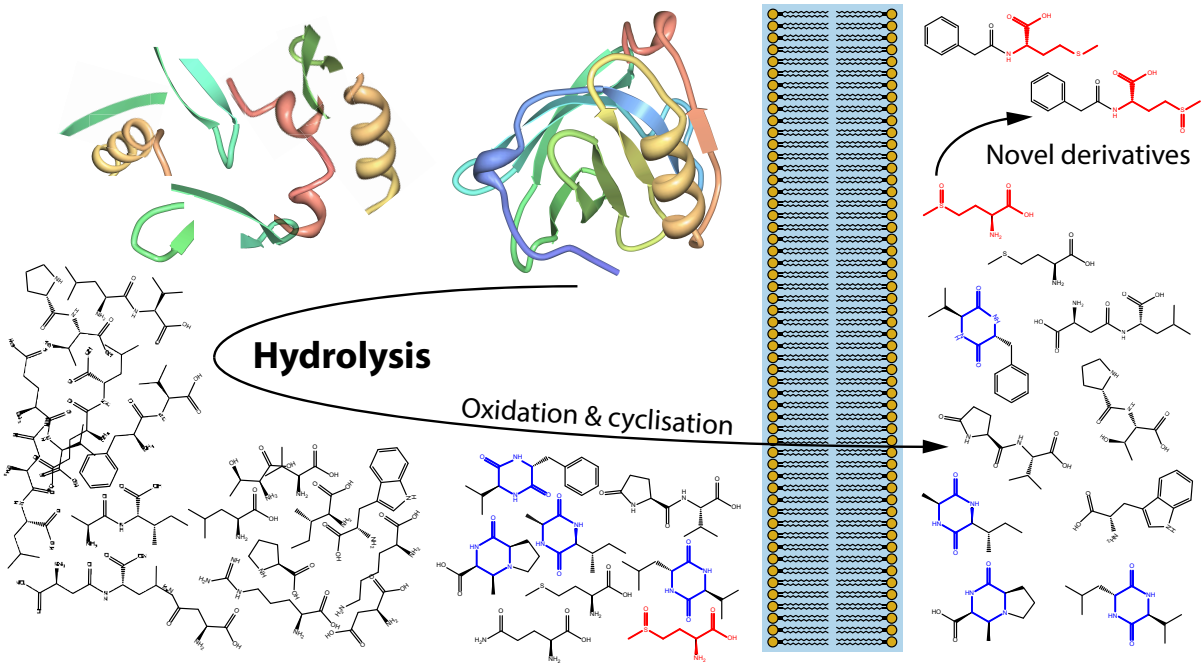
Figure 5.4. Kinetic plasma profiles of plasma insulin (A), total free fatty acids (B) and eicosapentaenoic acid (C) after intake of casein (●), cod (●), gluten (●) and whey (●). Error bars represent the standard errors of the mean (± SEM).

Our study suggests – as has also been suggested by others – that increased circulation of insulinotropic whey derived AAs are the source of this effect. Branched-chain amino acids (BCAAs), and particularly leucine, has been suggested as the cause of this effect [89]. All of the BCAAs were indeed higher for the WI meal (CAS: 19.4 %, COD: 17.8 %, GLU: 14.9

%, WI: 25.6 %). If total BCAAs were in themselves causing increased insulin secretion it is, however, then curious why insulin levels should not be markedly lower for the GLU meal that contained the lowest amount of BCAAs. It is possible that biological effects are amplified by the synergistic effects of AA combinations [90] or whey specific peptides (maybe BCAA containing), not measureable with our method.

We also found decreased levels of a number of fatty acids after the WI meal, likely as a consequence of the exaggerated insulin response. While the kinetic profile of some of the fatty acids (see Figure 5.4) where consistent with the profile of total free fatty acids measured in the first study, a number of other fatty acids did not show the initial decrease observed for total free fatty acids.

5.2 WHEY FRACTIONS



This study was conducted to find evidence for a possible mechanism behind the beneficial effects attributed to whey protein and to investigate if the effect could be isolated to a subfraction of whey.

Therefore a study using the same design as described in section 5.1.1 was conducted that compared whey isolate (WI) to products with enhanced proportions of α-lactalbumin (ALPH) and caseinoglycomacropeptide (CGMP), respectively. In additional, a whey hydrolysate (WH) product was included in the study.

This meal study was originally conducted to evaluate postprandial lipaemia in obese non-diabetic subject following intake of the different subfractions of whey protein. Outcome assessment on classical clinical variables showed less postprandial suppression of free fatty acids and a larger induction of net incremental area under the curve (iAUC) at 30 min of insulin for the WH meal compared to the other three meals, however, no difference could be found on the primary outcome variable, plasma triglycerides [91].

In the light of these results we focused on the compounds that distinguish WH from the other three meals. The results are summaries below.

5.2.1 RESULTS

All identified compounds are AAs or are associated with catabolism of specific AAs. A depiction of the pathway connections between all identified metabolites can be found in Figure 5.5.

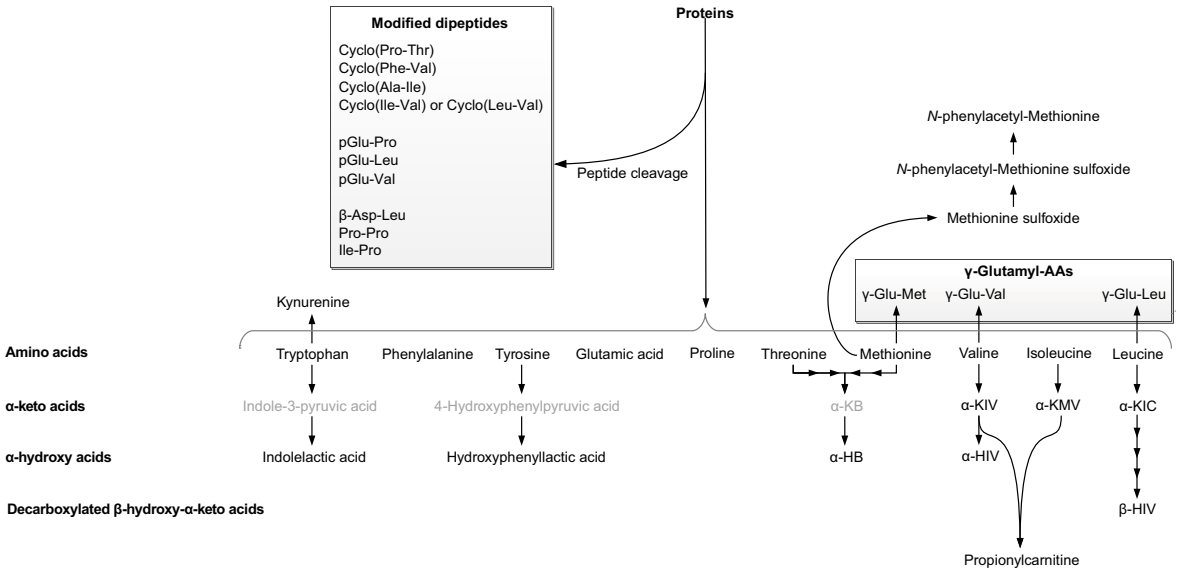


Figure 5.5. Pathway connection between metabolites found to be significantly different between meals. In grey are proposed intermediates not observed in this study. pGlu: pyroglutamyl, α-KB: α-ketobutyric acid, α-HB: α-hydroxybutyric acid, α-KIV: α-ketoisovaleric acid, α-HIV: α-hydroxyisovaleric acid, α-KMW: α-keto-3-methylvaleric acid, α-KIC: α-ketoisocaproic acid, β-HIV: β-hydroxyisovaleric acid.

One of the main results was that the plasma levels of each individual AA was highly correlated to the intake of this same AA, such that the higher the amount of an AA in the meal the higher the resulting postprandial plasma levels. This again raises the questions if a particularly high or low level of a single or few AAs are the cause of the effects observed for whey protein. As described in the previous section, BCAAs have previously been speculated to be the cause of the difference observed in insulin response between a whey and other protein sources, but the same argument has been made for a WH as compared to a WI product [89]. However, this study shows that plasma levels of none of the BCAAs are elevated for the WH meal compared to the other meals, as expected since the AA composition of WH is similar to that of WI (small differences arise because WH is made from whey concentrate, a product less concentrated in whey).

We hypothesize [92] instead that the effect of WH could be connected to a number of compounds we find exclusively for the WH meal.

Methionine sulfoxide (MetSO) was found at highly elevated levels in the plasma exclusively after ingestion of the WH meal (see Figure 5.6). MetSO is an oxidation product of methionine and it thus appears that the manufacturing process of WH caused oxidation of methionine which was not the case for the production of the other protein products. The WH is pasteurized at 85 °C and undergoes enzymatic hydrolysis for 12 hours at 53 °C [93]. These elevated temperatures and a slightly alkaline environment during hydrolysis likely contribute to the oxidation of methionine.

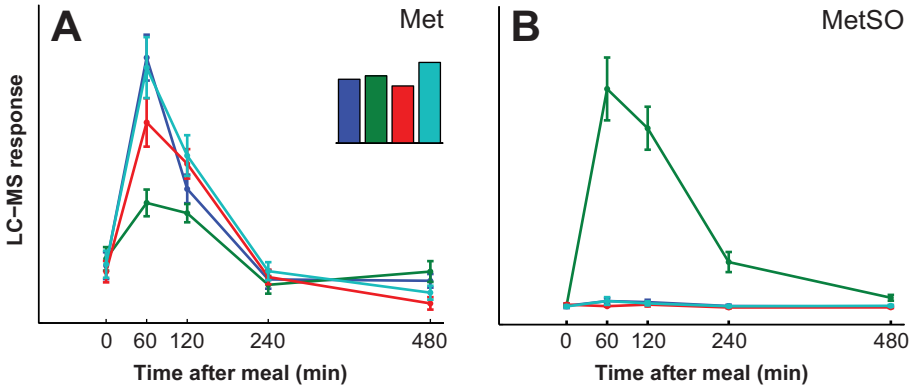


Figure 5.6. Kinetic profiles methionine (A) and methionine sulfoxide (B) after intake of caseinoglycomacropeptide (●), whey concentrate hydrolysate (●), α-lactalbumin (●) and whey isolate (●). Error bars represent the standard errors of the mean (± SEM). The bar plots represent the relative content of the amino acids in each test meal.

As a consequence of these elevated levels of MetSO we also found two MetSO metabolites, *N*-phenylacetyl-methionine (PAM) and *N*-phenylacetyl-methionine sulfoxide (PAMSO), elevated after the WH meal (see Figure 5.7). The discovery of these compounds was highly surprising since neither have previously been reported in biological systems.

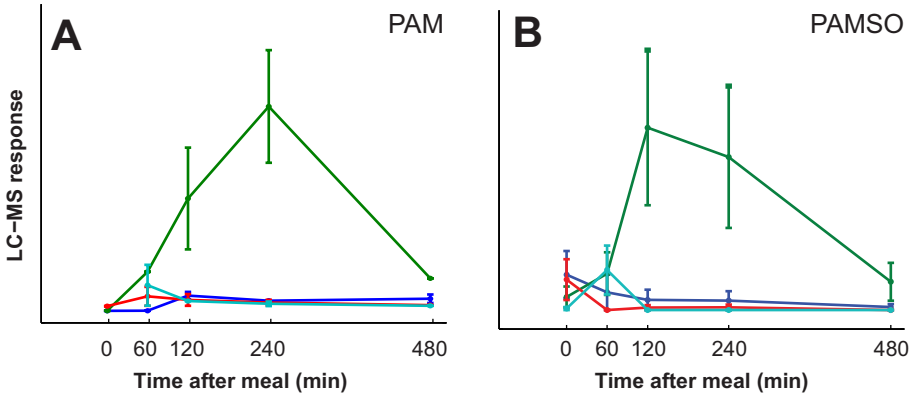


Figure 5.7. Kinetic profiles *N*-phenylacetyl-methionine (A) and *N*-phenylacetyl-methionine sulfoxide (B) after intake of caseinoglycomacropeptide (●), whey concentrate hydrolysate (●), α-lactalbumin (●) and whey isolate (●). Error bars represent the standard errors of the mean (± SEM).

In the literature I found that only one *N*-phenylacetyl amino acid conjugate was known, namely *N*-phenylacetyl-*glutamine* (PAG). PAG is formed by transfer of phenylacetyl from phenylacetyl-CoA to glutamine by glutamine *N*-phenylacetyltransferase (GPAT, EC 2.3.1.14) [94] (see Figure 5.8A) and PAG is a well-known [95] minor pathway of N-excretion in humans and is highly increased in uremia. PAG has also recently been detected using a metabolomics approach [96].

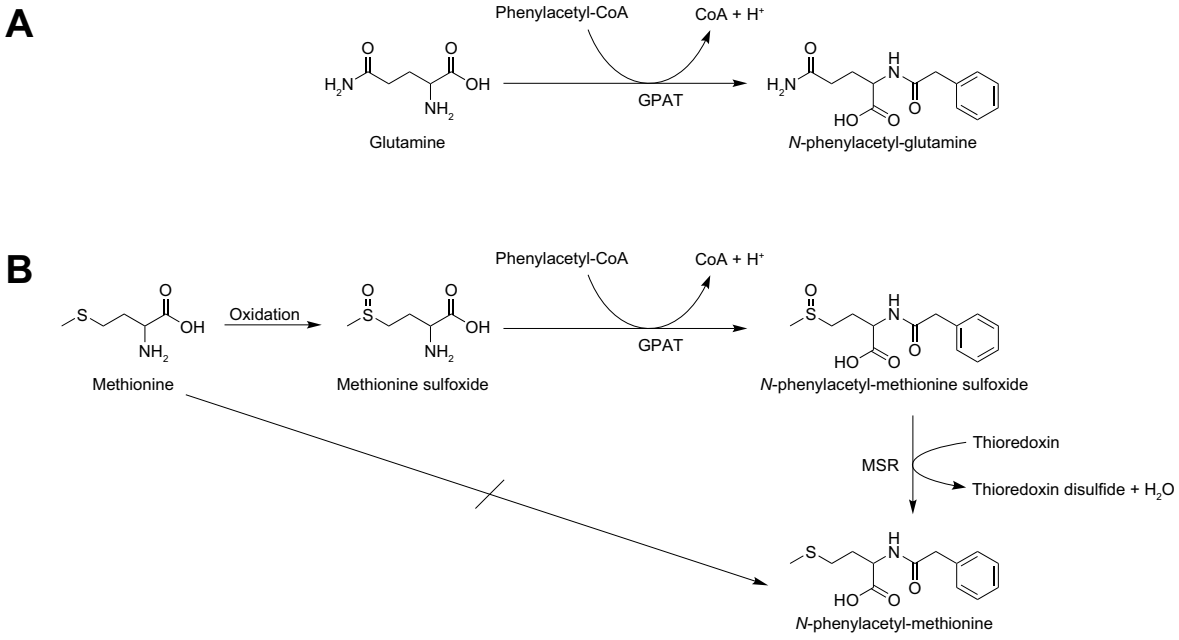


Figure 5.8. Known pathway of *N*-phenylacetyl-glutamine (A) and proposed pathway for *N*-phenylacetyl-methionine and *N*-phenylacetyl-methionine sulfoxide (B). MSR: peptide methionine sulfoxide reductase. GPAT: glutamine *N*-phenylacetyltransferase.

We propose [92] that GPAT can also utilize MetSO, but not methionine, as a substrate (see Figure 5.8B). If GPAT accepted methionine, then PAM would not exclusively be found elevated after the WH meal. Only as a consequence of elevated MetSO levels do the levels of PAMSO and PAM increase. Considering that glutamine has a higher structural similarity to MetSO than methionine this is reasonable. Following PAMSO generation by this pathway, PAM must be created by reduction of PAMSO; possibly by peptide methionine sulfoxide reductase (EC 1.8.4.11), known to reduce methionine sulfoxide in peptides.

MetSO has previously been shown to bind to the glutamine binding site of glutamate synthase (EC 2.6.1.53) [97] and increased levels of MetSO thus appear to potentially have biological implications. The consequences of elevated levels of PAM and PAMSO for the time being regrettably remain unknown. Future studies should confirm the pathway *in vitro* and animal models may be used to reveal any biological effect. Since MetSO occurs naturally in human plasma then PAM and PAMSO probably exist at very low levels too. For most of the non-WH samples these compounds could not, however, be detected at all with our instrument.

Another group of compounds elevated after the WH meal was a series of compounds of cyclic dipeptides. These compounds were found also to be present in the WH product and hence of exogenous origin. The identified dipeptides can be categorized in two groups, 2,5-diketopiperazines (DKPs) and pyroglutamyls.

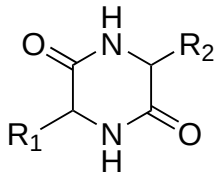


Figure 5.9. Structure of 2,5-diketopiperazines. R_1 and R_2 denote the side-chain of the amino acid moieties.

DKPs are formed from peptides by reaction between the free alpha amine in one AA and the alpha carboxylic acid in the adjacent AA. This results in a central diketopiperazine ring with two AA side-chains bound to the central ring (see Figure 5.9). DKPs are regarded as secondary metabolites and generated as a side-product of terminal peptide cleavage [98]. We identified four DKPs namely cyclo(Pro-Thr), cyclo(Phe-Val) and cyclo(Ala-Ile) and finally

a compound identified as either cyclo(Leu-Val) or cyclo(Ile-Val) which could not be discriminated by retention time nor fragmentation.

All DKPs identified in this study have previously been found in different organisms and in various food stuffs. DKPs have also been investigated in chicken essence and it was found that the concentration was determined not only by the relative ease of formation and stability but also by the thermal processing used [99]. As for MetSO we believe that the manufacturing of the WH caused the formation of the DKPs.

We detected the DKPs at very low levels in the plasma regardless of food intake but the levels are highly elevated after intake of WH. The plasma levels are essentially unaffected during the intervention for the other meals (see Figure 5.10). The inter-person variation and random experimental noise levels are high. Nonetheless T_{max} appears to be approximately 2 h for all four DKPs and they had not returned to baseline at the end of the intervention (8 h).

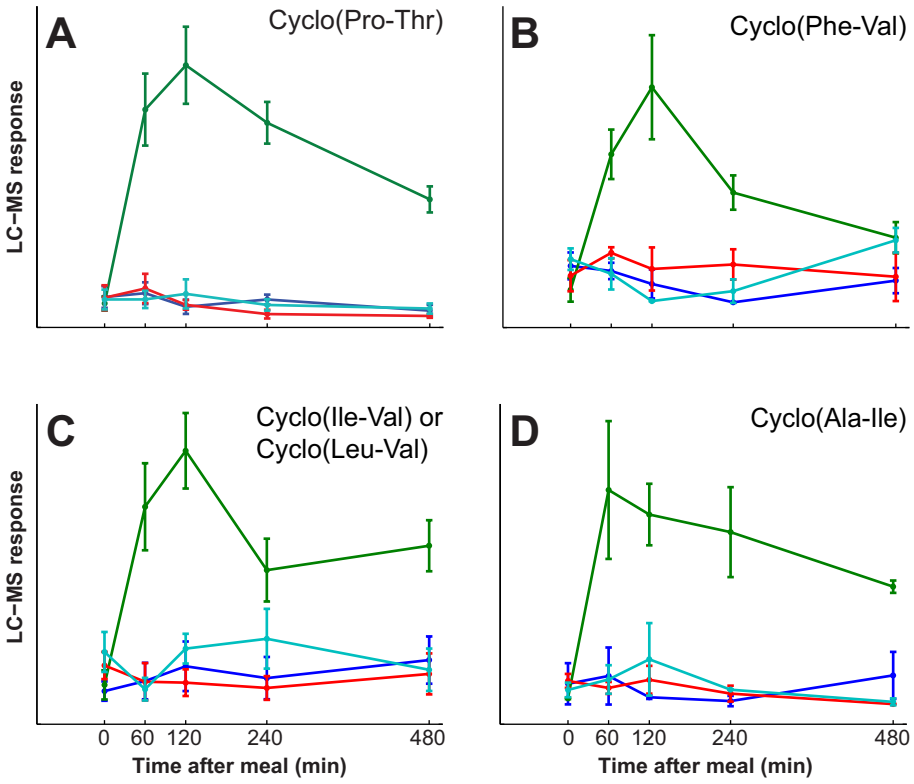


Figure 5.10. Kinetic profiles of cyclo(proline-threonine) (A), cyclo(phenylalanine-valine) (B), cyclo(isoleucine-valine) or cyclo(leucine-valine) (C) and cyclo(alanine-isoleucine) (D) after intake of caseinoglycomacropeptide (●), whey concentrate hydrolysate (●), α-lactalbumin (●) and whey isolate (●). Error bars represent the standard errors of the mean (± SEM).

This is, to my knowledge, the first report of DKPs being naturally present in human plasma or other mammalian plasma, whereas they have been studied as candidates for new pharmaceuticals. Nonetheless, hepatic enzymes that can metabolize DKPs are known to be present in mammals [100], suggesting that they are common human metabolites; likely just at such low concentrations that they hitherto have not been identified.

The endogenous functions of DKPs and the specific DKPs found in this study are largely unexplored and unknown. Conversely, biological activity of possible pharmaceutical interest has been extensively studied due to desirable physiochemical properties such as resistance to proteolysis, mimicking of peptide pharmacophores, controllable stereochemistry, conformational rigidity and donor and acceptor properties. In addition they share a common scaffold ideal for combinatorial chemistry [98]. In such a study cyclo(Phe-Val), also identified in our study, has been shown to have such exiting properties as antimalarial effect [101] and weak antibacterial activity [102], in addition to stimulating spinal cord repair [103]. In relation to our study the most interesting activity that has been reported for DKPs is probably the antihyperglycaemic effect reported for cyclo(His-Pro) in animal models leading to decreased postprandial glycaemia [104,105]. Even though cyclo(His-Pro) was not detected in this study similar effects of the DKPs identified could conceivably be involved in the difference observed for the WH meal compared to the other meals in regards to free fatty acids and insulin.

The second group of cyclic dipeptides consists of three pyroglutamyls: pGlu-Leu, pGlu-Val and pGlu-Pro. Pyroglutamyls are created from glutamic acid by reaction of the side-chain carboxylic acid with the alpha amine in the same glutamic acid moiety to create a 5-oxoproline ring structure.

Like for the DKPs there exists a basal level of the pyroglutamyls in the plasma of the subjects that rise markedly after intake of the WH meal – again apparently caused by the manufacturing process. It has previously been demonstrated that pyroglutamyls can be formed spontaneously simply by heating [106].

The three pyroglutamyls identified in this study have also previously been reported in different organisms and in foods such as cheese and wheat gluten [107,108]. In the cases of the latter two it could be speculated that the presence there is also a product of the manufacturing process.

In the body pyroglutamyls are probably formed as artifacts of protein hydrolysis. Little is known about the function of pyroglutamyls but the observation that enzymatic synthesis of pyroglytamyl takes places [109] and that pyroglutamyl-peptidases are present in many organisms [110] suggests that the residue may have biological and physiological functions. This is further supported by the fact that some bioactive peptides contain an *N*-terminal pyroglutamyl, most likely created by posttranslational modification [109].

T_{max} was 4 hours for pGlu-Pro, 2-4 hours for pGlu-Val and 2 hours for pGlu-Leu. Uniquely the plasma levels of pGlu-Leu start rising for all non-WH meals after 4 hours and peak at 8 hours or later (see Figure 5.11). This indicates that this particular compound can also be generated by the gut microbiota in the large intestine and is subsequently absorbed into the blood stream.

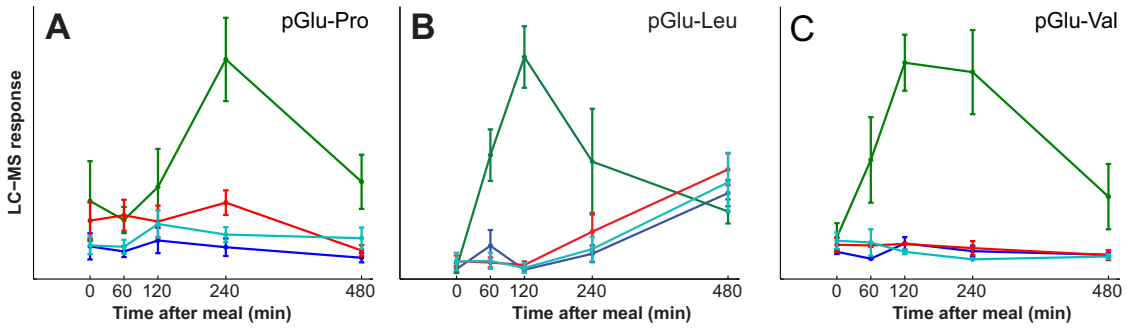


Figure 5.11. Kinetic profiles of pyroglutamyl-proline (A), pyroglutamyl-leucine (B) and pyroglutamyl-valine (C) after intake of caseinoglycomacropeptide (●), whey concentrate hydrolysate (●), α-lactalbumin (●) and whey isolate (●). Error bars represent the standard errors of the mean (± SEM).

Related in origin to the cyclic dipeptides are β-aspartates (also called isoaspartates). The process leading to β-aspartate is recognized as one of the most dominant causes of spontaneous protein damage under mild conditions [111]. β-aspartates are thought to be mainly created from aspartic acid through a succinimide intermediate [112], though it can also arise from deamidation of asparagine residues [113]. In this study we found a single linear β-aspartate i.e. β-Asp-Leu to also be highly elevated after the WH meal. T_{max} was 4 h (see Figure 5.12).

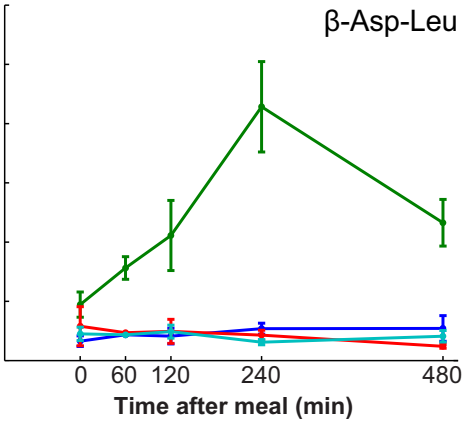


Figure 5.12. Kinetic profiles of β-aspartyl-leucine after intake of caseinoglycomacropeptide (●), whey concentrate hydrolysate (●), α-lactalbumin (●) and whey isolate (●). Error bars represent the standard errors of the mean (± SEM).

β-aspartyl dipeptides have previously been reported in human urine [114]. It has been reported that generation of β-aspartates in certain proteins in pharmaceuticals lead to potently increased immunogenicity [111]. It would clearly be of great concern if immunologically active β-aspartate containing peptides would enter the blood stream and the artificial generation should be avoided to limit health-risks, especially since hydrolyzed milk products are often prescribed to milk-intolerant infants. Immunologically active β-aspartates have, however, not been shown in dietary sources. On the contrary hydrolysis is generally considered to reduce antigenicity of milk proteins [115], as small molecules usually are not able to trigger an immune response.

Many of the peptides identified contain at least one BCAA and all contained at least one AA with a hydrophobic side-chain. This is likely not due to any endogenous biological process favoring these compounds but because the enzymes Alcalase and Neutrase were used for hydrolysis [93]. Both of these enzymes cleave preferentially next to AAs with hydrophobic side-chains.

Despite the possible candidates we have identified for the effects of WH additional studies will be required to substantiate these hypotheses. The differences observed on clinical parameters are small and it is entirely possible that none of the compounds we have identified are relevant for the effect. Effects of macromolecules, such as lipoproteins and hormones,

will not be capture by our metabolomics approach. It is also possible that our method does not have sufficient precision to reveal the relevant marker molecules.

No difference was found between the subfractions of whey. This could be because no difference exists such that the effects are not related to specific proteins but to shared properties such as solubility. It could also be because products with enhanced amounts of α-lactalbumin and caseinoglycomacropeptide were used, as opposed to pure fractions; this might have caused the effects to be obscured by biological and analytical variance.

5.3 DO OVERWEIGHT AND DIABETIC VOLUNTEERS RESPOND DIFFERENTLY?

A study identical in design to the one described in the previous section was performed with diabetics instead of obese non-diabetics. The clinical results have been described by Mortensen *et al.* [116]. The clinical results were very similar with the exception that the effect of WH on insulin response was still significant considering the iAUC of the whole study period, whereas for the obese there was only a significant difference for iAUC up to 30 min.

To investigate possible differences between obese and diabetics we carried out a metabolomics fingerprinting analysis identical to the one described for the study in the previous section. The statistical analysis resulted in 79 and 90 features, in positive and negative mode respectively, with different response between any two meals ($q < 0.05$). The features are reported in Table 5.1 and Table 5.2.

The compounds found at different levels between the test meals were almost identical. An additional cyclic dipeptide, cyclo(Ala-Leu) was found to be elevated after the WH meal, while cyclo(Phe-Val), pGlu-Pro, Ile-Pro, γ-glutamyl-valine, propionylcarnitine, and α-hydroxyisovaleric acid were not significantly different between meals for the diabetic subjects. I ascribe this difference only to the high experimental variation.

For the compounds with the same effect in both studies the analyses of two independent studies strengthen the validity of the results.

Unfortunately, the studies were not analyzed at the same time on the LC-MS instrument. This has the effect of the absolute responses not being directly comparable due to different sensitivity of the instrument and due to differences in ionization. This precludes comparison, for example, of baseline values between obese and diabetics which could otherwise have been valuable data.

Whether the difference in sensitivity only changes the scale relatively or also creates an offset between the scales is unknown. This also makes any comparison of the *relative* response spurious.

I attempted to estimate the half-life of selected compounds based on the kinetic profile and compared the two studies. The results were not consistent in either direction. Whether this is caused by the low temporal resolution of the kinetic profiles, the difference in the LC-MS response, or reflects actual biological conditions cannot be determined.

I also compared the relative increase in plasma concentrations for the identified compounds but found no differences between obese and non-diabetics.

Lastly, I investigated any correlation between plasma levels of any feature and the clinical baseline parameters such as BMI, waist circumference, hip circumference, fat free mass, fat percentage, blood pressure, urine albumin to creatinine ratio, hemoglobin, hemoglobin A1c, fasting plasma glucose, alanine transaminase, creatinine, triglycerides, cholesterol, LDL and HDL. Many features with a high correlation coefficient to a clinical parameter were found. However, upon investigating the distribution of the correlation coefficients it was clear that the correlation coefficients were normally distributed, as would be expected for random data. There was no over-representation of comparisons with high correlation coefficients (similar to the overrepresentation of low p-values, see section 3.4) and there was hence no indication that any correlation was not simply by chance. Since there are only eleven participants the correlation coefficients are based on only eleven data points, which is likely insufficient to reveal any true correlation. The low number of participants also makes other approaches, such as stratifying, senseless.

Table 5.1. Positive mode features significantly different between meals.

Compound	T_r	Formula	*m/z*	Calc. *m/z*	Suggestion ion	P-value (rank)[1]	Meal differentiation[2]
Cyclo(Pro-Thr)[4]	1.40		199.1084			2.1E-39 (1)	↑WH
Tryptophan[3]	1.52	$C_{11}H_{12}N_2O_2$	144.0808	144.0808	$[M-NH_3-CO_2+H]^+$	3.7E-38 (2)	ALPH=WI>WH>CGMP
			146.0601	146.0600	$[M-NH_3-CO-CH_2+H]+$	3.7E-37 (3)	
			159.0920	159.0917	$[M-HCOOH+H]^+$	7.2E-37 (4)	
			188.0708	188.0706	$[M-NH_3+H]^+$	2.8E-35 (5)	
			132.0811	132.0808	$[M-NH_3-CO-CO+H]^+$	8.8E-33 (6)	
			170.0609	170.0600	$[M-NH_3-H_2O+H]^+$	2.1E-32 (8)	
			205.0980	205.0972	**$[M+H]^+$**	1.1E-31 (9)	
			118.0668	118.0651	$[M-NH_3-CO-CO-CH_2+H]^+$	1.5E-26 (12)	
			409.1877	409.1870	$[2M+H]^+$	2.0E-22 (17)	
			130.0656	130.0651	$[M-NH_3-CO_2-CH_2+H]^+$	1.4E-14 (34)	
			245.1299	245.1285	$[M+(CH_3)_2CO-H_2O+H]^+$	4.0E-08 (55)	
			817.3615	817.3668	$[4M+H]^+$	9.1E-05 (70)	
Methionine Sulfoxide[3]	0.61	$C_5H_{11}NO_3S$	188.0362	188.0352	$[M+Na]^+$	1.3E-32 (7)	↑WH
			166.0534	166.0533	**$[M+H]^+$**	3.0E-20 (20)	
			210.0180	210.0171	$[M+2Na-H]^+$	2.2E-08 (52)	
	1.53		201.1271			4.2E-28 (10)	↑WH
Methionine[3]	0.94	$C_5H_{11}NO_2S$	104.0532	104.0529	$[M-HCOOH+H]^+$	5.8E-27 (11)	↓WH
			150.0586	150.0583	**$[M+H]^+$**	5.8E-26 (14)	
			133.0320	133.0318	$[M-NH_3+H]^+$	1.3E-24 (15)	
			87.0266	87.0263	$[M-NH_3-HCOOH+H]^+$	4.6E-21 (19)	
			102.0550	102.0550	$[M-CH_4S+H]^+$	4.3E-19 (23)	
			61.0118	61.0107	$[M-NH_3-C_2H_4-CO_2+H]^+$	2.1E-17 (27)	
Isoleucine[3,5]	1.37	$C_6H_{13}NO_2$	69.0702	69.0699	$[M-NH_3-HCOOH+H]^+$	5.3E-26 (13)	↑CGMP
	3.88		258.2802			1.6E-22 (16)	ALPH=WI>WH=CGMP
Cyclo(Ala-Leu)[4]	1.85	$C_9H_{16}N_2O_2$	185.1288	185.1285	**$[M+H]^+$**	7.1E-22 (18)	↑WH

Table continued on next page…

Table continued from previous page…

Tyrosine[3]	1.11	$C_9H_{11}NO_3$	91.0542	91.0543	$[M-NH_3-CO_2-CH_2O+H]^+$	1.6E-19 (21)	↓CGMP
			182.0812	182.0812	**$[M+H]^+$**	2.2E-19 (22)	
			119.0497	119.0492	$[M-NH_3-HCOOH+H]^+$	1.4E-18 (24)	
			147.0443	147.0441	$[M-NH_3-H_2O+H]^+$	6.0E-15 (31)	
			123.0440	123.0441	$[M-NH_3-CO-CH_2+H]^+$	1.4E-14 (35)	
			165.0541	165.0546	$[M-NH_3+H]^+$	2.9E-13 (41)	
			95.0493	95.0492	$[M-NH_3-CO-CO-CH_2+H]^+$	2.0E-09 (47)	
			136.0752	136.0757	$[M-HCOOH+H]^+$	2.3E-09 (48)	
γ-glutamyl-leucine[4]	1.61	$C_{11}H_{20}N_2O_5$	261.1439	261.1445	**$[M+H]^+$**	3.8E-18 (25)	WI>WH>ALPH>CGMP
			132.1016	132.1019	$[M-glu+H]^+$	8.6E-17 (28)	
			244.1210	244.1180	$[M-NH_3+H]^+$	1.6E-05 (65)	
Kynurenine[3]	1.41	$C_{10}H_{12}N_2O_3$	192.0660	192.0655	$[M-NH_3+H]^+$	4.7E-18 (26)	WI=ALPH>WH=CGMP
			94.0653	94.0652	$[M-NH_3-CO-CO-CO-CH_2+H]^+$	4.1E-13 (42)	
			136.0748	136.0757	$[M-NH_3-CO-CO+H]^+$	7.2E-12 (43)	
			209.0926	209.0921	**$[M+H]^+$**	3.8E-10 (46)	
			150.0557	150.0550	$[M-C_2H_5NO+H]^+$	1.1E-08 (51)	
Phenylalanine[3]	1.43	$C_9H_{11}NO_2$	103.0541	103.0543	$[M-NH_3-HCOOH+H]^+$	2.7E-15 (29)	WI=ALPH>WH>CGMP
			120.0811	120.0808	$[M-HCOOH+H]^+$	6.3E-15 (32)	
			107.0497	107.0492	$[M-C_2H_5NO+H]^+$	1.3E-14 (33)	
			131.0552	131.0492	$[M-NH_3-H_2O+H]^+$	1.7E-14 (36)	
			149.0607	149.0597	$[M-NH_3+H]^+$	2.0E-14 (37)	
			166.0860	166.0863	**$[M+H]^+$**	3.0E-10 (45)	
			93.0701	93.0699	$[M-NH_3-CO-CO+H]^+$	9.0E-09 (50)	
			331.1659	331.1652	$[2M+H]^+$	3.5E-07 (59)	
Threonine[3]	0.60	$C_4H_9NO_3$	164.0300	164.0294	$[M+2Na-H]^+$	4.4E-15 (30)	↑CGMP
			180.0045	180.0033	$[M+Na+K-H]^+$	6.8E-08 (56)	
			74.0601	74.0601	$[M-HCOOH+H]^+$	2.1E-06 (64)	
γ-glutamyl-methionine[4]	1.40		279.1013	279.1009	**$[M+H]^+$**	3.9E-14 (38)	WI=ALPH>CGMP>WH
β-Asp-Leu[4]	1.54	$C_{10}H_{18}N_2O_5$	247.1320	247.1289	**$[M+H]^+$**	6.0E-14 (39)	↑WH
	1.56		171.1139			7.7E-14 (40)	↑WH
Cyclo(Ile-Val) or cyclo(Leu-Val)[4]	2.30		213.1598	213.1598	**$[M+H]^+$**	8.2E-11 (44)	↑WH

Table continued on next page…

Table continued from previous page…

	3.53		230.2491			3.0E-09 (49)	↑WI
Proline[3]	0.66	$C_5H_9NO_2$	70.0651	70.0651	$[M\text{-}HCOOH+H]^+$	2.2E-08 (53)	↑CGMP
			116.0702	116.0706	**$[M+H]^+$**	1.1E-07 (57)	
	1.64		135.0559			3.6E-08 (54)	ALPH>WI>WH>CGMP
pGlu-Leu[4]	1.90	$C_{11}H_{18}N_2O_4$	243.1320	243.1340	**$[M+H]^+$**	1.8E-07 (58)	↑WH
Cyclo(Ala-Ile)[4]	1.79	$C_9H_{16}N_2O_2$	185.1292	185.1285	**$[M+H]^+$**	4.1E-07 (60)	↑WH
	1.38		281.1132			4.9E-07 (61)	↑WI
	0.85		347.0199			8.3E-07 (62)	↑ALPH
	1.34		447.1079			1.9E-06 (63)	WH>ALPH>CGMP>WI
	0.70		112.0764			2.1E-05 (66)	↑ALPH
	5.42		356.1855			5.2E-05 (67)	↑WI
	2.66		268.1031			5.7E-05 (68)	↑WH
	5.59		561.1679			8.0E-05 (69)	Uninterpretable
	1.50		367.1509			1.1E-04 (71)	ALPH>WI>WH=CGMP
	1.45		201.1213			1.6E-04 (72)	WI>ALPH>CGMP=WH
	5.25		291.2695			2.4E-04 (73)	Uninterpretable
	1.34		129.1025			3.5E-04 (74)	WH=WI>ALPH=CHMP
	1.35		722.0206			4.7E-04 (75)	↑WI
	0.79		160.1324			5.2E-04 (76)	Uninterpretable
	5.92		788.6648			5.8E-04 (77)	Uninterpretable
	4.93		109.1012			5.9E-04 (78)	Uninterpretable
	3.21		181.1223			6.2E-04 (79)	↑WH

[1]P-values (without correction) for the difference between meals (models described under statistics). In parenthesis is given the rank from lowest p-value (rank 1) to highest. In bold are the lowest ranked feature for each feature group (compound).
[2]The relative plasma levels are indicated. For the sake of simplicity the relationship might not in all cases reflect statistically significant differences. In some cases, especially for higher ranking features, interpretation can be obscured by experimental noise and inter-individual variation.
[3]Retention time and fragments confirmed by authentic commercial standards
[4]Retention time and fragments confirmed by synthesized standards
[5]It has been previously demonstrated that the 69.0706 fragment is much higher in isoleucine **[117–119]** and with our method standards showed it to be approximately 50 times higher for isoleucine than for leucine and is thus practically isoleucine specific. The monoisotopic peak and other fragments are nonspecific and must be considered a mixture of Leu and Ile.

Table 5.2. Negative mode features significantly different between meals.

Compound	T_r	Formula	*m/z*	*Calc. m/z*	Suggestion ion	P-value (rank)[1]	Meal differentiation[2]
Tryptophan[3]	1.52	$C_{11}H_{12}N_2O_2$	203.0814	203.0826	**[M-H]$^-$**	2.9E-38 (1)	ALPH=WI>WH>CGMP
			407.1714	407.1725	[2M-H]$^-$	2.6E-32 (2)	
			159.0919	159.0927	[M-CO_2-H]$^-$	1.1E-26 (4)	
			611.2621	611.2624	[3M-H]$^-$	4.1E-25 (5)	
			142.0656	142.0662	[M-NH_3-CO_2-H]$^-$	7.5E-18 (8)	
			239.0595	239.0593	[M+Cl]$^-$	1.0E-16 (10)	
			868.2649	868.2637	[4M+Fe^{3+}-4H]$^-$	1.1E-16 (11)	
			74.0247	74.0247	[M-C_9H_7N-H]$^-$	8.8E-14 (12)	
			429.1532	429.1522	[2M+Na-2H]$^-$	2.9E-10 (28)	
			186.0558	186.0560	[M-NH_3-H]$^-$	1.5E-09 (31)	
			445.1310	445.1283	[2M+K-2H]$^-$	3.7E-07 (48)	
			815.3496	815.3522	[4M-H]$^-$	3.4E-06 (52)	
			664.1739	664.1737	[3M+Fe^{3+}-4H]$^-$	1.1E-04 (66)	
			116.0490	116.0505	[M-NH_3-CO-CO-CH_2-H]$^-$	6.6E-04 (88)	
	3.23		257.1503			1.0E-28 (3)	↑WH
Tyrosine[3]	1.11	$C_9H_{11}NO_3$	163.0392	163.0400	[M-NH_3-H]$^-$	2.0E-19 (6)	↓CGMP
			180.0653	180.0666	**[M-H]$^-$**	5.0E-13 (13)	
			383.1217	383.1224	[2M+Na-2H]$^-$	7.0E-11 (26)	
			361.1399	361.1405	[2M-H]$^-$	1.3E-05 (56)	
	2.21		342.2397			6.1E-18 (7)	WI>WH>CGMP=ALPH
γ-glutamyl-leucine[4]	1.61	$C_{11}H_{20}N_2O_5$	128.0348	128.0353	[M-Leu-H]$^-$	6.9E-17 (9)	WI>WH>ALPH>CGMP
			259.1293	259.1299	**[M-H]$^-$**	8.8E-13 (15)	
			241.1168	241.1194	[M-H_2O-H]$^-$	2.8E-09 (34)	
Threonine[3]	0.60	$C_4H_9NO_3$	176.0086	176.0096	[M+NaCl-H]$^-$	5.8E-13 (14)	↑CGMP
			118.0501	118.0509	**[M-H]$^-$**	2.2E-10 (27)	
			74.0249	74.02473	[M-C_2H_4O-H]$^-$	3.8E-08 (42)	
Phenylalanine[3]	1.43	$C_9H_{11}NO_2$	164.0692	164.0717	**[M-H]$^-$**	1.8E-12 (16)	WI=ALPH>WH>CGMP
			147.0443	147.0451	[M-NH_3-H]$^-$	1.6E-11 (23)	
			72.0093	72.0091	[M-C_7H_8-H]$^-$	2.5E-08 (41)	
			351.1323	351.1326	[2M+Na-2H]$^-$	7.5E-07 (51)	
pGlu-Leu[4]	1.89	$C_{11}H_{18}N_2O_4$	241.1201	241.1194	**[M-H]$^-$**	2.2E-12 (17)	↑WH
	1.92		404.0406			2.4E-12 (18)	↑CGMP

Table continued on next page…

Table continued from previous page…

Methionine sulfoxide[3]	0.61	$C_5H_{11}NO_3S$	164.0377	164.0387	**[M-H]**$^-$	3.2E-12 (19)	↑WH
Methionine[3]	0.96	$C_5H_{11}NO_2S$	148.0433	148.0437	**[M-H]**$^-$	3.3E-12 (20)	↓WH
Hydroxyphenyllactic acid[3]	1.57	$C_9H_{10}O_4$	181.0496	181.0506	**[M-H]**$^-$	5.4E-12 (21)	WI-ALPH>WH>CGMP
			163.0390	163.0400	[M-H_2O-H]$^-$	1.6E-09 (32)	
N-phenylacetyl-methionine sulfoxide	1.83	$C_{13}H_{17}NO_4S$	282.0793	282.0805	**[M-H]**$^-$	5.7E-12 (22)	↑WH
N-phenylacetyl-Methionine[3]	2.65	$C_{13}H_{17}NO_3S$	266.0854	266.0856	**[M-H]**$^-$	2.0E-11 (24)	↑WH
			148.0428	148.0437	[M-phenylacetyl-H]$^-$	4.3E-08 (43)	
	1.40		161.9859			3.1E-11 (25)	WI>ALPH=CGMP>WH
			82.0295			3.4E-09 (36)	
	1.91		314.0465			3.1E-10 (29)	↑CGMP
pGlu-Val[3]	1.57	$C_{10}H_{16}N_2O_4$	227.1039	227.1037	**[M-H]**$^-$	4.3E-10 (30)	↑WH
	1.92		572.1549			1.6E-09 (33)	↑CGMP
	1.45		225.0821			2.9E-09 (35)	↑WH
α-ketoisovaleric acid[3,5,6]	1.51	$C_5H_8O_3$	115.0392	115.0400	**[M-H]**$^-$	4.5E-09 (37)	↑CGMP
	1.48		437.1597			7.5E-09 (38)	↑WH
Kynurenine[3]	1.41	$C_{10}H_{12}N_2O_3$	190.0500	190.0509	[M-NH_3-H]$^-$	1.9E-08 (39)	WI=ALPH>WH=CGMP
	1.92		478.0696			1.9E-08 (40)	↑CGMP
α-keto-3-methylvaleric acid[3,5]	1.88	$C_6H_{10}O_3$	358.0365		Unknown	5.6E-08 (44)	↑CGMP
			281.1001	281.1006	[2M+Na-2H]$^-$	7.8E-08 (46)	
			442.0930	442.0947	[3M+Fe^{3+}-4H]$^-$	9.5E-08 (47)	
			129.0544	129.0557	**[M-H]**$^-$	5.9E-05 (62)	
	1.83		164.0365			7.2E-08 (45)	↑WH
α-ketoisocaproic acid[3,5]	1.98	$C_6H_{10}O_3$	442.0927	442.0947	[3M+Fe^{3+}-4H]$^-$	5.9E-07 (49)	WI>WH>ALPH>CGMP
			129.0543	129.0557	**[M-H]**$^-$	2.1E-05 (59)	
Indolelactic acid[3]	2.17	$C_{11}H_{11}NO_3$	204.0657	204.0666	**[M-H]**$^-$	7.4E-07 (50)	WI=ALPH>WH=CGMP
	1.92		488.0984			3.7E-06 (53)	↑CGMP
	2.30		144.0446			4.3E-06 (54)	↓CGMP
Glutamic acid[3,7]	0.61	$C_5H_9NO_4$	146.0454	146.0459	**[M-H]**$^-$	6.8E-06 (55)	↑WH
	1.79		188.0913			1.4E-05 (57)	↑WH
γ-glutamyl-methionine[4]	1.41	$C_{10}H_{18}N_2O_5S$	277.0862	277.0863	**[M-H]**$^-$	1.5E-05 (58)	WI>CGMP=ALPH>WH
	1.77		243.1732			3.0E-05 (60)	↑WI
α-hydroxybutyric acid[3] or α-hydroxyisobutyric acid[3]	1.40	$C_4H_8O_3$	57.0352	57.03456	[M-HCOOH-H]$^-$	4.9E-05 (61)	CGMP>WH>WI>ALPH

Table continued on next page…

Table continued from previous page…

	5.68		772.5313			8.1E-05 (63)	Uninterpretable
	1.89		209.1151			9.8E-05 (64)	↑CGMP
	5.15		441.1034			1.1E-04 (65)	↑WI
	5.36		778.5595			1.5E-04 (67)	↑CGMP
	1.81		212.0010			1.9E-04 (68)	↓CGMP
	0.65		160.0586			2.0E-04 (69)	WI>ALPH=WH>CGMP
	1.65		151.0431			2.5E-04 (70)	↑WI
	1.91		394.0122			2.6E-04 (71)	Uninterpretable
	1.98		358.0357			2.7E-04 (72)	↓CGMP
	5.75		984.5838			2.7E-04 (73)	Uninterpretable
Valine[3]	0.83	$C_5H_{11}NO_2$	116.0706	116.0717	**[M-H]$^-$**	2.7E-04 (74)	↑CGMP
β-hydroxyisovaleric acid[3]	1.48	$C_5H_{10}O_3$	117.0549	117.0557	**[M-H]$^-$**	2.8E-04 (75)	Uninterpretable
	1.38		279.0987			2.8E-04 (76)	↑WI
	4.62		89.0235			2.9E-04 (77)	Uninterpretable
	1.90		242.0627			3.1E-04 (78)	↑CGMP
	1.72		591.2221			3.1E-04 (79)	Uninterpretable
	0.62		102.0551			3.9E-04 (80)	↑WH
	0.72		689.0007			4.7E-04 (81)	↑WH
	0.57		390.8914			5.3E-04 (82)	↑WH
	1.95		121.0290			5.7E-04 (83)	WI>WH=ALPH>CGMP
	0.94		357.0027			5.8E-04 (84)	↑WH
	2.05		192.0663			6.1E-04 (85)	↑WI
	5.04		464.3175			6.1E-04 (86)	↑CGMP
	5.52		384.9353			6.3E-04 (87)	Uninterpretable
	1.43		265.0040			7.8E-04 (89)	↑WI
	1.38		142.0504			9.6E-04 (90)	WI?WH>ALPH=CGMP

[1]P-values (without correction) for the difference between meals (models described under statistics). In parenthesis is given the rank from lowest p-value (rank 1) to highest. In bold are the lowest ranked feature for each feature group (compound).
[2]The relative plasma levels are indicated. For the sake of simplicity the relationship might not in all cases reflect statistically significant differences. In some cases, especially for higher ranking features, interpretation can be obscured by experimental noise and inter-individual variation.
[3]Retention time and fragments confirmed by authentic commercial standards
[4]Retention time and fragments confirmed by synthesized standards
[5]No fragments could be obtained by MS/MS.
[6]Isomers α-ketovaleric acid, levulinic acid and methylacetoacetic acid excluded from comparison of retention time.
[7]Retention time confirmed with authentic standard. Due to the very early elution of this compound unambiguous identification is not feasible.

Chapter 6

CONCLUSION

We did not succeed in finding highly specific exposure markers of whey as the effects were confined to modifying levels of endogenous metabolites. We did, however, show effect modification and temporal changes of markers of gastric emptying and a number of amino acids and fatty acids when compared to other protein sources.

Our investigation was not able to find any differences between the whey subfractions. Whey hydrolysate, on the other hand, contained unusual cyclic dipeptides which are, however, unlikely to be whey specific exposure markers but rather a result of the hydrolysis process. We hypothesize that these cyclic dipeptides may be causing the hypoglycaemic effects observed for the whey hydrolysate. In addition, we found that the manufacturing process for the hydrolysate caused methionine oxidation products, which were metabolized endogenously to novel metabolites.

A number of exposure and effect markers of fish intake were furthermore identified.

The finding that whey caused slower gastric emptying is in contrast to previous findings suggesting that whey is cleared faster than other proteins. Paradoxically, we also found disproportionately elevated levels and shorter T_{max} of some aromatic and branched-chain amino acids following the whey meal. This suggests that whey affects absorption of amino acids in a way independent from, or at least not wholly controlled by, gastric emptying. In addition, we find that whey caused decreased levels of a number of fatty acids due to increased insulin levels, which in turn is likely induced by the exaggerated amino acid levels.

We showed that the process of compound identification can be rationalized considerably by automatically determine the mass features most likely to represent the molecular ion, obtain an MS/MS spectrum of the same ion and then use a combination of mass spectral databases and *in silico* fragmentation (MetFusion) to create a ranked list of compound candidates. This list of candidates could be further reduced by filtering candidates with unrealistic *in silico* predicted retention times. We also showed that if the nature of the compound and the

spectrometer allows obtaining a spectrum with a reasonable number of fragments then the correct compound is in most cases highly ranked in the candidate list provided by MetFusion and a considerable amount of manual work is therefore saved.

Chapter 7

PERSPECTIVES

The developed semi-automatic pipeline for identification point to a future where the man-hours currently devoted to compound identification can be reduced dramatically. Further automation with improved ability to discriminate molecules is already in the works e.g. in the form of methods to analyze deep spectral trees.

Fully automatic identification is, however, a goal not in reach in the near future. One aspect gaining little attention is better use and sharing of retention time information. We showed that retention time mapping can be done quite accurately. Hence retention time prediction based on database values and retention time mapping could potentially have the ability to discriminate compounds. This possibility has not been exploited yet.

In the spirit of the hypothesis generating approach we employed, the meal studies put forth a number of questions that need to be answered before the effects of whey can be fully understood. It needs to be established which role the meal matrix plays in relation to the gastric emptying after intake of whey. It should also be established, for example by labelling studies, if the exaggerated plasma levels of amino acids after whey intake is purely due to a higher degree of absorption or if increased protein breakdown contribute to the high plasma levels. Studies with a higher temporal sampling resolution are also needed to accurately calculate AUC such that it can be assessed if the observed effects are a matter of temporal changes or if for example more branched-chain amino acids reach the blood stream.

In regards to the study of whey subfractions it is unfortunate that the study was not conducted with pure fractions. The modestly subfraction-enhanced products might have been too similar to make it possible to discriminate a biological difference that could have furthered the understanding of the mechanisms governing the effect of whey. Lastly the little known cyclic dipeptides need to be investigated further as they are currently primary candidates for the source of the effects of whey hydrolysate.

REFERENCES

1. Gapminder. Life expectancy and GDP per capita [Internet]. [cited 2013 May 15]. Available from: http://www.gapminder.org/data/

2. Edwards RD. Public transit, obesity, and medical costs: assessing the magnitudes. Prev Med. 2008 Jan;46(1):14–21.

3. Centers for Disease Control and Prevention. Long-term trends in diabetes [Internet]. [cited 2013 May 15]. Available from: http://www.cdc.gov/diabetes/statistics/slides/long_term_trends.pdf

4. Jia H, Lubetkin EI. Trends in Quality-Adjusted Life-Years Lost Contributed by Smoking and Obesity. Am J Prev Med. 2010 Feb;38(2):138–44.

5. Das UN. Obesity: Genes, brain, gut, and environment. Nutrition. 2010 May;26(5):459–73.

6. Astrup A. The Role of Dietary Fat in Obesity. Semin Vasc Med. 2005 Feb;5(1):40–7.

7. Skov AR, Toubro S, Rønn B, Holm L, Astrup A. Randomized trial on protein vs carbohydrate in ad libitum fat reduced diet for the treatment of obesity. Int J Obes Relat Metab Disord J Int Assoc Study Obes. 1999 May;23(5):528–36.

8. Pal S, Ellis V, Dhaliwal S. Effects of Whey Protein Isolate on Body Composition, Lipids, Insulin and Glucose in Overweight and Obese Individuals. Br J Nutr. 2010;104(05):716–23.

9. Luhovyy BL, Akhavan T, Anderson GH. Whey Proteins in the Regulation of Food Intake and Satiety. J Am Coll Nutr. 2007 Dec 1;26(6):704S–712.

10. Holmer-Jensen J, Mortensen LS, Astrup A, de Vrese M, Holst JJ, Thomsen C, et al. Acute differential effects of dietary protein quality on postprandial lipemia in obese non-diabetic subjects. Nutr Res. 2013 Jan;33(1):34–40.

11. Marshall K. Therapeutic applications of whey protein. Altern Med Rev J Clin Ther. 2004 Jun;9(2):136–56.

12. Keogh JB, Woonton BW, Taylor CM, Janakievski F, Desilva K, Clifton PM. Effect of glycomacropeptide fractions on cholecystokinin and food intake. Br J Nutr. 2010 Jul;104(2):286–90.

13. Astrup A, Cosentino LM, Grunwald GK, Storgaard M, Saris W, Melanson E, et al. The Role of Dietary Fat in Body Fatness: Evidence from a Preliminary Meta-Analysis of Ad Libitum Low-Fat Dietary Intervention Studies. Br J Nutr. 2000;83(Supplement S1):S25–S32.

14. WHO International Programme on Chemical Safety. Environmental Health Criteria 155: Biomarkers and Risk Assessment: Concepts and Principles [Internet]. 1993. Available from: Retrieved from http://www.inchem.org/documents/ehc/ehc/ehc155.htm

15. Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. Trends Biotechnol. 1998 Sep 1;16(9):373–8.

16. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol. 2004 May 1;22(5):245–52.

17. Nicholson JK, Lindon JC, Holmes E. "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica. 1999 Jan;29(11):1181–9.

18. Fiehn O. Combining Genomics, Metabolome Analysis, and Biochemical Modelling to Understand Metabolic Networks. Int J Genomics. 2001;2(3):155–68.

19. Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrom Rev. 2007;26(1):51–78.

20. Hyotylainen T, Wiedmer S, editors. Chromatographic Methods in Metabolomics. The Royal Society of Chemistry; 2013.

21. Barri T, Dragsted LO. UPLC-ESI-QTOF/MS and multivariate data analysis for blood plasma and serum metabolomics: effect of experimental artefacts and anticoagulant. Anal Chim Acta. 2013 Mar 20;768:118–28.

22. Barri T, Holmer-Jensen J, Hermansen K, Dragsted LO. Metabolic fingerprinting of high-fat plasma samples processed by centrifugation- and filtration-based protein precipitation delineates significant differences in metabolite information coverage. Anal Chim Acta. 2012 Mar 9;718:47–57.

23. Holčapek M, Jirásko R, Lísa M. Recent developments in liquid chromatography–mass spectrometry and related techniques. J Chromatogr A. 2012 Oct 12;1259:3–15.

24. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Anal Chem. 2006 Feb 1;78(3):779–87.

25. Pluskal T, Castillo S, Villar-Briones A, Orešič M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics. 2010 Jul 23;11(1):395.

26. Frederiksen RB. Optimize Peak Detection & Integration with ApexTrack/Processing Theory [Internet]. Helsinki, Finland; 2011 [cited 2013 Nov 22]. Available from: http://www.waters.com/webassets/cms/library/docs/local_seminar_presentations/FI_NUT2011_A4_Optimize_Peak_Detection_and_Integration_ApexTrack_RBF.pdf

27. Castillo S, Gopalacharyulu P, Yetukuri L, Orešič M. Algorithms and tools for the preprocessing of LC–MS metabolomics data. Chemom Intell Lab Syst. 2011 Aug 15;108(1):23–32.

28. Gürdeniz G. Nutritional metabolomics: Data Handling Strategies – examples using metabolic states and trans-fat exposures. Department of Nutrition, Exercise and Sports, Faculty of Sciences, University of Copenhagen; 2012.

29. Gürdeniz G, Kristensen M, Skov T, Dragsted LO. The Effect of LC-MS Data Preprocessing Methods on the Selection of Plasma Biomarkers in Fed vs. Fasted Rats. Metabolites. 2012 Jan 18;2(4):77–99.

30. Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics. 2008 Nov 28;9(1):504.

31. Wang S-Y, Kuo C-H, Tseng YJ. Batch Normalizer: A Fast Total Abundance Regression Calibration Method to Simultaneously Adjust Batch and Injection Order Effects in Liquid Chromatography/Time-of-Flight Mass Spectrometry-Based Metabolomics Data and Comparison with Current Calibration Methods. Anal Chem. 2013 Jan 15;85(2):1037–46.

32. Redestig H, Fukushima A, Stenlund H, Moritz T, Arita M, Saito K, et al. Compensation for Systematic Cross-Contribution Improves Normalization of Mass Spectrometry Based Metabolomics Data. Anal Chem. 2009 Oct 1;81(19):7974–80.

33. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, et al. Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. Anal Chem. 2006 Jan 1;78(2):567–74.

34. Van der Greef J, Smilde AK. Symbiosis of chemometrics and metabolomics: past, present, and future. J Chemom. 2005 May;19(5-7):376–86.

35. Zou H, Hastie T, Tibshirani R. Sparse Principal Component Analysis. J Comput Graph Stat. 2006;15(2):265–86.

36. Rago D. Biomarker identification in metabolomics of dietary studies on apple and apple products [PhD thesis]. Department of Nutrition, Exercise and Sports, Faculty of Sciences, University of Copenhagen; 2013.

37. Gürdeniz G, Hansen L, Rasmussen MA, Acar E, Olsen A, Christensen J, et al. Patterns of time since last meal revealed by sparse PCA in an observational LC–MS based metabolomics study. Metabolomics. 2013 Oct 1;9(5):1073–81.

38. Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers R-JAN, van der Greef J, Timmerman ME. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics. 2011 Mar 16;21(13):3043 –3048.

39. Rago D, Mette K, Gürdeniz G, Marini F, Poulsen M, Dragsted LO. A LC–MS metabolomics approach to investigate the effect of raw apple intake in the rat plasma metabolome. Metabolomics. 2013 Dec 1;9(6):1202–15.

40. Berk M, Montana G. A Skew-t-Normal Multi-Level Reduced-Rank Functional PCA Model with Applications to Replicated `Omics Time Series Data Sets. arXiv:11120152 [Internet]. 2011 Dec 1 [cited 2012 Jul 30]; Available from: http://arxiv.org/abs/1112.0152

41. Dunn OJ. Multiple Comparisons among Means. J Am Stat Assoc. 1961;56(293):52–64.

42. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat. 1979 Jan 1;6(2):65–70.

43. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988 Dec 1;75(4):800–2.

44. Šidák Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. J Am Stat Assoc. 1967;62(318):626–33.

45. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995 Jan 1;57(1):289–300.

46. Benjamini Y, Hochberg Y. On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. J Educ Behav Stat. 2000 Mar 20;25(1):60–83.

47. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika. 2006 Sep;93(3):491 –507.

48. Benjamini Y, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing under Dependency. Ann Stat. 2001;29(4):1165–88.

49. Schweder T, Spjøtvoll E. Plots of P-values to evaluate many tests simultaneously. Biometrika. 1982 Dec 1;69(3):493–502.

50. Stanstrup J, Gerlich M, Dragsted LO, Neumann S. Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. Anal Bioanal Chem. 2013 Jun 1;405(15):5037–48.

51. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2011. Available from: http://www.R-project.org/

52. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. Anal Chem. 2012 Jan 3;84(1):283–9.

53. PLS Toolbox. Manson, USA.: Eigenvector Research, Inc.; 2010.

54. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, et al. HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res. 2013 Jan;41(Database issue):D801–807.

55. Smith C, Maille G, Want E, Qin C, Trauger S, Brandon T, et al. METLIN: A Metabolite Mass Spectral Database. Ther Drug Monit Dec 2005. 2005;27(6):747–51.

56. Kruve A, Kaupmees K, Liigand J, Oss M, Leito I. Sodium adduct formation efficiency in ESI source. J Mass Spectrom. 2013;48(6):695–702.

57. Nitrogen rule [Internet]. Wikipedia Free Encycl. 2013 [cited 2013 Oct 10]. Available from: http://en.wikipedia.org/w/index.php?title=Nitrogen_rule&oldid=544288681

58. Kind T, Fiehn O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics. 2007 Mar 27;8(1):105.

59. Böcker S, Letzel MC, Lipták Z, Pervukhin A. SIRIUS: decomposing isotope patterns for metabolite identification. Bioinformatics. 2009 Jan 15;25(2):218–24.

60. Senior JK. Partitions and Their Representative Graphs. Am J Math. 1951 Jul;73(3):663.

61. Kind T, Fiehn O. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics. 2006 Apr 28;7:234.

62. Gerlich M, Neumann S. MetFusion: integration of compound identification strategies. J Mass Spectrom. 2013;48(3):291–8.

63. Venables WN, Ripley BD. Modern Applied Statistics with S. Springer; 2002.

64. Oberacher H. Wiley Registry of Tandem Mass Spectral Data, MS for ID. Hoboken: John Wiley & Sons Inc.; 2011.

65. Yang X, Neta P, Simón-Manso Y, Kilpatrick L, Liang Y, Stein SE. Building a High Quality and Comprehensive Tandem Mass Spectral Library (Poster) [Internet]. Available from: http://www.nist.gov/mml/bmd/data/tandemmass-speclib.cfm

66. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010;45(7):703–14.

67. Kind T, Liu K-H, Lee DY, DeFelice B, Meissen JK, Fiehn O. LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat Methods. 2013 Aug;10(8):755–8.

68. Degroeve S, Martens L. MS2PIP: a tool for MS/MS peak intensity prediction. Bioinformatics. 2013 Sep 27;btt544.

69. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics. 2010 Mar 22;11(1):148.

70. Williams AJ. Chemspider: A Platform for Crowdsourced Collaboration to Curate Data Derived From Public Compound Databases. In: Ekins S, Hupcey MAZ, Williams AJ, editors. Collab Comput Technol Biomed Res. John Wiley & Sons, Inc.; 2011. p. 363–86.

71. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012 Jan;40(Database issue):D109–114.

72. Bolton EE, Wang Y, Thiessen PA, Bryant SH. Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities. In: Ralph A. Wheeler and David C. Spellmeyer, editor. Annu Rep Comput Chem. Elsevier; 2008. p. 217–41.

73. Peironcely JE, Rojas-Chertó M, Tas A, Vreeken R, Reijmers T, Coulier L, et al. Automated Pipeline for De Novo Metabolite Identification Using Mass-Spectrometry-Based Metabolomics. Anal Chem. 2013 Apr 2;85(7):3576–83.

74. Gerlich M. Kombinierte Strategien zur verbesserten Metabolit-Identifikation mittels Massenspektrometriedaten [PhD thesis]. [Leibniz Institute of Plant Biochemistry (IPB), Halle]: Martin-Luther-Universität Halle-Wittenberg; 2013.

75. Vaughan AA, Dunn WB, Allwood JW, Wedge DC, Blackhall FH, Whetton AD, et al. Liquid Chromatography–Mass Spectrometry Calibration Transfer and Metabolomics Data Fusion. Anal Chem. 2012 Nov 20;84(22):9848–57.

76. Boswell PG, Schellenberg JR, Carr PW, Cohen JD, Hegeman AD. Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles. J Chromatogr A. 2011 Sep 23;1218(38):6742–9.

77. Boswell PG, Schellenberg JR, Carr PW, Cohen JD, Hegeman AD. A study on retention "projection" as a supplementary means for compound identification by liquid chromatography–mass spectrometry capable of predicting retention with different gradients, flow rates, and instruments. J Chromatogr A. 2011 Sep 23;1218(38):6732–41.

78. Menikarachchi LC, Cawley S, Hill DW, Hall LM, Hall L, Lai S, et al. MolFind: A Software Package Enabling HPLC/MS-Based Identification of Unknown Chemical Structures. Anal Chem. 2012 Nov 6;84(21):9388–94.

79. Hall LM, Hall LH, Kertesz TM, Hill DW, Sharp TR, Oblak EZ, et al. Development of Ecom50 and Retention Index Models for Nontargeted Metabolomics: Identification of 1,3-Dicyclohexylurea in Human Serum by HPLC/Mass Spectrometry. J Chem Inf Model. 2012 May 25;52(5):1222–37.

80. Creek DJ, Jankevics A, Breitling R, Watson DG, Barrett MP, Burgess KEV. Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. Anal Chem. 2011 Nov 15;83(22):8703–10.

81. Scheder R. monoProc: strictly monotone smoothing procedure [Internet]. 2007. Available from: http://CRAN.R-project.org/package=monoProc

82. Tihanyi K, Vastag M. Solubility, Delivery and ADME Problems of Drugs and Drug-Candidates. Bentham Science Publishers; 2011.

83. Neumann S, Thum A, Böttcher C. Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data. Metabolomics. 2013 Mar 1;9(1):84–91.

84. Sumner L, Amberg A, Barrett D, Beale M, Beger R, Daykin C, et al. Proposed minimum reporting standards for chemical analysis. Metabolomics. 2007;3(3):211–21.

85. Salek RM, Steinbeck C, Viant MR, Goodacre R, Dunn WB. The role of reporting standards for metabolite annotation and identification in metabolomic studies. GigaScience. 2013 Oct 16;2(1):13.

86. Heading RC, Nimmo J, Prescott LF, Tothill P. The dependence of paracetamol absorption on the rate of gastric emptying. Br J Pharmacol. 1973 Feb;47(2):415–21.

87. Holst JJ. The Physiology of Glucagon-like Peptide 1. Physiol Rev. 2007 Oct 1;87(4):1409–39.

88. Boirie Y, Dangin M, Gachon P, Vasson M-P, Maubois J-L, Beaufrère B. Slow and fast dietary proteins differently modulate postprandial protein accretion. Proc Natl Acad Sci. 1997 Dec 23;94(26):14930–5.

89. Nilsson M, Holst JJ, Björck IM. Metabolic effects of amino acid mixtures and whey protein in healthy subjects: studies using glucose-equivalent drinks. Am J Clin Nutr. 2007 Apr 1;85(4):996–1004.

90. Floyd JC Jr, Fajans SS, Pek S, Thiffault CA, Knopf RF, Conn JW. Synergistic effect of certain amino acid pairs upon insulin secretion in man. Diabetes. 1970 Feb;19(2):102–8.

91. Holmer-Jensen J, Hartvigsen ML, Mortensen LS, Astrup A, Vrese M de, Holst JJ, et al. Acute differential effects of milk-derived dietary proteins on postprandial lipaemia in obese non-diabetic subjects. Eur J Clin Nutr. 2012;66(1):32–8.

92. Stanstrup J, Rasmussen JE, Ritz C, Holmer-Jensen J, Hermansen K, Dragsted LO. Intakes of whey protein hydrolysate and whole whey proteins are discriminated by LC–MS. Metabolomics.

93. Calbet JAL, Holst JJ. Gastric emptying, gastric secretion and enterogastrone response after administration of milk proteins or their peptide hydrolysates in humans. Eur J Nutr. 2004 Jun;43(3):127–39.

94. Moldave K, Meister A. Synthesis of Phenylacetylglutamine by Human Tissue. J Biol Chem. 1957 Nov 1;229(1):463–76.

95. Zimmerman L, Egestad B, Jörnvall H, Bergström J. Identification and determination of phenylacetylglutamine, a major nitrogenous metabolite in plasma of uremic patients. Clin Nephrol. 1989 Sep;32(3):124–8.

96. Roux A, Xu Y, Heilier J-F, Olivier M-F, Ezan E, Tabet J-C, et al. Annotation of the Human Adult Urinary Metabolome and Metabolite Identification Using Ultra High Performance Liquid Chromatography Coupled to a Linear Quadrupole Ion Trap-Orbitrap Mass Spectrometer. Anal Chem. 2012 Aug 7;84(15):6429–37.

97. Trotta PP, Platzer KEB, Haschemeyer RH, Meister A. Glutamine-Binding Subunit of Glutamate Synthase and Partial Reactions Catalyzed by This Glutamine Amidotransferase. Proc Natl Acad Sci U S A. 1974 Nov;71(11):4607–11.

98. Martins MB, Carvalho I. Diketopiperazines: biological activity and synthesis. Tetrahedron. 2007 Oct 1;63(40):9923–32.

99. Chen Y-H, Liou S-E, Chen C-C. Two-step mass spectrometric approach for the identification of diketopiperazines in chicken essence. Eur Food Res Technol. 2004;218(6):589–97.

100. Delaforge M, Bouillé G, Jaouen M, Jankowski CK, Lamouroux C, Bensoussan C. Recognition and oxidative metabolism of cyclodipeptides by hepatic cytochrome P450. Peptides. 2001 Apr;22(4):557–65.

101. Pérez-Picaso L, Olivo HF, Argotte-Ramos R, Rodríguez-Gutiérrez M, Rios MY. Linear and cyclic dipeptides with antimalarial activity. Bioorg Med Chem Lett. 2012 Dec 1;22(23):7048–51.

102. Cabrera GM, Butler M, Rodriguez MA, Godeas A, Haddad R, Eberlin MN. A Sorbicillinoid Urea from an Intertidal Paecilomyces marquandii. J Nat Prod. 2006 Dec 1;69(12):1806–8.

103. Wong JWJ, McPhail LT, Brastianos HC, Andersen RJ, Ramer MS, O'Connor TP. A novel diketopiperazine stimulates sprouting of spinally projecting axons. Exp Neurol. 2008 Dec;214(2):331–40.

104. Hwang IK, Go VLW, Harris DM, Yip I, Kang KW, Song MK. Effects of cyclo (his–pro) plus zinc on glucose metabolism in genetically diabetic obese mice. Diabetes Obes Metab. 2003;5(5):317–24.

105. Song MK, Hwang IK, Rosenthal MJ, Harris DM, Yamaguchi DT, Yip I, et al. Anti-Hyperglycemic Activity of Zinc Plus Cyclo (His-Pro) in Genetically Diabetic Goto-Kakizaki and Aged Rats. Exp Biol Med. 2003 Dec 1;228(11):1338–45.

106. Kasai T, Nishitoba T, Sakamura S. Transformation of Glutamyl Dipeptides by Heating in Aqueous Solution. Agric Biol Chem. 1983;47(11):2647–9.

107. Li H, Dang HT, Li J, Sim CJ, Hong J, Kim D-K, et al. Pyroglutamyl dipeptides and tetrahydro-β-carboline alkaloids from a marine sponge Asteropus sp. Biochem Syst Ecol. 2010 Oct;38(5):1049–51.

108. Schlichtherle-Cerny H, Amadò R. Analysis of Taste-Active Compounds in an Enzymatic Hydrolysate of Deamidated Wheat Gluten. J Agric Food Chem. 2002 Mar 1;50(6):1515–22.

109. Awadé AC, Cleuziat P, Gonzalès T, Robert-Baudouy J. Pyrrolidone carboxyl peptidase (Pcp): An enzyme that removes pyroglutamic acid (pGlu) from pGlu-peptides and pGlu-proteins. Proteins. 1994;20(1):34–51.

110. Dando PM, Fortunato M, Strand GB, Smith TS, Barrett AJ. Pyroglutamyl-peptidase I: cloning, sequencing, and characterisation of the recombinant human enzyme. Protein Expr Purif. 2003 Mar;28(1):111–9.

111. Aswad DW, Paranandi MV, Schurter BT. Isoaspartate in peptides and proteins: formation, significance, and analysis. J Pharm Biomed Anal. 2000 Jan;21(6):1129–36.

112. Buchanan DL, Haley EE, Markiw RT. Occurrence of β-Aspartyl and γ-Glutamyl Oligopeptides in Human Urine*. Biochemistry (Mosc). 1962 Jul 1;1(4):612–20.

113. Violand BN, Schlittler MR, Toren PC, Siegel NR. Formation of isoaspartate 99 in bovine and porcine somatotropins. J Protein Chem. 1990 Feb 1;9(1):109–17.

114. Tanaka T, Nakajima T. Isolation and Identification of Urinary β-Aspartyl Dipeptides and Their Concentrations in Human Urine. J Biochem (Tokyo). 1978 Sep 1;84(3):617–25.

115. Boza JJ, Jiménez J, Martínez O, Suárez MD, Gil A. Nutritional Value and Antigenicity of Two Milk Protein Hydrolysates in Rats and Guinea Pigs. J Nutr. 1994 Oct 1;124(10):1978–86.

116. Mortensen LS, Holmer-Jensen J, Hartvigsen ML, Jensen VK, Astrup A, de Vrese M, et al. Effects of different fractions of whey protein on postprandial lipid and hormone responses in type 2 diabetes. Eur J Clin Nutr. 2012 Jul;66(7):799–805.

117. Armirotti A, Millo E, Damonte G. How to Discriminate Between Leucine and Isoleucine by Low Energy ESI-TRAP MSn. J Am Soc Mass Spectrom. 2007 Jan;18(1):57–63.

118. Casetta B, Tagliacozzi D, Shushan B, Federici G. Development of a Method for Rapid Quantitation of Amino Acids by Liquid Chromatography-Tandem Mass Spectrometry (LC-MSMS) in Plasma. Clin Chem Lab Med. 2000 May;38(5):391–401.

119. Squire NL, Beranová Š, Wesdemiotis C. Tandem mass spectrometry of peptides. III—differentiation between leucine and isoleucine based on neutral losses. J Mass Spectrom. 1995 Oct 1;30(10):1429–34.

APPENDIX I

PAPER I

Stanstrup J, Gerlich M, Dragsted LO, Neumann S

Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data

Anal Bioanal Chem. 2013; 405(15):5037–5048.

APPENDIX II

PAPER II

Stanstrup J, Schou SS, Holmer-Jensen J, Hermansen K, Dragsted LO

Whey protein delays gastric emptying and suppresses plasma fatty acids and their metabolites compared to casein, gluten and fish protein

(submitted)

APPENDIX III

PAPER III

Stanstrup J, Rasmussen JE, Ritz C, Holmer-Jensen J, Hermansen K, Dragsted LO.

Intakes of whey protein hydrolysate and whole whey proteins are discriminated by LC–MS.

Metabolomics. 2013.

PhD Thesis 2014
ISBN 978-87-7611-704-7

Jan Stanstrup

Metabolomics Investigation of Whey Intake:
Discovery of Markers and Biological Effects Supported by a Computer-Assisted Compound Identification Pipeline

SL grafik