Technische Universität München

TUM School of Life Sciences

Optimization of Nuclear Magnetic Resonance based Metabolomics workflows for human large-cohort analysis

Kristina Elisa Haslauer

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Wilfried Schwab

Prüfer*innen der Dissertation:

1. apl. Prof. Dr. Philippe Schmitt-Kopplin

2. Prof. Dr. Nicole Strittmatter

Die Dissertation wurde am 02.06.2023 bei der Technischen Universität München eingereicht und

durch die TUM School of Life Sciences am 30.10.2023 angenommen.

This page intentionally left blank

Acknowledgements

I would like to thank my supervisor Prof. Dr. Philippe Schmitt-Kopplin, who gave me the opportunity to do this work in his research group. I thank you for the freedom you gave me in this research and for your help, guidance and support.
Thank you, Silke, for supervising my PhD and introducing me to NMR and a thank you to Marianna for statistical support and open ears throughout the years. A special thank you to small Philippe for all the discussions, proof-reading and
Many thanks to the research group Analytical BioGeoChemistry of the Helmholtz Zentrum München. All of you made my PhD experience special to me, both scientifically and personally. I would like to thank you all for the discussions, your help in many aspects and the fruitful working atmosphere. Thanks for all the good times inside and outside of the lab.

Finally, I would like to thank my family. You have always supported me and for that I am grateful to all of you.
And especially Joe, thank you for putting up with me all these years. You are my rock and I'm beyond grateful for your love and support.

This page intentionally left blank

Abstract

Metabolic diseases are an increasing threat to the western society. Therefore, the metabolomics research branch evolves as a key technology. The germ of the idea is the qualitative observation of the metabolome, which is defined as the sum of all metabolites in a test matrix, relative to each other. Since the metabolome is directly linked to the observable phenotype, a direct link between changes in phenotype and metabolic signature can be established through the simultaneous study of these two variables. The great benefit is that metabolic signatures can be used to predict changes in the phenotype even before observable changes occur. This relatively new research field aims to gain a better understanding of disease related risk factors and development. A fundamental understanding of underlying mechanisms allows the identification of disease risks and of potential targets for new therapeutics. Furthermore, the discovery of early disease markers and their establishment in screenings allows early treatment, which might reduce harmful health outcomes for the population.

In the field of metabolomics, two subdisciplines have evolved: mass spectrometry-based metabolomics and the approach based on nuclear magnetic resonance spectroscopy (NMR). This thesis focuses on the latter.

As of today, there is little consensus in the scientific community regarding best practice guidelines and gold-standards for the metabolomics workflow. A classic NMR metabolomics workflow consists of sample preparation, data acquisition, data processing, statistical analysis and metabolite identification. These steps were investigated in this thesis specifically for the case of urine NMR metabolomics. The largest gaps were identified in the areas of sample preparation and data processing. Subsequently, guidelines and methods to improve the actual workflows were developed.

The first part of the thesis is describing a hydrogen deuterium exchange in creatinine for the use of deuterated buffer systems. As creatinine is a commonly used measure for normalization to account for urine dilution, especially in medical investigations, inaccurate results may be generated. To facilitate the use of recorded datasets, a correction factor was introduced.

The second part of the thesis focussed on data analysis, specifically the extraction of spectral information from 1-dimensional NMR spectra. Frequently used approaches were critically investigated and a novel algorithm was developed and introduced, which significantly reduces the spectral noise from large datasets.

In the frame of this work, it was shown that the NMR-based metabolomics approach is an effective method for investigating the influence of various factors on the human organism. The need for the development of a universal standard procedure for sample preparation and analysis was also demonstrated and suggestions were made to optimise this.

Zusammenfassung

Stoffwechselkrankheiten stellen eine zunehmende Bedrohung für die westliche Gesellschaft dar, daher entwickelt sich der Forschungszweig der Metabolomik zu einer Schlüsseltechnologie. Grundlage der Idee ist die qualitative Betrachtung des Metaboloms, das als Summe aller Metaboliten in einer Testmatrix relativ zueinander definiert ist. Da das Metabolom direkt mit dem beobachtbaren Phänotyp assoziiert ist, kann durch die gleichzeitige Untersuchung dieser beiden Variablen eine direkte Verbindung hergestellt werden. Der innovative Ansatz besteht darin, dass metabolische Signaturen zur Vorhersage von Veränderungen des Phänotyps verwendet werden können, noch bevor beobachtbare Veränderungen auftreten. Dieser relativ neue Forschungszweig zielt auf ein besseres Verständnis der krankheitsbezogenen Risikofaktoren und der Krankheitsentwicklung ab. Ein grundlegendes Verständnis der den Krankheiten zugrundeliegenden Mechanismen ermöglicht die Identifizierung von Risiken und potenziellen Zielen neuer Therapeutika. Darüber hinaus ermöglicht die Entdeckung von frühen Krankheitsmarkern und deren Monitoring in Vorsorgeuntersuchungen eine frühzeitige Behandlung und Krankheitsprävention.

Die Metabolomik wird derzeit von zwei analytischen Methoden dominiert, der Massenspektrometrie (MS) und der Kernspinresonanzspektroskopie (NMR). Die vorliegende Arbeit befasst sich mit letzterer Methodik.

Bis heute gibt es in der wissenschaftlichen Gemeinschaft wenig Konsens über -Richtlinien und Gold-Standards für einen NMR basierten Metabolomik-Workflow. Ein klassischer NMR-Metabolomik-Ablauf besteht aus Probenvorbereitung, Datenerfassung, Datenverarbeitung, statistischer Analyse und Metabolit-Identifizierung. Diese Schritte wurden in der vorliegenden Arbeit speziell für den Fall der Urin-NMR-Metabolomik untersucht. Die größten Lücken wurden in den Bereichen der Probenvorbereitung und der Datenverarbeitung festgestellt. Es wurden Richtlinien und Methoden zur Verbesserung der aktuellen Arbeitsabläufe entwickelt.

Der erste Teil der Arbeit beschreibt einen Wasserstoff-Deuterium-Austausch in Kreatinin, der bei Verwendung von deuterierten Puffersystemen auftreten kann. Da Kreatinin ein häufig verwendeter Parameter für die Normalisierung von Urinproben ist, insbesondere bei medizinischen Untersuchungen, können ungenaue Ergebnisse entstehen. Um dennoch die Auswertung und Interpretation bereits erstellter Datensätze zu ermöglichen, wurde eine Korrekturgleichung eingeführt.

Der zweite Teil der Arbeit befasst sich mit der Datenanalyse, insbesondere mit der Extraktion von spektralen Informationen aus 1-dimensionalen NMR-Spektren. Häufig verwendete Ansätze wurden kritisch untersucht und ein innovativer Algorithmus wurde entwickelt und vorgestellt, der das spektrale Rauschen in großen Datensätzen deutlich reduziert.

Im Rahmen dieser Arbeit wurde gezeigt, dass der NMR-basierte Metabolomik-Ansatz eine effektive Methode ist, um den Einfluss verschiedener Faktoren auf den menschlichen Organismus zu untersuchen. Die Notwendigkeit der Entwicklung eines allgemein akzeptierten Standardverfahrens für die Probenvorbereitung und -analyse wurde ebenfalls aufgezeigt und es wurden Vorschläge zur Optimierung dieses Verfahrens gemacht.

Table of contents

This page intentionally left blank

List of publications

Haslauer, Kristina E., et al. "Guidelines for the Use of Deuterium Oxide (D2O) in 1H NMR Metabolomics." Analytical Chemistry 91.17 (2019): 11063-11069.

Haslauer, Kristina E., Philippe Schmitt-Kopplin, and Silke S. Heinzmann. "Data processing optimization in untargeted metabolomics of urine using voigt lineshape model non-linear regression analysis." Metabolites 11.5 (2021): 285.

Abbreviations

1D	1-dimensional
2D	2-dimensional
ANOVA	Analysis Of Variance
COMETS	Consortium Of Metabolomics Studies
COPD	Chronic Obstructive Pulmonary Disease
COSY	Correlation Spectroscopy
COW	Correlation Optimized Warping
DEPT	Distortionless Enhancement by Polarization Transfer
DNA	Deoxyribonucleic Acid
DSS	Sodium trimethylsilylpropanesulfonate
DTW	Dynamic Time Warping
ESI	Electrospray Ionization
FID	Free Induction Decay
FT	Fourier Transformation
GC	Gas Chromatography
HCA	Hierarchical Cluster Analysis
HMBC	Heteronuclear Multiple Bond Correlation
HMDB	Human Metabolome Database
HSQC	Heteronuclear Single Quantum Coherence
ICA	Individual Component Analysis
IG	Inverse Gated
Jres	J-resolved Spectroscopy
LC	Liquid Chromatography
m/z	Mass to Charge
mRNA	Messenger RNA

MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance Spectroscopy
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Enhancement Spectroscopy
OPLS-DA	Orthogonal Partial Least Squares Discriminant Analysis
PAH	Polycyclic Aromatic Hydrocarbons
PCA	Principal Component Analysis
PLS-DA	Partial Least Squares - Discriminant Analysis
PQN	Probabilistic Quotient Normalization
RF	Random Forrest
RNA	Ribonucleic Acid
RSPA	Recursive Segment Wise Peak Alignment
SOM	Self-Organizing Maps
SOP	Standard Operating Procedure
STOCSY	Statistical Total Correlation Spectroscopy
SVM	Support Vector Machines
TMAO	Trimethylamine-N-oxide
TMS	Tetramethylsilane
TOCSY	Total Correlation Spectroscopy
TSP	3-(Trimethylsilyl)propionic-2,2,3,3 acid
uv	unit variance

Chapter 1|

1. General Introduction & Methods

1.1 Metabolomics

The concept that individuals have metabolic profiles was introduced as early as the late 1940s and early 1950s by Roger Williams and his group, who applied paper chromatographic methods to determine individual metabolic excretion patterns of several analytes in urine and saliva [1]. Although these results were promising, the lack of analytical methods to determine individual metabolite levels at a sufficient level was a huge drawback. The rise of advanced technologies in the second half of the 20th century again brought up this research area. Horning et al. introduced the term 'metabolic profile' in 1971, where they demonstrated the applicability of gas chromatography coupled to mass spectrometry to analyze multiple components in human samples. The group suggested the use of those metabolic profiles to determine abnormal conditions, analysis of drug metabolism or the effect of drugs on metabolic pathways [2].

The scientific disciplines 'Metabolomics' and 'Metabonomics' were defined by Nicholson and Fiehn around the turn of the millennium [3–5]. Although the two terms are often used interchangeably today, the original definitions differ significantly. Whereas metabolomics aims to analyze the relative changes in metabolite abundance in comparative studies and identify those [5], metabonomics, however, is defined as 'the quantitative measurement of the metabolic response to pathophysiological stimuli or genetic modifications' [4,3].

The metabolome itself refers to the totality of all small molecules synthesized or metabolized by a biological system [6] and therefore is extremely complex. The qualitative or quantitative longitudinal observations of the metabolome can detect biological changes in the study group at an early stage due to perturbations compared to their basal excretion [7]. Furthermore, case-control studies can provide information about the development of the disease or specific disease markers [8].

1.1.1.1 The role of Metabolomics in Systems Biology

The well-known systems theory is the underlying theory of systems biology, which declares that the behavior of a system is more than the sum of its components. Rather than that, the overall behavior of the system is significantly influenced by interactions between the parts of the system. The aim to describe biological systems in such a holistic manner requires a collaborative synergism between several scientific disciplines, such as biology, computer science and bioinformatics. It allows the understanding and prediction of how biological systems change over time or under varying conditions and opens up new possibilities to develop solutions to current health and environmental issues [7].

The scientific branch of systems biology is commonly known as omics, including genomics, transcriptomics, proteomics and metabolomics as main sub-disciplines [9]. A schematic depiction of the omics cascade is shown in Figure 1.

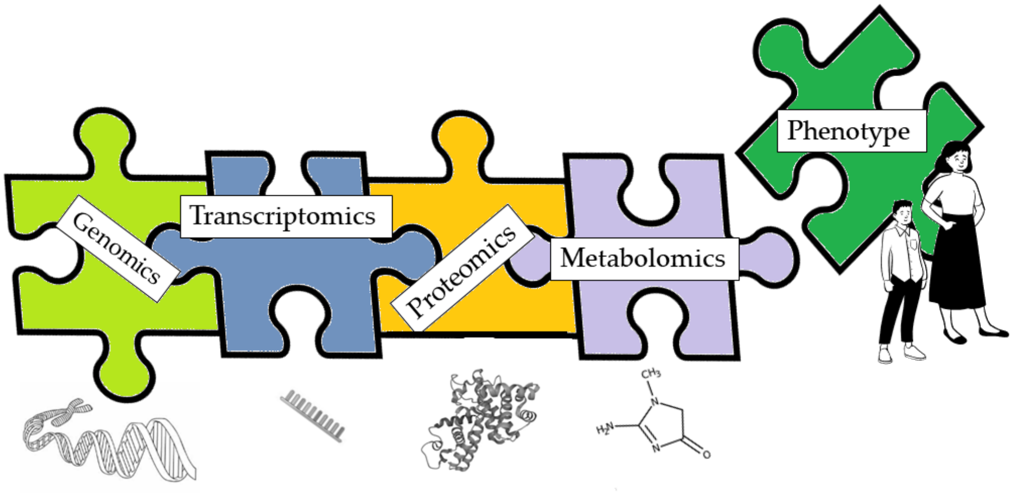


Figure 1: The omics cascade as an entity describing the response of a biological system to genetic and environmental influences

Genomics, as the first level of the omics cascade, studies the structure, function, evolution and editing of genomes and the complete genetic information of an organism. In humans the genome consists of about 3.2 billion base pairs. It can be understood as an instruction manual for all essential parts for a human existence, the building blocks, reproduction, repair mechanisms and the functional assembly. Therefore, the first part of the omics cascade gives a holistic overview of what could possibly happen in the biological system [10–12]. The human genome project paved the way to understand the role of genes in the formation of human phenotypes and the individual risks for certain conditions [13]. Genomics aims to identify genetic variants associated with a certain disease, the effects of a specific treatment or prediction of future conditions [14]. Until today, thousands of genetic variants have been linked to common diseases [15,16], such as cancer [17] and type II diabetes [18]. Genomics is the most established discipline in the omics field [19].

Genetic information encodes proteins and regulatory components that are necessary for the entire life span of the system. It is neither necessary nor energy-efficient for the organism to translate all the information into readable molecules at any given time. To move another step toward phenotype, it may be useful to focus on the fragments of genetic information that are being read at a given time.

The subsequent part of the omics cascade, transcriptomics, aims to study the translated ribonucleic acid (RNA) profile within the biological system. This RNA profile describes the parts of the genome actively expressed at the investigated time point. The transcriptome can be examined either qualitatively to determine which transcripts are present, to identify novel splice sites or sites for RNA editing, or to quantitatively determine the amount of each transcript [19]. The functional RNA molecules consist of protein-coding mRNAs and non-coding RNAs, which do not encode proteins but have regulatory functions. The advent of new technologies allowed large scale transcriptomics studies which revealed that only ~3% of the genome encodes proteins, whereas up to 80% is transcribed [20]. Since then, several studies showed the essential role of non-coding RNA in physiological processes, such as cell differentiation [21,22], neurogenesis [23] and endocrine regulation [24]. Transcriptomics data describes what appears to be happening in the biological system at given time points [25]. This research field has been broadly applied across diverse areas of biomedical research, such as diagnosis and profiling [26]. Alternative splicing patterns are of great interest in human health and disease, as 15-60% of known disease-causing mutations affect splicing [27,28]. Alterations in splicing may cause the disease directly or modify the severity of the disease or it can also be linked to disease susceptibility [28]. Frequently, transcripts contain alternative exons which increase the diversity and enables higher complexity encoded in the genome [29].

Once mature mRNA is generated, the protein-coding snippets are then translated into proteins by decoding the amino acid sequence determined by the order of nucleic acids and further posttranslational modifications, such as phosphorylation, glycosylation or methylation [30]. These modifications play a crucial role in cell signaling, the maintenance of cell structure, enzyme regulations and protein turnover [31]. Proteomics aims to analyze and quantify the composition of proteins, interaction, and abundance. This part of the omics cascade identifies and describes the functional molecules responsible for any biochemical process of a system and its response to internal and external stimuli [32].

As depicted in Figure 1, metabolomics is the ultimate stage of the omics cascade providing information about the functional readout of a biological system. Contrary to genes, mRNAs and proteins, downstream metabolites serve as markers for biochemical activity and hence are strongly linked to the observable phenotype [33]. Although it is the closest to the phenotype, this discipline emerged rather late. Thus, at this stage of research, no single instrument best practice guideline exists.

1.1.2 Metabolomics and the Human Urine Metabolome

Body fluids frequently used in metabolomics-based studies are urine, blood serum or plasma, saliva, tissue or stool homogenates, or cerebrospinal fluid [34].

In Figure 2 the frequency of publications containing the keyword 'metabolomics' and the above-mentioned biological sample types in the title is shown. Most publications (66 %) focus on blood metabolomics analyzing the non-cellular compounds as either serum or plasma. Serum remains after the blood was allowed to clot, whereas in plasma the clotting is prevented by adding an anticoagulant such as e.g., heparin. During the process of coagulation, platelets release chemical substances into the serum, as for example proinflammatory cytokines or sphingosine-1-phosphate [35–37]. Serum and plasma metabolomics have been applied in colorectal cancer research [38], diagnosis of hepatocarcinoma [39], lung cancer [40] and renal cell carcinoma [41].

Urine has been a favorable bio fluid in life science and medicine for decades, as it is generally sterile and easy to obtain. Although often considered a waste product, urine has considerable value in diagnostics. Through the easy accessibility, urine has been valued as early as in the ancient Egypt times for medical purposes. Hippocrates, one of the most outstanding personalities in the history of medicine, supported the technique of uroscopy, in which urine samples were examined for color, smell, sediment and particles for diagnostic purposes [42]. Since then, the progress in analytical and microbiological methods as well as profiling techniques allows urine examination more detailed and informative [43].

Unlike blood samples, urine is more susceptible to diet and diurnal variation, but it plays an important role in acquiring metabolite data. In particular, urine is the matrix of choice in some patient populations, such as young children [44]. Metabolomics studies based on urine as sample matrix account for 18 % of publications listed on PubMed published in 2022.

Tissue metabolomics account for 9 % of those publications. Although the analysis of tissue specimen is generally more invasive compared to body fluids, the main interest lies in organ specificity. As the origin of the specimen is localized close to the main disease progress, tissue metabolomics are considered more sensitive and therefore may provide a robust method for biomarker discovery. The sample preparation workflow of tissue metabolomics can include lyophilization, homogenization and extraction. Several different tissue types have been investigated, such as brain, kidney, esophagus, skin wound tissue or ovarian tissue [45].

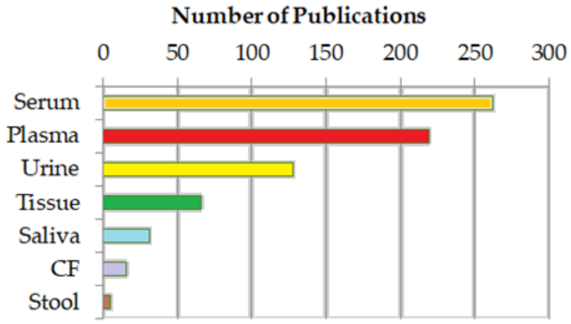


Figure 2: PUBMED query on the likelihood of occurrence of the keyword 'Metabolomics' in combination with frequently used biological matrices in the title of publications in 2022

Since salivary metabolomics is a very new field of research, this part is relatively small compared to the already established biofluids. Nevertheless, this area is very promising. The very simple extraction and general availability allow the collection of large sample quantities [46]. Saliva metabolic profiles of smokers and nonsmokers have been analyzed and smoking related perturbation were found [47]. Also the detection of oral cancer via saliva metabolomics was performed successfully [48].

The human cerebrospinal fluid metabolome has been described in 2008 by Wishart et al. Their attempt was to describe a baseline metabolome in healthy individuals and identify the best suited analytical technique to analyze the matrix. They identified cerebrospinal fluid as information rich and therefore valuable biofluid for metabolomics [49]. Nevertheless, cerebrospinal fluid is hard to obtain and therefore leaves the researcher with a small number of samples. This might explain the small percentage of metabolomics studies performed on cerebrospinal fluid until today.

When diseases affecting the lower digestive tract are to be investigated, stool is the matrix of choice in metabolomics approaches. Studies investigating perturbations in the stool metabolome have been performed for inflammatory bowel disease, Chron's disease and ulcerative colitis [50]. Although stool samples contain a lot of information about the patient and his gut microbiome, stool is a very complex test matrix due to its inhomogeneity and variability and is therefore rarely used compared to other samples.

The different types of bio samples provide different biochemical information and must be selected according to the specific research question. In the context of this thesis the focus is set on urine as sample matrix.

Contrary to other bio fluids, such as blood, urine has no homeostatic mechanism. Due to this, urine composition can be very diverse without harming the body and therefore is a valuable source for early biomarker discovery [51]. Several diseases remain silent until the late phase, when irreversible progression is made [52]. Urine has proven to be a valuable sample type to screen for disease specific signatures in an early stage. For example, a metabolomics approach has been used to investigate metabolic profiles and biomarkers for chronic obstructive pulmonary disease (COPD), which is an increasing health concern. McClay et al. found, that an urinary metabolomics approach is an effective diagnostic tool and could therefore be used for early screening [53].

Similarly, Matsumura et al. found biomarker for the diagnosis of lung cancer with excellent sensitivity and specificity (93% and 94%) [54].

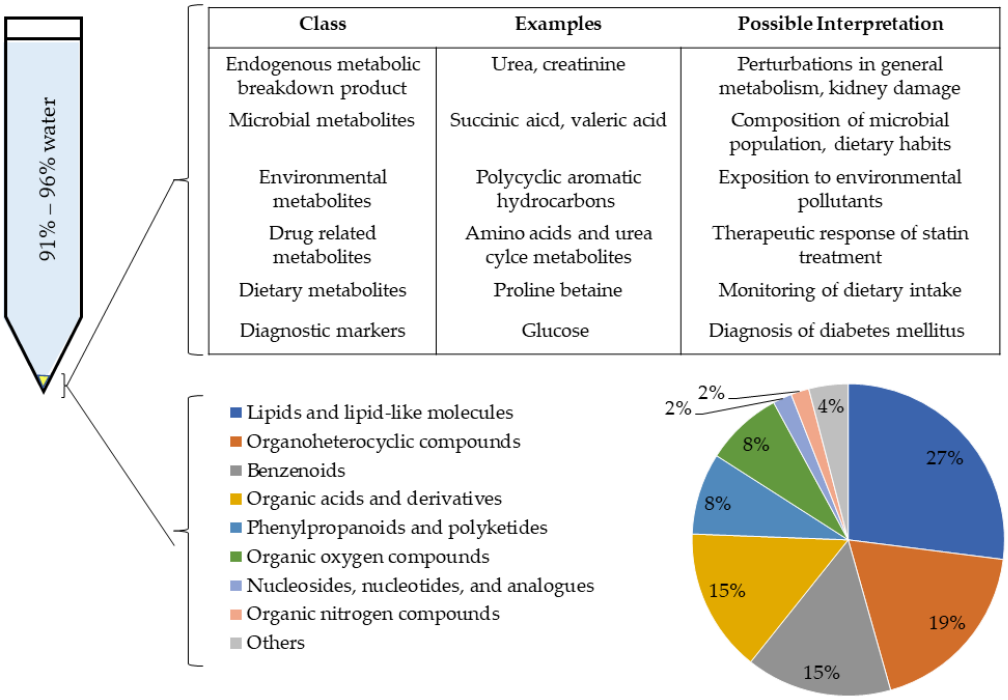
Additionally, to the possible simplification of diagnosis, such approaches enable the identification of metabolic pathways which are involved in disease progression. This may present new possibilities in identification of potential drug targets in treatment and lead to a better understanding of disease development, which is a basis for prevention measures [55].

In mammals, urine is produced by the kidneys via extraction of soluble wastes from the bloodstream. The excretory function consists of glomerular filtration, tubular reabsorption und secretion [56]. An average adult generates between 1.5 and 2 liters of urine per day [57]. Therefore, sampling is relatively easy and non-invasive with simultaneous high information gain.

Contrary to other omics disciplines where often a near-complete coverage of the genomic or proteomic information can be reached, most of the urinary metabolites could not have been identified until today [57]. Progress is made towards extension of reference databases, such as the human metabolome database (HMDB)[58]. Introduced as early as 2007, the HMDB is considered the standard reference database for human metabolomics studies nowadays and covers more than 110 000 fully annotated metabolites by 2018 [59].

The human urine metabolome is very complex and diverse, containing amino acids, organic acids, nucleosides, and carbohydrates among other classes [57]. Also, xenobiotics, such as drugs, pollutants, cosmetics and their metabolites, represent a branch of metabolomics research interests [60,61]. The sheer diversity of possible structures and metabolites explains the meagre coverage of identification.

In Figure 3 the chemical composition of urine is pictured. The individual components can be classified according to their chemical composition, respectively chemical superclasses, as applied in the HMDB [57]. As shown in the figure, the compounds can also be classified according to their origin and potential information. The concentration of urinary creatinine and the urine-to-plasma ratio of urea for example have been shown to indicate kidney problems [62,63]. Jain et al. found out, that microbial metabolites in urine provide a functional read-out of the status of the gut microbiome and probands diet. Thus, the analysis of microbial metabolites in urinary metabolomics allows a linkage between the metabolic phenotype and microbial population [64].



Class	Examples	Possible Interpretation
Endogenous metabolic breakdown product	Urea, creatinine	Perturbations in general metabolism, kidney damage
Microbial metabolites	Succinic aicd, valeric acid	Composition of microbial population, dietary habits
Environmental metabolites	Polycyclic aromatic hydrocarbons	Exposition to environmental pollutants
Drug related metabolites	Amino acids and urea cylce metabolites	Therapeutic response of statin treatment
Dietary metabolites	Proline betaine	Monitoring of dietary intake
Diagnostic markers	Glucose	Diagnosis of diabetes mellitus

Figure 3: Typical composition of human urine

Environmental pollutants are a major risk to human health, such as the harmful polycyclic aromatic hydrocarbons (PAH). PAHs are known to be ubiquitous in the environment. It is also known that PAHs have toxic, mutagenic, carcinogenic properties. To study their biological effects in human populations, those environmental pollutants are of major interest in metabolomics studies. The goal of such studies is to link the environmental exposure to specific phenotypes and to gain information about potentially affected pathways [65].

Monitoring xenobiotics and drug-related metabolites allows evidence to be gathered on the biochemical pathways that a drug of interest affects in the human body. Studies also have shown that the monitoring of metabolites related to a specific drug can detect good responders to the treatment. For example was the increase of dibasic acids positively correlated with the response to simvastatin, a drug which is used to lower blood cholesterol and prevent heart diseases [66,67].

In summary, urine as a test substance can answer a wide range of possible scientific questions. Since urine is widely available and very stable and less complex to process compared to blood derivatives, many metabolomics studies are performed using this matrix.

1.1.3 Conceptual Approaches of Metabolomics

Metabolomics approaches can be performed either targeted or untargeted with a vital difference in the concept. Untargeted metabolic profiling is executed with no a priori selection of metabolites or the knowledge of their identity, whereas a targeted method relies on a selection of specific metabolites or metabolite classes prior analysis. The principles of both methods are visually compared in Figure 4.

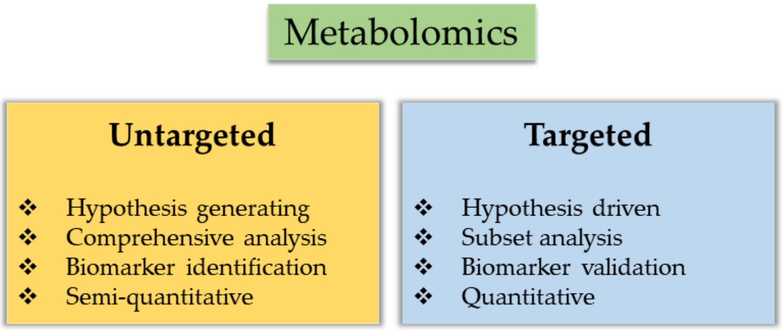


Figure 4: Principles of targeted and untargeted metabolomics

Untargeted metabolomics can be considered as hypothesis generating approach with the aim to measure and compare as many signals as possible across a sample set to allow a comprehensive analysis. This top-down technique is generating complex and large amounts of data which make the subsequent analysis a demanding task [68]. Nevertheless, this method enables to approach new scientific questions. Untargeted metabolomics can complement clinical research by biomarker discovery [69], disease early onset research [70,71] and precision medicine profiling [72–74]. The main difficulty besides the extensive computational effort is the bottleneck of metabolite identification, which is described in detail in section 1.2.7.

In contrast, targeted metabolomics is a hypothesis driven approach to investigate the effects of a treatment, diet, or environment on levels of a priori defined metabolites or spectral features. Contrary to untargeted metabolomics, where semi-quantification of interesting metabolites is generally the end point, targeted metabolomics usually includes an absolute quantification of the investigated metabolites. This difference derives from the varied experimental setup. If the analyte identity is known, and with that the chemical structure, the experimental design can be optimized towards these metabolites [75,76].

Both approaches can also be designed to be built on one another starting with a hypothesis generating global untargeted approach to identify metabolites or clusters of interest and subsequently examine this set of substances with a targeted technique.

Commercially available platforms, such as Bruker's IVDr software systems combine both approaches within one measurement. For each sample an untargeted profile is generated and subsequently a set of pre-defined metabolites is fully quantified [77].

1.1.4 Human Metabolomics Studies

Animal models have been the first choice to study molecular pathways of diseases, as it allowed researchers to investigate organisms under very defined conditions and at low costs. The environment, diet and individual factors can be highly controlled in animal studies. The strong level of standardization reduces the inter-individual variation drastically and enables better identification of potential effects. Especially when the subject of interest is a specific tissue or organ, e.g., liver or brain, animal studies are advantageous, as the desired sample can be taken and examined after the animal has been euthanized. Furthermore, the bureaucratic burden is significantly lower compared to human studies. Although animal models are very efficient in many cases, they have significant disadvantages. Differences between the model species, e.g., rodents, which are frequently used in animal studies, and humans are a major disadvantage. The clean laboratory facilities in which those animals grow up lead to immature immune systems and vastly different microbiomes. These effects reduce the translatability between animal models and humans [78–80]. Furthermore, the ethical aspect plays an important role and pushes the development of alternative study designs recently.

Human studies can be divided into two subgroups, which are observational studies and intervention studies. Interventional studies are usually performed with a small number of participants and indicated when a certain intervention can be performed, e.g. a new treatment versus placebo. The gold standard of interventional studies is the randomized controlled trial. The biggest benefit of intervention studies is that all other variables can be controlled, especially if the volunteers are unaware of the treatment they are getting. Intervention studies are limited in time, as it is not possible to monitor the study participants over a longer period of time. Thus, only short-term outcomes can be monitored. Furthermore, for ethical reasons, it is impossible to carry out an intervention study for some questions, for example when investigating the effects of exposure to environmental toxins.

Fortunately, the progression in data handing allows processing large amounts of data in short periods of time and therefore allowing the analysis of large-scale data from observational studies. Observational studies can be divided into three sub-groups, cohort studies, case-control studies, and cross-sectional studies. Cohort studies, also known as ecological studies, are designed to compare clusters of participants. The aim is to find factors which correlate with e.g., the risk of disease development. In case-control studies participants are selected based on their medical status, i.e., diseased or healthy. An example of a case-control study is analyzing the correlation between smoking habit and lung cancer. A major disadvantage of case-control studies is the potential of a recall bias [81]. Cross-sectional studies, also referred to as prevalence studies, assess data of a population at a specific timepoint [82].

Contrary to small-scale animal models and human interventional studies the conditions of large-scale observational studies are a lot less controllable. Likely sources of variance in urine composition in large scale human studies are e.g., diet, drug-intake, environmental influences, or exercise [83,84]. To account for these variances, food frequency questionnaires or diet diaries are frequently applied [85,86]. Nevertheless, these approaches are limited by misreporting and recall bias [87]. Large scale cohort studies, such as the single cohorts combined in the Consortium of Metabolomics Studies (COMETS) have applied questionnaires assessing smoking status, alcohol intake, body mass index, waist circumference, leisure-time physical activity and educational levels besides diet. Also, clinical measures were collected for some of the cohorts, such as blood pressure, fasting glucose or lipoproteins [88]. All these complementary measures were recorded to account for sources of variability in the subsequent data analysis. Additionally to variable sources of variance longitudinal observation of study cohorts faces other disadvantages. Exemplary, incomplete or interrupted follow-up samples of individuals are a major problem [89]. Furthermore, the researcher needs to be aware of logistical issue need, such as a constant cooling pipeline or batch effects.

In summary, large-scale human metabolomics studies are an excellent approach to study environmental effects on the human metabolism, the detection of biomarkers or disease development. Nevertheless, researchers need to be aware of limitations and challenges in subsequent data analysis.

1.2 Analytical Methods

The field of metabolomics is dominated by two main analytical methods, mass-spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR). MS is a technique based on the generation of ions by a variety of methods obtaining spectral data from the mass-to-charge ratios (m/z) of compounds, respectively their fragments, and their relative abundance. Ionization may be performed thermally, by application of electric fields or through the impact of energetic electrons [90]. To reduce the total numbers of analytes ionized and detected at one time point different chromatographic methods are applied, such as liquid chromatography (LC) or gas chromatography (GC) [91,92].

In Table 1 NMR and MS are contrasted for several aspects relevant in metabolomics approaches. Compared to MS, NMR is typically 10 to 100 times less sensitive. NMR techniques detect metabolites with concentration > 1 μM, whereas typical LC-MS can detect metabolites > 10 to 100 nM [93,94].

Table 1: Comparison of NMR and MS in metabolomics applications, adapted from Emwas et al. [93]

	NMR	**MS**
Sensitivity	Relatively to MS low	High
Reproducibility	High	Relatively to NMR low
Matrix effects	Low	High
Sample preparation	Low	High
Sample recovery	Nondestructive, therefore high	Destructive, therefore low
Selectivity	Nonselective	Selective
Quantification	Inherently quantitative	Internal standards necessary for quantification

Despite the higher sensitivity, MS approaches have remarkable disadvantages compared to NMR based methods. MS is known to be less reproducible and matrix effects are a common issue. Matrix effects occur as signal suppression or enhancement in the presence of sample matrix components [95].

Sample preparation is a relevant point when comparing the two methods, as metabolomics studies usually require a large number of samples to be processed. In addition to the error-proneness of complex preparation procedures, a time factor must also be considered. Depending on the specific approach, the separation and purification prior to introduction in the mass analyzer can be time consuming and complex. NMR spectroscopy usually requires no or si mple sample preparation, whereas MS measurements need more elaborate processing.

Since NMR, unlike MS, is a non-destructive method, it is usually possible to measure individual samples again after a longer period of time if they are stored adequately. This intrinsic property means that faulty measurements can be detected and replaced afterwards.

Furthermore, the selectivity of GC/LC-MS to different classes of analytes requires a sophisticated set-up to allow a maximum coverage of metabolites [96]. Contrary, NMR is known to be highly quantitative and reproducible. Moreover, NMR is not selective to compound classes [97,98]. These properties give a solid basis for analysis of a broad range of analytes under different conditions and the quantitative nature of NMR produces data suitable for multivariate statistical analysis [99]. The selectivity of GC/LC-MS to different classes of analytes requires a sophisticated set-up to allow a maximum coverage of metabolites [96].

Another intrinsic property of NMR spectroscopy is the quantitative nature of the measurement. In contrast to MS, NMR measurements do not require individual internal standards for specific constituents if the relaxation time is sufficient. Due to varying ionization efficiency, a structurally similar internal standard, usually a stable isotope labeled standard, is essential for quantification in MS [100]. Therefore, a quantification of various compounds in complex biological samples requires a profound knowledge of the matrix to establish an appropriate mixture of internal standards [101].

In summary, it can be said that both methods have fundamental advantages and disadvantages. In general, these specific properties make NMR more suitable for untargeted questions, while MS is a good tool for targeted questions.

1.2.1 Fundamental Principles of Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance (NMR) is a spectroscopic method based on the magnetic properties of nuclei. The phenomenon of nuclear magnetic resonance was discovered as early as the 1940s by Purcell and Bloch [102,103]. The method relies on the fact, that many nuclei have spins, such as ^{1}H, ^{13}C, ^{15}N or ^{19}F to name a few. The nuclear spin (I) is a form of angular momentum carried by atomic nuclei and can be described using quantum numbers. Atoms with an even number of protons and neutrons have a spin equivalent to zero, atoms with an uneven number have a non-zero spin. Most nuclei relevant in a biological context have the spin ½, such as ^{1}H and ^{13}C. The atoms with a spin different from zero have a magnetic moment μ, described by

Eq. 1

$$\mu = g\,I$$

with g being the gyromagnetic ratio, the ratio between the magnetic moment to the angular momentum, which is specific for each nucleus. The magnetic moment forces the nuclei to precess around the external magnetic field B_0 with a characteristic frequency, the Larmor frequency.

Eq. 2

$$\omega_L = \gamma B_0$$

For spin ½, only two energy levels exist. In an external magnetic field (B_0), the rotation axes of nuclei are forced to align parallel (α state, low energy) or antiparallel (β state, high energy) to the external magnetic field direction (z plane) by their magnetic moment. The nuclei in a sample are distributed among the different energy levels, where the number N in the respective energy level can be described by the Boltzmann distribution,

Eq. 3

$$\frac{N_{upper}}{N_{lower}} = e^{-\gamma_N Hh/kT}$$

where h is Planck's constant, $H(B)$ is the external magnetic field strength, k is the Boltzmann constant and T is temperature. By exposing the nuclei to a second oscillating magnetic field in the form of radiofrequency pulses corresponding to the Larmor frequency of a nuclei, energy can be transferred into the spin system, which changes the state of the system (often by rotation of 90° into the horizontal xy plane). After the pulse energy is introduced, the system relaxes back into its equilibrium state inducing weak currents in the probe coils. This resonance signal, also known as free induction decay (FID), is recorded by the spectrometer as a function of time. The FID is a complex pattern describing the exponential decay, which is relatively challenging to interpret. By performing a Fourier Transformation (FT), the FID is converted from the time domain in the frequency domain producing the actual NMR spectrum. The magnitude of a resonance is displayed along the frequency axis. NMR spectrometers are classified by their magnetic field strength, ranging from 7.05 T to 23.49 T. Also, devices with the same nominal magnetic field strength vary in their actual values. To determine the spectrometers operating parameters, the strength of it is denoted as the frequency of the water protons, which is around 300 MHz for 7.05 T magnet and 1 GHz for 23.49 T magnet.

This is the so-called frequency of the spectrometer. To enable a comparison between samples recorded on different devices, the chemical shift (δ) scale is used. The scale is expressed as parts per million (ppm), which is independent of the spectrometer frequency.

Eq. 4

$$\delta = \frac{frequency\ of\ signal\ (MHz) - freuquency\ of\ standard\ (MHz)}{frequncy\ of\ spectrometer}\ x\ 10^6$$

As the chemical shift scale is an arbitrary scale, a reference standard must be used. The most common standards are trimethylsilylpropionat (TSP) in aqueous solutions and tetramethylsilane (TMS) in organic solvents. The reference standard is always denoted as $\delta = 0\ ppm$. The applied frequency increases from left to right, thus the left side of the spectrum is the low field, and the right side is the high field region. Although in a one-dimensional experiment only one sort of nuclei is observed (mostly ^{1}H or ^{13}C in metabolomics), the nuclei differ in their resonance frequencies. This is caused by the local chemical environment of a nucleus, which influences the exact magnetic field experienced by a particular nucleus. The electrons surrounding the nucleus are in motion and thus creating their own magnetic fields. These fields counteract the magnetic field generated by the high-frequency pulse and thus reduce the field to which the core is subjected. Therefore, the electrons are shielding the nucleus and the energy between the spin states is decreasing, which results in a smaller chemical shift. The different electron densities around the observed nuclei make NMR very useful in structure determination and the distinction of molecules within a complex matrix [104–106].

1.2.2 Sample Preparation for Urine NMR Metabolomics

To acquire high quality NMR spectra, care needs to be taken in every step of the analysis including sampling and sample preparation. Using urine as test substance these steps are relatively straight forward. Urinary excretion was found to vary throughout the days, exemplary levels of creatinine, mannitol, dimethylamine, 1-methylnicotinamide, xylose, acetone, transaconitate and phenylalanine are different between samples collected in the morning versus afternoon [107,108]. Therefore 24 h urine sampling is preferred. After sampling, the urine must be stored at -80 °C to avoid any microbial or chemical alterations in the matrix [109–111].

Since the subsequent analysis is a bottleneck in NMR metabolomics approaches, it is of immense importance to generate high-quality spectra. An overview of the quality criteria and the sample preparation factors influencing them is shown in Table 2.

In order to propose standard operating procedures (SOPs) for metabolomics, Bernini et al. investigated different pre-analytical treatments [112]. To obtain high quality spectra, homogenous samples without debris are required. Samples can be either centrifuged or filtered to remove debris. Care must be taken to ensure that the timing and intensity of centrifugation are identical, as deviations from a standard protocol may alter the metabolic profile. Detailed information can be found in the publication by Bernini et al. [112].

The pH value of human urine can range from 5 to 8, depending on an individual’s acid-base status [113–115]. Variation in pH is strongly affecting chemical shifts for some metabolites with ionizable groups [116]. Metabolites of this group, such as citric acid, hippuric acid, dimethylamine and some amino acids are major components in human urine. To ensure high quality data for latter analysis, buffer systems (e.g. a K_2HPO_4/NaH_2PO_4) need to be added to maintain a constant pH. Additionally to pH, the presence of ionic species such as Ca^{2+} and Mg^{2+} affect peak shifts [116]. Efforts were made to overcome this issue by adding EDTA [117] or KF [118].

Table 2: Criteria for spectral quality, their influence factors and crucial steps

Criteria	Influence factors	Handling
Linewidth	Sample homogeneity	Remove debris via centrifugation or filtration [112]
Peak shift	pH and salt content	pH buffering [111]

In Table 3 a review of published methods for urine NMR metabolomics is shown. It becomes clear that the methods used in this field differ greatly from one another and that no established standard procedure is generally used. A major advantage of NMR spectroscopy is the reproducibility of the results.

Table 3: Comparison of sample preparation and measurement methods

Paper	Preservative	Urine : buffer ratio	Buffer	Final D20 concentration	Autosampler and storage temperature	Temperature
[119]	0.2 mM NaN_3	9:1	1.5 M KH_2PO_4 in D_2O	10 % (v/v)	Autosampler at 6°C.	300K
[120]	0.2 mM NaN_3	9:1	1.5 M K_2HPO_4 in H_2O	no information	-	300 K
[121]	1 mM NaN_3	2:1	0.24 M Na_2HPO_4	6.66 % (v/v)	Autosampler, no information about temperature	300K
[122]	0.57 mM NaN_3	2.5:1	4.5 M KF and PO_4 in 100% D_2O	28.57% (v/v)	Autosampler at 4°C	300 K
[123]	0.5 mM NaN_3	3:1	1.5 M PO_4 buffer in 100% D_2O	25% (v/v)	-	298,15 K

However, because the sample preparation and in some cases also the measurement differ so greatly, individual study results cannot be directly compared with each other. As early as 2007, in the initial state of the research field, Lauridsen et al. addressed this issue and published recommendations for sample preparation and measurement based on a stability study they had conducted [111]. While investigating the consequences of freeze-drying and reconstitution in deuterium on the metabolic profile, they observed the effect of deuteration of creatinine and the associated shift in creatinine resonance at a chemical shift of 4.05 ppm. Although this problem is well known, the extent to which the proportion of deuterated buffer, temperature, and time elapsed between sample preparation and the actual measurement affect the level of the creatinine signal has never been investigated. Since creatinine is an important parameter in the field of urine metabolomics, this topic was systematically investigated in the first part of this dissertation.

1.2.3 1D-Proton-NMR in Metabolomics Approaches

As metabolomics is fundamentally based on the relative comparison of individual spectra to each other, high-quality spectra are a fundamental step in the process. To take a high-resolution spectrum, a stable and homogenous magnetic field is required. Even the superconducting magnets used in NMR spectrometers experience fluctuations in the magnetic field, additional fluctuations arise from environmental effects. To keep the magnetic field stable, a lock signal is used. A possible drift in the magnetic field is monitored by continuously measuring the absorption of the solvents deuterium signal and fixing this signal to a predefined frequency [104]. Without drift compensation, frequencies at which sample signals appear would be expanded resulting in peak broadening. Sodium azide is often added as bacteriostatic preservative to avoid microbial degradation [109]. In Table 3 frequently used concentrations are shown.

In NMR spectroscopy, the chemical shift is the relative distance of a resonance line of the sample from the resonance line of an arbitrarily chosen standard to which the chemical shift 0 ppm is assigned. The chemical shift, which is independent of the magnetic field strength of the spectrometer used, is given in ppm. In urine metabolomics commonly trimethylsilylpropionic acid (TSP) is used as such standard, other reference standards are sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) or trimethylsilane (TMS) in organic solvents [121,124].

Metabolomics approaches typically aim to detect minor biological changes in metabolite composition and concentration so that additional variance should be reduced. Therefore, study design, sample storage and preparation are crucial steps in a metabolomics study [125]. Nevertheless, challenges remain in reproducible analysis and processing of the acquired data.

Figure 5 shows the schematic sequence of the elementary steps necessary before recording an NMR spectrum. After the introduction of the sample into the spectrometer, it must be ensured that the temperature of the sample is equilibrated. If the sample has not reached the temperature equilibrium when the following parameters are determined or the measurement has already started, interfering artefacts may occur. Firstly, the magnet needs to be altered to compensate for environmental and sample effects. This procedure is called shimming. Most instruments are equipped with an automated gradient shim, which adjusts the magnetic field in a decent manner, still manual shimming is required to eliminate inhomogeneity. This is performed by stepwise adjustment of currents in the shim coils and observation of the peak shape from the internal reference standard. A sufficient shim is reached, once the line width at half of the peak amplitude is below 1 Hz and the overall peak shape is symmetric [126,124].

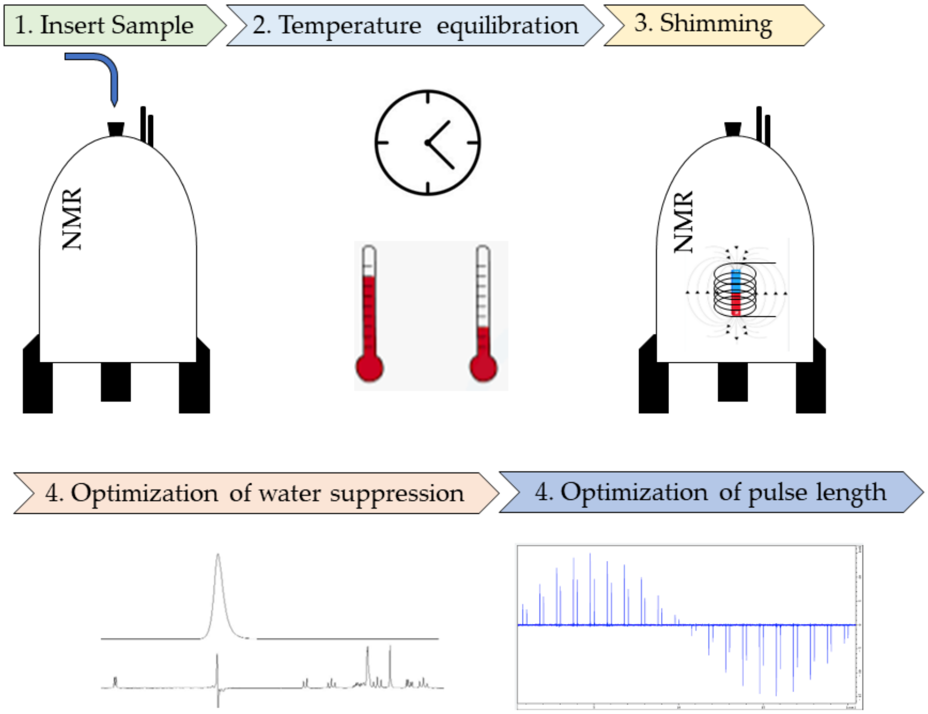

1. Insert Sample
2. Temperature equilibration
3. Shimming
NMR
NMR
4. Optimization of water suppression
4. Optimization of pulse length

Figure 5: Depiction of steps required to be performed prior to an NMR experiment to obtain high quality spectra

In contrast to sample preparation, for which there is no consensus, this exists for experimental setup. The commonly used pulse sequence for ^{1}H spectra is the 1D nuclear Overhauser enhancement spectroscopy (NOESY) presat [121]. As biological samples and especially urine is constituted mainly from water, the optimization of water suppression is a key factor in the experimental setup to ensure reproducibility among samples [127]. The excessive presence of ^{1}H atoms in the sample overwhelm the available dynamic range determined by the instrument and therefore this resonance needs to be suppressed [128]. Water suppression in NMR metabolomics must be feasible in a reasonable amount of time and result in quantitative as well as reproducible data. The 1D-NOESY sequence offers those features with generally little optimization effort. During the relaxation delay, a long low power pulse is applied at the frequency of the signal to be suppressed, which will saturate the unwanted resonance and greatly reduce its intensity. Some parameters, such as the offset o1, need to be adjusted carefully to reach high quality results. This parameter defines the center point of the recorded spectrum [128].

Another crucial point is the pulse width and the corresponding power. Prior to applying a radiofrequency pulse, most nuclear spins are aligned parallel to the magnetic field designated as z-axis. The application of such a pulse will rotate the bulk magnetization by a specific angle, depending on the intensity of this pulse. As resonance is measured in the xy-plane, an angle of 90° is resulting in the maximum signal.

Besides those parameters, several others need to be adjusted by the spectroscopist, such as acquisition time, relaxation delay, spectral width and necessary transients. Nevertheless, in NMR approaches these key factors play an important role in data interpretability. As large sets of samples are analyzed and the individual samples may vary in their ionic strength, attention must be put in considerations about these parameters The effects of such variations on metabolite identification and quantification have been studied and found to significantly affect the results. The importance of a well thought-out and perfectly adapted parameter set is indisputable [129,130].

1.2.4 Spectral processing, normalization and scaling

In order to obtain high-quality spectra, there are other points that must be considered in addition to sample preparation and the actual measurement. These include the steps of spectra processing. Figure 6 shows the individual steps schematically.

In NMR spectroscopy the signals are generated by the non-equilibrium nuclear spin magnetization as a function of time, referred to as free induction decay (FID). To convert the data from the time dimension to the frequency dimension, a fourier transformation (FT) is performed. To eliminate possible artefacts and confounding factors, some operations can be carried out before the data is transformed. This enables a high quality of the resulting spectra. One possible manipulation is the application of an apodization function. These functions are also called window functions and are typically applied to the FID to emphasize regions of the FID over another. The FID is multiplied with this function, which leads to a reduction of truncation artifacts at the outer ends of the signals and enhances spectral quality. A frequently used function is a decaying exponential function, which is multiplied with the FID. By using this operation, the weighting of the signal at the beginning of the signal recording is increased compared to the end. As the signal to noise ratio of the FID decreases towards the end of the recording, multiplying by the decaying exponential function generates a better signal to noise ratio. This transformation is accompanied by a reduction in resolution. A sensible compromise between resolution and sensitivity must be found here, which is applied to all spectra of the data set to be analyzed and generates an optimal spectrum quality.

Another frequently used manipulation is zero filling. Here, non-informative values with amplitude zero are appended to the end of the FID. Since this increases the absolute number of data points, it also increases the digital resolution [104].

After Fourier transformation, the spectrum needs to be phased correctly to result in positive peaks. The spectrometer measures the time dependent voltage, which is proportional to the magnetization, on two orthogonal axes with one of the voltages being notated as 'real part' and the other as 'imaginary part'. Both signals together are recorded as FID. The phase of this function depends on the value at timepoint zero, which should reach a maximum for the real part. In case of a phase offset, and with that a non-maximum value at timepoint zero, this can be corrected afterwards to obtain peaks with full absorption character. Practically, this means that all parts of the peak appear above the baseline rather than below [104].

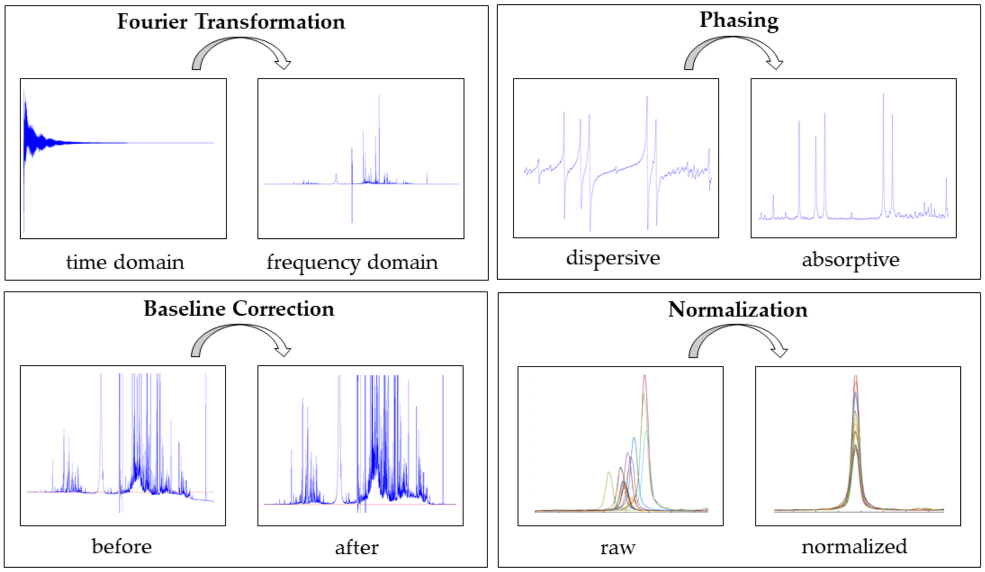


Figure 6: Overview of necessary steps in spectral pre-processing

To make spectra comparable, the resonance frequencies are not given as absolute values (in Hz), but according to the general conventions always as values relative to a common standard. As already mentioned in the previous paragraph, TSP is often used in metabolomics as an internal reference. Regardless of the strength of the magnetic field, the frequency of the internal reference is simply defined as the zero point and the frequencies of the other resonances are given according to how many parts-per-million they are away from the reference standard [104].

The quantitative character of NMR spectra is based on the fact that the peak heights or areas are proportional to the concentration. In order to optimally determine these peak heights or areas, baselines that are as flat as possible are essential. Baseline outliers are mainly caused by erroneous values in the first data points of the FID. The result of these low-frequency modulations is a runaway baseline [131]. There are two categories of methods to correct this error. The correction can be done directly in the time domain, in other words by reconstructing the FID, or by constructing a baseline in the frequency domain. This is then subtracted from the spectrum resulting in a flat baseline [132].

After all these technical deficiencies have been dealt with as optimally as possible, there is still a biological component in the case of urine metabolomics that should not be underestimated. As already mentioned in some places, the theory of metabolomics is based on the identification of differences between samples and the visualization of these differences through intensity comparisons. A major advantage of urine as a test matrix is the lack of homeostatic regulation of the individual components. However, this also means that urine can have very different concentrations. Individual drinking behavior can lead to urine samples differing up to 15-fold in concentration [133–135].

To reduce this unwanted variance from dilution while preserving the wanted variation introduces the need of reliable data normalization and scaling. The possibilities of normalization can basically be divided into two categories. On the one hand, there are methods that use a metabolite that is excreted as constantly as possible as the characteristic value of the dilution factor. In the context of urine metabolomics, this mainly includes creatinine [136,137].

The second category is based on the assumption that the total amount of excreted substances is relatively constant. These include the frequently used probabilistic quotient normalization [138], total area normalization [139] and quantile normalization [140].

The normalization on stable endogenous metabolites relies on the assumption, that under specific circumstances the excretion of the used metabolite is stable. Creatinine levels are described to be relatively stable over a period of 24 h in healthy individuals [141–143], however factors such as acute infections, injury, severe emotional stress or exercise can affect the extraction levels and therefore lead to false results if metabolite-creatinine ratios used for analysis [144].

Furthermore, sample storage and preparation may alter urinary creatinine levels [108].

Probabilistic quotient normalization (PQN) is based on the hypothesis, that the variations in concentration which are of interest for the scientific questions only affect parts of the spectrum, whereas the dilution affects the whole spectrum. Therefore, the PQN approach calculates fold-changes between every feature of the spectrum and the corresponding feature of a reference spectrum. The mean value of all fold-changes is then used as normalization factor for the whole spectrum. This normalization method is robust against large changes in few metabolites [138].

Total area normalization computes a factor from the accumulated sum of all features in one spectrum. In this method, alterations of highly concentrated metabolites may affect the normalization factor and therefore influence the latter analysis [139].

Quantile normalization forces an identical peak intensity distribution through the dataset. To normalize the spectra to each other, each vector of features is sorted and then the arithmetic mean of the distribution is calculated. Following that, the mean of identical quantiles is calculated and assigned to all features realizing the corresponding quantile. This approach can be problematic with highly variable metabolites, as these may differ strongly between samples [139,145].

This incomplete listing and brief discussion of normalization techniques reflects the complexity of metabolomics data processing.

Because of the necessity to detect minimal variations in highly noise-prone datasets, it is important to be extremely precise and thoughtful. Each step must be carefully considered with respect to the scientific question. It is therefore very difficult to establish a standard procedure, which must be considered on a case-by-case basis [146].

1.2.5 Peak alignment and data reduction

As abovementioned, proton resonances can be, besides other reasons, affected by variations in ionic strength. Despite thoughtful adjustment of experimental parameters, these factors may influence the resulting spectra [147]. The major effect of these variations is the so-called positional noise, which is a drift of resonance positions along the chemical shift axis [148,149]. In order to provide a good foundation for the subsequent statistical analyses, a correction of the position noise must be carried out. An optimal result is desired, as the often subtle differences of the metabolic fingerprint should be detected [150].

Over the decades several strategies emerged targeting this issue, including but not limited to different warping approaches such as dynamic time warping (DTW)[151] or correlation optimized warping (COW) [152] and recursive segment-wise peak alignment (RSPA)[153].

Warping techniques are based on expanding and contracting the x-axis to make the spectrum, or its peaks, as similar as possible to a reference spectrum. Since both expansion and contraction may be required within the individual regions of a spectrum, the spectra are divided into individual segments. These are then either stretched or compressed according to the reference spectrum [150,148,154].

The RSPA approach segmentates the spectrum and iteratively reduces segment size and shifts peaks within these segments until a sufficient correlation to the reference spectrum segment is reached. The alignment process is always a balancing act to find the optimum where slightly shifted peaks from the same substance are shifted to the same position without mixing spatially close peaks from different substances [155,153].

After the position noise has been carried out as best as possible, some kind of binning approach usually follows.

Binning, also known as bucketing, describes the segmentation of a whole spectrum into discrete frequency bands, typically with a size between 0.04 and 0.001 ppm [156,157,149,158]. These buckets are then integrated and the numerical values are used for further analysis.

Depending on the extent of the peak drift, it can happen that not all peaks could be shifted directly over each other after an alignment. One tries to compensate for this error by binning, since smaller or larger residual shifts are eliminated depending on the width of the bin as the values within the bins collapse.

However, depending on the extent of peak drift and the size of the buckets, peaks may be either spread across several buckets between samples or too many signals are included in one bucket. If too many peaks fall into a bin, the total variation may mask biologically relevant variation within that bin [150].

This problem is addressed by intelligent binning approaches. Intelligent binning tries to set the binning boundaries sensibly, for example at the local minima between two adjacent peaks. The aim of this approach is that large peaks, and therefore also wider peaks, are not pulled apart and are distributed over several bins, and at the same time smaller peaks are not lost in bins that are too large [159–161].

Additionally to other spectral manipulations, sub-spectral filtering can improve data quality. This filtering technique eliminates non-informative regions, e.g. the water signal region or areas, where no signals occur [162]. An inclusion of such noisy regions in multivariate analysis has been shown to have a negative impact on model performance [163].

On top of the position noise, the overlapping of the signals also limits the interpretability considerably [164]. In Figure 7 a common urine spectrum is shown including some peaks annotated. Especially in the region between 3 and 4 ppm massive signal overlap occurs. Contrary, between signals in the area from 0.9 to 2 ppm non-informative baseline is present.

To improve the performance of the analysis, it is advisable to exclude these non-informative regions from the statistics. To achieve this goal, one can use some intrinsic properties of the NMR spectra. A substance is determined in an NMR spectrum mainly by two characteristics, these are the chemical shift and the splitting pattern.

If it is known which substance is being searched for, the corresponding splitting pattern can be targeted specifically in the regions where the substance resonates. The ratios of the individual peak heights and multiplets to each other can be used to extract peaks from overlapping regions.

Among the metabolomics community a variety of methods and tools are applied to perform this kind of data reduction.

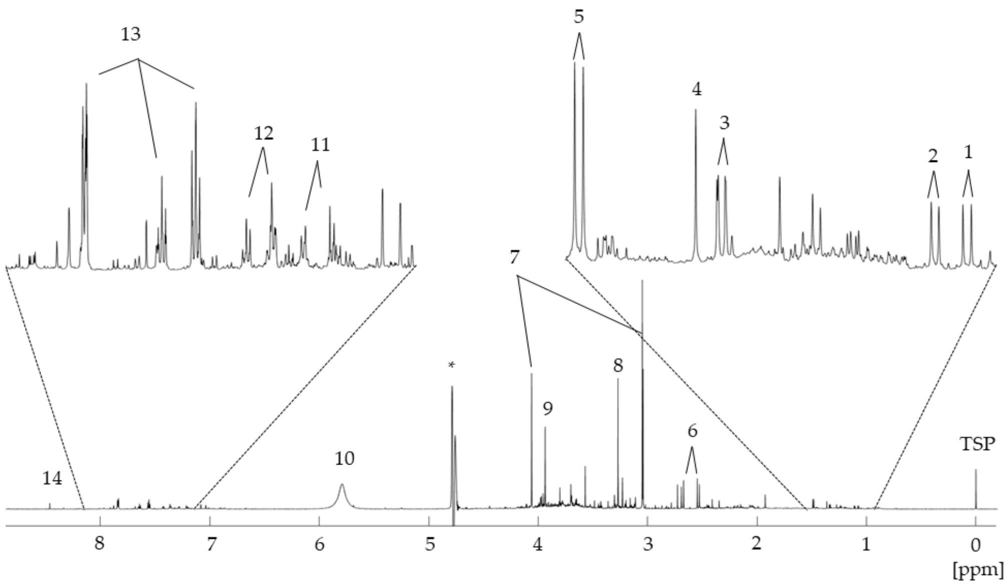


Figure 7: typical urine spectrum recorded on a 800 MHz spectrometer, 1: 3-Aminoisobutyric acid, 2: 4-deoxythreonic acid, 3: lactate, 4: 3-hydroxyisobutyrate, 5: alanine, 6: citrate, 7: creatinine, 8: trimethylamine-N-oxide, 9: creatine, 10: urea, 11: p-cresol-sulfate

Several peak fitting algorithms, such as BATMAN [165], BAYESIL [166] or NMRProcFlow [167] are freely available and often used within the research community.

BATMAN is a frequently used tool for quantification of metabolites in complex biological matrices by deconvolution and integration of peaks. The applied Bayesian model relies on extensive prior information about the metabolites, such as the expected chemical shift, multiplicity, J-coupling constants, and intensity ratios and fits Lorentzian shaped peaks into the spectra based on this information. BATMAN can be downloaded free of charge as R-package [168,165].

BAYESIL offers a similar approach as web tool but includes besides deconvolution and integration based on a reference library also most necessary preprocessing steps, such as Fourier transformation, phasing, chemical shift referencing and baseline correction. Therefore, it offers the user a complete data processing tool with little user input[166].

The interactive 1D-^{1}H-NMR processing tool NMRProcFlow is also open-source software including Fourier transformation, baseline correction, chemical shift referencing, several alignment algorithms, options for equidistant bucketing and intelligent bucketing and normalization. The tool also provides options for quantification using an external metabolite library [167].

There is also commercial software from companies that provide not only the evaluation algorithms but also corresponding satellite databases. One example is the NMR Suite Software Package (Chenomx Inc., Edmonton, Canada) The advantage here is that a large amount of data can be accessed without much preparation and precise results can be generated quickly. At the expense of convenience is the flexibility of the analyses. If a certain metabolite is not in the database, it cannot be qualified with the given workflow.

Another disadvantage is that usually such software solutions cannot be applied to existing data sets unless the conditions specified by the company regarding sample preparation and measurement parameters are met.

Although the mentioned applications provide easy to use applications for metabolite quantification, the major drawback is the necessity of external metabolite libraries. This key problem precludes the application of those reliable quantification tools in untargeted metabolomics. Biomarker discovery therefore remains incredibly challenging, as only full spectra analysis (spectra without any form of data reduction) or binned spectra are applicable for this research question, where spectral noise hampers the subsequent data analysis.

The aim of this work was to fill this gap and to develop an approach that allows untargeted evaluation while keeping noise to a minimum. Untargeted approaches are indispensable, especially for diseases whose origins and mechanisms have not yet been fully elucidated by research. This branch of research enables the detection and early recognition of diseases, as well as the discovery of risk factors.

1.2.6 Statistical Tools in NMR Metabolomics

Methodologies applied in metabolomics approaches are often adapted from earlier omics techniques. Dependent of the method of choice, some preliminary considerations are required. Several statistical methods assume the data to be normally distributed with a constant variance, however, metabolomics data often has skewed distributions across samples, resulting in heteroscedastic data. Logarithmic transformation is a frequently used method to approximate the data to a normal distribution and therefore utilize it for subsequent analysis [169].

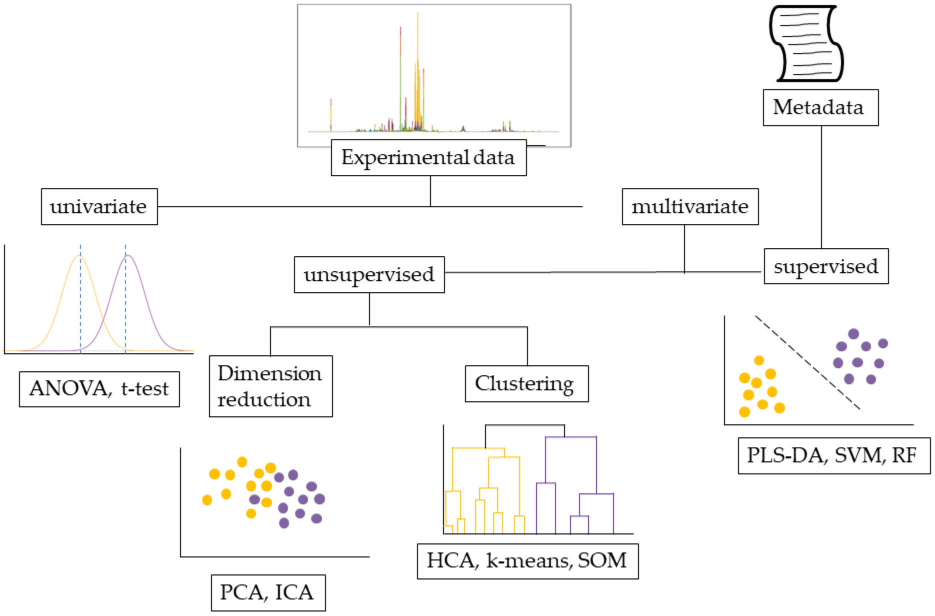


Figure 8: Schematic overview of strategies for univariate and multivariate statistical analysis with frequently used examples for each category

In general, all techniques can be classified into two main groups, the multivariate strategies and univariate strategies. A schematic overview can be found in Figure 8.

Univariate strategies, such as analysis of variance (ANOVA) or a t-test, test the effects of e.g., a medical treatment on individual metabolites and thus are suitable for hypothesis testing. Multivariate strategies, however, aim to utilize dependency structures between metabolites and therefore are suitable for hypothesis generation [169].

Multivariate techniques themselves can be further classified into unsupervised and supervised methods. Supervised methods include additional information about the expected phenotype groups within the dataset.

In unsupervised methods no additional information about underlying treatment or intervention groups is included. The group separations are entirely data driven and therefore less prone to overfitting.

One of the most frequently used methods in this group is principal component analysis (PCA). PCA reduces dimensions by projecting a large dataset into a smaller one which still contains most of the information. The PCA approach determines a new coordinate system in a least squares sense, where the new dimensions include the most variance within the dataset. The principal components are the eigenvectors of the initial covariance matrix, thus they can either be calculated by eigenvector decomposition of the covariance matrix or a singular value decomposition of the data matrix. Generally, the first few principal components contain the necessary information, which are used as a starting point for data analysis [170,171]. Individual component analysis (ICA) is closely related to PCA aiming to detect independent components in the data and has been shown to perform well in metabolomics approaches [172].

Cluster analysis represents another unsupervised multivariate method with the most prominent method being hierarchal cluster analysis (HCA) [155,173]. Clustering methods aims to identify hierarchical groups in the original dataset according to intrinsic similarities of their features, which are visualized as a dendrogram. These nested clusters are determined by the chosen similarity metric, which is generally any measure of distance such as Euclidian distance, for example. Furthermore, a linkage function needs to be set with single linkage, complete linkage and average linkage being the most common [131]. Other frequently used applications are k-means clustering [174,175] and self-organizing maps (SOM) [176,177]

Popular tools in the category of supervised multivariate analysis tools are partial least squares discriminant analysis (PLS-DA) [178], orthogonal projection to latent structures discriminant analysis (OPLS-DA) [179], random forest (RF) [180] and support vector machines (SVM) [181].

PLS-DA is a linear classification model with descriptive and predictive properties. This approach relates the data matrix to the response variable, such as the class affiliation, by weighting the initial features corresponding to their discriminating ability. The resulting model can be either used to determine the variables with maximum predictive ability or to predict class affiliations of unknown samples [178,171].

The extension of PLS-DA, the OPLS-DA method, enhances the discriminating ability of the model by splitting the variance within the data into the between-group variation and an orthogonal part, the within-groups variation. This cleavage enables an easier interpretability and thus is widely used in metabolomics approaches, where intra-group variation is generally relatively strong [182,179].

Although less frequently used, RF is a useful addition to the analysis toolbox. It is a combination of decision trees to reach best outcomes, thus the class selected by most trees. A random forest model also enables to determine the feature importance via the Gini index. This index measures the degree of probability that a particular feature is wrong when it is randomly chosen [183,180,184].

SVM are supervised learning techniques used for classification in metabolomics approaches. The SVM algorithm uses the classified dataset to detect a hyperplane with the best separation ability between two groups. The best separation is reached, when the distance to the nearest group member is largest [181,185].

Contrary to the above-mentioned hypothesis generating tools in metabolomics analysis, hypothesis testing techniques are of central importance. ANOVA and t-test based methods are deployed if the question needs to be answered weather profiles in metabolite excretion differ significantly between treatment groups [131]. T-tests are generally suitable for two groups, whereas ANOVA is the method of choice for larger group assignments. For both techniques sever tests can be applied based on the research question and data structure (e.g., paired vs. unpaired samples) [171]. It must be noted that the resulting values form these methods need to be corrected for multiple hypothesis testing. Frequently used methods are the correction methods of Bonferroni and Benjamini and Hochberg [186–188].

1.2.7 Metabolite identification

At the preliminary end of the untargeted analysis, one or more NMR signals were identified that turned out to be interesting parameters in the context of the scientific question. Identifying these signals is a major challenge. The identification is important because it allows the biological plausibility to be tested. It can happen that signals look very promising, but in retrospect it turns out that these signals are only caused by technical differences, e.g. the sampling time differs between healthy and sick patients. Furthermore, knowledge of the metabolite is important for the follow-up. Usually, untargeted approaches are used to identify potentially interesting biomarkers, which are then tested in subsequent targeted studies. In order to optimally adapt these studies to the analyte, for example with respect to sample stability and storage, the analyte must be known.

The procedure is a little simpler if one already has an idea of what the analyte might be. This can be based on previous experience and/or characteristic signals. In this case, it is simply a matter of gathering enough evidence to confirm the identity of the signal. This can be done, for example, by adding a pure substance and observing the resulting signal increase in one-dimensional spectra, comparing the signals in two-dimensional spectra with those of a pure standard or match them with databases. The significance of the options has to be assessed individually. To obtain a high level of security, several of the above-mentioned options may need to be combined.

If no potential candidates are known, a complete identification must be carried out. This is usually much more complex and time consuming. NMR signals, especially the signals in two-dimensional spectra, already provide some clues to the chemical structure of the molecules. In complex mixtures, however, it is often impossible to assign the signals of dingle resonances in the two-dimensional range due to overlap or low concentrations. It is often necessary to apply chromatographic techniques to separate the mixture in advance and enrich the analytes to obtain sufficient concentrations to generate a signal. It can also be advantageous, if strengths from different techniques are used synergistically to obtain meaningful results. The results of NMR spectroscopy are usually complemented by orthogonal methods such as MS, infrared and ultraviolet spectroscopy [189].

This is usually done by considering the respective properties of the methods together, for example the multiplicities, the mass of the molecule and the fragmentation pattern. Newer approaches aim to combine different methods with computer-based methods. Such a hybrid approach combining NMR with MS was introduced by Bingol et al. 2015, which is termed SUMMIT MS/NMR. This technique omits the purification step; instead, all masses are assigned to their possible chemical formulas and NMR spectra are then predicted. The extent of the matches then allows conclusions to be made about the identity of individual signals [190].

The identification of individual substances from complex mixtures is a challenging field of research. The main problems here are the generally low concentration and the overlapping of signals. These disadvantages occur both in NMR spectroscopy and in the use of mass-based methods.

2-dimensonal (2D) NMR techniques, together with databases and statistical approaches, can aid to address this problem. 2D spectra resolve the resonances by extension into a second dimension according to another physical property. This solution solves the problem of peak overlap in many cases and offers new information about the metabolite of interest. In the following, selected 2D experiments will be described and their usefulness for metabolite identification will be explained.

Correlation spectroscopy (COSY) is often applied to detect through bond coupling between coupled nuclei, as it is a very simple and fast 2D NMR experiment, which is also easy to interpret [191,192]. The COSY experiment is based on the transfer polarization by a mixing pulse between directly J-coupled spins and thus providing information about the direct environment of the resonance proton.

Total Correlation Spectroscopy (TOCSY) is an extension of the COSY experiment creating correlations between all protons in a given spin system, not restricted to only germinal and vicinal protons. Heteroatoms, such as oxygen, disrupt the TOCSY transfer. The number of observable transfer steps can be adjusted by the mixing time [193,194,104].

Another one and often underrated experiment is the 2D J-resolved spectroscopy experiment (Jres). The Jres, as other 2D experiment, simplifies the initial spectrum by the distribution into another dimension, but instead of couplings to other resonances, the Jres separates the scalar couplings of a resonances multiplet into the second dimension. The multiplicity is displayed along f1 axis and chemical shift along f2, which allows the assignment of resonances in crowded regions to specific multiplet [195].

The $^{1}H,^{13}C$ Heteronuclear Single Quantum Coherence (HSQC) Spectroscopy experiment maps the proton resonances and those of the carbon atoms where the protons are directly attached to. This approach is an inverse detection method where the magnetization from the sensitive proton nucleus is transferred to the less sensitive carbon nucleus, which leads to drastically reduced acquisition times compared to direct methods such as $^{1}H,^{13}C$ COSY [189].

Additional to the HSQC experiment, the Heteronuclear Multiple Bond Correlation (HMBC) Spectroscopy reveals correlations between heteroatoms separated by two or more bonds. Here the single bond correlation is eliminated by application of a low pass filter only allowing smaller J-couplings. This experiment is extremely useful for assignment of quaternary and carbonyl carbons, which cannot be detected by HSQC [196].

Although this variety of 2D NMR experiments provides complementary information about the resonance of interest, for most metabolites a complete assignment and identification is not feasible.

Additional information can be provided in some cases through the application of Statistical Total Correlation Spectroscopy (STOCSY) [197]. This computational approach enables the simplified assignment of resonances deriving from one metabolite through correlations among the signals. A pseudo 2D spectrum is created displaying the correlation of intensities over the spectrum. This method can be applied to aid metabolite identification without often time-consuming experiments, but it can only be applied on a large enough dataset, as otherwise the correlations would not be detected properly.

For final identification of a metabolite, the confirmation either via spiking or via comparison with reference spectra is needed. If the intrinsic properties of the metabolite, such as multiplicity, chemical shift and coupling pattern align with spectra of a pure compound, the analyst can confirm the metabolite identity. Several open access databases exist, such as the human metabolome database (HMDB) [58], the NMRShiftDB [198] or the MetaboMiner database [147]. HMDB is the largest database containing authentic NMR spectra for biofluid interpretation. As of March 24th 2023, the database has a total of 253,245 metabolite entries and contains 242,268 NMR spectra (1D and 2D) for a total of 12,345 compounds. From these numbers it can be deduced that only 5% of the metabolites have also recorded the corresponding NMR spectra. These databases are expanding both in quality and quantity of reference spectra, nevertheless they only cover all small number of metabolites completely.

As described above, the responsibility for the correct labeling of metabolites lies with the users. Since there are no universally accepted principles there is a lack of quality control. The research group around Sumner already proposed in 2007 to agree on minimal reporting standards within the community [199]. They suggested a rather vague classification into 4 groups, which is shown in the following table.

Table 4: The 4 Levels of metabolite identification proposed by the metabolomics standards initiative [199]

Level 1	**Identified Compound**	At least two orthogonal data in direct comparison with spiking of an authentic reference standard.
Level 2	**Putatively Annotated Compound**	No reference standards are used, annotation is based on spectral data and/or the similarity to spectra in data repositories.
Level 3	**Putatively Characterized Compound Class**	Based on physicochemical or spectral properties the compound can be assigned to a chemical class of compounds.
Level 4	**Unknown Compound**	Only spectral data describes the compound, otherwise it is unidentified and unclassified.

In this classification scheme, many points remain very vague, for example, no statement is made as to what extent the metabolite to be identified must correspond to the reference standard. Nevertheless, this categorization must be understood as an important basis for establishing generally applicable standards within the community.

Since this categorization is still not very widespread, more effort must be invested in general acceptance and usage.

1.3 Thesis Structure and Objective

The aim of this work was to investigate the scientific possibilities in the research field of NMR metabolomics and to contribute to the current state of the art by carefully complementing the existing methods and techniques. After a thorough literature review, sample preparation, data acquisition, data processing, statistical analysis and metabolite identification were identified as key issues within the workflow (see Figure 9). Before this work began, the current literature, a summary of which can be found in the previous sections, was examined and the points on which there has been little or no focus were identified. The research revealed that there is certainly room for improvement at each of the individual sticking points, but the points of sample preparation and data processing have received the least consensus so far. Figure 9 shows each of the key points as hurdles that must be overcome in the context of a NMR metabolomics research project. The size of the hurdle is symbolic of the amount of work that the author believes is necessary in the respective areas to achieve a general consensus and standardized conditions within the community.

As two separate key points could have been identified, a research question was defined for each point.

First research question

How does the amount of deuterium and the storage condition of the finished urine samples before and during measurement influence the spectrum, how can this effect be described and what are the possibilities to avoid this?

In chapter 2 the influence of sample preparation on interpretability of results in NMR metabolomics datasets is described on the example of hydrogen-deuterium exchange in creatinine, one of the major metabolites in human urine. The time and concentration dependency were systematically investigated. As the metabolomics community was not aware of this effect, several public available metabolomics datasets were affected from the transformation of the creatinine resonance. Therefore, a correction equation was introduced to allow the post-analysis adjustment.

Second Research Question

Is there a way to reduce the noise within the spectral data in a way that improves the subsequent statistical analysis without losing the untargeted character?

In chapter 3, the performance of established and frequently used data processing methods was evaluated and the strong influence of noisy data is shown. To improve data quality and simplify the analysis, a peak fitting algorithm based on a Voigt lineshape was developed. Contrary to current techniques, the introduced algorithm does not rely on a reference database, which enables the use for untargeted analysis. The applicability was shown on a real dataset and systematically compared with other methods, such as full spectra analysis and equidistant binning. It could have been shown, that using Voigt fitted data as input layer for unsupervised (PCA) and supervised (OPLS-DA) analysis improves the descriptive and predictive ability in untargeted NMR metabolomics approaches.

The appendix contains the original publication of chapter 2 and 3 including the supplementary information.

Summarizing, this thesis pointed out the necessity of coherent sample preparation, experimental setup, and data processing in the field of NMR metabolomics applying different kinds of analytical and statistical approaches with the focus on NMR.

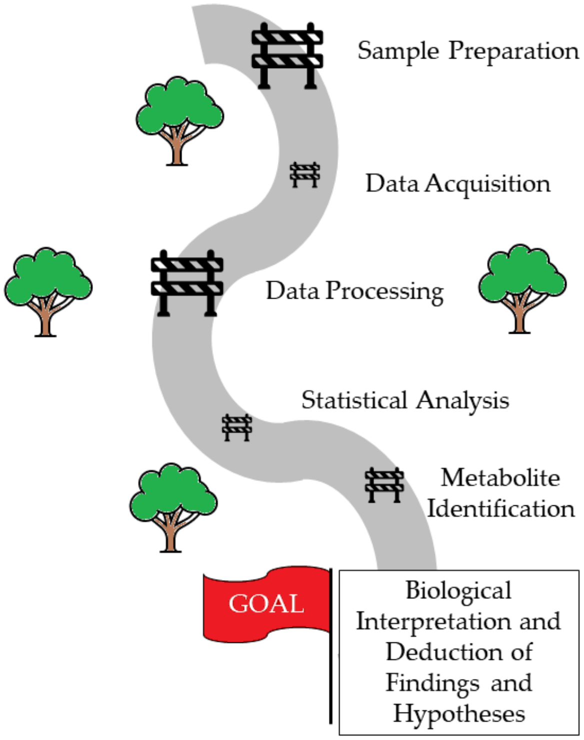

Sample Preparation
Data Acquisition
Data Processing
Statistical Analysis
Metabolite Identification
GOAL
Biological Interpretation and Deduction of Findings and Hypotheses

Figure 9: Roadmap for a metabolomics workflow

This page intentionally left blank

Chapter 2|

Guidelines for the Use of Deuterium Oxide (D_2O) in ^1H NMR Metabolomics

Abstract

In metabolomics, nuclear magnetic resonance (NMR) spectroscopy allows to identify and quantify compounds in biological samples. The sample preparation generally requires only few steps; however, an indispensable factor is the addition of a locking substance into the biofluid sample, such as deuterium oxide (D_2O). While creatinine loss in pure D2O is well-described, the effects of different D_2O concentrations on the signal profile of biological samples are unknown. In this work, we investigated the effect of D_2O levels in the NMR buffer system in urine samples, in dependence on dwell time and temperature exposition. We reveal a decrease of the urinary creatinine peak area up to 35% after 24 h of dwell time at room temperature (RT) using 25% (v/v) D_2O, but only 4% loss using 2.5% D_2O. ^1H, inverse-gated (IG) ^{13}C, DEPT-HSQC NMR, and mass spectrometry (MS) experiments confirmed a proton−deuterium (H/D) exchange at the CH_2. This leads to underestimation of creatinine levels and has an extensive effect when creatinine is used for normalization. This work offers a sample stability examination, depending on the D_2O concentration, dwell time, and temperature and enables a method to correct for the successive loss. We propose an equation to correct the creatinine loss for samples prepared with various D_2O concentrations and storage temperatures for dwell times up to 24 h. The correction function was validated against an external data set with n = 26 samples. To ensure sufficient creatinine stability in future studies, we suggest

that a maximum of 10% D_2O should be used at 4 °C or 2.5% D_2O at RT, respectively.

This chapter was published as Haslauer, K. E., Hemmler, D., Schmitt-Kopplin, P., & Heinzmann, S. S. (2019). Guidelines for the Use of Deuterium Oxide (D2O) in 1H NMR Metabolomics. *Analytical chemistry*, *91*(17), 11063-11069.

Candidate's contributions: K.E. Haslauer designed the research, performed the NMR experiments and analyzed the data. K.E. Haslauer prepared the figures, wrote and revised the manuscript.

Chapter 3|

Data Processing Optimization in Untargeted Metabolomics of Urine Using Voigt Lineshape Model Non-Linear Regression Analysis

Abstract

Nuclear magnetic resonance (NMR) spectroscopy is well-established to address questions in large-scale untargeted metabolomics. Although several approaches in data processing and analysis are available, significant issues remain. NMR spectroscopy of urine generates information-rich but complex spectra in which signals often overlap. Furthermore, slight changes in pH and salt concentrations cause peak shifting, which introduces, in combination with baseline irregularities, un-informative noise in statistical analysis. Within this work, a straight-forward data processing tool addresses these problems by applying a non-linear curve fitting model based on Voigt function line shape and integration of the underlying peak areas. This method allows a rapid untargeted analysis of urine metabolomics datasets without relying on time-consuming 2D-spectra based deconvolution or information from spectral libraries. The approach is validated with spiking experiments and tested on a human urine ^1H dataset compared to conventionally used methods and aims to facilitate metabolomics data analysis.

This chapter was published as Haslauer, K. E., Schmitt-Kopplin, P., & Heinzmann, S. S. (2021). Data Processing Optimization in Untargeted Metabolomics of Urine Using Voigt Lineshape Model Non-Linear Regression Analysis. *Metabolites*, *11*(5), 285.

Candidate's contributions: K.E. Haslauer designed the research, performed the experiments and analyzed the data. K.E. Haslauer prepared the figures, wrote and revised the manuscript.

This page intentionally left blank

Chapter 4|

Concluding Discussion and Outlook

This thesis reports on methodological tropics in the research field of NMR metabolomics. Analytical methods and data analysis strategies were developed to enable and simplify the comprehensive study of large cohort datasets. The introduced approaches were shown to provide solutions for known (chapter 3) and unknown (chapter 2) pitfalls in this research field. The progress towards a suitable workflow for NMR metabolomics research is of great importance and one of the fundamentals towards the understanding of influence factors of the human metabolism.

First Research Question

How does the amount of deuterium and the storage condition of the finished urine samples before and during measurement influence the spectrum, how can this effect be described and what are the possibilities to avoid this?

Conclusion

Chapter 2 started during examination of a NMR metabolomics dataset and some remeasurements, where a decrease in the CH_2 resonance of creatinine ($\delta = 4.06$ ppm) was observable over time with a simultaneous rise of a triplet slightly upfield. ($\delta = 4.04$ ppm). The research question for this work was the examination and reveal of the underlying mechanisms causing this time dependent transformation. The assignment of the triplet resonance was performed via different NMR experiments, such as inverse gated (IG) ^{13}C measurement with proton decoupling using a WALTZ-16 sequence to eliminate a Nuclear Overhauser effect (NOE) and multiplicity edited ^{1}H-^{13}C-HSQC (DEPT-HSQC). The selection of those experiments was shown to be appropriate for the hypothesis testing of a proton-deuterium (H/D) exchange, as an IG-^{13}C experiment is able to conflate the shift changes and observable in 1D experiments with the multiplicity. Also peaks for a double H/D exchange were observable, which are not detected in a 1D-^{1}H experiment. The characteristic splitting patterns are due to different spin systems and proton decoupling. Splitting of resonances is caused by the influence of the small magnetic fields produced by the spin of nuclei. The number of splitting is determined by the number of neighboring nuclei following the NI+1 rule with N is the number of neighboring protons and I is the nuclear spin quantum number. Hydrogen is a spin ½ nucleus, whereas deuterium is a spin 1 nucleus resulting in a triplet for a H/D exchange on one position and a quintet for a H/D exchange of both protons at the (3,4)-position of creatinine. This splitting pattern could have been verified with NMR and further confirmed with high-resolution electrospray ionization–mass spectrometry (ESI-MS) as an orthogonal method. The MS spectrum affirmed the hypothesis of a H/D exchange in creatinine through the presence of all three mass-to-charge rations (m/z) (m/z 114.069 for $[C_4H_7N_3O+H]^+$, 115.076 for $[C_4H_6DN_3O+H]^+$, and 116.081 for $[C_4H_5D_2N_3O+H]^+$). The extent of exchange was systematically investigated for the dependence on time, temperature and D_2O concentration and a sampling handing guideline was introduced to avoid such conversions. Additionally, a correction equation was proposed to recalculate initial creatinine levels based on the CH_2 to CHD ratio. The equation then was successfully applied to correct deuterium dependent creatinine loss in a test dataset.

Implications of research

The fact that creatinine is excreted at a relatively constant rate in healthy individuals is widely accepted, especially in the medical field. Therefore, when testing parameters from urine, creatinine is often used for normalization in order to compensate for the dilution effect of hydration. When such an assumption is considered to be valid, little emphasis is placed on continuous testing of the accuracy of this assumption. Because there is usually no possibility of using an internal standard in such applications, errors in normalization are rarely detected.

The characterization of creatinine deuteration and the determination of factors influencing this transition raises awareness among the NMR metabolomics community towards such seldom noticed but frequently occurring issues. It is of immense importance that such phenomena, once observed, are investigated and communicated within the research community to enable continuous improvement of techniques and applications.

In order to make already measured data sets, in which this error occurred, nonetheless usable, the dynamics of deuteration were investigated with respect to the influence factors of time, temperature and final concentration, and a correction equation was introduced. Here it was very important not to include such parameters as time and temperature in the equation, as these are usually not known and therefore not applicable.

The research work presented in chapter 2 contributes both as a practical guide and as a reminder for the ongoing improvement of the research field of untargeted metabolomics. Various extraordinary reviews have already pointed to the lack of consensus regarding sample preparation within the research community [200,111,125]. This work clearly expresses the importance of this topic.

Second Research Question

Is there a way to reduce the noise within the spectral data in a way that improves the subsequent statistical analysis without losing the untargeted character?

Conclusion

The aim of the work published in chapter 3 was to establish an algorithm to select peaks above a user defined signal to noise ratio and approximate the underlying peak area sufficiently. Especially data processing and there within peak fitting and deconvolution has been neglected for a long time for untargeted approaches. This may mainly be caused by the challenging nature of this task, as peak shapes are influenced by many aspects. The peak widths differ among the resonances within a spectrum and their affinity for peak broadening is also metabolite specific. Peak broadening can be caused by chemical exchange, ionic strength or paramagnetic compounds in the sample. Theoretically, NMR peaks have a Lorentzian lineshape, but due to the peak broadening Gaussian lineshapes occur. The ratio of Lorentzian and Gaussian components can vary between exclusively Lorentzian and exclusively Gaussian. The Voigt lineshape is a convolution of Lorentzian and Gaussian shapes and was described as appropriate approximation for quantification in NMR [201,202].

It has been shown, that although without a priori knowledge of multiplicity, ratio between resonances and expected chemical shift, a peak fitting based on a Voigt line shape provides semi-quantitative data within acceptable deviations.

Using a peak-fitting approach based on a least-squares approximation of the Voigt line model, along with aligning the peak integrals to a reference spectrum, can achieve this goal. This type of peak alignment reduces the distortion in data analysis caused by noise by eliminating the non-informative regions. This is done automatically and thus sets identical benchmarks for all peaks. Where previously there was a lot of room for interpretation by the analyst, this newly developed algorithm allows general and, above all, comparable standards to be set. The quality parameters provided, such as the sum of squares of the residuals, can then be used to assess the quality of the generated data and adjust the settings accordingly if necessary.

Implications of research

With the introduction of an untargeted peak fitting algorithm, the naturally occurring noise in NMR metabolomics datasets can be drastically reduced and therefore subsequent statistical analysis can be enhanced. It was shown that the application of such a noise reduced dataset for supervised and unsupervised statistical methods improves both, the ability of the models to distinguish between groups of the study cohort and the predictive ability for assignment of unknown samples. Through the data reduction and simplification of data, the operator's task is simplified as well and allows an extended level of comparability and standardization compared to other methods.

Because the algorithm determines the quality of the individual fit, individual results can be better classified and compared.

It can be observed that many research groups are working on the extension and improvement of database driven (semi) quantification tools. This effort is motivated by the fact that for further development of NMR-based metabolomics, it is essential to generate numerical values that are as precise as possible and reflect the content of the respective metabolite.

Conversely, the major disadvantage is that only metabolites included in the database can be screened. Depending on the database used, there is a greater or lesser risk that the biologically interesting changes will go undetected if that metabolite is not reported in the library. Especially when referring to untargeted metabolomics, which is meant to be used to generate hypotheses, it is obvious that these methods are out of question.

The algorithm presented here closes this gap by providing a solid compromise between absolute undirected binning or full spectra analysis and targeted quantification tools.

A further development could be thought in the combination of the database supported methods with the presented algorithm. One possibility would be to weight the resulting numerical data according to confidence levels and thus create stepwise gradations between e.g. *quantified with reference database*, *good fit*, *medium fit* and *poor fit*. In this way, a pre-selection of results could take place (e.g. based on p-values and ranked by confidence levels). This combination of the originally two different approaches allows simultaneous quantitative and targeted evaluation without losing potentially interesting parameters.

Outlook

Metabolomics is a relatively new research discipline and although the potential information is extremely promising, the community lacks standard operation procedures or on agreed best practice advice. On one hand, this exploratory nature allows the researchers the academic freedom to design their very own workflows and analytical techniques, on the other side this lack of agreement opens the chances for pitfalls. As each research group follows their own practice, almost no external revision happens, and analytical and technical procedures are not scrutinized.

At this point, it should be emphasized how important a certain standardization within a research area is for the success of this field [203,204].

A couple of comments and suggestions have already been made about the importance of standardization and harmonization of scientific practice. The main issue is that in our now highly interconnected world, the expectations for scientific data in terms of usability and comparability have increased. This is partly due to the endeavors of the Open Data movements, which want to make science more inclusive as well as more sustainable [205]. By providing data sets in public repositories, these resources can be reused for other purposes and to investigate further research questions. Furthermore, this free availability of data also allows the use by scientists from regions where little government funding is available for the purchase of often highly expensive analytical equipment. To make this possible, harmonization within the community is important, as this is the only way to enable meaningful exchange and reuse and pooling of data sets.

Another reason why standardization is necessary is the potential to detect errors. Through a consensus on best practice, every scientist in the field would be able to unambiguously interpret the raw data and, if necessary, indicate errors. And even though this correction is often unpleasant, it provides the basis for self-critical, self-correcting, and continuously evolving science. These are the fundamental principles of science that need to be preserved and promoted.

The present work has contributed to this further development in two ways. Firstly, through concrete proposals that enable standardization, and secondly, through the publication of such work, an increasing awareness of these concerns is being created. Especially in the early phase of a research field, exchange and discourse within the research community is a very valuable asset.

Regarding the general outlook of the research direction of NMR metabolomics, it can be summarized that the discipline of NMR metabolomics has become popular as it offers the possibilities to answer crucial questions of the nowadays scientific community in the biomedical and pharmaceutical field. Metabolomics approaches may contribute to the discovery of new diagnostic biomarkers for diseases as well as revealing the underlying metabolic alterations. This information can help to understand the dynamics and evolution of some of the most dramatic and complex diseases of this century, such as diabetes and coronary diseases. The fundamental understanding of a disease and the underlying mechanisms how external factors influence the human body is crucial for early prevention [206–208]. This understanding will also be important for the further development of the personalized medicine approach.

The scope of this technology has already reached the point where industrial health tech companies, such as the Nightingale Health Plc (Finland) or lifespin GmbH (Germany), are looking to harness this approach to enable more advanced healthcare.

In order to fully exploit the possibilities of metabolomics, especially by means of NMR, several steps are still necessary. Overall, the research field is currently in a very explorative initial stage, in which the individual steps of the workflow, the possible areas of application and evaluation methods are to be generated and discussed.

In this work, after extensive literature research, the two points identified by the author as the most important key points were addressed. Nevertheless, the other parts of the workflow must also be carefully examined. Critical points of the workflow, starting at the beginning of sample collection, are described in the following.

Sampling time, frequency, storage conditions and on-site preparations (e.g. addition of bacteriostatic agents) vary strongly between studies and are often not reported in detail [125]. A systematic review of the best-practice sampling method, which finds a good compromise between maximizing the scientific significance and the feasibility of the study, could provide information here and could serve as a gold standard for further studies.

There is a basic agreement on the measurement method as such, as already described in the introduction. Nevertheless, it would be useful to introduce a certain quality assurance, e.g. the assessment of the mean peak width and other NMR parameter. The obligatory reporting of such standard parameter sets allows the comparison of NMR data, the estimation of the dispersion range of different parameters in comparison with other laboratories and facilitates the repetition of experiments.

For the subsequent data processing, including baseline correction, phasing and shift referencing, guidelines have already been published [162]. To put it in general terms, it would be very desirable to agree on a minimum level of reporting of these values. Often, these processing steps are not or only partially reported in publications [209].

The following step in the metabolomics workflow represents a greater hindrance in the sense of harmonization.

Methods of statistical analysis are difficult to standardize since they must always be adapted to the individual case. Nevertheless, a reporting standard would also have to be introduced, covering topics such as the handling of missing values, weightings and statistical significance. In particular, the well-known problem of p-hacking is an issue when analyzing large amounts of data, as is the case with metabolomics approaches.

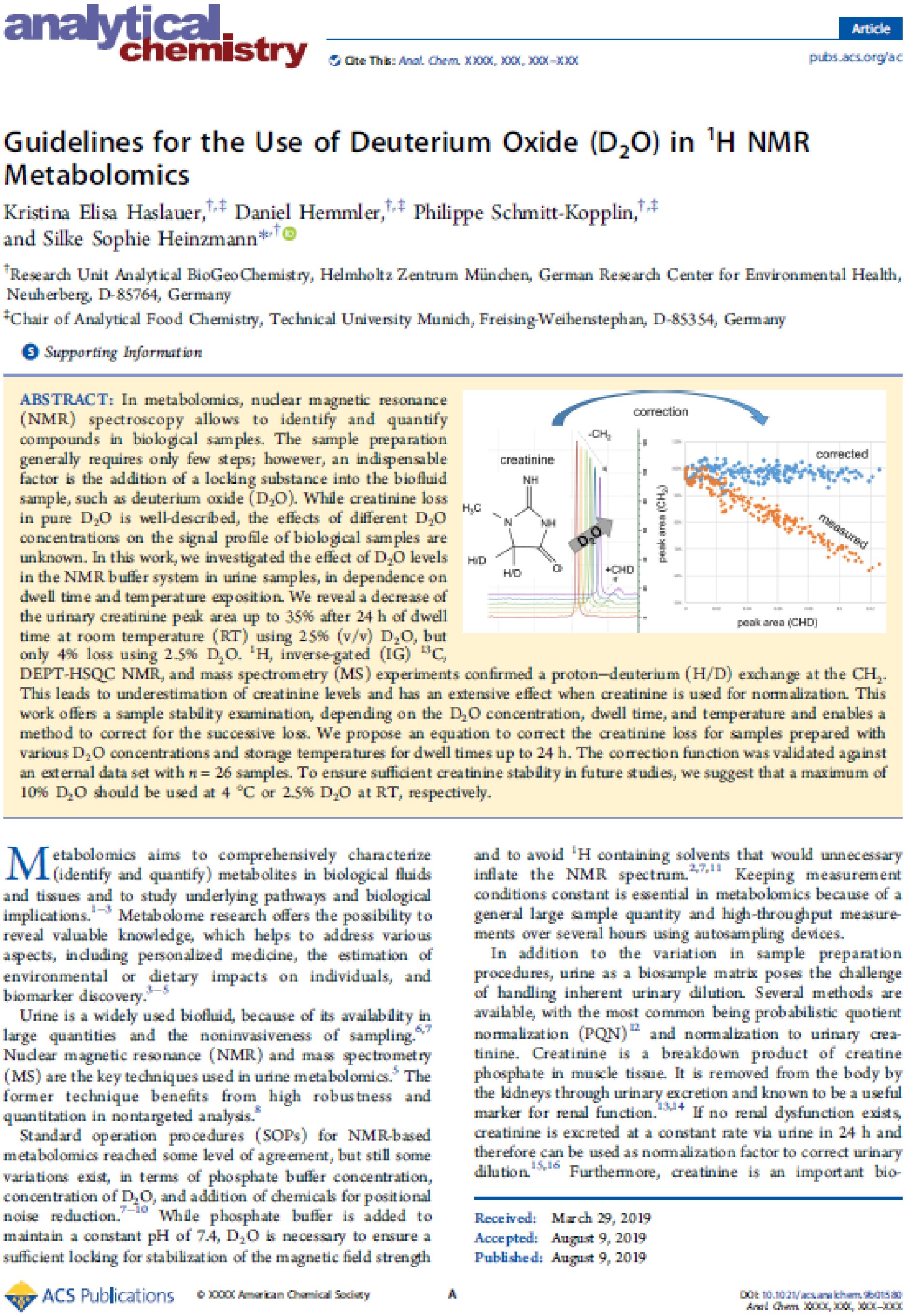
Metabolite identification is likely to be the most developed part of metabolomics workflows at this point in time, since, as written in the introduction, the foundations have already been laid and proposals have been made to standardize reporting standards and to meet certain scientific standards.

In summary, there is still a lot of work to be done. However, the relatively short existence of this research area also provides many opportunities for scientists to shape and contribute to forming the standards. In conclusion, this effort will be worthwhile since this technique will be another piece of the puzzle in systems biology, and we will be able to continue to deepen our understanding of the human organism. Generation of a fundamental and constantly expanding knowledge of the human organism, pathogenesis and prevention is the great challenge of our time.

This page intentionally left blank

A. Appendix Chapter 2

A. 1 Original Publication



analytical chemistry

Article

Cite This: Anal. Chem. XXXX, XXX, XXX−XXX

pubs.acs.org/ac

Guidelines for the Use of Deuterium Oxide (D_2O) in ^{1}H NMR Metabolomics

Kristina Elisa Haslauer,[†,‡] Daniel Hemmler,[†,‡] Philippe Schmitt-Kopplin,[†,‡] and Silke Sophie Heinzmann[*,†]

[†]Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, D-85764, Germany

[‡]Chair of Analytical Food Chemistry, Technical University Munich, Freising-Weihenstephan, D-85354, Germany

Supporting Information

ABSTRACT: In metabolomics, nuclear magnetic resonance (NMR) spectroscopy allows to identify and quantify compounds in biological samples. The sample preparation generally requires only few steps; however, an indispensable factor is the addition of a locking substance into the biofluid sample, such as deuterium oxide (D_2O). While creatinine loss in pure D_2O is well-described, the effects of different D_2O concentrations on the signal profile of biological samples are unknown. In this work, we investigated the effect of D_2O levels in the NMR buffer system in urine samples, in dependence on dwell time and temperature exposition. We reveal a decrease of the urinary creatinine peak area up to 35% after 24 h of dwell time at room temperature (RT) using 25% (v/v) D_2O, but only 4% loss using 2.5% D_2O. ^{1}H, inverse-gated (IG) ^{13}C, DEPT-HSQC NMR, and mass spectrometry (MS) experiments confirmed a proton−deuterium (H/D) exchange at the CH_2. This leads to underestimation of creatinine levels and has an extensive effect when creatinine is used for normalization. This work offers a sample stability examination, depending on the D_2O concentration, dwell time, and temperature and enables a method to correct for the successive loss. We propose an equation to correct the creatinine loss for samples prepared with various D_2O concentrations and storage temperatures for dwell times up to 24 h. The correction function was validated against an external data set with $n = 26$ samples. To ensure sufficient creatinine stability in future studies, we suggest that a maximum of 10% D_2O should be used at 4 °C or 2.5% D_2O at RT, respectively.

Metabolomics aims to comprehensively characterize (identify and quantify) metabolites in biological fluids and tissues and to study underlying pathways and biological implications.[1−3] Metabolome research offers the possibility to reveal valuable knowledge, which helps to address various aspects, including personalized medicine, the estimation of environmental or dietary impacts on individuals, and biomarker discovery.[3−5]

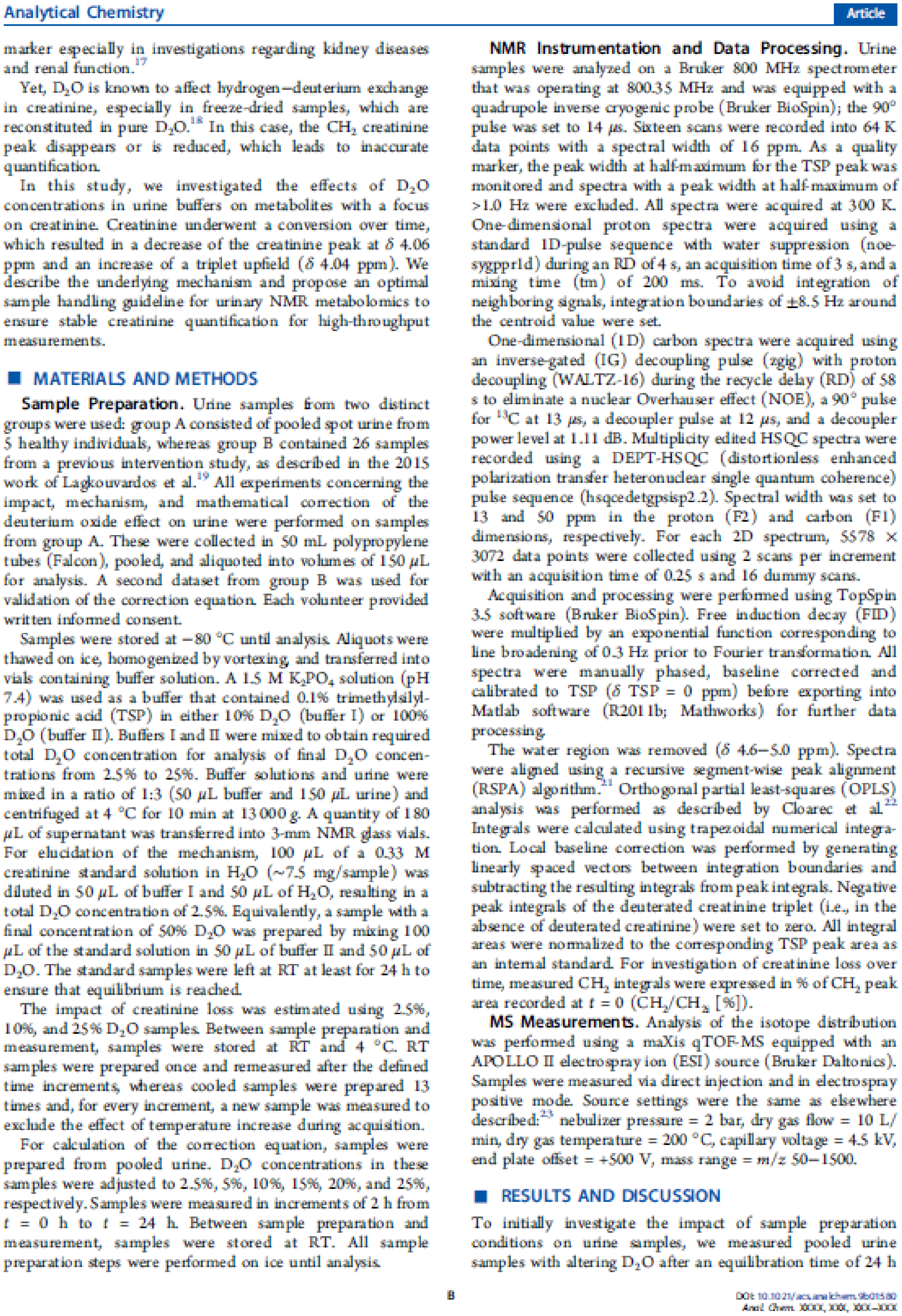
Urine is a widely used biofluid, because of its availability in large quantities and the noninvasiveness of sampling.[6,7] Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are the key techniques used in urine metabolomics.[5] The former technique benefits from high robustness and quantitation in nontargeted analysis.[8]

Standard operation procedures (SOPs) for NMR-based metabolomics reached some level of agreement, but still some variations exist, in terms of phosphate buffer concentration, concentration of D_2O, and addition of chemicals for positional noise reduction.[7−10] While phosphate buffer is added to maintain a constant pH of 7.4, D_2O is necessary to ensure a sufficient locking for stabilization of the magnetic field strength and to avoid ^{1}H containing solvents that would unnecessary inflate the NMR spectrum.[2,7,11] Keeping measurement conditions constant is essential in metabolomics because of a general large sample quantity and high-throughput measurements over several hours using autosampling devices.

In addition to the variation in sample preparation procedures, urine as a biosample matrix poses the challenge of handling inherent urinary dilution. Several methods are available, with the most common being probabilistic quotient normalization (PQN)[12] and normalization to urinary creatinine. Creatinine is a breakdown product of creatine phosphate in muscle tissue. It is removed from the body by the kidneys through urinary excretion and known to be a useful marker for renal function.[13,14] If no renal dysfunction exists, creatinine is excreted at a constant rate via urine in 24 h and therefore can be used as normalization factor to correct urinary dilution.[15,16] Furthermore, creatinine is an important bio-

Received: March 29, 2019
Accepted: August 9, 2019
Published: August 9, 2019

DOI: 10.1021/acs.analchem.9b01580
Anal. Chem. XXXX, XXX, XXX−XXX



marker especially in investigations regarding kidney diseases and renal function.[17]

Yet, D_2O is known to affect hydrogen−deuterium exchange in creatinine, especially in freeze-dried samples, which are reconstituted in pure D_2O.[18] In this case, the CH_2 creatinine peak disappears or is reduced, which leads to inaccurate quantification.

In this study, we investigated the effects of D_2O concentrations in urine buffers on metabolites with a focus on creatinine. Creatinine underwent a conversion over time, which resulted in a decrease of the creatinine peak at δ 4.06 ppm and an increase of a triplet upfield (δ 4.04 ppm). We describe the underlying mechanism and propose an optimal sample handling guideline for urinary NMR metabolomics to ensure stable creatinine quantification for high-throughput measurements.

■ MATERIALS AND METHODS

Sample Preparation. Urine samples from two distinct groups were used: group A consisted of pooled spot urine from 5 healthy individuals, whereas group B contained 26 samples from a previous intervention study, as described in the 2015 work of Lagkouvardos et al.[19] All experiments concerning the impact, mechanism, and mathematical correction of the deuterium oxide effect on urine were performed on samples from group A. These were collected in 50 mL polypropylene tubes (Falcon), pooled, and aliquoted into volumes of 150 μL for analysis. A second dataset from group B was used for validation of the correction equation. Each volunteer provided written informed consent.

Samples were stored at −80 °C until analysis. Aliquots were thawed on ice, homogenized by vortexing, and transferred into vials containing buffer solution. A 1.5 M K_2PO_4 solution (pH 7.4) was used as a buffer that contained 0.1% trimethylsilylpropionic acid (TSP) in either 10% D_2O (buffer I) or 100% D_2O (buffer II). Buffers I and II were mixed to obtain required total D_2O concentration for analysis of final D_2O concentrations from 2.5% to 25%. Buffer solutions and urine were mixed in a ratio of 1:3 (50 μL buffer and 150 μL urine) and centrifuged at 4 °C for 10 min at 13 000 g. A quantity of 180 μL of supernatant was transferred into 3-mm NMR glass vials. For elucidation of the mechanism, 100 μL of a 0.33 M creatinine standard solution in H_2O (~7.5 mg/sample) was diluted in 50 μL of buffer I and 50 μL of H_2O, resulting in a total D_2O concentration of 2.5%. Equivalently, a sample with a final concentration of 50% D_2O was prepared by mixing 100 μL of the standard solution in 50 μL of buffer II and 50 μL of D_2O. The standard samples were left at RT at least for 24 h to ensure that equilibrium is reached.

The impact of creatinine loss was estimated using 2.5%, 10%, and 25% D_2O samples. Between sample preparation and measurement, samples were stored at RT and 4 °C. RT samples were prepared once and remeasured after the defined time increments, whereas cooled samples were prepared 13 times and, for every increment, a new sample was measured to exclude the effect of temperature increase during acquisition.

For calculation of the correction equation, samples were prepared from pooled urine. D_2O concentrations in these samples were adjusted to 2.5%, 5%, 10%, 15%, 20%, and 25%, respectively. Samples were measured in increments of 2 h from t = 0 h to t = 24 h. Between sample preparation and measurement, samples were stored at RT. All sample preparation steps were performed on ice until analysis.

NMR Instrumentation and Data Processing. Urine samples were analyzed on a Bruker 800 MHz spectrometer that was operating at 800.35 MHz and was equipped with a quadrupole inverse cryogenic probe (Bruker BioSpin); the 90° pulse was set to 14 μs. Sixteen scans were recorded into 64 K data points with a spectral width of 16 ppm. As a quality marker, the peak width at half-maximum for the TSP peak was monitored and spectra with a peak width at half-maximum of >1.0 Hz were excluded. All spectra were acquired at 300 K. One-dimensional proton spectra were acquired using a standard 1D-pulse sequence with water suppression (noesygppr1d) during an RD of 4 s, an acquisition time of 3 s, and a mixing time (tm) of 200 ms. To avoid integration of neighboring signals, integration boundaries of ±8.5 Hz around the centroid value were set.

One-dimensional (1D) carbon spectra were acquired using an inverse-gated (IG) decoupling pulse (zgig) with proton decoupling (WALTZ-16) during the recycle delay (RD) of 58 s to eliminate a nuclear Overhauser effect (NOE), a 90° pulse for ^{13}C at 13 μs, a decoupler pulse at 12 μs, and a decoupler power level at 1.11 dB. Multiplicity edited HSQC spectra were recorded using a DEPT-HSQC (distortionless enhanced polarization transfer heteronuclear single quantum coherence) pulse sequence (hsqcedetgpsisp2.2). Spectral width was set to 13 and 50 ppm in the proton (F2) and carbon (F1) dimensions, respectively. For each 2D spectrum, 5578 × 3072 data points were collected using 2 scans per increment with an acquisition time of 0.25 s and 16 dummy scans.

Acquisition and processing were performed using TopSpin 3.5 software (Bruker BioSpin). Free induction decay (FID) were multiplied by an exponential function corresponding to line broadening of 0.3 Hz prior to Fourier transformation. All spectra were manually phased, baseline corrected and calibrated to TSP (δ TSP = 0 ppm) before exporting into Matlab software (R2011b; Mathworks) for further data processing.

The water region was removed (δ 4.6−5.0 ppm). Spectra were aligned using a recursive segment-wise peak alignment (RSPA) algorithm.[21] Orthogonal partial least-squares (OPLS) analysis was performed as described by Cloarec et al.[22] Integrals were calculated using trapezoidal numerical integration. Local baseline correction was performed by generating linearly spaced vectors between integration boundaries and subtracting the resulting integrals from peak integrals. Negative peak integrals of the deuterated creatinine triplet (i.e., in the absence of deuterated creatinine) were set to zero. All integral areas were normalized to the corresponding TSP peak area as an internal standard. For investigation of creatinine loss over time, measured CH_2 integrals were expressed in % of CH_2 peak area recorded at t = 0 (CH_2/CH_{2_0} [%]).

MS Measurements. Analysis of the isotope distribution was performed using a maXis qTOF-MS equipped with an APOLLO II electrospray ion (ESI) source (Bruker Daltonics). Samples were measured via direct injection and in electrospray positive mode. Source settings were the same as elsewhere described:[23] nebulizer pressure = 2 bar, dry gas flow = 10 L/min, dry gas temperature = 200 °C, capillary voltage = 4.5 kV, end plate offset = +500 V, mass range = m/z 50−1500.

■ RESULTS AND DISCUSSION

To initially investigate the impact of sample preparation conditions on urine samples, we measured pooled urine samples with altering D_2O after an equilibration time of 24 h after buffer contact. In Figure 1, an overlap of six urine spectra with altering D_2O concentrations shows a clear decrease in creatinine CH_2 peak intensity after 24 h and an increase in an upfield triplet, depending on D_2O concentration. At a D_2O concentration of 2.5%, no triplet was observed, i.e., the D_2O concentration was too low to induce an effect. To systematically investigate further effects of D_2O over time at RT, besides creatinine, we analyzed samples under the two extreme D_2O concentrations (2.5% and 25%) every 2 h for 24 h, taking 2.5% D_2O as a control. No other signals were found (threshold correlation coefficients of R^2 > 0.5). Yet, several urine metabolites are known to be susceptible to proton−deuterium exchange, such as histidine,[24] which was not seen here. Our results suggest that, under the sample preparation conditions of 25% D_2O within 24 h, no other metabolites are affected by the H-D exchange. Therefore, the following evaluation focuses on the observed effects on creatinine.

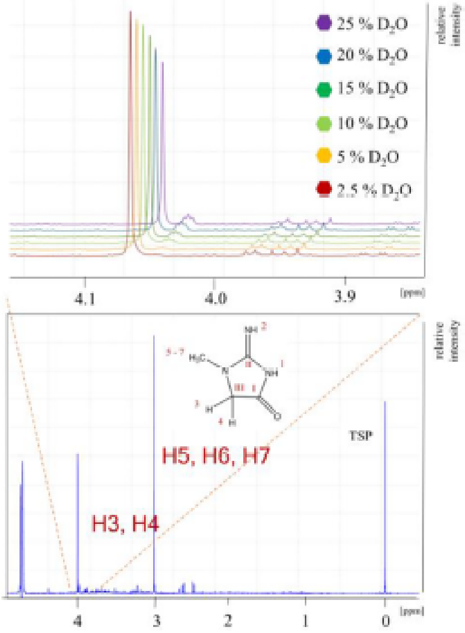


Figure 1. Proton spectra (−0.1−5.5 ppm) of creatinine standard in H_2O/D_2O and buffer with peak annotation to creatinine structure and enlargement of CH creatinine peak area as stacked plot with D_2O concentrations ranging from 2.5% (red) to 25% (purple).

The main issue with a decrease in creatinine peak area is derived from the usage of creatinine as a normalization factor and its utilization as an important marker for renal activity. To circumvent this issue, alternatively to the CH_2 peak, the CH_3 moiety could serve for creatinine quantification. The standard deviation of peak area of CH_3 was very low (<2%). However, a 2D-HSQC spectrum of a QC sample from a clinical study (for details, see the 2018 work of Gil el al.[20]) revealed overlap in the CH_3 peak area but not for the CH_2 peak (see Figure S1A in the Supporting Information). This overlap is derived from 1,1-dimethylbiguanide (metformin). Metformin is a first-line medication for type-2 diabetes. Type-2 diabetes had a global total prevalence of 8.4% in 2014, ranging from 7.3% to 13.7%, depending on the region,[25] and is therefore expected to cause substantial problems, especially in epidemiological studies or studies that include diabetes patients. Selected ^1H NMR spectra of type-2 diabetes patients highlight this problem (Figure S1B). For these spectra, the CH_3/CH_2 peak integral has a standard deviation of 30%. Therefore, we concluded that the CH_3 peak is not suitable for creatinine quantification.

Elucidation of H/D-Exchange Mechanism. As suggested by Leibfritz et al.,[18] we hypothesized the cause of this creatinine conversion to arise from a H/D exchange. We examined the underlying mechanism by a combination of (A) solvent-suppressed ^1H NMR for the quantitative estimation of creatinine degradation, (B) inverse-gated (IG) ^{13}C NMR to study changes in the chemical shift due to proton−deuterium exchange and splitting patterns from carbon-deuterium bonds. (C) Multiplicity edited $^1H-^{13}C$-HSQC (DEPT-HSQC) were recorded to link the features revealed by the individual experiments together. This confirmed a D_2O-dependent proton−deuterium exchange at the (3,4)-position (see Figure 2). Neither the addition of potassium fluoride (KF) nor the variance of the phosphate concentration influenced proton−deuterium exchange. However, as expected, the proton−deuterium exchange did not occur in the complete absence of phosphate (data not shown). A decrease of the CH_2 creatinine peak occurs simultaneously with the increase of the monodeuterated (CHD) peak. (IG) ^{13}C spectra allowed us to quantitatively study carbon nucleotides without NOE and uncover a triplet for monodeuterated (III*) ^{13}C and a quintet for polydeuterated (III**) ^{13}C. To investigate the extent of CHD and CD_2 formation under relevant operating conditions, we recorded an IG ^{13}C spectrum of human urine with 25% D_2O. As expected, monodeuteration occurred, but the formation of double deuteration was below a S/N ratio of 3 (see Figure S4 in the Supporting Information).

Pattern splitting occurred because of different nuclear spin systems and proton decoupling ($2NI + 1$, with $I(H) = 1/2$, $I(D) = 1$, and N being the number of nuclei, no splitting for protons), resulting in a singlet for CH_2, a triplet for CHD, and a quintet for CD_2, respectively. Equivalent splitting patterns were found for CH_2 and CHD peaks in DEPT-HSQC-spectra (Figure 2), including a multiplicity inversion for the single resonating proton in the CHD peak.

To confirm the elucidated mechanism, high-resolution electrospray ionization−mass spectrometry (ESI-MS) was used as an orthogonal analytical method to NMR. The proton−deuterium exchange was verified for the 50% D_2O stored for 48 h, after applying positive ESI mode (Figure 3). The spectrum clearly shows the presence of all three states (m/z 114.069 for $[C_4H_7N_3O+H]^+$, 115.076 for $[C_4H_6DN_3O+H]^+$, and 116.081 for $[C_4H_5D_2N_3O+H]^+$. As expected, the H/D-exchange did not occur at the CH_3 of creatinine (δ 3.05) signal in creatinine.

Impact of the H/D Exchange on the Creatinine CH_2 Peak Area under Different Conditions. The described H/D-exchange leads to a loss in CH_2 creatinine peak area. In this work, we investigated to which extent sample preparation (i.e., D_2O concentration of the buffer) and measurement conditions (i.e., temperature during dwell time) affect the resulting peak area. Six different conditions were examined regarding their CH_2 peak area stability over 24 h. We chose

DOI: 10.1021/acs.analchem.9b01580
Anal. Chem. XXXX, XXX, XXX−XXX

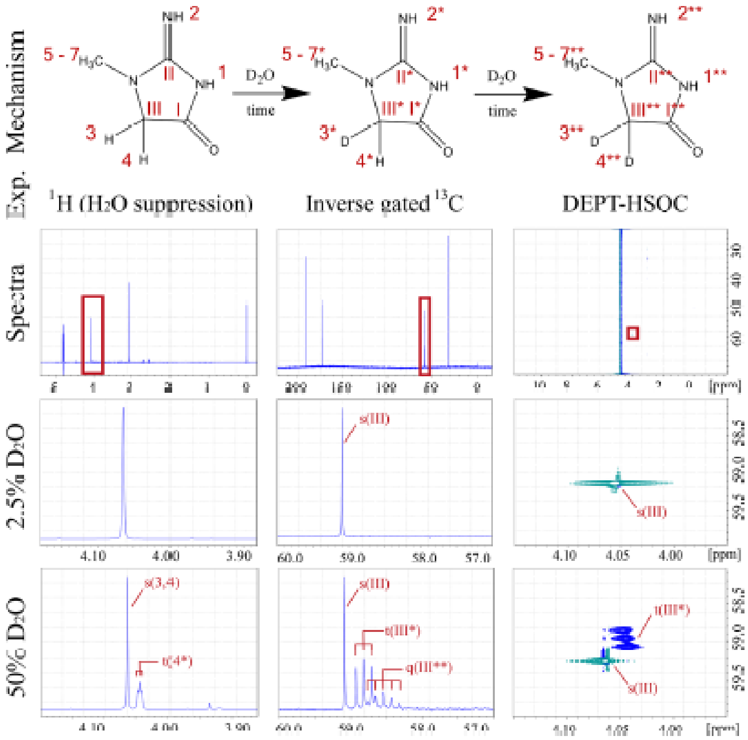


Figure 2. Mechanism of H/D exchange in creatinine with annotation of protons and carbons; ^{1}H, (IG) ^{13}C, and DEPT-HSQC spectra with enlargements of the creatinine peak region and assignments of peaks for 2.5% and 50% D_2O samples 48 h after buffer addition.

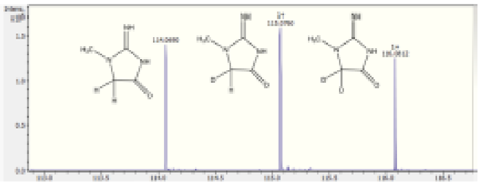


Figure 3. Positive ESI-MS spectrum; enlargement of creatinine including annotation of the different deuteration states (*m/z* 114.069 for $[C_4H_7N_3O+H]^+$, 115.076 for $[C_4H_6DN_3O+H]^+$, and 116.081 for $[C_4H_5D_2N_3O+H]^+$.

three different D_2O concentrations: 2.5% D_2O as minimal D_2O concentration, 10% D_2O as recommended in widely used urine NMR protocols,[10] and 25% because this sample preparation (≥25%) was used in several previously published studies.[7,26] Samples were kept at RT and 4 °C to cover the conditions of an availability of a cooled autosampler versus analysis at RT.

Prior to the analysis of creatinine conversion, a general estimation of accuracy and robustness was performed, resulting in a relative standard deviation (RSD) of <1% for multiple measurements of the same sample (n = 24) and up to ~10% variation for measurements of identical samples prepared multiple times and measured on different days (n = 24). This originates from various impact factors on the overall technical error (sample preparation, analytical error, spectral processing, and peak integration variability). Since temperature-controlled time-course measured samples (i.e., 4 °C) were individually prepared and 25 °C samples were prepared only once, the results of the cooled samples are expected to result in a larger inherent variability. Considering this variability, we set an acceptance level for values to be true to ±5% of the CH_{2i} peak area (t = 0).

In Figure 4, we show the impact of D_2O concentrations at RT and 4 °C for dwell time up to 24 h. At RT (Figure 4A), only samples containing 2.5% D_2O are sufficiently stable to allow 24 h of measurements, whereas 10% and 25% D_2O show losses up to 14% and 35% of the initial peak area, with losses of >5% after 4 and 0 h. For cooled samples (Figure 4B), the decrease in peak area is of lesser extent, but still significant: 2.5% and 10% D_2O concentration showed to be sufficiently stable for 24 h, whereas samples containing 25% D_2O showed significant decrease after 8 h. In summary, the availability of a 4 °C cooled autosampling device allows for the use of 10% D_2O, while analysis at RT needs minimization of the D_2O content to no more than 2.5%.

Correction Equation to Compensate Creatinine Loss. In order to use datasets that were analyzed under suboptimal conditions, we went on to investigate the possibility of

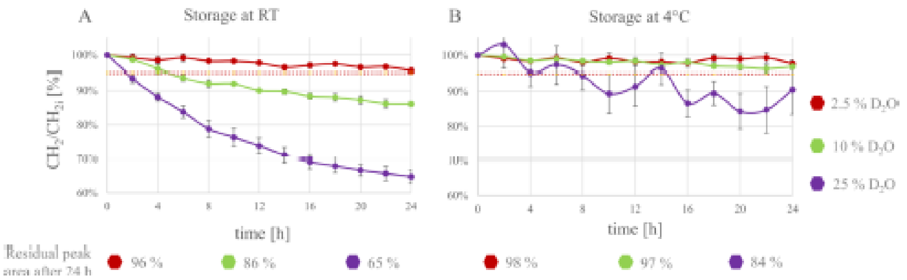


Figure 4. Ratio of measured creatinine CH_2 integral area over time relative to initial creatinine CH_2 for D_2O concentration of 2.5%, 10% and 25% with σ-error bars, dashed lines indicate acceptance limits (100% ± 5%) over 24 h. Data points represent mean values from $n = 4$ measurements, standard deviations are shown as error bars; residual creatinine peak areas are shown for all conditions. (A) Samples at RT show a strong decrease in the creatinine CH_2 peak area. The strongest effect is visible for 25% D_2O, and only the condition 2.5% D_2O is relatively stable. (B) Storage temperature at 4 °C allows stability of the CH_2 creatinine signal for both 2.5% and 10% D_2O.

correcting creatinine, based on the remaining creatinine CH_2 singlet peak and the emerging CHD triplet.

Using the complete dataset from group A ($n = 214$), we found a linear correlation ($R^2 = 0.94$) between the relative change in CH_2 and CHD peak integrals to the initial CH_{2i} integral (Figure 5). The equation, as obtained by linear fitting (Figure 5), can be converted and utilized to estimate initial values at $t = 0$ (CH_{2i}), based on CHD and CH_2 peak areas:

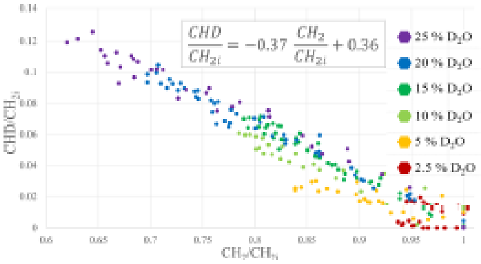


Figure 5. Linear correlation of CHD and CH_2 peak areas after normalization to CH_{2i} peak area at $t = 0$ h, color-coded by D_2O concentration; the equation shows linear approximation ($R^2 = 0.94$).

$$CH_{2i} \approx 2.8CHD + CH_2$$

This equation allows one to estimate the initial creatinine concentration in already analyzed samples, based on the peak integral of the residual creatinine peak (CH_2) and its conversion product, the emerged deuterated creatinine peak (CHD). We hypothesized the empirically found factor of monodeuterated peak area results from two aspects: (1) the relaxation time of hydrogen neighbored to deuterium is larger than hydrogen alone and (2) the CHD peak originated from one instead of two hydrogen atoms since deuterium is ^1H NMR invisible.

Indeed, an inversion−recovery T_1 experiment revealed that T_1 relaxation times change from 2.0 s for undeuterated creatinine to 5.8 s for monodeuterated creatinine (see Figures S2 and S3 in the Supporting Information). This results in a significant loss in signal intensity when recycle delays and acquisition times are kept rather short. This signal loss can be corrected by applying the formula for the compensation factor used in the 2017 work of Maitre et al.,[27] which results in a factor of 1.4. Together with the stoichiometric correction of the number of hydrogen atoms, this explains the factor of 2.8 presented herein.

Application of the Correction Equation. We applied this correction to the training dataset used for calculation of the equation ($n = 214$) and an independent test dataset ($n = 26$) in order to compare the gained improvement for creatinine quantification (see Figure 6).

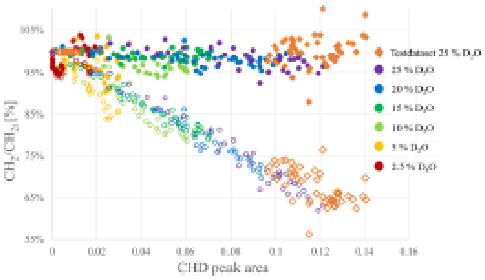


Figure 6. Comparison of dataset A with $n = 214$ samples for 24 h measurements with (filled circles) and without (empty circles) application of correction including graphical distribution plotted over CHD peak area and for the independent test dataset B with $n = 26$ samples for $t = 0$, 12, and 24 h with 25% D_2O with (filled rhombus) and without (empty rhombus) correction.

Table 1 shows creatinine peak areas before and after application of correction for datasets A and B. Remarkably, the result was achieved for different D_2O concentration and independent of time. No systematic error toward D_2O concentration was observed in dataset A. This allows application of the correction for different D_2O concentrations in the buffer and without knowledge of the dwell time (i.e., sample preparation to time of analysis). Potential variation can be introduced by independent overlays of signals in the region of the triplet area. This result also suggests that other degradation and conversion effects are negligible under the investigated conditions of up to 24 h dwell time, RT, and a maximum of 25% D_2O as mean values remain within the accepted error level.

DOI: 10.1021/acs.analchem.9b01580
Anal. Chem. XXXX, XXX, XXX−XXX

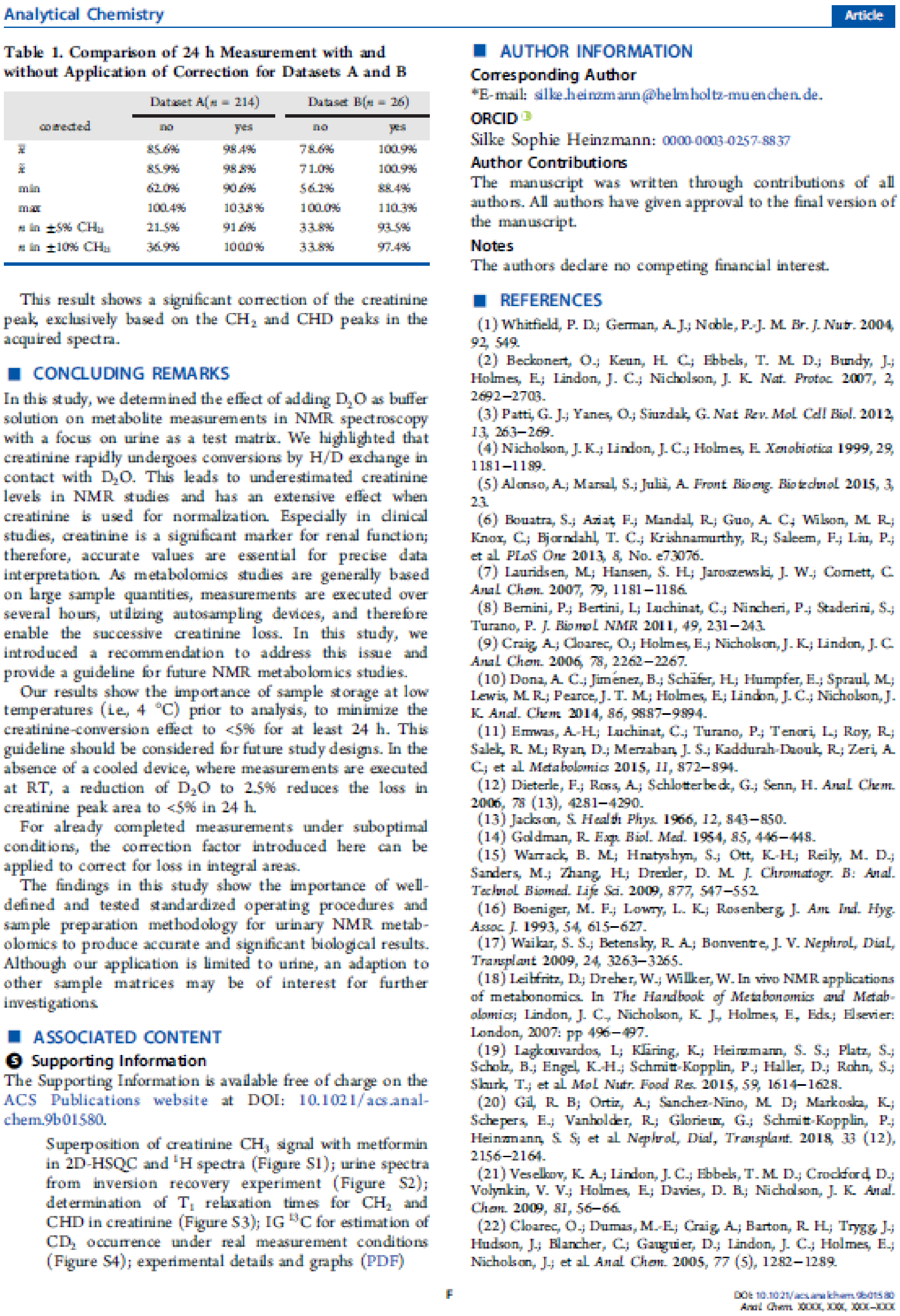


Table 1. Comparison of 24 h Measurement with and without Application of Correction for Datasets A and B

	Dataset A(n = 214)		Dataset B(n = 26)	
corrected	no	yes	no	yes
$\bar{x}$	85.6%	98.4%	78.6%	100.9%
$\tilde{x}$	85.9%	98.8%	71.0%	100.9%
min	62.0%	90.6%	56.2%	88.4%
max	100.4%	103.8%	100.0%	110.3%
n in ±5% CH_2	21.5%	91.6%	33.8%	93.5%
n in ±10% CH_2	36.9%	100.0%	33.8%	97.4%

This result shows a significant correction of the creatinine peak, exclusively based on the CH_2 and CHD peaks in the acquired spectra.

CONCLUDING REMARKS

In this study, we determined the effect of adding D_2O as buffer solution on metabolite measurements in NMR spectroscopy with a focus on urine as a test matrix. We highlighted that creatinine rapidly undergoes conversions by H/D exchange in contact with D_2O. This leads to underestimated creatinine levels in NMR studies and has an extensive effect when creatinine is used for normalization. Especially in clinical studies, creatinine is a significant marker for renal function; therefore, accurate values are essential for precise data interpretation. As metabolomics studies are generally based on large sample quantities, measurements are executed over several hours, utilizing autosampling devices, and therefore enable the successive creatinine loss. In this study, we introduced a recommendation to address this issue and provide a guideline for future NMR metabolomics studies.

Our results show the importance of sample storage at low temperatures (i.e., 4 °C) prior to analysis, to minimize the creatinine-conversion effect to <5% for at least 24 h. This guideline should be considered for future study designs. In the absence of a cooled device, where measurements are executed at RT, a reduction of D_2O to 2.5% reduces the loss in creatinine peak area to <5% in 24 h.

For already completed measurements under suboptimal conditions, the correction factor introduced here can be applied to correct for loss in integral areas.

The findings in this study show the importance of well-defined and tested standardized operating procedures and sample preparation methodology for urinary NMR metabolomics to produce accurate and significant biological results. Although our application is limited to urine, an adaption to other sample matrices may be of interest for further investigations.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b01580.

Superposition of creatinine CH_3 signal with metformin in 2D-HSQC and ^1H spectra (Figure S1); urine spectra from inversion recovery experiment (Figure S2); determination of T_1 relaxation times for CH_2 and CHD in creatinine (Figure S3); IG ^{13}C for estimation of CD_2 occurrence under real measurement conditions (Figure S4); experimental details and graphs (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: silke.heinzmann@helmholtz-muenchen.de.

ORCID

Silke Sophie Heinzmann: 0000-0003-0257-8837

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Whitfield, P. D.; German, A. J.; Noble, P.-J. M. *Br. J. Nutr.* 2004, *92*, 549.

(2) Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Nat. Protoc.* 2007, *2*, 2692–2703.

(3) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* 2012, *13*, 263–269.

(4) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* 1999, *29*, 1181–1189.

(5) Alonso, A.; Marsal, S.; Julià, A. *Front. Bioeng. Biotechnol.* 2015, *3*, 23.

(6) Bouatra, S.; Aziat, F.; Mandal, R.; Guo, A. C.; Wilson, M. R.; Knox, C.; Bjorndahl, T. C.; Krishnamurthy, R.; Saleem, F.; Liu, P.; et al. *PLoS One* 2013, *8*, No. e73076.

(7) Lauridsen, M.; Hansen, S. H.; Jaroszewski, J. W.; Cornett, C. *Anal. Chem.* 2007, *79*, 1181–1186.

(8) Bernini, P.; Bertini, I.; Luchinat, C.; Nincheri, P.; Staderini, S.; Turano, P. *J. Biomol. NMR* 2011, *49*, 231–243.

(9) Craig, A.; Cloarec, O.; Holmes, E.; Nicholson, J. K.; Lindon, J. C. *Anal. Chem.* 2006, *78*, 2262–2267.

(10) Dona, A. C.; Jiménez, B.; Schäfer, H.; Humpfer, E.; Spraul, M.; Lewis, M. R.; Pearce, J. T. M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* 2014, *86*, 9887–9894.

(11) Emwas, A.-H.; Luchinat, C.; Turano, P.; Tenori, L.; Roy, R.; Salek, R. M.; Ryan, D.; Merzaban, J. S.; Kaddurah-Daouk, R.; Zeri, A. C.; et al. *Metabolomics* 2015, *11*, 872–894.

(12) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. *Anal. Chem.* 2006, 78 (13), 4281–4290.

(13) Jackson, S. *Health Phys.* 1966, *12*, 843–850.

(14) Goldman, R. *Exp. Biol. Med.* 1954, *85*, 446–448.

(15) Warrack, B. M.; Hnatyshyn, S.; Ott, K.-H.; Reily, M. D.; Sanders, M.; Zhang, H.; Drexler, D. M. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* 2009, *877*, 547–552.

(16) Boeniger, M. F.; Lowry, L. K.; Rosenberg, J. *Am. Ind. Hyg. Assoc. J.* 1993, *54*, 615–627.

(17) Waikar, S. S.; Betensky, R. A.; Bonventre, J. V. *Nephrol, Dial, Transplant.* 2009, *24*, 3263–3265.

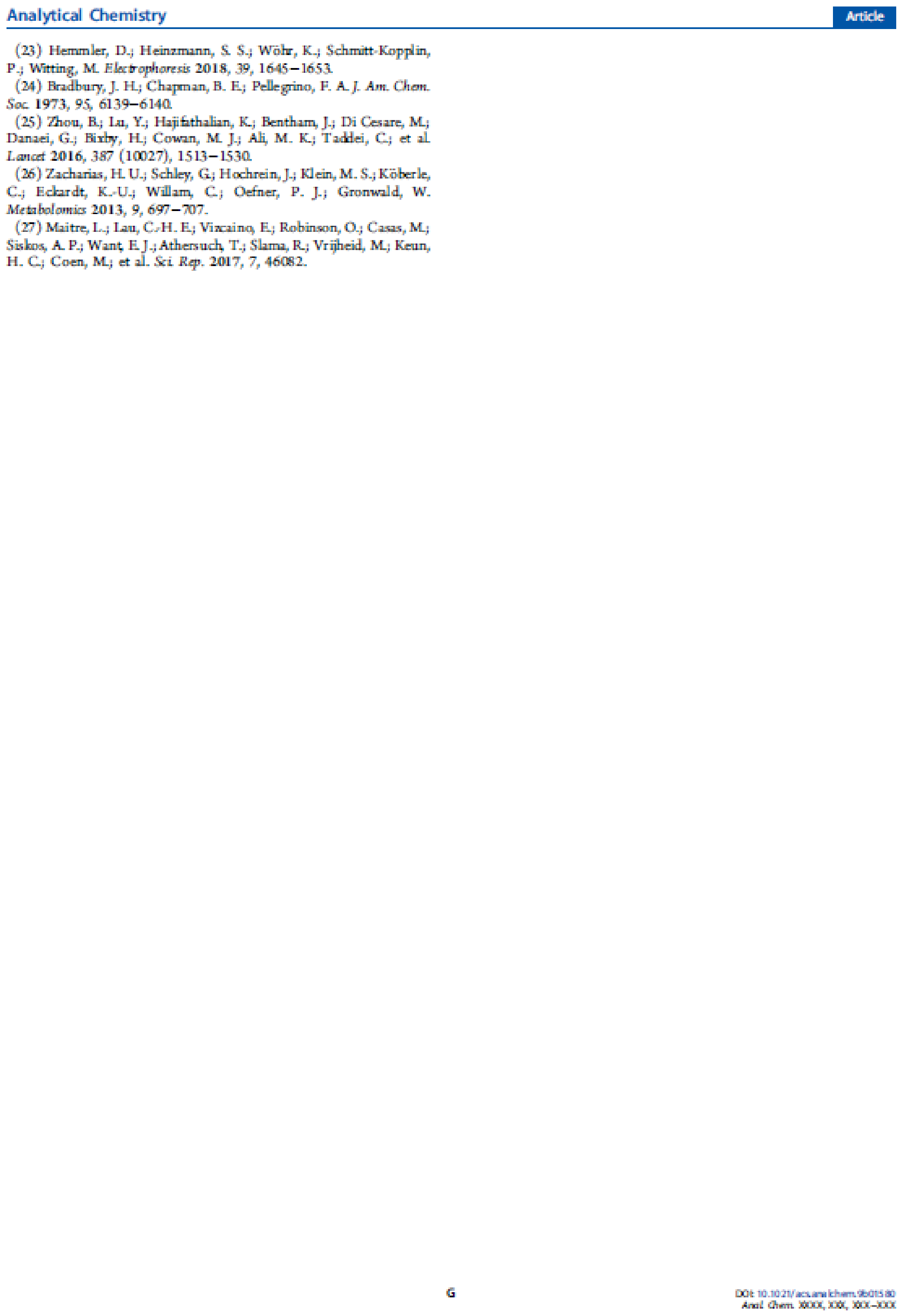
(18) Leibfritz, D.; Dreher, W.; Willker, W. In vivo NMR applications of metabonomics. In *The Handbook of Metabonomics and Metabolomics*; Lindon, J. C., Nicholson, K. J., Holmes, E., Eds.; Elsevier: London, 2007: pp 496–497.

(19) Lagkouvardos, I.; Kläring, K.; Heinzmann, S. S.; Platz, S.; Scholz, B.; Engel, K.-H.; Schmitt-Kopplin, P.; Haller, D.; Rohn, S.; Skurk, T.; et al. *Mol. Nutr. Food Res.* 2015, *59*, 1614–1628.

(20) Gil, R. B; Ortiz, A.; Sanchez-Nino, M. D; Markoska, K.; Schepers, E.; Vanholder, R.; Glorieux, G.; Schmitt-Kopplin, P.; Heinzmann, S. S; et al. *Nephrol, Dial, Transplant.* 2018, 33 (12), 2156–2164.

(21) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M. D.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. *Anal. Chem.* 2009, *81*, 56–66.

(22) Cloarec, O.; Dumas, M.-E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J.; et al. *Anal. Chem.* 2005, 77 (5), 1282–1289.



(23) Hemmler, D.; Heinzmann, S. S.; Wöhr, K.; Schmitt-Kopplin, P.; Witting, M. *Electrophoresis* 2018, *39*, 1645−1653.

(24) Bradbury, J. H.; Chapman, B. E.; Pellegrino, F. A. *J. Am. Chem. Soc.* 1973, *95*, 6139−6140.

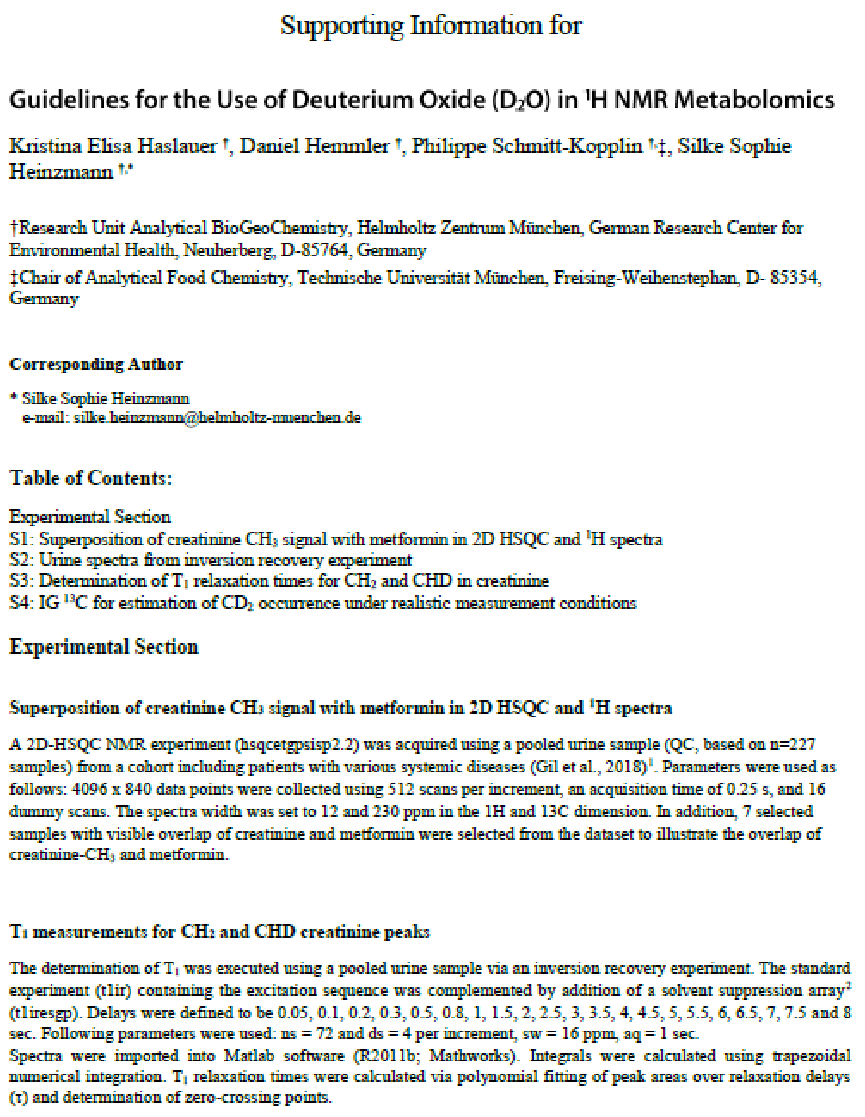
(25) Zhou, B.; Lu, Y.; Hajifathalian, K.; Bentham, J.; Di Cesare, M.; Danaei, G.; Bixby, H.; Cowan, M. J.; Ali, M. K.; Taddei, C.; et al. *Lancet* 2016, *387* (10027), 1513−1530.

(26) Zacharias, H. U.; Schley, G.; Hochrein, J.; Klein, M. S.; Köberle, C.; Eckardt, K.-U.; Willam, C.; Oefner, P. J.; Gronwald, W. *Metabolomics* 2013, *9*, 697−707.

(27) Maitre, L.; Lau, C.-H. E.; Vizcaino, E.; Robinson, O.; Casas, M.; Siskos, A. P.; Want, E. J.; Athersuch, T.; Slama, R.; Vrijheid, M.; Keun, H. C.; Coen, M.; et al. *Sci. Rep.* 2017, *7*, 46082.

DOI: 10.1021/acs.analchem.9b01580

A. 2 Supplementary Information



Supporting Information for

Guidelines for the Use of Deuterium Oxide (D_2O) in ^1H NMR Metabolomics

Kristina Elisa Haslauer [†], Daniel Hemmler [†], Philippe Schmitt-Kopplin [†,‡], Silke Sophie Heinzmann [†,*]

†Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, D-85764, Germany

‡Chair of Analytical Food Chemistry, Technische Universität München, Freising-Weihenstephan, D- 85354, Germany

Corresponding Author

* Silke Sophie Heinzmann
e-mail: silke.heinzmann@helmholtz-muenchen.de

Table of Contents:

Experimental Section

Superposition of creatinine CH_3 signal with metformin in 2D HSQC and ^1H spectra

A 2D-HSQC NMR experiment (hsqcetgpsisp2.2) was acquired using a pooled urine sample (QC, based on n=227 samples) from a cohort including patients with various systemic diseases (Gil et al., 2018)[1]. Parameters were used as follows: 4096 x 840 data points were collected using 512 scans per increment, an acquisition time of 0.25 s, and 16 dummy scans. The spectra width was set to 12 and 230 ppm in the 1H and 13C dimension. In addition, 7 selected samples with visible overlap of creatinine and metformin were selected from the dataset to illustrate the overlap of creatinine-CH_3 and metformin.

T_1 measurements for CH_2 and CHD creatinine peaks

The determination of T_1 was executed using a pooled urine sample via an inversion recovery experiment. The standard experiment (t1ir) containing the excitation sequence was complemented by addition of a solvent suppression array[2] (t1iresgp). Delays were defined to be 0.05, 0.1, 0.2, 0.3, 0.5, 0.8, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5 and 8 sec. Following parameters were used: ns = 72 and ds = 4 per increment, sw = 16 ppm, aq = 1 sec.
Spectra were imported into Matlab software (R2011b; Mathworks). Integrals were calculated using trapezoidal numerical integration. T_1 relaxation times were calculated via polynomial fitting of peak areas over relaxation delays (τ) and determination of zero-crossing points.

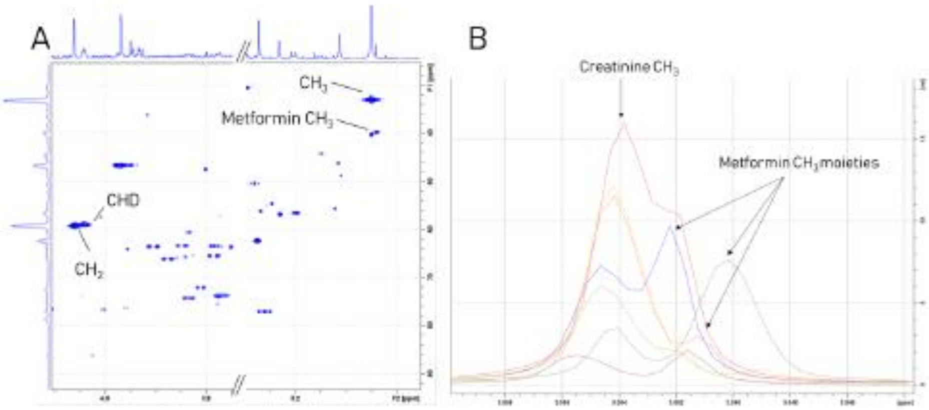


Figure S1: (A) 2D-HSQC from a QC sample highlighting the metformin-overlap of the creatinine-CH$_3$, while the creatinine-CH$_2$ and creatinine-CHD show little to no overlap with other signals. (B) Overlap of selected urine samples from a chronic kidney diseases (CKD) study containing metformin with annotation of creatinine CH$_3$ and metformin CH$_3$ moieties.

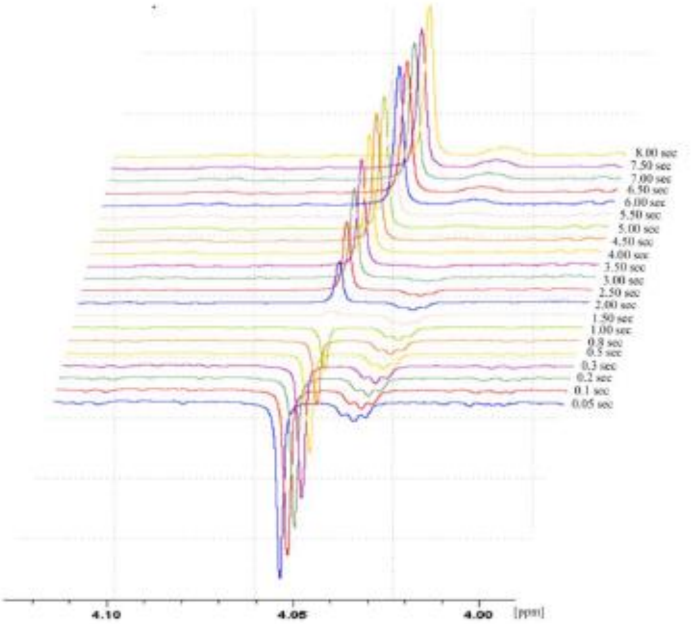


Figure S2: Stacked plot for inversion recovery experiment for urine sample with τ ranging from 0.05 sec to 8.5 sec, enlarged in CH$_2$/CHD peak area; zero crossing of CH$_2$ protons at ~ 1.5 sec; CHD protons between 3.5 sec – 4.5 sec

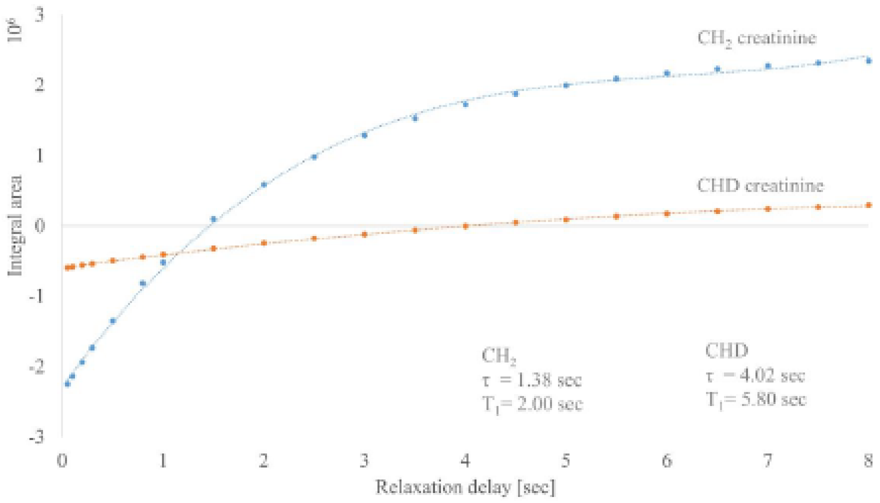
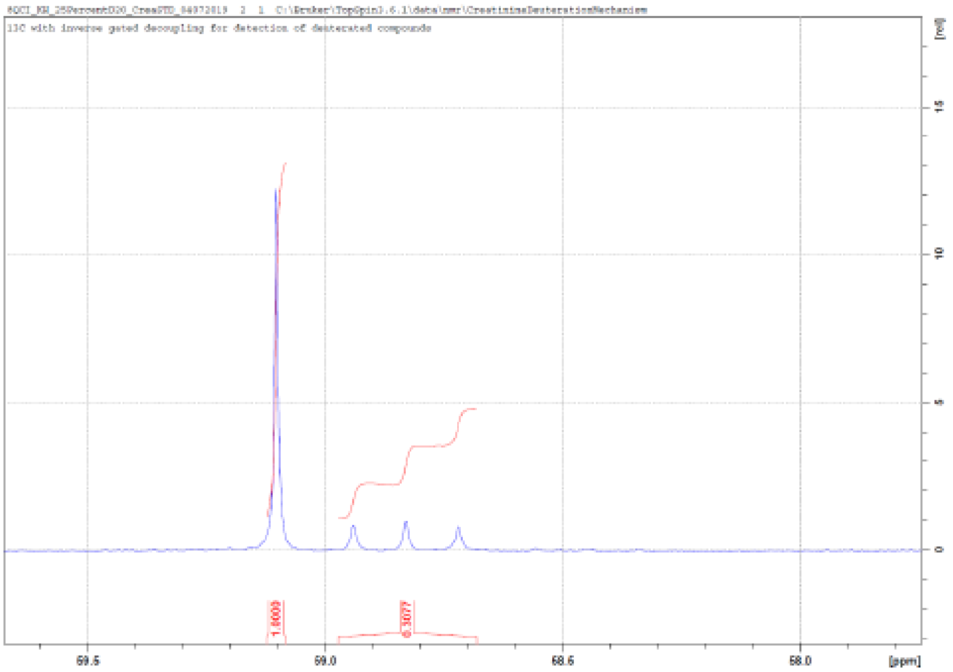
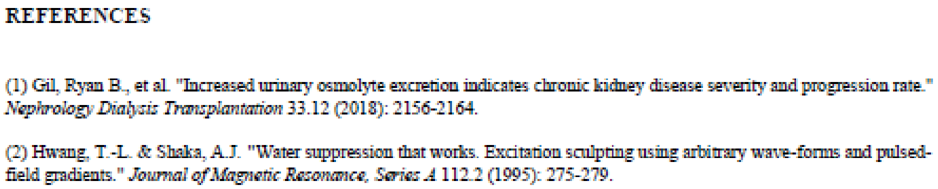


Figure S3: Integral areas over varying relaxation delays (τ) for CH$_2$ and CHD creatinine peaks


BQCI_KH_25PercentD2O_CreaSTD_04072019 2 1 C:\Bruker\TopSpin3.6.1\data\nmr\CreatinineDeuterationMechanism

13C with inverse gated decoupling for detection of deuterated compounds

Figure S4: IG ^{13}C for estimation of CD$_2$ occurrence under realistic measurement conditions: A urine sample containing 25% D$_2$O was analyzed after a dwell time of 24 h. The singlet of creatinine-CH$_2$ and the triplet of CHD are clearly visible, the quintet of creatinine-CD$_2$ is below S/N. Peak area integration was performed in TopSpin 3.6.1.



REFERENCES

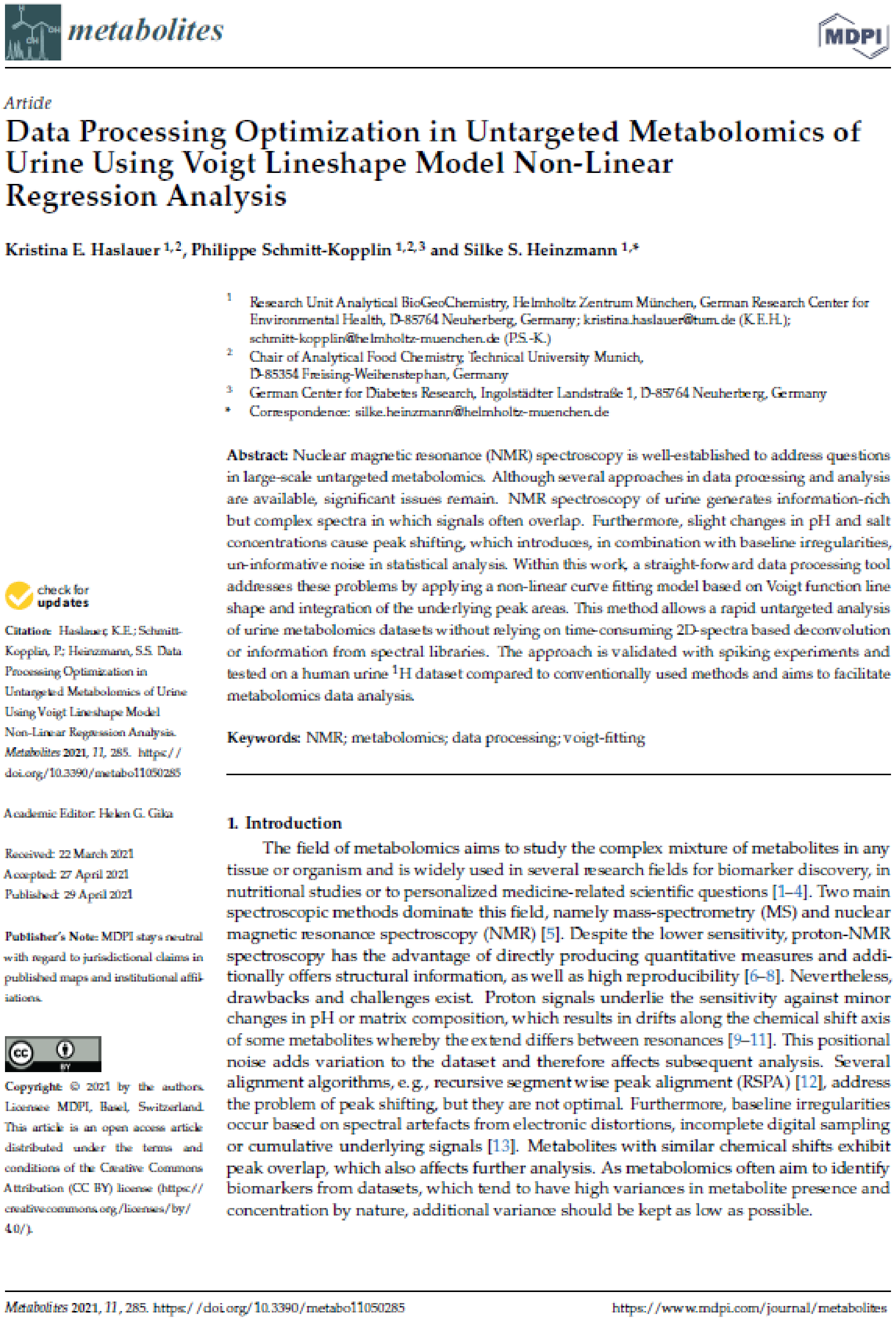
(1) Gil, Ryan B., et al. "Increased urinary osmolyte excretion indicates chronic kidney disease severity and progression rate." *Nephrology Dialysis Transplantation* 33.12 (2018): 2156-2164.

(2) Hwang, T.-L. & Shaka, A.J. "Water suppression that works. Excitation sculpting using arbitrary wave-forms and pulsed-field gradients." *Journal of Magnetic Resonance, Series A* 112.2 (1995): 275-279.

This page intentionally left blank

B. Appendix Chapter 3

B. 1 Original Publication



metabolites

MDPI

Article

Data Processing Optimization in Untargeted Metabolomics of Urine Using Voigt Lineshape Model Non-Linear Regression Analysis

Kristina E. Haslauer [1,2], **Philippe Schmitt-Kopplin** [1,2,3] **and Silke S. Heinzmann** [1,*]

[1] Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, German Research Center for Environmental Health, D-85764 Neuherberg, Germany; kristina.haslauer@tum.de (K.E.H.); schmitt-kopplin@helmholtz-muenchen.de (P.S.-K.)

[2] Chair of Analytical Food Chemistry, Technical University Munich, D-85354 Freising-Weihenstephan, Germany

[3] German Center for Diabetes Research, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

[*] Correspondence: silke.heinzmann@helmholtz-muenchen.de

Abstract: Nuclear magnetic resonance (NMR) spectroscopy is well-established to address questions in large-scale untargeted metabolomics. Although several approaches in data processing and analysis are available, significant issues remain. NMR spectroscopy of urine generates information-rich but complex spectra in which signals often overlap. Furthermore, slight changes in pH and salt concentrations cause peak shifting, which introduces, in combination with baseline irregularities, un-informative noise in statistical analysis. Within this work, a straight-forward data processing tool addresses these problems by applying a non-linear curve fitting model based on Voigt function line shape and integration of the underlying peak areas. This method allows a rapid untargeted analysis of urine metabolomics datasets without relying on time-consuming 2D-spectra based deconvolution or information from spectral libraries. The approach is validated with spiking experiments and tested on a human urine ^{1}H dataset compared to conventionally used methods and aims to facilitate metabolomics data analysis.

Keywords: NMR; metabolomics; data processing; voigt-fitting

check for updates

Citation: Haslauer, K.E.; Schmitt-Kopplin, P.; Heinzmann, S.S. Data Processing Optimization in Untargeted Metabolomics of Urine Using Voigt Lineshape Model Non-Linear Regression Analysis. *Metabolites* 2021, *11*, 285. https://doi.org/10.3390/metabo11050285

Academic Editor: Helen G. Gika

Received: 22 March 2021
Accepted: 27 April 2021
Published: 29 April 2021

1. Introduction

The field of metabolomics aims to study the complex mixture of metabolites in any tissue or organism and is widely used in several research fields for biomarker discovery, in nutritional studies or to personalized medicine-related scientific questions [1–4]. Two main spectroscopic methods dominate this field, namely mass-spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR) [5]. Despite the lower sensitivity, proton-NMR spectroscopy has the advantage of directly producing quantitative measures and additionally offers structural information, as well as high reproducibility [6–8]. Nevertheless, drawbacks and challenges exist. Proton signals underlie the sensitivity against minor changes in pH or matrix composition, which results in drifts along the chemical shift axis of some metabolites whereby the extend differs between resonances [9–11]. This positional noise adds variation to the dataset and therefore affects subsequent analysis. Several alignment algorithms, e.g., recursive segmentwise peak alignment (RSPA) [12], address the problem of peak shifting, but they are not optimal. Furthermore, baseline irregularities occur based on spectral artefacts from electronic distortions, incomplete digital sampling or cumulative underlying signals [13]. Metabolites with similar chemical shifts exhibit peak overlap, which also affects further analysis. As metabolomics often aim to identify biomarkers from datasets, which tend to have high variances in metabolite presence and concentration by nature, additional variance should be kept as low as possible.

Various tools have been published, which circumvent these drawbacks and facilitate data analyses, utilizing defined metabolite libraries and fitting peaks, according to their pre-defined multiplicities and characteristics within defined matrices [14–16]. A comprehensive overview can be found in Bingol et al. (2018) [5]. These methods have been shown to produce reliable and quantitative results, but rely on databases, which are often limited to a specific biofluid, and fail to extract unknown informative features. Non-commercial untargeted approaches are made up from two main strategies, full spectra analysis, which uses all points of the spectrum and various binning methods, where equidistant binning with a binsize of $0.01 - 0.001$ ppm is prevalent [17]. Both methods are affected by peak shifting, baseline influence and signal overlap, which adds uninformative noise. Furthermore, full spectra analysis results in large datasets which are bulky to process. Binning has the advantage of a reduction in dimensionality, which speeds up analysis, but limits the ability of detecting metabolites of interest as some peaks may shift between bins through the dataset. In particular, binning either sums up all data points within a certain bin or determines the area under the curve (AUC), significant changes in minor peaks may be covered by general variance caused by baseline differences or signal overlap. To address these issues an easy-to-use and straightforward processing step is introduced, which is based on a peak-picking algorithm followed by a Voigt lineshape model fitting. In theory, NMR peaks are Lorentzian. However, slight variations in peak linewidth (e.g., due to shimming imperfection) lead to random error in the Lorentzian model. To account for this issue, a Voigt lineshape model, which is a convolution of Lorentzian and Gaussian shapes, has shown to be more accurate [18,19]. As both binning and full spectra analysis are widely used methods for NMR metabolomics processing, the performance of the Voigt fitting workflow is validated by comparison to these methods. The introduced workflow aims to provide an enhanced processing method that extracts information from NMR spectra without limitations set by the necessity of pre-defined databases.

To overcome these drawbacks we introduce an untargeted workflow for complex NMR spectra, which consists of 6 main steps that are shown in Figure 1. As with the input information, the workflow uses aligned, normalized NMR spectra and a reference spectrum (e.g., quality control or mean spectrum (`mean (x)`)). First, a peak picking approach is performed on the full dataset for every single spectrum by finding all local maxima. This is followed by an optional noise reduction step, where all peaks with a net intensity between the local maximum and the neighboring minimum are discarded. Therefore, an adjustment for the noise level, especially in regions with a baseline above zero, as well as in overlapping peak regions is achieved. In the next step, the non-linear peak fitting algorithm constructs Voigt line-shaped approximated peaks to the experimental data by optimizing amplitudes, peak maxima, the ratio between Gaussian and Lorentzian and peak width. Peak fitting is based on the `lsqcurvefit` function inbuilt in MATLAB, employing a trust-region-reflective algorithm. In the following step, the AUC of fitted peaks are calculated over a defined integration range (i.e., multiple of optimized peak width). The chemical shift (i.e., their local maxima) of these peak integrals vary slightly, even in aligned datasets. Therefore, peak shifts are adjusted to the reference spectrum by an alignment step that iterates through every processed spectrum to find peaks within a user defined peak shift window. The generated dataset of integrated peaks can now be further reduced by applying a frequency filter to exclude peaks that are present in less than a set percentage in the dataset. Finally, the workflow gives as output: a list of peak integrals, a plot of each fitted spectrum and quality metrics, such as the residual sum of squares and the standard error of fit for the fitting parameters. These metrics, as well as the graphical results (see Figure 2) allow a quality assessment of the obtained data and consider it for further processing, e.g., applying weighting function.

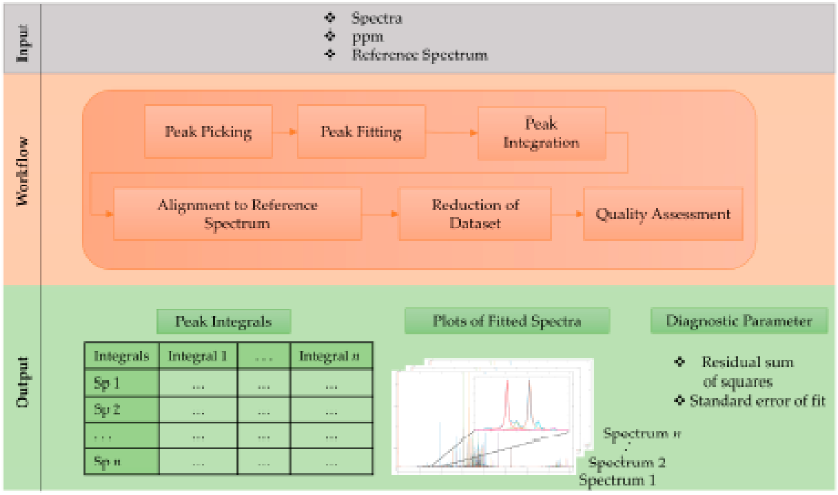


Figure 1. Step-by-step schematic workflow description for non-linear peak fitting based on Voigt line shape model.

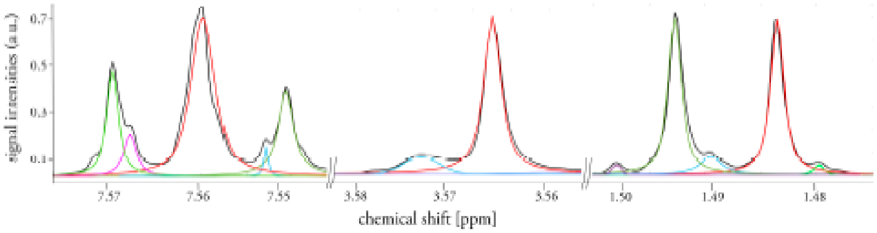


Figure 2. Typical fit results for an exemplary urine spectrum in three regions where signals overlap and/or small peaks are present; initial spectrum is shown as black line, fitted peaks are depicted in colored lines.

2. Results

Error Estimation over Matrices

Efficient metabolomics analysis aims to uncover patterns and trends within the data. However, in NMR metabolomics analysis, such trends are often covered by background noise and peak shifts. The comparability of the introduced approach with conventionally used methods, full spectra (i.e., peak height) and binned data analysis (i.e., AUC of spectral bin), is shown using a standard addition of three metabolites (Alanine, Caffeine and Nicotinamide) with three spiked concentrations in four different urine samples, which results in 12 data points. The data are used to calculate a standard curve for every method. These equations were used to re-calculate the concentration for all 12 individual values. Boxplots (see Figure 3) are employed to illustrate the error proneness for all three methods sorted by the respective metabolite including standard errors. Averaging the mean standard errors over the three investigated metabolites for every method gives total mean relative standard errors ($\overline{\text{RSE}}$) (13.31% for full spectra analysis, 11.02% for binned data and 7.33% using Voigt fitted data). The metabolites differ, shown in this study, in their chemical shift, their tendency to shift and/or overlap, thus, the large span of relative errors (see Figure 3) is somewhat expectable. Overall, these results indicate that applying the Voigt fitting algorithm does not artificially increase the variation in comparison to full spectra and binned data analysis. The main advantage of the Voigt function integral lies in the removal of background noise, illustrated by large improvements in RSE for the overlapped signal of caffeine and similar RSE for relatively large and/or non-overlapped peaks, such as alanine and nicotinamide and is therefore applicable for usage in metabolomics approaches.

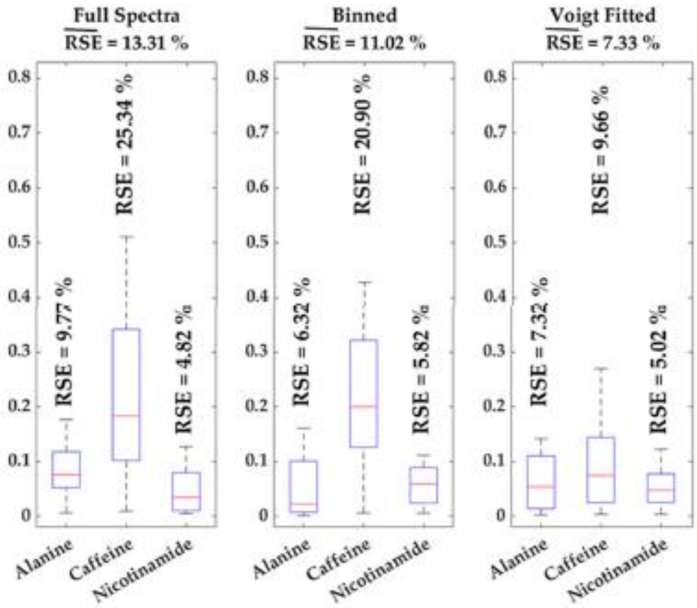


Figure 3. Boxplots of standard errors of relative quantification for all three spiked metabolites and methods, individual relative standard errors (RSE) are given as well as the mean RSE ($\overline{\text{RSE}}$) for each method.

The publicly available dataset MTBLS1 [20] from the MetaboLights repository [21] was processed using the full spectra, binned data and Voigt fitted data approach. The MTBLS1 study contains 132 spectra of human urine samples from patients with Type 2 diabetes mellitus (T2DM) and a control group. A principal component analysis (PCA) was performed to determine the areas of highest variance using the different data processing methods as input data. In Figure 4A–C scores plots of the first two principal components (PC1 and PC2) are shown for all three methods, which are colored according to their groups (T2DM/Control). Both full spectra and binned data scores plots fail to separate healthy and diseased individuals. Using Voigt fitted peak integrals as the input data for PCA, a separation can be observed between patients with type 2 diabetes mellitus (T2DM) and the control group, which was intuitively expected. The loadings plot of full spectra analysis (Figure 4D) shows that the majority of variance arises from high amplitudes in the upfield region ($\delta < 1$ ppm), around the residual water signal (δ~4.7 ppm) and in the very downfield region ($\delta > 8.5$ ppm). In these regions, generally, few or no peaks occur in urine samples and they are mainly dominated by bare baseline. Similar results are observed for binned spectra, where high variations in uninformative regions also dominate the principal components (Figure 4E). The Voigt peak fitting approach reduces the spectral data to informative peak areas. Here, the loadings for PC2 (Figure 4F) show high variance of urinary glucose levels between patients and the control group (ratio between mean relative intensities (a.u.): 2.09 T2DM/Control, which is expected in an unmedicated cohort

Intriguingly, this obvious information could not be extracted from PCA loadings of the full spectra and binned data analysis, as it was covered by background noise.

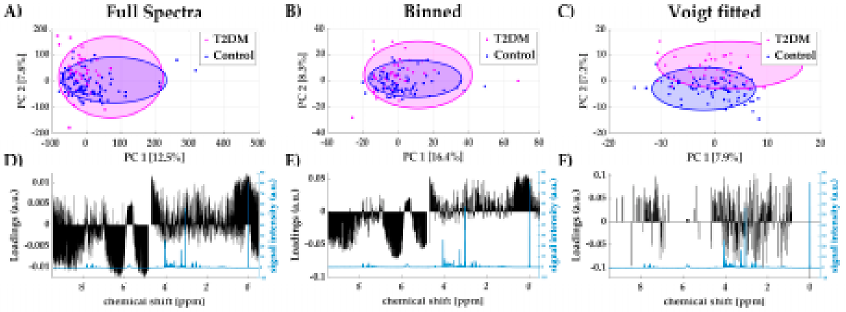


Figure 4. Scores plots of PC1 and PC2 using full spectra (**A**), binned data (**B**) and Voigt fitted data (**C**) including 95%-confidence ellipses for each group (Type 2 diabetes mellitus (T2DM) and control); loadings plot for PC1 for all three methods (black) with reference spectrum (blue) (**D**–**F**).

In summary, these results show that using Voigt fitted peak integrals instead of the whole spectrum (as is or binned) allows a crucial reduction of noise, and thus, facilitate the unsupervised data analysis.

Supervised methods, such as orthogonal projection on latent structures (OPLS) [22], aims to separate the total variation within a dataset into a predictive (i.e., information related to the sample class) and an orthogonal (i.e., unrelated) component. This method is generally accepted to exclude non-informative noise and thus uncover the relevant information related to the sample class. In Figure 5A–C OPLS discriminant analysis scores plots are shown including their R^2 and Q^2 values. Although all three methods yield a valid model to distinguish diabetic and non-diabetic individuals, both R^2 and Q^2 are higher using the Voigt fitted dataset. Furthermore, the loadings plots of the predictive component (Figure 5D,E) still show a considerable influence of non-informative regions (~0 ppm, ~5 ppm, >8.5 ppm).

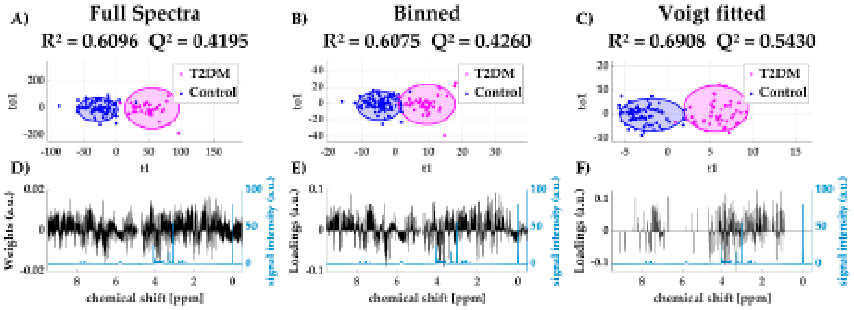


Figure 5. Scores plots of predictive and orthogonal variation using full spectra (**A**), binned data (**B**) and Voigt fitted data (**C**) including 95%-confidence ellipses for each group (Type 2 diabetes mellitus (T2DM) and control); loadings plot for first predictive component for all three methods (black) with reference spectrum (blue) (**D**–**F**).

Overall, these results indicate that the noise reduction achieved by applying the introduced peak fitting using a Voigt approximation enables a more convenient analysis of NMR metabolomics datasets. Through the reduction of the dataset, a yet inevitable visual inspection of results becomes more simple and false positive results caused by baseline differences are reduced. Furthermore, the impact of different data analysts is largely reduced.

3. Discussion

The field of untargeted NMR metabolomics became increasingly important over the past few years. However, effective and reliable data processing remains a bottleneck. The majority of studies published in the field of untargeted metabolomics rely on, either full spectra analysis or on different binning methods. Although NMR is generally highly reproducible, minor changes in baseline intensities may occur due to accumulation of underlying signals, as well as line broadening due to inhomogeneity of the magnetic field. Both conventionally used methods are limited in their ability to compensate for this non-informative variance. Nevertheless, a reduction of this noise is a crucial aspect in uncovering relevant variance and allow identification of biomarkers. The introduced approach aims to improve the efficiency of untargeted NMR metabolomics data analysis by using a peak fitting approach, based on a Voigt line shape model approximation in a least square sense, along with alignment of peak integrals to a reference spectrum. Peak fitting reduces the noise driven bias by reduction of the data. Regions containing mere baseline or very small peaks below a defined S/N ratio are excluded from further analysis, and thus, reduce the influence of the measurement error, which is usually relatively large for small values, and the irrelevant variation within the data. A comparison of all three data processing methods (full spectra, binning and Voigt fitting) demonstrates the reduced extend of noise influence of analysis performed using Voigt fitted data compared to conventionally used processing methods for both unsupervised and supervised analysis methods. A significant influence of noise within the first principal components is a well-known feature and generally accepted as fact within the NMR metabolomics community. An orthogonal PLS is typically the method of choice to segregate this noise from the biological variation of interest. Although an orthogonal filter is applied, a significant influence of non-informative variance is demonstrated using conventional data processing. Several research articles have been published, optimizing both integration and data reduction in various biofluids. From these, several approaches need input data, such as a predefined target list or spectral libraries and deliver a targeted metabolomics output, as reviewed by Bingol et al. [5], while our approach remains untargeted. Other workflows, such as SigMa [23] require extensive compound libraries. Applied to serum and plasma samples, Takis et al. [24] introduced a deconvolution-free integration method, SMolESY, which enables a suppression of the macromolecular background, which is particularly important in blood samples. Its application to urine samples remains unclear, as plasma, unlike urine, does not face extensive peak shifting. Our project contributes to the continuous progress in the field of optimized data processing in untargeted NMR metabolomics.

Voigt-fitting decreases the chance of detecting false positive markers by general data reduction and simplifies the interpretation, and analysis of loadings, respectively weights. Nevertheless, thresholds for S/N and frequency filter must be adjusted carefully to avoid rejection of relevant signals. The Voigt fitting approach was developed and tested on human urine samples as representative biofluid for complex mixtures. However, this method can also be adapted and optimized for other biological matrices.

The relevance of improved data processing methods is clearly supported by the comparison of performance of data processing methods in this work. Peak fitting using a Voigt line shape model has been demonstrated to enhance the power of statistical analysis in contrast to conventionally used methods. The used script is written in MATLAB R2020a and can be obtained for implementation by contacting the corresponding author.

4. Materials and Methods

4.1. Study Cohort

For illustration of performance improvement the fitting approach in comparison with full spectra and binning approach the MTBLS1 dataset (raw spectra) from the MetaboLight repository [21]. The MTBLS1 dataset consists of 48 samples from unmedicated patients with Type 2 diabetes mellitus (T2DM) and 84 samples from healthy individuals as control group. The study was conducted to examine urinary metabolic changes in patients with T2DM in comparison to the control group. Details about sampling, sample preparation, acquisition along with main findings are available in the original manuscript [20].

4.2. Validation Dataset

The error estimation of the three tested methods was calculated using four different urine samples each spiked with L-alanine, Caffeine and Nicotinamide in three concentrations by comparing the results to peak height in full spectra analysis and AUC in binned spectra analysis. L-alanine was used because its resonance appears in a non-crowded region and shows a distinct doublet as easy-to-integrate standard. Caffeine has resonances in a crowded region where baseline effects do occur (3–4 ppm) and Nicotinamide causes resonances in the downfield area to comprehensively cover the whole spectrum. A stock solution of 1 mg mL^{-1} H_2O was prepared. A total of 135 µL urine was combined with either 5, 10 or 15 µL stock solution resulting in an addition of 5, 10 and 15 µg standard. The samples were then filled up to a total volume of 150 µL, 50 µL 1.5 M K_2PO_4-buffer (pH 7.4) containing 0.1% Trimethylsilylpropionic acid (TSP) in 100% D_2O was added, samples were thoroughly vortexed and centrifuged at 4 °C for 10 min at 12,700× *g*. A volume of 180 µL of supernatant was transferred into 3-mm NMR tubes. Samples were measured immediately after preparation.

4.3. NMR Data Acquisition and Processing

The samples were analyzed on a Bruker 800 MHz spectrometer operating at 800.35 MHz equipped with a quadrupole inverse cryogenic (QCI) probe probe (Bruker BioSpin, Rheinstetten, Germany). A total of 256 scans were recorded into 64 K datapoints with a spectral with of 16 ppm and a 90° pulse of 13 µs. All spectra were acquired at 300 K using a standard 1D-pulse sequence with water suppression (noesygppr1d) during an recycle delay of 4 s, an acquisition time of 3 s, and a mixing time (tm) of 200 ms. Spectra were manually phased and baseline corrected in TopSpin 3.6.1 (Bruker BioSpin, Rheinstetten, Germany).

4.4. Data Processing

Spectra were imported into Matlab software (R2020a; Mathworks) for data processing with a resolution of 2.5×10^{-4} ppm, resulting in 44,001 data points per spectrum (−1 to 10 ppm). The water region was removed (δ 4.70–4.85 ppm). Spectra were aligned using a recursive segment-wise peak alignment (RSPA) algorithm [12], probabilistic quotient normalization was used to account for biological variation in urine dilution [25]. To compare the performance of the here introduced approach, two conventionally used processing methods (full spectra analysis and binning of spectra) were used as state of the art reference for untargeted metabolomics [4,26]. For full spectra analysis the data matrix was used as is after water removal and alignment resulting in a 132 × 43,400 matrix. Binning was performed by dividing every spectrum in equidistant buckets with a bin width of 0.01 ppm and determining the area under the curve (AUC) for every bin by trapezoidal integration. The resulting data matrix has a size of 132 × 1085. Peak fitting was performed using the above described workflow and is resulting in a 132 × 432 data matrix. A threshold was set to a minimum of 30% abundance through the samples with a signal to noise (S/N) ratio above 5.

Principal component analysis (PCA) was performed in Matlab software (R2020a; Mathworks) using unit variance (UV) scaling prior to analysis.

Orthogonal projection on latent structures (OPLS) discriminant analysis was performed according to the method described in Cloarec et al. (2005) [27].

Author Contributions: Conceptualization, K.E.H.; methodology, K.E.H.; software, K.E.H.; validation, K.E.H. and S.S.H.; formal analysis, K.E.H.; investigation, K.E.H.; resources, P.S.-K.; data curation, K.E.H.; writing—original draft preparation, K.E.H.; writing—review and editing, P.S.-K. and S.S.H.; visualization, K.E.H.; supervision, P.S.-K. and S.S.H.; project administration, P.S.-K. and S.S.H.; funding acquisition, P.S.-K. and S.S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG) Research Unit 5042 miTarget: "The microbiome as a therapeutic target in inflammatory bowel diseases".

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board as stated Salek et al [20].

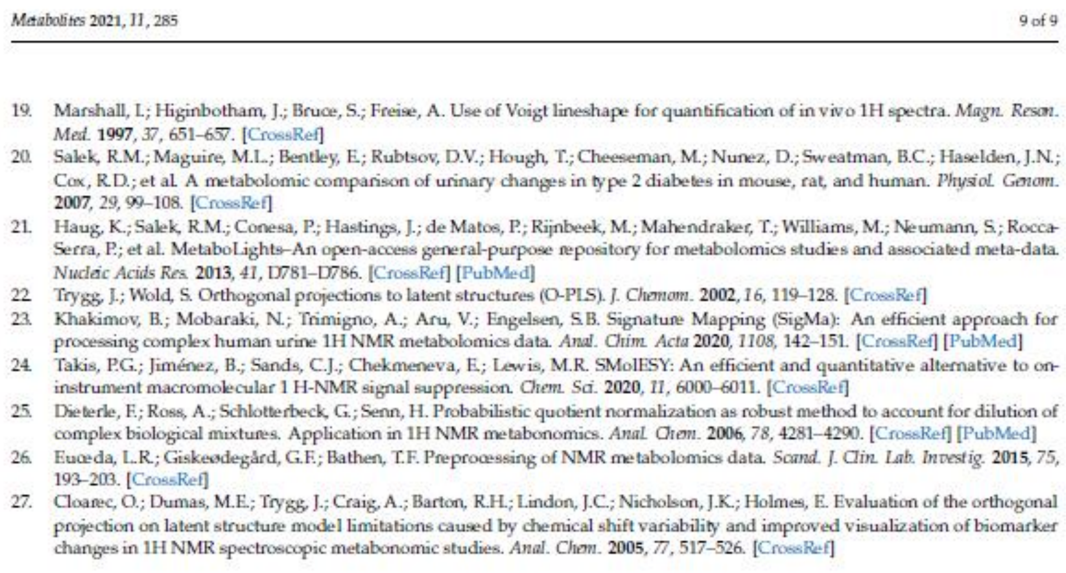
Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www.ebi.ac.uk/metabolights/MTBLS1/descriptors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Everett, J.R. NMR-based pharmacometabonomics: A new paradigm for personalised or precision medicine. *Prog. Nucl. Magn. Reson. Spectrosc.* **2017**, *102–103*, 1–14. [CrossRef]
2. Trimigno, A.; Khakimov, B.; Savorani, F.; Tenori, L.; Hendrixson, V.; Čivilis, A.; Glibetic, M.; Gurinovic, M.; Pentikäinen, S.; Sallinen, J.; et al. Investigation of Variations in the Human Urine Metabolome amongst European Populations: An Exploratory Search for Biomarkers of People at Risk-of-Poverty. *Mol. Nutr. Food Res.* **2019**, *63*, 1800216. [CrossRef] [PubMed]
3. Ussher, J.R.; Elmariah, S.; Gerszten, R.E.; Dyck, J.R. The Emerging Role of Metabolomics in the Diagnosis and Prognosis of Cardiovascular Disease. *J. Am. Coll. Cardiol.* **2016**, *68*, 2850–2870. [CrossRef] [PubMed]
4. Heinzmann, S.S.; Brown, I.J.; Chan, Q.; Bictash, M.; Dumas, M.-E.; Kochhar, S.; Stamler, J.; Holmes, E.; Elliott, P.; Nicholson, J.K. Metabolic profiling strategy for discovery of nutritional biomarkers: Proline betaine as a marker of citrus consumption. *Am. J. Clin. Nutr.* **2010**, *92*, 436–443. [CrossRef] [PubMed]
5. Bingol, K. Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods. *High Throughput* **2018**, *7*, 9. [CrossRef]
6. Alonso, A.; Marsal, S.; Julià, A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Front. Bioeng. Biotechnol.* **2015**, *3*, 23. [CrossRef]
7. Fan, T.W.-M.; Lane, A.N. Applications of NMR spectroscopy to systems biochemistry. *Prog. Nucl. Magn. Reson. Spectrosc.* **2016**, *92–93*, 18–53. [CrossRef]
8. Nagana Gowda, G.A.; Raftery, D. Can NMR solve some significant challenges in metabolomics? *J. Magn. Reson.* **2015**, *260*, 144–160. [CrossRef]
9. Beneduci, A.; Chidichimo, G.; Dardo, G.; Pontoni, G. Highly routinely reproducible alignment of 1H NMR spectral peaks of metabolites in huge sets of urines. *Anal. Chim. Acta* **2011**, *685*, 186–195. [CrossRef]
10. Nicholson, J.K.; Wilson, I.D. High resolution proton magnetic resonance spectroscopy of biological fluids. *Prog. Nucl. Magn. Reson. Spectrosc.* **1989**, *21*, 449–501. [CrossRef]
11. Takis, P.G.; Schäfer, H.; Spraul, M.; Luchinat, C. Deconvoluting interrelationships between concentrations and chemical shifts in urine provides a powerful analysis tool. *Nat. Commun.* **2017**, *8*, 1662. [CrossRef]
12. Veselkov, K.A.; Lindon, J.C.; Ebbels, T.M.D.; Crockford, D.; Volynkin, V.V.; Holmes, E.; Davies, D.B.; Nicholson, J.K. Recursive Segment-Wise Peak Alignment of Biological 1H NMR Spectra for Improved Metabolic Biomarker Recovery. *Anal. Chem.* **2009**, *81*, 55–66. [CrossRef] [PubMed]
13. Emwas, A.-H.; Roy, R.; McKay, R.T.; Ryan, D.; Brennan, L.; Tenori, L.; Luchinat, C.; Gao, X.; Zeri, A.C.; Gowda, G.A.N.; et al. Recommendations and Standardization of Biomarker Quantification Using NMR-Based Metabolomics with Particular Focus on Urinary Analysis. *J. Proteome Res.* **2016**, *15*, 360–373. [CrossRef] [PubMed]
14. Hao, J.; Astle, W.; de Iorio, M.; Ebbels, T.M.D. BATMAN—An R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* **2012**, *28*, 2088–2090. [CrossRef]
15. Lefort, G.; Liaubet, L.; Canlet, C.; Tardivel, P.; Père, M.-C.; Quesnel, H.; Paris, A.; Iannuccelli, N.; Vialaneix, N.; Servien, R. ASICS: An R package for a whole analysis workflow of 1D 1H NMR spectra. *Bioinformatics* **2019**, *35*, 4356–4363. [CrossRef] [PubMed]
16. Hao, J.; Liebeke, M.; Astle, W.; De Iorio, M.; Bundy, J.G.; Ebbels, T.M. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat. Protoc.* **2014**, *9*, 1416–1427. [CrossRef]
17. Zacharias, H.U.; Altenbuchinger, M.; Gronwald, W. Statistical Analysis of NMR Metabolic Fingerprints: Established Methods and Recent Advances. *Metabolites* **2018**, *8*, 47. [CrossRef]
18. Marshall, I.; Bruce, S.D.; Higinbotham, J.; MacLullich, A.; Wardlaw, J.M.; Ferguson, K.J.; Seckl, J. Choice of spectroscopic lineshape model affects metabolite peak areas and area ratios. *Magn. Reson. Med.* **2000**, *44*, 646–649. [CrossRef]



19. Marshall, I.; Higinbotham, J.; Bruce, S.; Freise, A. Use of Voigt lineshape for quantification of in vivo 1H spectra. *Magn. Reson. Med.* **1997**, *37*, 651–657. [CrossRef]
20. Salek, R.M.; Maguire, M.L.; Bentley, E.; Rubtsov, D.V.; Hough, T.; Cheeseman, M.; Nunez, D.; Sweatman, B.C.; Haselden, J.N.; Cox, R.D.; et al. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiol. Genom.* **2007**, *29*, 99–108. [CrossRef]
21. Haug, K.; Salek, R.M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; et al. MetaboLights–An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **2013**, *41*, D781–D786. [CrossRef] [PubMed]
22. Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*, 119–128. [CrossRef]
23. Khakimov, B.; Mobaraki, N.; Trimigno, A.; Aru, V.; Engelsen, S.B. Signature Mapping (SigMa): An efficient approach for processing complex human urine 1H NMR metabolomics data. *Anal. Chim. Acta* **2020**, *1108*, 142–151. [CrossRef] [PubMed]
24. Takis, P.G.; Jiménez, B.; Sands, C.J.; Chekmeneva, E.; Lewis, M.R. SMolESY: An efficient and quantitative alternative to on-instrument macromolecular 1 H-NMR signal suppression. *Chem. Sci.* **2020**, *11*, 6000–6011. [CrossRef]
25. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.* **2006**, *78*, 4281–4290. [CrossRef] [PubMed]
26. Euceda, L.R.; Giskeødegård, G.F.; Bathen, T.F. Preprocessing of NMR metabolomics data. *Scand. J. Clin. Lab. Investig.* **2015**, *75*, 193–203. [CrossRef]
27. Cloarec, O.; Dumas, M.E.; Trygg, J.; Craig, A.; Barton, R.H.; Lindon, J.C.; Nicholson, J.K.; Holmes, E. Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in 1H NMR spectroscopic metabonomic studies. *Anal. Chem.* **2005**, *77*, 517–526. [CrossRef]

This page intentionally left blank

Bibliography

[1] Roger J. Williams, Individual Metabolic Patterns and Human Disease: An Exploratory Study Utilizing Predominantly Paper Chromatographic Methods, University of Texas Publications (1951) 204 pp.

[2] E.C. Horning, M.G. Horning, Metabolic Profiles: Gas-Phase Methods for Analysis of Metabolites, Clinical Chemistry 17 (1971) 802–809.

[3] J.K. Nicholson, J.C. Lindon, E. Holmes, 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, Xenobiotica; the fate of foreign compounds in biological systems 29 (1999) 1181–1189.

[4] J.K. Nicholson, Wilson I.D., Understanding 'Global' Systems Biology: Metabonomics and the Continuum of Metabolism, Nature Reviews Drug Discovery 2 (2003) 668–676.

[5] O. Fiehn, Metabolomics — the link between genotypes and phenotypes, Functional genomics (2002) 155–171.

[6] S. Oliver, Systematic functional analysis of the yeast genome, Trends in Biotechnology 16 (1998) 373–378.

[7] B. Karahalil, Overview of Systems Biology and Omics Technologies, Current medicinal chemistry 23 (2016) 4221–4230.

[8] G.G. Harrigan, R. Goodacre, Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis, Springer US, Boston, MA, 2003.

[9] K. Dettmer, B.D. Hammock, Metabolomics-a new exciting field within the "omics" sciences., Environmental health perspectives 112 (2004) A396-A397.

[10] C. Town (Ed.), Functional Genomics, Springer Netherlands, Dordrecht, 2002.

[11] S. Gravel, B.M. Henn, R.N. Gutenkunst, A.R. Indap, G.T. Marth, A.G. Clark, F. Yu, R.A. Gibbs, C.D. Bustamante, Demographic history and rare allele sharing among human populations, Proceedings of the National Academy of Sciences of the United States of America 108 (2011) 11983–11988.

[12] R. Kandpal, B. Saviola, J. Felton, The era of 'omics unlimited, BioTechniques 46 (2009) 351-2, 354-5.

[13] F.S. Collins, V.A. McKusick, Implications of the Human Genome Project for medical science, JAMA 285 (2001) 540–544.

[14] S. Calvo, M. Jain, X. Xie, S.A. Sheth, B. Chang, O.A. Goldberger, A. Spinazzola, M. Zeviani, S.A. Carr, V.K. Mootha, Systematic identification of human mitochondrial disease genes through integrative genomics, Nature genetics 38 (2006) 576–582.

[15] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, T.A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, PNAS 106 (2009) 9362–9367.

[16] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, J.H. Cho, A.E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C.N. Rotimi, M. Slatkin, D. Valle, A.S. Whittemore, M. Boehnke, A.G. Clark, E.E. Eichler, G. Gibson, J.L. Haines, T.F.C. Mackay, S.A. McCarroll, P.M. Visscher, Finding the missing heritability of complex diseases, Nature 461 (2009) 747–753.

[17] F. Martínez-Jiménez, F. Muiños, I. Sentís, J. Deu-Pons, I. Reyes-Salazar, C. Arnedo-Pac, L. Mularoni, O. Pich, J. Bonet, H. Kranas, A. Gonzalez-Perez, N. Lopez-Bigas, A compendium of mutational cancer driver genes, Nature reviews. Cancer 20 (2020) 555–572.

[18] I. Barroso, Genetics of Type 2 diabetes, Diabetic medicine : a journal of the British Diabetic Association 22 (2005) 517–535.

[19] Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease, Genome Biol 18 (2017) 83.

[20] An integrated encyclopedia of DNA elements in the human genome, 2015.

[21] J.R. Alvarez-Dominguez, Z. Bai, D. Xu, B. Yuan, K.A. Lo, M.J. Yoon, Y.C. Lim, M. Knoll, N. Slavov, S. Chen, C. Peng, H.F. Lodish, L. Sun, De Novo Reconstruction of Adipose Tissue Transcriptomes Reveals Long Non-coding RNA Regulators of Brown Adipocyte Development, Cell Metabolism 21 (2015) 764–776.

[22] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat Biotechnol 28 (2010) 511–515.

[23] S.-Y. Ng, G.K. Bogu, B.S. Soh, L.W. Stanton, The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis, Molecular cell 51 (2013) 349–359.

[24] M. Knoll, H.F. Lodish, L. Sun, Long non-coding RNAs as regulators of the endocrine system, Nat Rev Endocrinol 11 (2015) 151–160.

[25] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, T. Shafee, Transcriptomics technologies, PLoS computational biology 13 (2017) e1005457.

[26] F. Ozsolak, P.M. Milos, RNA sequencing: advances, challenges and opportunities, Nat Rev Genet 12 (2011) 87–98.

[27] T.W. Nilsen, B.R. Graveley, Expansion of the eukaryotic proteome by alternative splicing, Nature 463 (2010) 457–463.

[28] G.-S. Wang, T.A. Cooper, Splicing in disease: disruption of the splicing code and the decoding machinery, Nat Rev Genet 8 (2007) 749–761.

[29] O. Kelemen, P. Convertini, Z. Zhang, Y. Wen, M. Shen, M. Falaleeva, S. Stamm, Function of alternative splicing, Gene 514 (2013) 1–30.

[30] B.T. Dye, B.A. Schulman, Structural mechanisms underlying posttranslational modification by ubiquitin-like proteins, Annual review of biophysics and biomolecular structure 36 (2007) 131–150.

[31] R. Wu, W. Haas, N. Dephoure, E.L. Huttlin, B. Zhai, M.E. Sowa, S.P. Gygi, A large-scale method to measure absolute protein phosphorylation stoichiometries, Nature methods 8 (2011) 677–683.

[32] M. Tyers, M. Mann, From genomics to proteomics, Nature 422 (2003) 193–197.

[33] G.J. Patti, O. Yanes, G. Siuzdak, Innovation: Metabolomics: the apogee of the omics trilogy, Nature reviews. Molecular cell biology 13 (2012) 263–269.

[34] S. Zhang, G.A. Nagana Gowda, T. Ye, D. Raftery, Advances in NMR-based biofluid analysis and metabolite profiling, The Analyst 135 (2010) 1490–1498.

[35] R.B. Schnabel, J. Baumert, M. Barbalic, J. Dupuis, P.T. Ellinor, P. Durda, A. Dehghan, J.C. Bis, T. Illig, A.C. Morrison, N.S. Jenny, J.F. Keaney, C. Gieger, C. Tilley, J.F. Yamamoto, N. Khuseyinova, G. Heiss, M. Doyle, S. Blankenberg, C. Herder, J.D. Walston, Y. Zhu, R.S. Vasan, N. Klopp, E. Boerwinkle, M.G. Larson, B.M. Psaty, A. Peters, C.M. Ballantyne, J.C.M. Witteman, R.C. Hoogeveen, E.J. Benjamin, W. Koenig, R.P. Tracy, Duffy antigen receptor for chemokines (Darc) polymorphism regulates

circulating concentrations of monocyte chemoattractant protein-1 and other inflammatory mediators, Blood 115 (2010) 5289–5299.

[36] Z. Yu, G. Kastenmüller, Y. He, P. Belcredi, G. Möller, C. Prehn, J. Mendes, S. Wahl, W. Roemisch-Margl, U. Ceglarek, A. Polonikov, N. Dahmen, H. Prokisch, L. Xie, Y. Li, H.-E. Wichmann, A. Peters, F. Kronenberg, K. Suhre, J. Adamski, T. Illig, R. Wang-Sattler, Differences between human plasma and serum metabolite profiles, PloS one 6 (2011) e21230.

[37] Yatomi, Yutaka, et al., Sphingosine 1-phosphate, a bioactive sphingolipid abundantly stored in platelets, is a normal constituent of human plasma and serum, The journal of biochemistry 121 (1997) 969–973.

[38] Y. Ma, P. Zhang, F. Wang, W. Liu, J. Yang, H. Qin, An integrated proteomics and metabolomics approach for defining oncofetal biomarkers in the colorectal cancer, Annals of surgery 255 (2012) 720–730.

[39] F. Chen, J. Xue, L. Zhou, S. Wu, Z. Chen, Identification of serum biomarkers of hepatocarcinoma through liquid chromatography/mass spectrometry-based metabonomic method, Anal Bioanal Chem 401 (2011) 1899–1904.

[40] N. Kumar, M. Shahjaman, M.N.H. Mollah, S.M.S. Islam, M.A. Hoque, Serum and Plasma Metabolomic Biomarkers for Lung Cancer, Bioinformation 13 (2017) 202–208.

[41] L. Lin, Z. Huang, Y. Gao, X. Yan, J. Xing, W. Hang, LC-MS based serum metabonomic analysis for renal cell carcinoma diagnosis, staging, and biomarker discovery, Journal of proteome research 10 (2011) 1396–1405.

[42] G. Echeverry, G.L. Hortin, A.J. Rai, Introduction to Urinalysis: Historical Perspectives and Clinical Application, in: The Urinary Proteome, Humana Press, 2010 pp. 1–12.

[43] G. Eknoyan, Looking at the urine: the renaissance of an unbroken tradition, American journal of kidney diseases : the official journal of the National Kidney Foundation 49 (2007) 865–872.

[44] N.J. Serkova, T.J. Standiford, K.A. Stringer, The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses, American journal of respiratory and critical care medicine 184 (2011) 647–655.

[45] M. Saoi, P. Britz-McKibbin, New Advances in Tissue Metabolomics: A Review, Metabolites 11 (2021).

[46] K. Tzimas, E. Pappa, Saliva Metabolomic Profile in Dental Medicine Research: A Narrative Review, Metabolites 13 (2023) 379.

[47] D.C. Mueller, M. Piller, R. Niessner, M. Scherer, G. Scherer, Untargeted metabolomic profiling in saliva of smokers and nonsmokers by a validated GC-TOF-MS method, Journal of proteome research 13 (2014) 1602–1613.

[48] S. Ishikawa, M. Sugimoto, K. Kitabatake, M. Tu, A. Sugano, I. Yamamori, A. Iba, K. Yusa, M. Kaneko, S. Ota, K. Hiwatari, A. Enomoto, T. Masaru, M. Iino, Effect of timing of collection of salivary metabolomic biomarkers on oral cancer detection, Amino acids 49 (2017) 761–770.

[49] D.S. Wishart, M.J. Lewis, J.A. Morrissey, M.D. Flegel, K. Jeroncic, Y. Xiong, D. Cheng, R. Eisner, B. Gautam, D. Tzur, S. Sawhney, F. Bamforth, R. Greiner, L. Li, The human cerebrospinal fluid metabolome, Journal of chromatography. B, Analytical technologies in the biomedical and life sciences 871 (2008) 164–173.

[50] D.G. Brown, S. Rao, T.L. Weir, J. O'Malia, M. Bazan, R.J. Brown, E.P. Ryan, Metabolomics and metabolic pathway networks from human colorectal cancers, adjacent mucosa, and stool, Cancer Metab 4 (2016) 11.

[51] Y. Gao, Urine-an untapped goldmine for biomarker discovery?, Science China. Life sciences 56 (2013) 1145–1146.

[52] A. Zhang, H. Sun, P. Wang, Y. Han, X. Wang, Future perspectives of personalized medicine in traditional Chinese medicine: a systems biology approach, Complementary therapies in medicine 20 (2012) 93–99.

[53] J.L. McClay, D.E. Adkins, N.G. Isern, T.M. O'Connell, J.B. Wooten, B.K. Zedler, M.S. Dasika, B.T. Webb, B.-J. Webb-Robertson, J.G. Pounds, E.L. Murrelle, M.F. Leppert, E.J.C.G. van den Oord, (1)H nuclear magnetic resonance metabolomics analysis identifies novel urinary biomarkers for lung function, Journal of proteome research 9 (2010) 3083–3090.

[54] K. Matsumura, M. Opiekun, H. Oka, A. Vachani, S.M. Albelda, K. Yamazaki, G.K. Beauchamp, Urinary volatile compounds as biomarkers for lung cancer: a proof of principle study using odor signatures in mouse models of lung cancer, PLOS ONE 5 (2010) e8819.

[55] A. Zhang, H. Sun, X. Wu, X. Wang, Urine metabolomics, Clinica chimica acta; international journal of clinical chemistry 414 (2012) 65–69.

[56] R.T. Krediet, Preservation of Residual Kidney Function and Urine Volume in Patients on Dialysis, Clinical journal of the American Society of Nephrology : CJASN 12 (2017) 377–379.

[57] S. Bouatra, F. Aziat, R. Mandal, A.C. Guo, M.R. Wilson, C. Knox, T.C. Bjorndahl, R. Krishnamurthy, F. Saleem, P. Liu, Z.T. Dame, J. Poelzer, J. Huynh, F.S. Yallou, N. Psychogios, E. Dong, R. Bogumil, C. Roehring, D.S. Wishart, The human urine metabolome, PloS one 8 (2013) e73076.

[58] D.S. Wishart, D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D.D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G.E. Duggan, G.D. MacInnis, A.M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B.D. Sykes, H.J. Vogel, L. Querengesser, HMDB: the Human Metabolome Database, Nucleic Acids Res 35 (2007) D521-6.

[59] D.S. Wishart, Y.D. Feunang, A. Marcu, A.C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, A. Scalbert, HMDB 4.0: the human metabolome database for 2018, Nucleic Acids Res 46 (2018) D608-D617.

[60] C. Chen, F.J. Gonzalez, J.R. Idle, LC-MS-based metabolomics in drug metabolism, Drug Metabolism Reviews 39 (2007) 581–597.

[61] C.H. Johnson, A.D. Patterson, J.R. Idle, F.J. Gonzalez, Xenobiotic metabolomics: major impact on the metabolome, Annual review of pharmacology and toxicology 52 (2012) 37–56.

[62] E. Tynkevich, M. Flamant, J.-P. Haymann, M. Metzger, E. Thervet, J.-J. Boffa, F. Vrtovsnik, P. Houillier, M. Froissart, B. Stengel, Decrease in urinary creatinine excretion in early stage chronic kidney disease, PLOS ONE 9 (2014) e111949.

[63] J. Liu, L. Bankir, A. Verma, S.S. Waikar, R. Palsson, Association of the Urine-to-Plasma Urea Ratio With CKD Progression, American journal of kidney diseases : the official journal of the National Kidney Foundation (2022).

[64] A. Jain, X.H. Li, W.N. Chen, An untargeted fecal and urine metabolomics analysis of the interplay between the gut microbiome, diet and human metabolism in Indian and Chinese adults, Sci Rep 9 (2019) 9191.

[65] P. Gao, E. Da Silva, L. Hou, N.D. Denslow, P. Xiang, L.Q. Ma, Human exposure to polycyclic aromatic hydrocarbons: Metabolomics perspective, Environment International 119 (2018) 466–477.

[66] M. Trupp, H. Zhu, W.R. Wikoff, R.A. Baillie, Z.-B. Zeng, P.D. Karp, O. Fiehn, R.M. Krauss, R. Kaddurah-Daouk, Metabolomics reveals amino acids contribute to variation in response to simvastatin treatment, PLOS ONE 7 (2012) e38386.

[67] D.S. Wishart, Applications of metabolomics in drug discovery and development, Drugs in R&D 9 (2008) 307–322.

[68] A. Alonso, S. Marsal, A. Julià, Analytical methods in untargeted metabolomics: state of the art in 2015, Front. Bioeng. Biotechnol. 3 (2015) 23.

[69] F.G. Pinto, I. Mahmud, T.A. Harmon, V.Y. Rubio, T.J. Garrett, Rapid Prostate Cancer Noninvasive Biomarker Screening Using Segmented Flow Mass Spectrometry-Based Untargeted Metabolomics, Journal of proteome research 19 (2020) 2080–2091.

[70] V. Mardegan, G. Giordano, M. Stocchero, P. Pirillo, G. Poloniato, E. Donadel, S. Salvadori, C. Giaquinto, E. Priante, E. Baraldi, Untargeted and Targeted Metabolomic Profiling of Preterm Newborns with EarlyOnset Sepsis: A Case-Control Study, Metabolites 11 (2021) 115.

[71] P. Reveglia, C. Paolillo, G. Ferretti, A. de Carlo, A. Angiolillo, R. Nasso, M. Caputo, C. Matrone, A. Di Costanzo, G. Corso, Challenges in LC-MS-based metabolomics for Alzheimer's disease early detection: targeted approaches versus untargeted approaches, Metabolomics 17 (2021) 78.

[72] R.D. Beger, W. Dunn, M.A. Schmidt, S.S. Gross, J.A. Kirwan, M. Cascante, L. Brennan, D.S. Wishart, M. Oresic, T. Hankemeier, D.I. Broadhurst, A.N. Lane, K. Suhre, G. Kastenmüller, S.J. Sumner, I. Thiele, O. Fiehn, R. Kaddurah-Daouk, Metabolomics enables precision medicine: "A White Paper, Community Perspective", Metabolomics 12 (2016) 149.

[73] D.S. Wishart, Emerging applications of metabolomics in drug discovery and precision medicine, Nat Rev Drug Discov 15 (2016) 473–484.

[74] C.B. Clish, Metabolomics: an emerging but powerful tool for precision medicine, Cold Spring Harb Mol Case Stud 1 (2015) a000588.

[75] K. Bingol, Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods, High-throughput 7 (2018).

[76] A.M. Weljie, J. Newton, P. Mercier, E. Carlson, C.M. Slupsky, Targeted profiling: quantitative analysis of 1H NMR metabolomics data, Analytical chemistry 78 (2006) 4430–4442.

[77] D.S. Wishart, NMR metabolomics: A look ahead, Journal of magnetic resonance (San Diego, Calif. : 1997) 306 (2019) 155–161.

[78] N.T. Doncheva, O. Palasca, R. Yarani, T. Litman, C. Anthon, M.A.M. Groenen, P.F. Stadler, F. Pociot, L.J. Jensen, J. Gorodkin, Human pathways in animal models: possibilities and limitations, Nucleic Acids Res 49 (2021) 1859–1871.

[79] D.G. Hackam, D.A. Redelmeier, Translation of research evidence from animals to humans, JAMA 296 (2006) 1731–1732.

[80] B.R. Berridge, Animal Study Translation: The Other Reproducibility Challenge, ILAR J 62 (2021) 1–6.

[81] S. Tenny, C.C. Kerndt, M.R. Hoffman, StatPearls. Case Control Studies, Treasure Island (FL), 2022.

[82] M.S. Thiese, Observational and interventional study design types; an overview, Biochemia Medica 24 (2014) 199–210.

[83] E.P. Rhee, R.E. Gerszten, Metabolomics and cardiovascular biomarker discovery, Clinical Chemistry 58 (2012) 139–147.

[84] W.B. Dunn, I.D. Wilson, A.W. Nicholls, D. Broadhurst, The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans, Bioanalysis 4 (2012) 2249–2264.

[85] A.J. Lloyd, N.D. Willis, T. Wilson, H. Zubair, E. Chambers, I. Garcia-Perez, L. Xie, K. Tailliart, M. Beckmann, J.C. Mathers, J. Draper, Addressing the pitfalls when designing intervention studies to discover and validate biomarkers of habitual dietary intake, Metabolomics 15 (2019) 72.

[86] L. Penn, H. Boeing, C.J. Boushey, L.O. Dragsted, J. Kaput, A. Scalbert, A.A. Welch, J.C. Mathers, Assessment of dietary intake: NuGO symposium report, Genes & nutrition 5 (2010) 205–213.

[87] S.A. Bingham, C. Gill, A. Welch, K. Day, A. Cassidy, K.T. Khaw, M.J. Sneyd, T.J. Key, L. Roe, N.E. Day, Comparison of dietary assessment methods in nutritional epidemiology: weighed records v. 24 h recalls, food-frequency questionnaires and estimated-diet records, The British journal of nutrition 72 (1994) 619–643.

[88] B. Yu, K.A. Zanetti, M. Temprosa, D. Albanes, N. Appel, C.B. Barrera, Y. Ben-Shlomo, E. Boerwinkle, J.P. Casas, C. Clish, C. Dale, A. Dehghan, A. Derkach, A.H. Eliassen, P. Elliott, E. Fahy, C. Gieger, M.J. Gunter, S. Harada, T. Harris, D.R. Herr, D. Herrington, J.N. Hirschhorn, E. Hoover, A.W. Hsing, M. Johansson, R.S. Kelly, C.M. Khoo, M. Kivimäki, B.S. Kristal, C. Langenberg, J. Lasky-Su, D.A. Lawlor, L.A. Lotta, M. Mangino, L. Le Marchand, E. Mathé, C.E. Matthews, C. Menni, L.A. Mucci, R. Murphy, M. Oresic, E. Orwoll, J. Ose, A.C. Pereira, M.C. Playdon, L. Poston, J. Price, Q. Qi, K. Rexrode, A. Risch, J. Sampson, W.J. Seow, H.D. Sesso, S.H. Shah, X.-O. Shu, G.C.S. Smith, U. Sovio, V.L. Stevens, R. Stolzenberg-Solomon, T. Takebayashi, T. Tillin, R. Travis, I. Tzoulaki, C.M. Ulrich, R.S. Vasan, M. Verma, Y. Wang, N.J. Wareham, A. Wong, N. Younes, H. Zhao, W. Zheng, S.C. Moore, The Consortium of Metabolomics Studies (COMETS): Metabolomics in 47 Prospective Cohort Studies, American journal of epidemiology 188 (2019) 991–1012.

[89] E.J. Caruana, M. Roman, J. Hernández-Sánchez, P. Solli, Longitudinal studies, Journal of thoracic disease 7 (2015) E537-40.

[90] A. El-Aneed, A. Cohen, J. Banoub, Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers, Applied Spectroscopy Reviews 44 (2009) 210–230.

[91] K. Dettmer, P.A. Aronov, B.D. Hammock, Mass spectrometry-based metabolomics, Mass spectrometry reviews 26 (2007) 51–78.

[92] S.G. Villas-Bôas, S. Mas, M. Akesson, J. Smedsgaard, J. Nielsen, Mass spectrometry in metabolome analysis, Mass spectrometry reviews 24 (2005) 613–646.

[93] A.-H. Emwas, R. Roy, R.T. McKay, L. Tenori, E. Saccenti, G.A.N. Gowda, D. Raftery, F. Alahmari, L. Jaremko, M. Jaremko, D.S. Wishart, NMR Spectroscopy for Metabolomics Research, Metabolites 9 (2019).

[94] A.M. Tsedilin, A.N. Fakhrutdinov, D.B. Eremin, S.S. Zalesskiy, A.O. Chizhov, N.G. Kolotyrkina, V.P. Ananikov, How sensitive and accurate are routine NMR and MS measurements?, Mendeleev Communications 25 (2015) 454–456.

[95] H. Trufelli, P. Palma, G. Famiglini, A. Cappiello, An overview of matrix effects in liquid chromatography-mass spectrometry, Mass spectrometry reviews 30 (2011) 491–509.

[96] Z. Pan, D. Raftery, Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics, Anal Bioanal Chem 387 (2007) 525–527.

[97] A.-H.M. Emwas, The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research, Methods in molecular biology (Clifton, N.J.) 1277 (2015) 161–193.

[98] N.V. Reo, NMR-based metabolomics, Drug and Chemical Toxicology 25 (2002) 375–382.

[99] D. Djukovic, G.A. Nagana Gowda, D. Raftery, Mass Spectrometry and NMR Spectroscopy–Based Quantitative Metabolomics, in: Proteomic and Metabolomic Approaches to Biomarker Discovery, Elsevier, 2013 pp. 279–297.

[100] J. Wieling, LC-MS-MS experiences with internal standards, Chromatographia 55 (2002) S107-S113.

[101] A.K. Boysen, K.R. Heal, L.T. Carlson, A.E. Ingalls, Best-Matched Internal Standard Normalization in Liquid Chromatography-Mass Spectrometry Metabolomics Applied to Environmental Samples, Analytical chemistry 90 (2018) 1363–1369.

[102] E.M. Purcell, H.C. Torrey, R.V. Pound, Resonance Absorption by Nuclear Magnetic Moments in a Solid, Phys. Rev. 69 (1946) 37–38.

[103] F. Bloch, Nuclear Induction, Phys. Rev. 70 (1946) 460–474.

[104] J. Keeler, Understanding NMR Spectroscopy, John Wiley & Sons, 2011.

[105] M.A. Hemminga, Introduction to NMR, Trends in Food Science & Technology 3 (1992) 179–186.

[106] V. Mlynárik, Introduction to nuclear magnetic resonance, Analytical biochemistry 529 (2017) 4–9.

[107] C.M. Slupsky, K.N. Rankin, J. Wagner, H. Fu, D. Chang, A.M. Weljie, E.J. Saude, B. Lix, D.J. Adamko, S. Shah, R. Greiner, B.D. Sykes, T.J. Marrie, Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles, Analytical chemistry 79 (2007) 6995–7004.

[108] E.J. Saude, D. Adamko, B.H. Rowe, T. Marrie, B.D. Sykes, Variation of metabolites in normal human urine, Metabolomics 3 (2007) 439–451.

[109] E.J. Saude, B.D. Sykes, Urine stability for metabolomic studies: effects of preparation and storage, Metabolomics 3 (2007) 19–27.

[110] M.J. Rist, C. Muhle-Goll, B. Görling, A. Bub, S. Heissler, B. Watzl, B. Luy, Influence of Freezing and Storage Procedure on Human Urine Samples in NMR-Based Metabolomics, Metabolites 3 (2013) 243–258.

[111] M. Lauridsen, S.H. Hansen, J.W. Jaroszewski, C. Cornett, Human urine as test material in 1H NMR-based metabonomics: recommendations for sample preparation and storage, Analytical chemistry 79 (2007) 1181–1186.

[112] P. Bernini, I. Bertini, C. Luchinat, P. Nincheri, S. Staderini, P. Turano, Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks, J Biomol NMR 49 (2011) 231–243.

[113] E. Martín Hernández, C. Aparicio López, G. Alvarez Calatayud, M.A. García Herrera, Litiasis vesical por ácido úrico en un niño con hipouricemia renal, Anales espanoles de pediatria 55 (2001) 273–276.

[114] R. Rylander, T. Remer, S. Berkemeyer, J. Vormann, Acid-base status affects renal magnesium losses in healthy, elderly persons, The Journal of nutrition 136 (2006) 2374–2377.

[115] A.A. Welch, A. Mulligan, S.A. Bingham, K.-T. Khaw, Urine pH is an indicator of dietary acid-base load, fruit and vegetables and meat intakes: results from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Norfolk population study, The British journal of nutrition 99 (2008) 1335–1343.

[116] C. Xiao, F. Hao, X. Qin, Y. Wang, H. Tang, An optimized buffer system for NMR-based urinary metabonomics with effective pH control, chemical shift consistency and dilution minimization, The Analyst 134 (2009) 916–925.

[117] V.M. Asiago, G.A. Nagana Gowda, S. Zhang, N. Shanaiah, J. Clark, D. Raftery, Use of EDTA to minimize ionic strength dependent frequency shifts in the 1H NMR spectra of urine, Metabolomics 4 (2008) 328–336.

[118] R.B. Gil, R. Lehmann, P. Schmitt-Kopplin, S.S. Heinzmann, (1)H NMR-based metabolite profiling workflow to reduce inter-sample chemical shift variations in urine samples for improved biomarker discovery, Anal Bioanal Chem 408 (2016) 4683–4691.

[119] A.C. Dona, B. Jiménez, H. Schäfer, E. Humpfer, M. Spraul, M.R. Lewis, J.T.M. Pearce, E. Holmes, J.C. Lindon, J.K. Nicholson, Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping, Analytical chemistry 86 (2014) 9887–9894.

[120] A. Vignoli, V. Ghini, G. Meoni, C. Licari, P.G. Takis, L. Tenori, P. Turano, C. Luchinat, High-Throughput Metabolomics by 1D NMR, Angewandte Chemie (International ed. in English) 58 (2019) 968–994.

[121] O. Beckonert, H.C. Keun, T.M.D. Ebbels, J. Bundy, E. Holmes, J.C. Lindon, J.K. Nicholson, Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts, Nat Protoc 2 (2007) 2692–2703.

[122] S.S. Heinzmann, M. Waldenberger, A. Peters, P. Schmitt-Kopplin, Cluster Analysis Statistical Spectroscopy for the Identification of Metabolites in 1H NMR Metabolomics, Metabolites 12 (2022) 992.

[123] V. Falaina, C. Fotakis, T. Boutsikou, T. Tsiaka, G. Moros, S. Ouzounis, V. Andreou, Z. Iliodromiti, T. Xanthos, Y. Vandenplas, N. Iacovidou, P. Zoumpoulakis, Urine Metabolomic Profile of Breast- versus Formula-Fed Neonates Using a Synbiotic-Enriched Formula, International journal of molecular sciences 23 (2022).

[124] A. Shimizu, M. Ikeguchi, S. Sugai, Appropriateness of DSS and TSP as internal references for (1)H NMR studies of molten globule proteins in aqueous media, J Biomol NMR 4 (1994) 859–862.

[125] A.-H. Emwas, C. Luchinat, P. Turano, L. Tenori, R. Roy, R.M. Salek, D. Ryan, J.S. Merzaban, R. Kaddurah-Daouk, A.C. Zeri, G.A. Nagana Gowda, D. Raftery, Y. Wang, L. Brennan, D.S. Wishart, Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: a review, Metabolomics 11 (2015) 872–894.

[126] C. Deborde, J.-X. Fontaine, D. Jacob, A. Botana, V. Nicaise, F. Richard-Forget, S. Lecomte, C. Decourtil, K. Hamade, F. Mesnard, A. Moing, R. Molinié, Optimizing 1D 1H-NMR profiling of plant samples for high throughput analysis: extract preparation, standardization, automation and spectra processing, Metabolomics 15 (2019) 28.

[127] N. Araníbar, K.-H. Ott, V. Roongta, L. Mueller, Metabolomic analysis using optimized NMR and statistical methods, Analytical biochemistry 355 (2006) 62–70.

[128] R.T. McKay, How the 1D-NOESY suppresses solvent signal in metabonomics NMR spectroscopy: An examination of the pulse sequence components and evolution, Concepts Magn. Reson. 38A (2011) 197–220.

[129] S. Sokolenko, R. McKay, E.J.M. Blondeel, M.J. Lewis, D. Chang, B. George, M.G. Aucoin, Understanding the variability of compound quantification from targeted profiling metabolomics of 1D-1H-NMR spectra in synthetic mixtures and urine with additional insights on choice of pulse sequences and robotic sampling, Metabolomics 9 (2013) 887–903.

[130] S.K. Bharti, R. Roy, Quantitative 1H NMR spectroscopy, TrAC Trends in Analytical Chemistry 35 (2012) 5–26.

[131] S. Ren, A.A. Hinzman, E.L. Kang, R.D. Szczesniak, L.J. Lu, Computational and statistical analysis of metabolomics data, Metabolomics 11 (2015) 1492–1513.

[132] Y. Xi, D.M. Rocke, Baseline correction for NMR spectroscopic metabolomics data analysis, BMC Bioinformatics 9 (2008) 324.

[133] R.A. Scott, L.J. Scott, R. Mägi, L. Marullo, K.J. Gaulton, M. Kaakinen, N. Pervjakova, T.H. Pers, A.D. Johnson, J.D. Eicher, A.U. Jackson, T. Ferreira, Y. Lee, C. Ma, V. Steinthorsdottir, G. Thorleifsson, L. Qi, N.R. van Zuydam, A. Mahajan, H. Chen, P. Almgren, B.F. Voight, H. Grallert, M. Müller-Nurasyid, J.S. Ried, N.W. Rayner, N. Robertson, L.C. Karssen, E.M. van Leeuwen, S.M. Willems, C. Fuchsberger, P. Kwan, T.M. Teslovich, P. Chanda, M. Li, Y. Lu, C. Dina, D. Thuillier, L. Yengo, L. Jiang, T. Sparso, H.A. Kestler, H. Chheda, L. Eisele, S. Gustafsson, M. Frånberg, R.J. Strawbridge, R. Benediktsson, A.B. Hreidarsson, A. Kong, G. Sigurðsson, N.D. Kerrison, J. Luan, L. Liang, T. Meitinger, M. Roden, B. Thorand, T. Esko, E. Mihailov, C. Fox, C.-T. Liu, D. Rybin, B. Isomaa, V. Lyssenko, T. Tuomi, D.J. Couper, J.S. Pankow, N. Grarup, C.T. Have, M.E. Jørgensen, T. Jørgensen, A. Linneberg, M.C. Cornelis, R.M. van Dam, D.J. Hunter, P. Kraft, Q. Sun, S. Edkins, K.R. Owen, J.R.B. Perry, A.R. Wood, E. Zeggini, J. Tajes-Fernandes, G.R. Abecasis, L.L. Bonnycastle, P.S. Chines, H.M. Stringham, H.A. Koistinen, L. Kinnunen, B. Sennblad, T.W. Mühleisen, M.M. Nöthen, S. Pechlivanis, D. Baldassarre, K. Gertow, S.E. Humphries, E. Tremoli, N. Klopp, J. Meyer, G. Steinbach, R. Wennauer, J.G. Eriksson, S. Männistö, L. Peltonen, E. Tikkanen, G. Charpentier, E. Eury, S. Lobbens, B. Gigante, K. Leander, O. McLeod, E.P. Bottinger, O. Gottesman, D. Ruderfer, M. Blüher, P. Kovacs, A. Tonjes, N.M. Maruthur, C. Scapoli, R. Erbel, K.-H. Jöckel, S. Moebus, U. de Faire, A. Hamsten, M. Stumvoll, P. Deloukas, P.J. Donnelly, T.M. Frayling, A.T. Hattersley, S. Ripatti, V. Salomaa, N.L. Pedersen, B.O. Boehm, R.N. Bergman, F.S. Collins, K.L. Mohlke, J. Tuomilehto, T. Hansen, O. Pedersen, I. Barroso, L. Lannfelt, E. Ingelsson, L. Lind, C.M. Lindgren, S. Cauchi, P. Froguel, R.J.F. Loos, B. Balkau, H. Boeing, P.W. Franks, A. Barricarte Gurrea, D. Palli, Y.T. van

der Schouw, D. Altshuler, L.C. Groop, C. Langenberg, N.J. Wareham, E. Sijbrands, C.M. van Duijn, J.C. Florez, J.B. Meigs, E. Boerwinkle, C. Gieger, K. Strauch, A. Metspalu, A.D. Morris, C.N.A. Palmer, F.B. Hu, U. Thorsteinsdottir, K. Stefansson, J. Dupuis, A.P. Morris, M. Boehnke, M.I. McCarthy, I. Prokopenko, An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans, Diabetes 66 (2017) 2888–2902.

[134] Y. Chen, G. Shen, R. Zhang, J. He, Y. Zhang, J. Xu, W. Yang, X. Chen, Y. Song, Z. Abliz, Combination of injection volume calibration by creatinine and MS signals' normalization to overcome urine variability in LC-MS-based metabolomics studies, Analytical chemistry 85 (2013) 7659–7665.

[135] S.L. Nam, A.P. de La Mata, R.P. Dias, J.J. Harynuk, Towards Standardization of Data Normalization Strategies to Improve Urinary Metabolomics Studies by GC×GC-TOFMS, Metabolites 10 (2020).

[136] Y. Wu, L. Li, Sample normalization methods in quantitative metabolomics, Journal of chromatography. A 1430 (2016) 80–95.

[137] D. Ryan, K. Robards, P.D. Prenzler, M. Kendall, Recent and potential developments in the analysis of urine: a review, Analytica chimica acta 684 (2011) 8–20.

[138] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics, Analytical chemistry 78 (2006) 4281–4290.

[139] I. Karaman, Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis, Advances in experimental medicine and biology 965 (2017) 145–161.

[140] B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics 19 (2003) 185–193.

[141] N. Knudsen, E. Christiansen, M. Brandt-Christensen, B. Nygaard, H. Perrild, Age- and sex-adjusted iodine/creatinine ratio. A new standard in epidemiological surveys? Evaluation of three different estimates of iodine excretion based on casual urine samples and comparison to 24 h values, Eur J Clin Nutr 54 (2000) 361–363.

[142] T.I. Justesen, J.L.A. Petersen, P. Ekbom, P. Damm, E.R. Mathiesen, Albumin-to-creatinine ratio in random urine samples might replace 24-h urine collections in

screening for micro- and macroalbuminuria in pregnant woman with type 1 diabetes, Diabetes Care 29 (2006) 924–925.

[143] P. Shaffer, THE EXCRETION OF KREATININ AND KREATIN IN HEALTH AND DISEASE, American Journal of Physiology-Legacy Content 23 (1908) 1–22.

[144] F. Spierto, W. Hannon, E. Gunter, S. Smith, Stability of urine creatinine, Clinica Chimica Acta 264 (1997) 227–232.

[145] S.M. Kohl, M.S. Klein, J. Hochrein, P.J. Oefner, R. Spang, W. Gronwald, State-of-the art data normalization methods improve NMR-based metabolomic analysis, Metabolomics 8 (2012) 146–160.

[146] A. Craig, O. Cloarec, E. Holmes, J.K. Nicholson, J.C. Lindon, Scaling and normalization effects in NMR spectroscopic metabonomic data sets, Analytical chemistry 78 (2006) 2262–2267.

[147] J. Xia, T.C. Bjorndahl, P. Tang, D.S. Wishart, MetaboMiner--semi-automated identification of metabolites from 2D NMR spectra of complex biofluids, BMC Bioinformatics 9 (2008) 507.

[148] J. Forshed, I. Schuppe-Koistinen, S.P. Jacobsson, Peak alignment of NMR signals by means of a genetic algorithm, Analytica chimica acta 487 (2003) 189–199.

[149] E. Holmes, P.J. Foxall, J.K. Nicholson, G.H. Neild, S.M. Brown, C.R. Beddell, B.C. Sweatman, E. Rahr, J.C. Lindon, M. Spraul, Automatic data reduction and pattern recognition methods for analysis of 1H nuclear magnetic resonance spectra of human urine from normal and pathological states, Analytical biochemistry 220 (1994) 284–296.

[150] L. Csenki, E. Alm, R.J.O. Torgrip, K.M. Aberg, L.I. Nord, I. Schuppe-Koistinen, J. Lindberg, Proof of principle of a generalized fuzzy Hough transform approach to peak alignment of one-dimensional 1H NMR data, Anal Bioanal Chem 389 (2007) 875–885.

[151] V. Pravdova, B. Walczak, D.L. Massart, A comparison of two algorithms for warping of analytical signals, Analytica chimica acta 456 (2002) 77–92.

[152] G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, J. Chemometrics 18 (2004) 231–241.

[153] K.A. Veselkov, J.C. Lindon, T.M.D. Ebbels, D. Crockford, V.V. Volynkin, E. Holmes, D.B. Davies, J.K. Nicholson, Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery, Analytical chemistry 81 (2009) 56–66.

[154] W. Wu, M. Daszykowski, B. Walczak, B.C. Sweatman, S.C. Connor, J.N. Haselden, D.J. Crowther, R.W. Gill, M.W. Lutz, Peak alignment of urine NMR spectra using fuzzy warping, Journal of chemical information and modeling 46 (2006) 863–875.

[155] E. Holmes, R.L. Loo, J. Stamler, M. Bictash, I.K.S. Yap, Q. Chan, T. Ebbels, M. de Iorio, I.J. Brown, K.A. Veselkov, M.L. Daviglus, H. Kesteloot, H. Ueshima, L. Zhao, J.K. Nicholson, P. Elliott, Human metabolic phenotype diversity and its association with diet and blood pressure, Nature 453 (2008) 396–400.

[156] E. Holmes, J.K. Nicholson, A.W. Nicholls, J.C. Lindon, S.C. Connor, S. Polley, J. Connelly, The identification of novel biomarkers of renal toxicity using automatic data reduction techniques and PCA of proton NMR spectra of urine, Chemometrics and Intelligent Laboratory Systems 44 (1998) 245–255.

[157] M. Spraul, P. Neidig, U. Klauck, P. Kessler, E. Holmes, J.K. Nicholson, B.C. Sweatman, S.R. Salman, R.D. Farrant, E. Rahr, C.R. Beddell, J.C. Lindon, Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples, Journal of Pharmaceutical and Biomedical Analysis 12 (1994) 1215–1225.

[158] H.U. Zacharias, M. Altenbuchinger, W. Gronwald, Statistical Analysis of NMR Metabolic Fingerprints: Established Methods and Recent Advances, Metabolites 8 (2018).

[159] P.E. Anderson, D.A. Mahle, T.E. Doom, N.V. Reo, N.J. DelRaso, M.L. Raymer, Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data, Metabolomics 7 (2011) 179–190.

[160] R.A. Davis, A.J. Charlton, J. Godward, S.A. Jones, M. Harrison, J.C. Wilson, Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform, Chemometrics and Intelligent Laboratory Systems 85 (2007) 144–154.

[161] B. Worley, R. Powers, Generalized adaptive intelligent binning of multiway data, Chemometrics and Intelligent Laboratory Systems 146 (2015) 42–46.

[162] A.-H. Emwas, E. Saccenti, X. Gao, R.T. McKay, V.A.P.M. Dos Santos, R. Roy, D.S. Wishart, Recommended strategies for spectral processing and post-processing of 1D 1H-NMR data of biofluids with a particular focus on urine, Metabolomics 14 (2018) 31.

[163] S. Halouska, R. Powers, Negative impact of noise on the principal component analysis of NMR data, Journal of Magnetic Resonance 178 (2006) 88–95.

[164] A.-H. Emwas, R. Roy, R.T. McKay, D. Ryan, L. Brennan, L. Tenori, C. Luchinat, X. Gao, A.C. Zeri, G.A.N. Gowda, D. Raftery, C. Steinbeck, R.M. Salek, D.S. Wishart, Recommendations and Standardization of Biomarker Quantification Using NMR-Based Metabolomics with Particular Focus on Urinary Analysis, Journal of proteome research 15 (2016) 360–373.

[165] J. Hao, M. Liebeke, W. Astle, M. de Iorio, J.G. Bundy, T.M.D. Ebbels, Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN, Nat Protoc 9 (2014) 1416–1427.

[166] S. Ravanbakhsh, P. Liu, T.C. Bjorndahl, T.C. Bjordahl, R. Mandal, J.R. Grant, M. Wilson, R. Eisner, I. Sinelnikov, X. Hu, C. Luchinat, R. Greiner, D.S. Wishart, Accurate, fully-automated NMR spectral profiling for metabolomics, PLOS ONE 10 (2015) e0124219.

[167] D. Jacob, C. Deborde, M. Lefebvre, M. Maucourt, A. Moing, NMRProcFlow: a graphical and interactive tool dedicated to 1D spectra processing for NMR-based metabolomics, Metabolomics 13 (2017) 36.

[168] J. Hao, W. Astle, M. de Iorio, T.M.D. Ebbels, BATMAN--an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model, Bioinformatics 28 (2012) 2088–2090.

[169] M.R. Viant, B.G. Lyeth, M.G. Miller, R.F. Berman, An NMR metabolomic investigation of early metabolic disturbances following traumatic brain injury in a mammalian model, NMR in biomedicine 18 (2005) 507–516.

[170] I. Jolliffe, Principal Component Analysis, in: B.S. Everitt, D.C. Howell (Eds.) Encyclopedia of Statistics in Behavioral Science, John Wiley & Sons, Ltd, Chichester, UK, 2005.

[171] J. Bartel, J. Krumsiek, F.J. Theis, Statistical methods for the analysis of high-throughput metabolomics data, Computational and structural biotechnology journal 4 (2013) e201301009.

[172] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, J. Selbig, Metabolite fingerprinting: detecting biological features by independent component analysis, Bioinformatics 20 (2004) 2447–2454.

[173] O. Beckonert, M. E. Bollard, T.M. Ebbels, H.C. Keun, H. Antti, E. Holmes, J.C. Lindon, J.K. Nicholson, NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches, Analytica chimica acta 490 (2003) 3–15.

[174] X. Li, X. Lu, J. Tian, P. Gao, H. Kong, G. Xu, Application of fuzzy c-means clustering in data analysis of metabolomics, Analytical chemistry 81 (2009) 4468–4475.

[175] J.A. Hageman, R.A. van den Berg, J.A. Westerhuis, H.C.J. Hoefsloot, A.K. Smilde, Bagged K-Means Clustering of Metabolome Data, Critical Reviews in Analytical Chemistry 36 (2006) 211–220.

[176] T. Kohonen, The self-organizing map, Proc. IEEE 78 (1990) 1464–1480.

[177] V.-P. Mäkinen, P. Soininen, C. Forsblom, M. Parkkonen, P. Ingman, K. Kaski, P.-H. Groop, M. Ala-Korpela, 1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death, Molecular Systems Biology 4 (2008) 167.

[178] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics 17 (2003) 166–173.

[179] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), Journal of Chemometrics 16 (2002) 119–128.

[180] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32.

[181] C.J. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery 2 (1998) 121–167.

[182] M. Bylesjö, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J. Trygg, OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification, Journal of Chemometrics 20 (2006) 341–351.

[183] T. Chen, Y. Cao, Y. Zhang, J. Liu, Y. Bao, C. Wang, W. Jia, A. Zhao, Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection, Evidence-based complementary and alternative medicine : eCAM 2013 (2013) 298183.

[184] B. Xi, H. Gu, H. Baniasadi, D. Raftery, Statistical analysis and modeling of mass spectrometry-based metabolomics data, Methods in molecular biology (Clifton, N.J.) 1198 (2014) 333–353.

[185] S. Mahadevan, S.L. Shah, T.J. Marrie, C.M. Slupsky, Analysis of metabolomic data using support vector machines, Analytical chemistry 80 (2008) 7562–7570.

[186] A. Camargo, F. Azuaje, H. Wang, H. Zheng, Permutation - based statistical tests for multiple hypotheses, Source code for biology and medicine 3 (2008) 15.

[187] J.M. Bland, D.G. Altman, Multiple significance tests: the Bonferroni method, BMJ 310 (1995) 170.

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

[189] A.C. Dona, M. Kyriakides, F. Scott, E.A. Shephard, D. Varshavi, K. Veselkov, J.R. Everett, A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments, Computational and structural biotechnology journal 14 (2016) 135–153.

[190] K. Bingol, L. Bruschweiler-Li, C. Yu, A. Somogyi, F. Zhang, R. Brüschweiler, Metabolomics beyond spectroscopic databases: a combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures, Analytical chemistry 87 (2015) 3864–3870.

[191] A. Le Guennec, I. Tea, I. Antheaume, E. Martineau, B. Charrier, M. Pathan, S. Akoka, P. Giraudeau, Fast determination of absolute metabolite concentrations by spatially encoded 2D NMR: application to breast cancer cell extracts, Analytical chemistry 84 (2012) 10831–10837.

[192] H.K. Kim, Y.H. Choi, R. Verpoorte, NMR-based metabolomic analysis of plants, Nat Protoc 5 (2010) 536–549.

[193] K. Bingol, L. Bruschweiler-Li, D.-W. Li, R. Brüschweiler, Customized metabolomics database for the analysis of NMR ^{1}H-^{1}H TOCSY and ^{13}C-^{1}H HSQC-TOCSY spectra of complex mixtures, Analytical chemistry 86 (2014) 5494–5501.

[194] J.L. Markley, R. Brüschweiler, A.S. Edison, H.R. Eghbalnia, R. Powers, D. Raftery, D.S. Wishart, The future of NMR-based metabolomics, Current Opinion in Biotechnology 43 (2017) 34–40.

[195] C. Ludwig, M.R. Viant, Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox, Phytochemical analysis : PCA 21 (2010) 22–32.

[196] P. Bernini, I. Bertini, C. Luchinat, S. Nepi, E. Saccenti, H. Schäfer, B. Schütz, M. Spraul, L. Tenori, Individual human phenotypes in metabolic space and time, Journal of proteome research 8 (2009) 4264–4271.

[197] O. Cloarec, M.-E. Dumas, A. Craig, R.H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J.C. Lindon, E. Holmes, J. Nicholson, Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets, Analytical chemistry 77 (2005) 1282–1289.

[198] C. Steinbeck, S. Krause, S. Kuhn, NMRShiftDB-constructing a free chemical information system with open-source components, Journal of chemical information and computer sciences 43 (2003) 1733–1739.

[199] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W.-M. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M.D. Reily, J.J. Thaden, M.R. Viant, Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI), Metabolomics 3 (2007) 211–221.

[200] C. Martias, N. Baroukh, S. Mavel, H. Blasco, A. Lefèvre, L. Roch, F. Montigny, J. Gatien, L. Schibler, D. Dufour-Rainfray, L. Nadal-Desbarats, P. Emond, Optimization of Sample Preparation for Metabolomics Exploration of Urine, Feces, Blood and Saliva in Humans Using Combined NMR and UHPLC-HRMS Platforms, Molecules (Basel, Switzerland) 26 (2021).

[201] S.D. Bruce, J. Higinbotham, I. Marshall, P.H. Beswick, An analytical derivation of a popular approximation of the Voigt function for quantification of NMR spectra, Journal of Magnetic Resonance 142 (2000) 57–63.

[202] I. Marshall, J. Higinbotham, S. Bruce, A. Freise, Use of Voigt lineshape for quantification of in vivo 1H spectra, Magnetic Resonance in Medicine 37 (1997) 651–657.

[203] A.W. Brown, K.A. Kaiser, D.B. Allison, Issues with data and analyses: Errors, underlying themes, and potential solutions, PNAS 115 (2018) 2563–2570.

[204] S. Hollmann, A. Kremer, Š. Baebler, C. Trefois, K. Gruden, W.R. Rudnicki, W. Tong, A. Gruca, E. Bongcam-Rudloff, C.T. Evelo, A. Nechyporenko, M. Frohme, D. Šafránek, B. Regierer, D. D'Elia, The need for standardisation in life science research - an approach to excellence and trust, F1000Research 9 (2020) 1398.

[205] R.R. Downs, Improving Opportunities for New Value of Open Data: Assessing and Certifying Research Data Repositories, Data Science Journal 20 (2021).

[206] T. Gebregiworgis, R. Powers, Application of NMR metabolomics to search for human disease biomarkers, Bentham Science Publishers, 2012.

[207] M. Mora-Ortiz, P. Nuñez Ramos, A. Oregioni, S.P. Claus, NMR metabolomics identifies over 60 biomarkers associated with Type II Diabetes impairment in db/db mice, Metabolomics 15 (2019) 89.

[208] G.D. Lewis, A. Asnani, R.E. Gerszten, Application of metabolomics to cardiovascular biomarker and pathway discovery, Journal of the American College of Cardiology 52 (2008) 117–123.

[209] W. Pathmasiri, K. Kay, S. McRitchie, S. Sumner, Analysis of NMR Metabolomics Data, Methods in molecular biology (Clifton, N.J.) 2104 (2020) 61–97.

This page intentionally left blank

List of figures

This page intentionally left blank