

---

# Computational Methods for Breath Metabolomics in Clinical Diagnostics

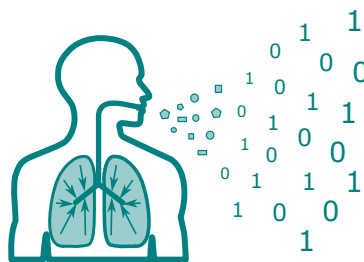
---

Dissertation  
zur Erlangung des Grades des  
Doktors der Naturwissenschaften der  
Naturwissenschaftlich-Technischen  
Fakultäten der  
Universität des Saarlandes

eingereicht von

Anne-Christin Hauschild, M.Sc.

Saarbrücken, 16. Juni 2016





Day of Colloquium	July 18th 2016
Dean of the Faculty	Univ.-Prof. Dr. Frank-Olaf Schreyer
Chair of the Committee	Prof. Dr. Verena Wolf
First reviewer	Prof. Dr. Volkhard Helms
Second reviewer	Prof. Dr. Jan Baumbach
Academic Assistant	Dr. Markus List





## Executive Summary

For a long time, human odors and vapors have been known for their diagnostic power. Therefore, the analysis of the metabolic composition of human breath and odors creates the opportunity for a non-invasive tool for clinical diagnostics. Innovative analytical technologies to capture the metabolic profile of a patient's breath are available, such as, for instance, the ion mobility spectrometry coupled to a multicapillary column. However, we are lacking automated systems to process, analyse and evaluate large clinical studies of the human exhaled air. To fill this gap, a number of computational challenges need to be addressed. For instance, breath studies generate large amounts of heterogeneous data that requires automated preprocessing, peak-detection and identification as a basis for a sophisticated follow up analysis. In addition, generalizable statistical evaluation frameworks for the detection of breath biomarker profiles that are robust enough to be employed in routine clinical practice are necessary. In particular since breath metabolomics is susceptible to specific confounding factors and background noise, similar to other clinical diagnostics technologies. Moreover, specific manifestations of disease stages and progression, may largely influence the breathomics profiles. To this end, this thesis will address these challenges to move towards more automatization and generalization in clinical breath research. In particular I present methods to support the search for biomarker profiles that enable a non-invasive detection of diseases, treatment optimization and prognosis to provide a new powerful tool for precision medicine.

## Kurzzusammenfassung

Seit jeher ist bekannt, dass Körpergeruch und der Atem Hinweise zu deren Gesundheitszustand liefern können. Eine Analyse der Atemluft auf molekularer Ebene verspricht daher neue Ansätze zur Diagnose spezifischer Krankheiten. Innovative Technologien wie die Ionen Mobilitäts Spectrometrie in Kombination mit einer Multikapilarsäule, erlauben erstmals hochauflösende metabolische Profile der Atemluft innerhalb kürzester Zeit zu erzeugen. Zur Zeit fehlen jedoch die notwendigen computergestützten Applikationen zur automatischen Organisation und Auswertung der generierten Daten. Eine besondere Herausforderung stellen dabei die großen Mengen heterogener klinischer und analytischer Daten und deren Verarbeitung. Ähnlich wie andere Hochdurchsatzverfahren unterliegt die Atemluft dem Einfluss von Hintergrundsignalen wie der Umgebungsluft oder anderen die Ergebnisse verzerrenden Faktoren, wie zum Beispiel Ernährung, Lebensgewohnheiten oder Medikation. Dies erfordert den Einsatz von modernen Methoden der Statistik und des maschinellen Lernens, um robuste und generalisierbare Krankheitsmarker zu identifizieren. Ein besonderer Augenmerk gilt hierbei auch Krankheiten deren metabolischer Fingerabdruck sich im Krankheitsverlauf drastisch verändern können. Das Ziel meiner Arbeit ist es Lösungen für die beschriebenen Probleme zu finden und damit die Suche nach praxistauglichen Krankheitsmarkern mit bioinformatischen Methoden zu unterstützen. Im Rahmen mehrerer Studien und Softwareprojekten wurden grundlegende Methodiken vorgestellt, evaluiert und etabliert, insbesondere im Hinblick auf die Entwicklung computergestützter Systeme zur automatischen Analyse von Atemluftdaten. Die vorgestellten Verfahren legen den Grundstein für die nicht invasive Detektion von Krankheiten, Optimierung und Prognose von Behandlungen und darüber hinaus für ein weiteres Werkzeug der personalisierten Medizin.



# Abstract

Odors and vapors of the body and breath have been known for their diagnostic power for millennia. More recent history confirmed this knowledge within clinical studies by successfully training dogs and mice to detect diseases, by sniffing specific volatile organic profiles. Like a vertebrate nose, there exist analytical technologies capable of capturing such metabolites. The science of analyzing the aggregation of all metabolites within the breath of an organism is called breathomics. The crucial task is to identify discriminating patterns that are predictive for certain diseases. Additionally, like other diagnostic technologies, breath is influenced by various sources of systematic or random noise. The field needs to move from separability to predictability by evolving from pilot studies to large scale screening studies. Therefore, there is a necessity for further standardization and automatization in managing, analyzing and evaluating this novel type of metabolomics data. In order to achieve this, several challenges remain to be addressed: data accumulation and heterogeneity; manual peak finding; unknown metabolites; robust statistics and biomarkers; background noise and confounding factors; heterogeneous diseases and disease stages; usability, maintainability, and re-usability.

In this thesis I will describe six projects that propose possible solutions to these challenges. (1) The IMSDB is the first functional and flexible comprehensive breathomics database. It provides flexible yet quick storage of heterogeneous clinical and large amounts of metabolic breath data. (2) A pilot study lays the foundations for a more robust and adequate prediction, evaluation and feature selection of breathomics data, by introducing established machine learning techniques to the field of breath analysis. (3) Further, the thesis presents the first qualitative analysis of the performance of automated peak detection methods and thereby proves their ability to compete with the manual gold standard. (4) The MIMA software tool enables the automated identification of the captured organic components by mapping different analytical technologies. (5) The Carotta software system provides a user friendly unsupervised learning platform, that enables easy discovery of hidden structures in metabolomic breath data such as disease subtypes or confounding factors. (6) Finally, I will introduce the first longitudinal modeling of breath metabolite behavior during the course of an evolving disease.

In conclusion, the aggregation of these projects builds the foundation for a more robust and standardized analysis schema, leading to more comparability and generalization of future breathomics studies. Moreover, it sets the basis for automated frameworks integrating the described tools and approaches into steps of a continuous breath analysis pipeline.



# Acknowledgements

First, I would like to express my gratitude to my supervisors Prof. Jan Baumbach and Prof. Volkhard Helms for their persistent support and guidance through out my thesis. I dedicate special thanks to Prof. Jan Baumbach and Prof. Jörg Ingo Baumbach for giving me the possibility to work on such and interesting and motivating topic and for involving me in many exiting projects. Further I want to thank them for their priceless scientific guidance and inspiring discussions in the fields of breath analysis and computational biology. To Prof. Jan Baumbach I am deeply grateful for the constant motivation and scientific advices and inspirations but also for creating such a fruitful research environment allowing me to cultivate and follow my own ideas.

My gratitude and deep appreciation to all my collaboration partners, namely Prof. Sandy P. Eckel at the Division of Biostatistics, University of Southern California (USC), Los Angeles, United States, Prof. Sasha Kreuer, Tobias Fink and Felix Maurer at the Department of Anesthesiology, Intensive Care, and Pain Therapy, Saarland University Medical Center, Homburg (Saar), Germany, and Prof. Sven Rahmann, Marianna D'Addario and Dr. Dominik Kopczynski at the Bioinformatics group at the technical university of Dortmund. Further I am very thankful for the fruitful collaborations with my Master and Bachelor students Till Schneider, Tobias Frisch and Rune Sostack Clausen.

I would like to thank all of my colleagues in Germany and Denmark for all the discussions and interesting conversations and for creating such a inspiring work environment. Thanks to all the former colleagues at the Korean Institute of Science and Technology Kathrin Eisinger, Sasidhar Maddula and Felix Maurer for the interesting discussions about breath analysis and technologies behind it. Moreover, thanks to all the members of the Baumbach groups, in Saarbrücken, Josh, Peng, Rachid, and Richard as well as in Odense, Anders, Christian, Diogo, Eudes, Lucas, Markus, Nicolas, Paolo, Richa, Richard, for all the exiting discussions about computational and statistical challenges we were facing.

There are countless friends that supported me during the course of my career in Germany and many that I made during my time in Denmark and I would like to name a few, Juliane, Daniel and Charlotte, Peter, Thomas, Alejandro, Markus, Fabian, Verena, Matthias, David and Elena, Marie-Pier, Manuela and Andre, Linda, Susanna, Sophia and Pancho, Liza and Eliot. You made the time of my PhD a lovely and enjoyable experience in my life even in stressful times.

A special thanks to all of you that motivated and supported me during the difficult time of Thesis writing.

I would also like to thank my dear ex-husband Markus, your encouragement and constant motivation that helped me through countless stressful moments throughout my PhD. I am happy to call you my friend.

Most importantly, I would like to thank my parents, Johann and Anette who have selflessly supported me throughout my entire life. for their financial and invaluable moral support. They gave me encouragement throughout my education even though they did not not always agree with my decisions.

# Contents

<b>Executive Summary</b>	<b>5</b>
<b>Kurzzusammenfassung</b>	<b>5</b>
<b>Summary</b>	<b>7</b>
<b>Acknowledgements</b>	<b>9</b>
<b>Content</b>	<b>11</b>
<b>1 Introduction</b>	<b>15</b>
1.1 Motivation . . . . .	15
1.2 Computational Challenges in Breathomics . . . . .	17
1.3 Key Objectives . . . . .	20
1.4 First Author Publications . . . . .	23
1.5 Other Publications . . . . .	23
1.6 Overview and Structure . . . . .	24
<b>2 Background and Related Work</b>	<b>27</b>
2.1 Analytical Technologies . . . . .	27
2.1.1 Ion Mobility Spectrometry . . . . .	27
2.1.2 Gas Chromatography / Mass Spectrometry . . . . .	30
2.2 Preprocessing of Breath Data . . . . .	30
2.2.1 RIP-detailing and Baseline Correction . . . . .	32
2.2.2 Denoising and Smoothing . . . . .	32
2.2.3 Peak Detection . . . . .	32
2.2.4 Data Integration . . . . .	34
2.3 Data Management . . . . .	35
2.4 Hypothesis Testing . . . . .	36
2.5 Multivariate Machine Learning . . . . .	39
2.5.1 Exploratory Analysis . . . . .	39
2.5.2 Unsupervised Statistical Learning . . . . .	41
2.5.3 Supervised Learning . . . . .	45
2.6 Longitudinal Analysis and Mixture Models . . . . .	51
2.7 Validation and Permutation . . . . .	52

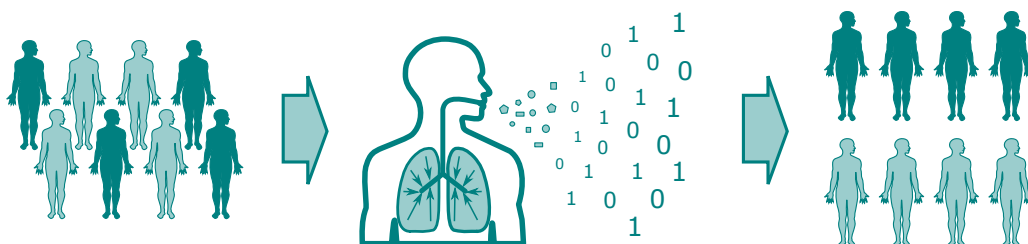
2.7.1	Traditional Conservative Validation . . . . .	53
2.7.2	Cross Validation . . . . .	53
2.7.3	Randomization and Permutation Tests . . . . .	54
2.8	Overview of Related Work in Breathomics . . . . .	54
2.8.1	Previous Breathomics Studies . . . . .	55
2.8.2	State of the Art Software Packages . . . . .	56
<b>3</b>	<b>Data Sets</b>	<b>59</b>
3.1	COPD and Lung Cancer Data . . . . .	59
3.2	GC/MS - MCC/IMS Comparison Data . . . . .	60
3.3	Anonymous Data . . . . .	61
3.4	Artificial Data . . . . .	61
3.5	Longitudinal Rat Breath Data . . . . .	62
3.6	Summary . . . . .	64
<b>4</b>	<b>Clinical Breath Data Management</b>	<b>65</b>
4.1	Requirements and State of the Art . . . . .	66
4.2	Methods . . . . .	68
4.2.1	Structure of MCC/IMS Data . . . . .	68
4.2.2	Structured Data Model . . . . .	69
4.3	Implementation . . . . .	72
4.3.1	Validation and Integration . . . . .	72
4.3.2	Data Retrieval and Reorganization into data sets . . . . .	72
4.4	Results and Discussion . . . . .	73
4.4.1	Database Upload . . . . .	74
4.4.2	Decision Tree Classification and Performance Evaluation . . . . .	74
4.5	Conclusion . . . . .	74
<b>5</b>	<b>Towards Robust Machine Learning in Breathomics</b>	<b>77</b>
5.1	Requirements and State of the Art . . . . .	78
5.2	Methods . . . . .	78
5.2.1	Data . . . . .	78
5.2.2	Overview . . . . .	79
5.2.3	Preprocessing . . . . .	80
5.2.4	Methods . . . . .	80
5.2.5	Statistical Analysis and Validation . . . . .	81
5.3	Results and Discussion . . . . .	82
5.4	Conclusion . . . . .	86
<b>6</b>	<b>Evaluating Automated Peak Detection</b>	<b>89</b>
6.1	Requirements and State of the Art . . . . .	90
6.2	Methods . . . . .	91
6.2.1	Preprocessing . . . . .	92
6.2.2	Peak Detection and Postprocessing . . . . .	92
6.2.3	Evaluation Methods . . . . .	93
6.3	Results and Discussion . . . . .	95



<i>CONTENTS</i>	13
6.3.1 Peak Position Comparison . . . . .	95
6.3.2 Evaluation via Statistical Learning . . . . .	96
6.4 Conclusion . . . . .	98
<b>7 Metabolite Identification with MIMA</b>	<b>101</b>
7.1 Requirements and State of the Art . . . . .	102
7.2 Implementation . . . . .	103
7.2.1 MCC/IMS - MCC/IMS Mapping . . . . .	103
7.2.2 MCC/IMS - GC/MS Mapping . . . . .	103
7.3 Results . . . . .	105
7.4 Discussion and Conclusion . . . . .	105
<b>8 Uncovering Hidden Structures with CAROTTA</b>	<b>109</b>
8.1 Requirements and State of the Art . . . . .	110
8.2 Structure and Implementation . . . . .	111
8.2.1 Visualization of Data and Results . . . . .	113
8.2.2 Modularity and Extendibility . . . . .	115
8.2.3 Import and Export . . . . .	116
8.2.4 Language and Packages . . . . .	116
8.3 Results and Discussion . . . . .	116
8.3.1 Results for Artificial Data Analysis . . . . .	116
8.3.2 Results for COPD Data Analysis . . . . .	117
8.3.3 Comparison to Existing Software . . . . .	120
8.4 Conclusion . . . . .	121
<b>9 Longitudinal Breath Analysis</b>	<b>123</b>
9.1 Requirements and State of the Art . . . . .	124
9.2 Methods . . . . .	125
9.2.1 Preprocessing . . . . .	125
9.2.2 Background Screening . . . . .	126
9.2.3 Model Selection . . . . .	128
9.2.4 Evaluation and Identification . . . . .	130
9.2.5 Implementation . . . . .	130
9.3 Results . . . . .	131
9.3.1 Peak Screening . . . . .	131
9.3.2 Model Selection . . . . .	131
9.3.3 Model Evaluation . . . . .	132
9.4 Discussion . . . . .	132
9.5 Conclusion . . . . .	135
<b>10 Discussion and Conclusion</b>	<b>137</b>
10.1 Data Accumulation and Heterogeneity . . . . .	137
10.2 Robust and Generalizable Statistical Analysis and Evaluation . . . . .	138
10.3 Peak Detection Evaluation . . . . .	140
10.4 Unknown Metabolites . . . . .	140
10.5 Background Noise and Confounding Factors . . . . .	141

10.6 Heterogeneous Diseases and Disease Stages . . . . .	142
10.7 Usability, Maintainability and Re-Usability . . . . .	142
<b>11 Conclusion</b>	<b>145</b>
<b>12 Outlook</b>	<b>147</b>
<b>Bibliography</b>	<b>149</b>
<b>Acronyms</b>	<b>170</b>
<b>List of Figures</b>	<b>173</b>
<b>List of Tables</b>	<b>180</b>
<b>Appendices</b>	<b>183</b>
<b>A IMSDB</b>	<b>185</b>

# Chapter 1



## Introduction

### 1.1 Motivation

Every organism emits a complex array of volatile and nonvolatile compounds which are defined by an individual's genetics, lifestyle and health (233). The sum of volatile organic compounds that are emitted by all living cells and tissues is called the volatolome (105; 199). Already the Greek and Chinese medicine used human odors and vapors for diagnostic purposes, for instance for diagnosing infectious diseases. In more recent history, odors have been used for recognizing gas gangrene on the battlefield and diabetic ketoacidosis in the emergency rooms (32).

However, over the last century, odors and vapors only played a minor role in clinical diagnostics. Contrarily, newly developed high throughput methods for Omics (Proteomics, Metabolomics, Transcriptomics, Genomics) allow analyzing human blood, urine, other body fluids or tissue samples and represent nowadays one of the most prominent approaches to detect diseases (27; 28). These innovations developed quickly and led to comprehensive analyses of diseases not only on the cellular but even molecular level. Nevertheless, many obstacles remain. Most of these techniques are either invasive and therefore inconvenient for the patients, or show low sensitivity or specificity. In addition, they are often expensive, difficult to interpret, or simply too time consuming for certain fast progressing diseases, such as sepsis (84).

Over the last decade the evaluation of odors and vapors in human breath has regained more and more attention, particularly in the diagnostics of pulmonary diseases. Its potential has been validated in various studies, as analyzing the performance of canines or rodents to scent lung or prostate cancer (74; 59). This indicates the existence of informa-

tive volatile excretions recognizable by these animals, implying that they are in principal detectable by modern analytical techniques. Observations like these, initiated the novel field of breathomics, defined as the metabolomics study of human exhaled air, which grows tremendously. One of the major goals in the field is the development of technologies to non-invasively “sniff biomarker molecules that are predictive for the biomedical fate of individual patients. The integration of breathomics approaches into the so-called personalized or precision medicine offer great hope to extend the therapeutic windows to earlier stages of disease progression and hopefully improve chances of survival and recovery.

Various analytical high-throughput and high-resolution technologies are now available, that are capable of overcoming some of the obstacles occurring in the analysis of exhaled air, as humidity or the variability of molecule density and composition. These advances of modern analytical techniques, especially in chromatography and spectrometry, now facilitates the replacement of human and animal senses with refined chemical measurements (75). Currently, the major spectrometric techniques employed in breathomics are gas chromatography/mass spectrometry (GC/MS) (141; 117; 158), sensor technology, such as the electronic nose (55; 68), proton transfer reaction mass spectrometry (PTR/MS) (25; 109) and multi capillary column coupled with an ion mobility spectrometer (231; 103; 24; 230; 216). In principal, all these approaches allow for many possible applications, especially in medicine (187; 20; 83) and biomedicine (147). They have been shown to identify medically relevant patterns in the spectrum of exhaled substances associated with certain diseases or disease progression stages. Consequently, they have the potential to enable early and fast diagnosis as well as therapy optimization.

MCC/IMS and GC/MS are two of the most widely used technologies for breathomics. Currently, GC/MS is the most popular analytical technique in breathomics, mainly owed to the availability of exhaustive mass profile databases that can be used for identification of the detected compounds. Due to its novelty, comparable databases for MCC/IMS are still in their infancies. However, there are a number of disadvantages of GC/MS in comparison to MCC/IMS, such as longer sample processing times (1–3 hours), equipment is more expensive, and requires expert knowledge to operate. In contrast to this, the MCC/IMS has many advantages, such as the short sampling time (about 10 s) and sample processing (about 5–10 min), as well as a robust and easy handling in every day practice. These characteristics make it well suited for large-scale screening studies and disease or drug monitoring (20; 83). Therefore, the majority of this work focuses on developing bioinformatics methods for the MCC/IMS technology.

While the analytical issues in breathomics are mostly solved, one faces the traditional biomarker research barrier: A lack of robustness and automatization in the sub-sequent computational analysis hinders the translation to the world outside of scientific laboratories. The field is lacking the computational infrastructure and customized systems for breath analysis. The required components comprise automated software for storage and processing, as well as the establishment of state of the art statistical analysis and evaluation for breathomics. The aim of this thesis is to develop bioinformatics methods and tools for computational breath analysis, with the goal of moving from biomarker discovery to validation, from separability to predictability, and from manual to automated analysis. This combination of analytical technologies and computational approaches has the potential to strongly contribute to the development of non-invasive biomedical decision

making.

The long term vision is to pave the way for novel, breathomics based diagnostic tools to complement the set of techniques for medical decision making which could revolutionize every days' clinical diagnostic procedures.

## 1.2 Computational Challenges in Breathomics

In order to achieve the transition from pilot studies to screening studies and, subsequently, clinical practice, a number of challenges have to be addressed.

### 1. Data Accumulation and Heterogeneity

Similarly to other omics technologies like genomics and transcriptomics, data quantity and complexity increases tremendously as experimentalists move from small pilot studies to large clinical trials (170). Additionally, data management systems need to adjust for the development of novel analytical instruments with increased quality and/or resolution, furthering the data growth even more. For instance, the change from PTR/MS to proton transfer reaction time of flight mass spectrometry (PTR/TOF/MS) led to a 10-fold increase in the number of detected compounds (ca. 200 to 2000) (144). This issue arises in particular in breath analysis applying MCC/IMS, which is fast and easy to use in clinical practice, leading to a rapid accumulation of huge amounts of data. Additionally, the data needs permanent storage for future reanalysis with improved statistical methods. Therefore, the automatization of storage and analysis is a key challenge in the field.

In addition, the accumulated clinical data is becoming increasingly complex and heterogeneous. Especially, the investigations of numerous diseases lead to a diverse set of clinical parameters. Furthermore, we need to deal with an increased number of studies dealing with time series data, and with the integration of various analytical techniques, such as GC/MS or PTR/MS. This fusion of information from different sources further increases complexity of the data. Therefore, an efficient, flexible, and robust data management system is needed as a basis to address the described challenge.

### 2. Robust and Generalizable Statistics and Biomarkers

One of the major goals in breathomics is to identify the smallest subset of exhaled compounds that can serve as potential biomarkers. Moreover, there is a necessity for statistical methodologies to model breath data in order to find these biomarkers and predict clinical outcomes, in diagnostics, and for the management of disease and drug response. These computational models need to provide high quality (e.g. high sensitivity as well as specificity), robustness against noise, and generalizability from the training set to novel samples. The field of supervised learning offers a variety of methods that serve this purpose. Previous studies focused on exploratory data analysis approaches, such as hypothesis tests, principal component analysis or linear or one dimensional relations evaluation to the clinical outcome.

These relatively simple inference statistics can distinguish between two groups with respect to a present data set in many cases. However, without proper evaluation, this does not allow for conclusions about the predictive power of a statistical model,

and therefore the predictability of clinical outcomes and diagnosis in real application might still be low. This holds true especially in pilot studies, when the number of study participants ( $N$ ) is small (e.g.,  $N < 50$  as in (31)) and the number of compounds ( $p$ ) is much larger than  $N$  ( $N \ll p$  as in (221; 85; 231; 103)). This constellation leads to immense problems of robustness and generalizability due to overfitting of the model, often referred to as the "curse of dimensionality". This can result in both, an overestimation of the prediction performance and false positives among the informative features, sometimes called "Voodoo Biomarkers" (156). Especially, in circumstances where prediction models are trained and tested on the same data set, the overfitted prediction models are not generalizable to the target population. Comparative investigations showed, that the combination of small sample sizes and a lack of reliability in evaluation can lead to a lack of concordance between studies, especially in terms of robust biomarker selection(173; 156).

A large body of bioinformatics and machine learning approaches solving the described problems exists, but have not been adapted for breathomics data. Therefore, more sophisticated and nonlinear statistical learning methods combined with advanced evaluation techniques need to be adapted accordingly. The main objective here is a more robust estimation of the accuracy as well as a robust feature set selection.

### 3. Manual Peak Finding

An important step in automatic high-throughput multi capillary column coupled with an ion mobility spectrometer (MCC/IMS) data analysis is the preprocessing and in particular the peak finding. Each so-called peak represents a specific molecule in the sampled air (see Background Section 2.2.3). Such an automatic peak detection method should be at least as accurate as a human annotation, but fast enough to cope with thousands of measurements. Several peak detection algorithms for MCC/IMS data have been proposed and described in the literature (180; 37; 49; 128). However, they were generally evaluated using intrinsic quality measures, i.e. criteria that can be derived from the data, such as goodness of fit, but have not been evaluated quantitatively nor qualitatively in large scale in a real world breathomics setting. This lack of qualitative and quantitative analysis prohibited a broad application of those methods in the practice of clinical pilot studies. Consequently, a proper evaluation of the existing peak detection algorithms is required in order to identify the most effective approach and generally establish more confidence in this automated methodology.

### 4. Unknown Metabolites

Although the MCC/IMS technology overcomes many practical problems in nowadays chase for biomedical markers, a remaining problem is to identify the molecular compounds detected by the device. Knowing the identity allows for further research of the metabolic pathways of the disease and further ensures the direct relation of marker and disease. This can rule out false positive features that correspond to confounding factors and increases the likelihood for an approval as clinical relevant indicator. Other more established analytical technologies, such as GC are combined with a mass spectrometer. This technology has been used for decades

and therefore provides specific fragment profiles enabling the identification of compounds using various commercial databases. The largest and most widely used is the NIST/EPA/NIH mass spectral library, which was developed by the NIST (NIST-library) containing several hundred thousand entries. In the past, most IMS research was funded by the military and focused on detection of drugs and explosives, and thus the actual component coordinates were therefore hardly published. Furthermore, the resolution of single IMS often does not allow for the separation of complex samples, such as human exhaled air (21). In contrast, the MCC/IMS technology has been developed more recently and allows for a better separation of the components. Thus, the MCC/IMS - molecule database consists of only a few hundred entries so far and is therefore unable to identify a large portion of compounds. Proper solutions to cope with this obstacle have to be found.

#### 5. Background Noise and Confounding Factors

Even more than for other human excretions used for biomarker discovery the metabolic patterns of the human exhaled air are influenced by various sources of disturbance originating from genetics, environment, nutrition or variation of the background (36). Especially, environmental influences like toothpaste (185; 186), medication (129) or disease variations (127), might give rise to hidden structures in the data, unrelated to the question under consideration and heavily conceal the important information. Furthermore, the previously described lack of reliable statistical evaluation of the data might lead to false positive biomarker candidates. Therefore, guidelines for thorough study design, a detailed documentation and an diligent processing and identification of confounders related metabolites is needed.

#### 6. Heterogeneous Diseases and Disease Stages

The focus of current breathomics research relies solely on snapshots of certain diseases. However, most diseases consist of various subtypes, depend on environmental influences or strongly evolve during their stages of development, e.g. cancer subtypes (231), subtypes of pulmonary diseases (88), sepsis (176) or lateral sclerosis (239). This may lead to largely neglecting potential biomarkers whose relative change over time is much more informative than a single observation. For instance, a very recent study of Langley *et al.* suggests that this might be the case for sepsis in primates (135). Additionally, in many circumstances, the metabolic background of diseases is unknown and hidden sub classes largely influence the susceptibility to medication. For this reason, advanced approaches uncovering so far unknown sub-structures in the data are crucial and methods accounting and adapting to present sub classes and evolving diseases are required.

#### 7. Usability, Maintainability, and Re-usability

Besides functionality, three of the main requirements for breathomics software tools especially for those developed for non-computer science experts, are usability, maintainability and re-usability. One of the main reasons for the tentative usage of modern bioinformatics methods in breathomics is the fact that most of the various software packages for more advanced analysis require expert knowledge in the area of statistics and often even expertise in programming. Popular examples are graphical tools, like Weka (99) and RapidMiner (157), or statistical learning environments,


like R (115). An application has a high usability if it meets the requirements of providing an intuitive and easy to access graphical user interface for target users and leads to a good overall user experience. Therefore, an adequate graphical interface should provide intuitive menus and wizards, views and perspectives for data visualization and an informative tutorial (204; 228). A system is maintainable, if changes and extensions are integrated relatively easily. Another main factor is the re-usability of software packages and the availability as open source software for future projects and groups, as well as the ability to reuse previous studies, analysis approaches, settings and results (144; 76).

In the following section, I will discuss the objectives and requirements to overcome the previously described challenges.

### 1.3 Key Objectives

The previous section highlighted different challenges in computational analyses in breathomics: (1) the fast data accumulation and heterogeneous clinical data, (2) the lack of an appropriate evaluation of peak detection algorithms, (3) the identification of unknown metabolites, (4) robust statistics and reliable biomarkers, (5) the consideration of background perturbations and confounding factors, (6) the heterogeneity of diseases and (7) the necessity of user-friendly software packages. In this section I will summarize the six key objectives addressed in this thesis to overcome the described challenges and necessary to support more automatization in MCC/IMS data assessment. Figure 1.1 depicts these requirements in a mind map.

**Clinical IMS Database** To address the problem of fast growing amounts of increasingly complex data in clinical studies, there is a necessity for a flexible and comprehensive centralized data repository, which is capable of gathering all kinds of related information. Additionally it requires an intuitive user interface that provides easy and quick access to the platforms' functionality: automated data integration and integrity validation, versioning and roll-back strategies, data retrieval as well as semi-automatic data mining and machine learning capabilities. Chapter 4 will present the IMSDB a database for clinical IMS data aiming at this challenge.

	<b>Summary:</b> <b>Challenge</b> Huge amounts of breathomics data and heterogeneous clinical factors.
	<b>Solution</b> Build a flexible and intuitive database management system, providing fast and easy access to the data.

**Supervised Learning** To demonstrate the benefits of the application of sophisticated statistical learning methods in breathomics a pilot study applying these techniques to a breath data set is required as a proof of concept. Moreover, a robust and generalizable model evaluation utilizing cross-validation ensures reliable performance and



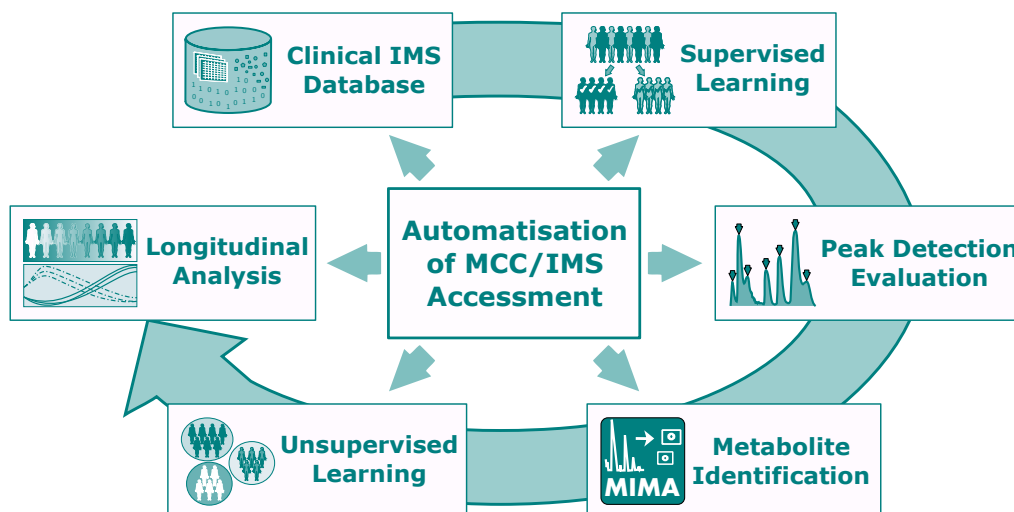


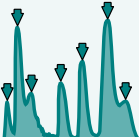


Figure 1.1: Key objectives of this thesis: (1) Clinical IMS Database; (2) Peak Detection Evaluation; (3) Metabolite Identification; (4) Supervised Learning; (5) Unsupervised Learning; (6) Longitudinal Analysis.

feature importance estimations. The study presented in Chapter 5 will show a thorough analysis of different linear and non-linear learning approaches on MCC/IMS breath data.


	<b>Summary:</b>	
	<b>Challenge</b>	Missing reliable statistical analysis and modeling.
	<b>Solution</b>	Application of sophisticated statistical learning methods and robust and generalizable evaluation.

**Peak Detection Evaluation** In order to overcome the common doubts in the quality and reliability of automated peak detection, we require a thorough evaluation of the peak detection methods. Chapter 6 describes a qualitative and quantitative analysis of show the power and accuracy of those methods and compare it to a manually generated gold standard.




**Summary:**  
**Challenge** Lack of trust in automated peak detection methods.  
**Solution** Reliable evaluation and comparison to a manually curated data sets.

**Metabolite Identification** A simple solution to speeding up the time consuming metabolite identification is the automated mapping of the results of parallel GC/MS and MCC/IMS measurements. The MIMA software tool depicted in Chapter 7 solves this issue.



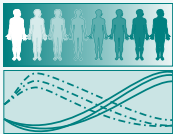
**Summary:**  
**Challenge** Tedious and time consuming manual peak identification.  
**Solution** Automated mapping of GC/MS and MC-C/IMS metabolite lists via MIMA.

**Uncovering Hidden Structures** The goal is to develop a user-friendly software toolbox supporting unsupervised learning methods specifically designed to identify hidden subgroups and characterize human breath profiles. It requires a modular structure to enable extendability on the one hand and nested analysis on the other hand. Particular emphasis lies on the visualization and export. The Carotta tool box is presented in Chapter 8. It addresses the problem of hidden substructures as disease subtypes and confounding metabolite clusters.



**Summary:**  
**Challenge** Unknown confounding factors and hidden structures.  
**Solution** Utilization of widespread unsupervised learning methods on MCC/IMS data and development of a user-friendly software package.

**Longitudinal Analysis** The rapid analysis utilizing MCC/IMS technology nowadays enables online monitoring (129). Therefore, future MCC/IMS studies require mature methods for longitudinal analysis. To establish such a sophisticated pipeline is the major requirement of the final study in this thesis. Chapter 9 aims at the challenge of proper modeling of the metabolite changes as well as identifying those showing the most divergent intensity course between certain groups.

	<b>Summary:</b> <b>Challenge</b> Focus on single time point assessment even in longitudinal data sets.
	<b>Solution</b> Establish advanced pipeline for longitudinal breath analysis.

## 1.4 First Author Publications

Hauschild, A. C.\*, Frisch, T.\*, Baumbach, J. I., and Baumbach, J. (2015). **Carotta: Revealing hidden confounder markers in metabolic breath profiles.** *Metabolites*, 5(2):344363. *\*Shared first author*

Maurer, F.\*, Hauschild, A. C.\*, Eisinger, K., Baumbach, J., Mayor, and Baumbach, J. I. (2014). **MIMA - A software for analyte identification in MC-C/IMS chromatograms by mapping accompanying GC/MS measurements.** *International Journal for Ion Mobility Spectrometry*, 17(2):95101. *\*Shared first author*

Schneider, T.\*, Hauschild, A. C.\*, Baumbach, J. I., and Baumbach, J. (2013). **An integrative clinical database and diagnostics platform for biomarker identification and analysis in ion mobility spectra of human exhaled air.** *Journal of Integrative Bioinformatics*, 10:733755. *\*Shared first author*

Hauschild, A. C., Kopczynski, D., DAddario, M., Baumbach, J. I., Rahmann, S., and Baumbach, J. (2013). **Peak detection method evaluation for ion mobility spectrometry by using machine learning approaches.** *Metabolites*, 3(2):277293

Hauschild, A. C., Schneider, T., Pauling, J., Rupp, K., Jang, M., Baumbach, J. I., and Baumbach, J. (2012). **Computational methods for metabolomic data analysis of ion mobility spectrometry data - reviewing the state of the art.** *Metabolites*, 2(4):733755

Hauschild, A. C., Baumbach, J. I., and Baumbach, J. (2012). **Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification.** *Genet. Molecular Research*, 11(3):27332744

## 1.5 Other Publications

Smolinska, A., Hauschild, A. C., Fijten, R., Dallinga, J., Baumbach, J., and van Schooten, F. (2014). **Current breathomics? A review on data pre-processing techniques and machine learning in metabolomics breath analysis.** *Journal of Breath Research*, 8(2):027105

List, M., Hauschild, A. C., Tan, Q., Kruse, T. A., Mollenhauer, J., Baumbach, J., and Batra, R. (2014). **Classification of breast cancer subtypes by combining**

- gene expression and DNA methylation data.** *Journal of Integrative Bioinformatics*, 11(2):236
- Kreuer, S.** , Hauschild, A. C., Fink, F., Baumbach, J. I., Maddula, S., and Volk, T. (2014). **Two different approaches for pharmacokinetic modeling of exhaled drug concentrations.** *Scientific Reports (Nature Publishing Group)*, 4
- Furtwängler, R.** , Hauschild, A. C., Hübel, J., Rakicioglou, H., Bödeker, B., Maddula, S., Simon, A., and Baumbach, J. I. (2014). **Signals of neutropenia in human breath?.** *International Journal for Ion Mobility Spectrometry*, 17(1):1923
- Fink, T.** , Wolf, A., Maurer, F., Albrecht, F. W., Heim, N., Wolf, B., Hauschild, A. C., Bödeker, B., Baumbach, J. I., Volk, T., Sessler, D. I., and Kreuer, S. (2014). **Volatile organic compounds during inflammation and sepsis in rats: A potential breath test using ion-mobility spectrometry.** *Anesthesiology*
- Eckel, S. P.** , Baumbach, J., and Hauschild, A. C. (2014). **On the importance of statistics in breath analysis - hope or curse?** *Journal of Breath Research*, 8(1):012001
- Barbosa, E.** , Röttger, R., Hauschild, A. C., Azevedo, V., and Baumbach, J. (2014). **On the limits of computational functional genomics for bacterial lifestyle prediction.** *Briefings in Functional Genomics*, 13(5):398408
- van Beek, J. H.** , Hauschild, A. C., Hettling, H., and Binsl, T. W. (2009). **Robust modelling, measurement and analysis of human and animal metabolic systems.** *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 367(1895):19711992

## 1.6 Overview and Structure

**Chapter 1** In this Chapter, the motivation for this thesis is presented and the various challenges are pointed out. It further gives a detailed description of the different objectives and requirements of this project.

**Chapter 2** This Chapter provides the reader with all basic definitions of the state of the art methodology used in different steps of breathomics as well as more sophisticated methods applied and adapted in this thesis. In the remainder, we describe data managements systems, preprocessing, peak detection as well as statistical analysis methods of the past and the future with respect to the usage in the breathomics community. Large sections of this Chapter are inspired by two review papers I contributed, Hauschild *et al.* 2012 and Smolinska *et al.* 2014 as well as other methodological papers (106; 72; 195; 103; 107; 108).

**Chapter 3** In this Chapter, the different data sets utilized in this thesis are introduced (106; 152; 195; 107; 103).

**Chapter 4** The IMSDB is described in this Chapter. It is a newly developed database for metabolomics data based on the MCC/IMS technology combined with sample related data, for instance clinical anamnesis data or other treatment details in animal experiments. This Chapter is written on the basis of the paper Schneider & Hauschild *et al.* in 2013 (195).

**Chapter 5** This Chapter presents a pilot study on robust machine learning techniques used to discriminate patients suffering pulmonary diseases, namely chronic obstructive pulmonary disease, and bronchial carcinoma, from healthy volunteers. We highlight the importance of the robust and generalizable evaluation of the performance as well as the feature selection. The content of this Chapter originates from the paper Hauschild *et al.* 2012 (103).

**Chapter 6** A sophisticated evaluation of the previously developed peak detection approaches is presented. It shows how modern machine learning techniques and robust performance assessment elucidate the strengths and weaknesses in comparison to the common manual peak finding. The Chapter summarizes the study of Hauschild *et al.* 2013 (107).

**Chapter 7** The MIMA software toolkit is introduced in this Chapter. At first, it describes the manual standard procedure, of MCC/IMS peak picking and mapping to GC/MS measurements. Secondly, the functionality and benefits of applying the MIMA software will be shown with the aid of an example data set. The MIMA paper has been published in 2014 (152) and is the basis of this Chapter.

**Chapter 8** This Chapter presents the CAROTTA software for unsupervised learning in breathomics data, published by Hauschild and Frisch *et al.* in 2015 (106). After describing the particular challenges of confounding factors in breath data, the concept of using unsupervised methods to tackle these issues and on a real world data set is given.

**Chapter 9** In this Chapter, the analysis of a longitudinal rat sepsis data utilizing linear mixed effect models is described. At first a mature screening of the non-relevant volatile metabolites is shown, followed by a description of the model selection and the final model evaluation. The manuscript of this study is in preparation.

**Chapter 10** Discussion and conclusion elaborates the powers and limitations of the presented projects and studies. In comparison to the previous Chapters which also contain a small discussion and conclusion, this Chapter focuses on the overall context of the entire thesis and the challenges described previously.

**Chapter 12** In the last Chapter, I will give an outlook on possible future developments and suggest follow-up studies to overcome some of the limitations outlined in the discussion Chapter.



## Chapter 2

# Background and Related Work

### 2.1 Analytical Technologies

Likewise the field of breathomics, analytical technologies for adequate analysis of exhaled air are rapidly evolving. The most widely applied spectrometric methods are gas chromatography/mass spectrometry (141; 117; 158), solid phase micro extraction/gas chromatography coupled with mass spectrometry (SPME-GC/MS) (141; 140; 50), electronic noses (55; 68; 69), proton transfer reaction mass spectrometry (25; 109) and multi capillary column coupled with an ion mobility spectrometer (171; 24; 19; 230; 216; 14). Real time analysis systems like PTR-MS and MCC/IMS provide the huge advantage making the pre-concentration step unnecessary (125). All of these technologies are non-invasive and can for instance provide early and fast diagnosis or therapy monitoring and can be used for the identification of disease-specific biomarkers (108). Due to the advantages described previously, this work will focus on the analysis of GC/MS and MCC/IMS technologies. They will be introduced in more detail in the following.

#### 2.1.1 Ion Mobility Spectrometry

The ion mobility spectrometer (IMS) technology was developed in the early 1970s and originally used for military applications (22; 110) and the detection of drugs or explosives, e.g., at airports. In combination with a multi capillary column (MCC) it is a well-known technology for detecting volatile organic compound (VOC) in air. The main analytical advantages of the MCC/IMS technique are the ability to handle the moisture in exhaled air and the high sensitivity (detection limit at nanograms to picograms per liter) compared to other spectrometric techniques (e.g., GC/MS). The carrier gas guides the analytes into a MCC, where the pre-separation takes place (20). After passing the column containing approximately 1000 parallel capillaries with an inner diameter of 40  $\mu\text{m}$ , they reach the ionization chamber. Here, the analytes become chemically ionized by collisions with the reactant ions, which are carrier gas molecules previously ionized by a radioactive ionization source (usually  $^{63}\text{Ni}$ ). An ion shutter opens cyclically, and the resulting ions enter the drift region. Similarly, to an *TOF*-MS, described later, the ions gain energy from an external electric field and are guided towards a Faraday-plate. Meanwhile the

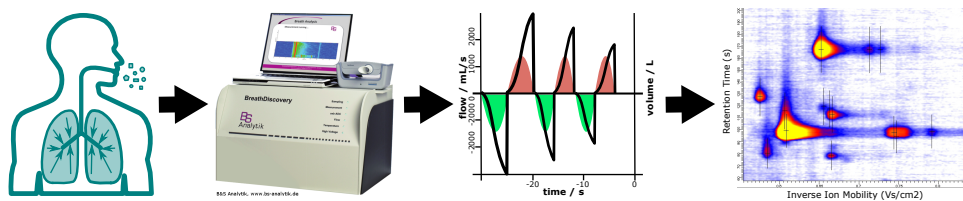


Figure 2.1: Sample and data flow: BioScout device; corresponding  $CO_2$  profile; resulting MCC/IMS chromatogram. The  $CO_2$  profile allows for precise analysis of air originating from certain parts of the respiratory system. For instance, solely end tidal air. Pictures of the BioScout device taken from the B&S Analytik website, <http://www.bs-analytik.de/>.

so-called drift gas will flow in the opposite direction and prevent neutral molecules from entering the drift region. During their flight, the ionized molecules collide with the neutral drift gas molecules and are thereby separated by mass, shape and polarity. The signal recorded by the Faraday plate is called ion mobility spectrum. The MCC/IMS technique has a number of advantages in performance, costs and applicability compared to the GC/MS. It has a low detection limit, measurements are inexpensive (<5 EUR) and fast (<5 min), and it can handle the moisture that comes with exhaled air, which makes it suitable for many medical (20; 187) and biomedical (147) applications as well as for process analysis (19). The MCC/IMS BioScout device (188; 21) built by B&S Analytik (Dortmund, Germany)<sup>1</sup> is a commercialized version of the technology and particularly designed for medical applications, see Figure 2.1.

An example of a MCC/IMS chromatogram recorded by the BioScout device is shown in Figure 2.2. The Y axis corresponds to the retention time of the MCC and the X axis to the reduced inverse ion mobility  $1/K_0$ . To describe the drift time independently of technical properties like drift tube length, drift gas flow or electric field strength, a normalized unit, called inverse reduced mobility ( $Vs/cm^2$ ), proportional to the drift time is used(209). The signal height corresponds to the signal strength detected by the Faraday plate in the IMS device (199).

**Data Format** The raw file yields 12,500 data points per single IMS spectrum at highest resolution, which equals to a sample rate of 250 kHz. Let  $T$  be the set of possible x-axis values (inverse reduced mobility or “drift times”) and  $R$  be the set of possible y-axis values (“retention times”).

Omitting meta-information, e.g. device adjustment parameters, we obtain an  $|R| \times |T|$  matrix  $S = (S(r, t))_{r \in R, t \in T}$ , which can be visualized as a heat map (Figure 2.2). A single IMS spectrum corresponds to a row of the matrix, while a column of the matrix is called contour line. The entire matrix is called IMS chromatogram (107).

A broader area of application led to new requirements in terms of tracking experimental conditions which may influence not only the measurement itself, but also its interpreta-

<sup>1</sup><http://www.bs-analytik.de/>



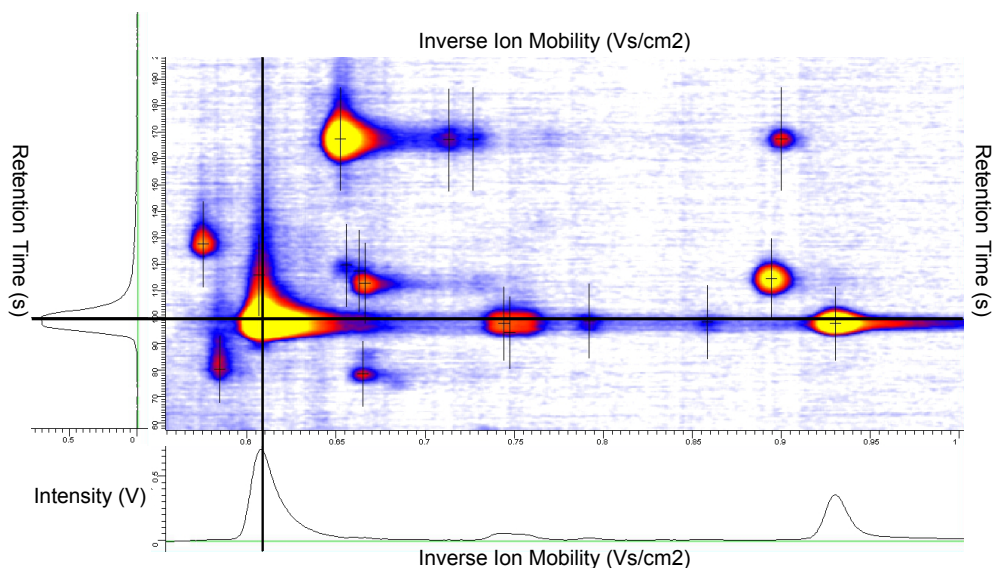


Figure 2.2: An example of MCC/IMS chromatogram. The X-axis corresponds to the reduced inverse ion mobility  $1/K_0$  (Vs/cm<sup>2</sup>) and it is proportional to the drift time (IMS), while the Y-axis corresponds to the retention time  $r$  (MCC) which is proportional to the substance's affinity for the stationary phase. The colors reflect the signal height: the yellow color for the highest signal and the white color the lowest [white < blue < purple < red < yellow].

tion and give vital information for further data processing, such as the combined analysis of data from different studies. Thus, a sophisticated uniform data format required the storage of not only the data itself but also experimental and technical conditions. The corresponding standard file format consists of a header and the data matrix. The header comprises all sampling conditions such as general information, sample information, IMS (device) information, external sampling control, and statistics. They also provide standard nomenclature rules and an extension that is dedicated to sensor-controlled sampling. For details, please refer to supplementary material of Vautz *et al.* (214). An further extension of this file format allows cross-linking MCC/IMS data to GC/MS data suggested by Maddula *et al.* (149).

**Applications in Breath Analysis** For decades, the application of ion mobility spectrometry was primarily focused on explosive and drug detection for the military and safety industry (75; 11). However, recent developments enable an expansion to other fields, such as forensics (119), food technology (120; 213; 217) and medicine (97; 172; 216; 230). The previously mentioned combination to chromatographic technologies for pre-separation further increased in specificity of the device (214; 36; 20). Thereby, the MCC/IMS has developed into an interesting non-invasive device for breath analysis in clinical diagnostic

since it is portable, sensitive, fast and inexpensive (117; 171; 16). See Section 2.8 for a comprehensive list of studies.

### 2.1.2 Gas Chromatography / Mass Spectrometry

The gas chromatography / mass spectrometry is one of the long-established analytical technologies in metabolomics, and besides MCC/IMS most commonly applied to measure volatile organic compounds in complex mixtures such as exhaled air (225; 158; 77). The system is comprised of two components to first separate and subsequently identify the different compounds in the sample, respectively. At first, the gas chromatographic column separates the composition of molecules based on interactions with the mobile phase and the stationary phase defined by the coating of the capillary column. These two phases define the nature of the separation, for instance most commonly by molecular mass and polarity of the compounds. Adjacently, the molecules enter the second component that consists of an ionization source and a mass spectrometer (MS). Various methods are available to ionize the molecules, such as electron ionization (EI) or chemical ionization (CI) (64). During the most widely applied EI, the molecules are bombarded with free electrons, causing a characteristic fragmentation pattern. Finally, the molecules enter the so called time-of-flight (ToF)-MS which measures the time necessary to travel through the electric field from the ionization source to a detector plate. The electric field transfers the same kinetic energy ( $E = \frac{1}{2}mv^2$ ) to all ions. Based on their velocity which is related to their mass, the ions with different mass-to-charge ratio ( $m/z$  values) are separated into groups or packets (199). According to several studies, the GC-ToF-MS is sensitive, robust and highly accurate method to analyze exhaled air samples (50; 77). It produces a specific mass fingerprint for each ionized compound leading to a high level of reproducibility (50; 225). These can be matched to spectral libraries of manufacturer-supplied software or huge libraries like the NIST-library<sup>2</sup> containing more than 200,000 compounds. This leads to a robust identification of compounds. Nevertheless, in contrast to the MCC/IMS the analysis is time consuming and online measurements are unfeasible, which in certain clinical application is mandatory.

## 2.2 Preprocessing of Breath Data

Both analytical techniques described before produce specific types of raw data. A proper preprocessing of this data is crucial to obtain a reliable data matrix in order apply the actual statistical analysis. Unreliable data will jeopardize the robustness and trustworthiness of the results of the multivariate analysis (garbage in >> garbage out). Currently many commercial and free tools and methods exist for preprocessing of MCC/IMS (35; 12; 49; 219; 128; 152) as well as GC/MS data (174; 143; 112; 111; ChromaTOF). Data preprocessing includes on various sub-steps, such as denoising (smoothing), baseline correction, alignment across all samples, peak picking and combining of the peaks originating from the same compound occurring in multiple samples. Finally, the peak areas matching different compounds are collected and converted into a data matrix for statistical analysis. In the following I will focus on preprocessing of MCC/IMS data.

<sup>2</sup>see <http://webbook.nist.gov/chemistry/>

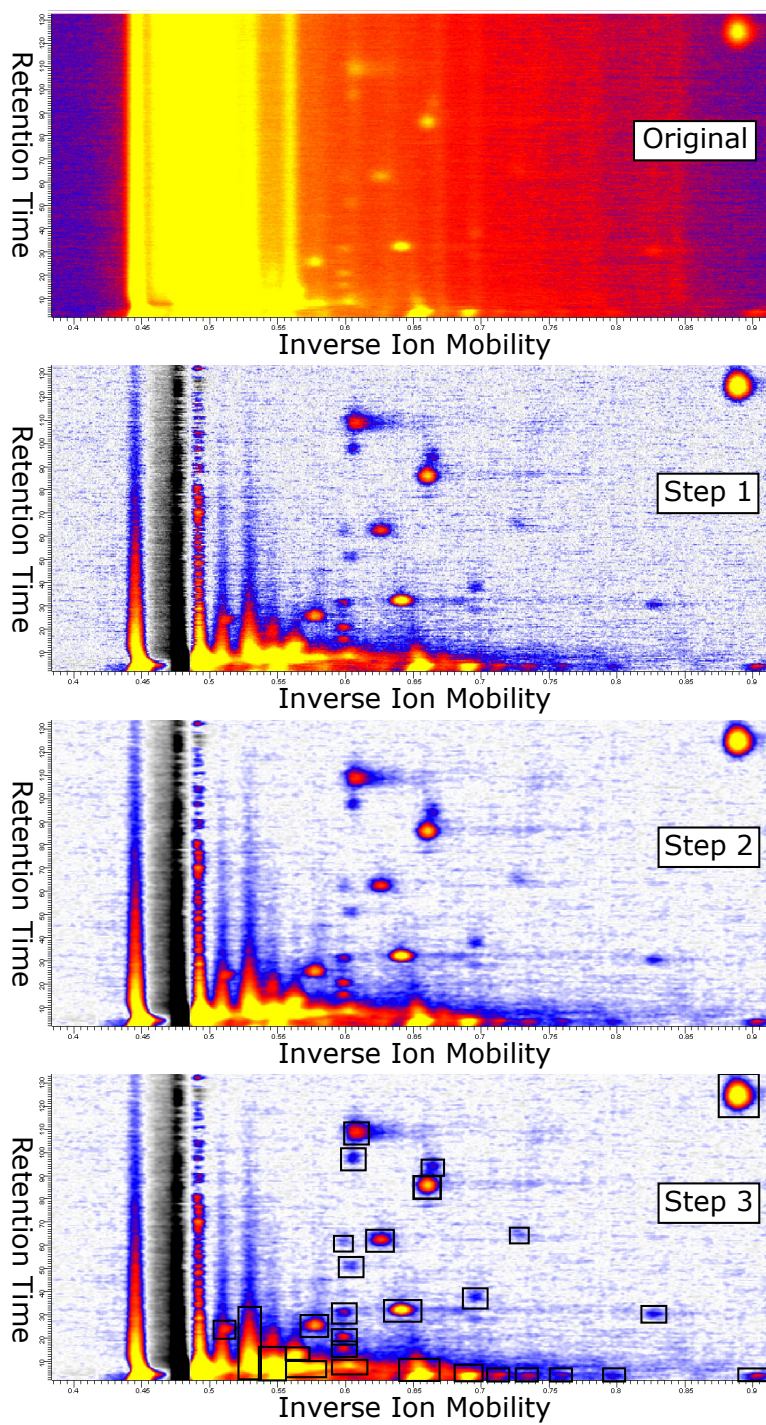


Figure 2.3: Example of a processing strategy of MCC/IMS data involving RIP-detailing (Step 1) and denoising and baseline correction (Step 2), peak picking (Step 3).

### 2.2.1 RIP-detailing and Baseline Correction

Every MCC/IMS chromatogram contains a characteristic structure named the RIP (199). This structure can be considered as a source of disturbance in a measurement, and it appears in the shape of a broad vertical line in each chromatogram at an inverse mobility of  $1/K_0=0.46$  Vs/cm<sup>2</sup>. The decreasing signal on the right side of the RIP is called RIP tailing. To remove the influence of this disturbance, RIP-detailing or RIP compensation approaches are applied as the first step of MCC/IMS data preprocessing (see Figure 2.2, Step 1). Several methods have been proposed for this purpose. For instance, fitting a log-normal function to the mean of all spectra and subtracted this function from each spectrum in the chromatogram (12). Another approaches achieved RIP de-tailing by subtracting the 25% quantile or 50% quantile (median) intensity determined for each  $1/K_0$  value from each IMS spectrum in the data matrix (37; 49). The consecutive step aims to improve the comparability of IMS chromatograms by correcting baseline. This is done by subtracting the mean intensity of pure noise region from all measured spectra (12). The noise region is chosen to be in the right or left upper corner of an MCC/IMS chromatogram.

### 2.2.2 Denoising and Smoothing

Besides RIP-detailing and baseline correction, the improve of the signal to noise ratio is a very important step of preprocessing. This is done by denoising and smoothing (Step 2 in Figure 2.2). One of the first approaches performed a multi-resolution analysis, which includes discrete wavelet transformations on different levels of resolution for both denoising and smoothing (12). Subsequently, the original chromatogram is reconstructed using the corresponding wavelet coefficients, which can be altered using hard and soft thresholding. To smooth the data, the coefficients corresponding to the high frequency regions are eliminated independently of the coefficient amplitude. Removing the low amplitude coefficients regardless of frequency, results in denoising. The resulting IMS-chromatogram is reduced by at least 75% of the original data with negligible loss of information. A more recent approach proposed a pipeline of filters: At first, the median filter is applied for denoising, followed by a Savitzky-Golay filter (49). The Savitzky-Golay filter smooths the spectrum by computing a weighted average across a defined window of data points (192). The final step smooths the data by using a two-dimensional Gaussian blur (163).

### 2.2.3 Peak Detection

The peak detection is the last and third step of data pre-processing procedure. Several automated strategies are available such as merged peak cluster localization (MPCL) (13), growing interval merging (GIM) (12), and wavelet-based multiscale peak detection (WBMPD) (12), water shed transformation (WST) (49) and peak model estimation (PME) (128). After introducing the a general definition of a MCC/IMS peak, the most widely used approaches for MCC/IMS peak detection and their key ideas are depicted.

### MCC/IMS Peak Definition

In each measurement, several regions with high signal values stand out; these regions called peaks are defined by the position and the intensity of the local maxima in this region. These parameters give information about a particular VOC and its concentration. In the literature we may find several mathematical formalizations of what a peak is, e.g., a parametric model describing the shape with statistical functions (220; 128). In this paper, we simply define a peak within one MCC/IMS measurement with three parameters: retention time  $r$ , inverse reduced mobility  $t$  and signal intensity  $s$ . In other words, a peak  $P$  is a triple  $P = (r, t, s)$ .

### Manual Peak Detection in VisualNow

The easiest and most intuitive way of peak detection is manual evaluation of a visualization of the measurement. The human eye and visual cortex is optimized for pattern recognition in 3D. Therefore one can immediately spot most of the peaks in the measurement (see Figure 2.2 Step 3). The VisualNow software allows to visualize the measurement and to pick regions that define a compound. While analyzing a whole set of measurements, this procedure has to be done for each of them. There are several drawbacks of this procedure. On the one hand it is time consuming and therefore inappropriate in a high-throughput context, on the other hand the results depend on a subjective assessment, and are therefore hardly reproducible. Nevertheless, it is still the state of the art for the evaluation of smaller MCC/IMS data sets. Thus, the manual created peak lists still remains the “gold standard” and is used in the majority of MCC/IMS studies.

### Automated Local Maxima Search

This approach identifies local maxima, which is the simplest computer-aided way to identify peaks. Most of the more complex methods, such as the below described, use this as initial procedure to find “seeds” for their algorithms. A point  $(r, t)$  is a local maximum if all 8 neighbors in the matrix have a lower intensity than the intensity at  $(r, t)$ . In addition, we call the neighborhood of a point  $(r, t)$  “significant” if its own intensity, that of its 8 neighbors, and that of  $A$  additional adjacent points, lie above a given threshold  $I$ . Here, we use intensity threshold  $I = 10$  and  $A = 2$  additional points. The local maxima search (LMS) reports all such local maxima that satisfy the “significant” neighborhood condition for these values.

### Merged Peak Cluster Localization

The merged peak cluster localization algorithm for MCC/IMS was first introduced by Bader *et al.* (13). The MPCL consists of two phases: (1) clustering and (2) merging. In the first phase, each data point in the chromatogram is assigned to one of the two classes, either peak or non-peak. A clustering method similar to the traditional k-means is used, based on the Euclidean distance metric of the intensity values. In the second phase, neighboring data points that belong to the peak-label and therefore the same peak are merged together. The simplicity of the method causes a few limitations such as the distinction of two neighboring peaks, where the signal intensity of the overlap is above the

peak-to-noise-threshold. Finally, each peak of the analyzed measurement is characterized by the centroid point, i.e. that data point, which has the smallest mean distance to all other points in the peak region (35). This approach is implemented in the commercial software package VisualNow (B&S Analytik, Dortmund, Germany), for instance.

### Water Shed Transformation

The WST approach for MCC/IMS peak detection mainly builds on the watershed algorithm (219). The IMS chromatogram is treated like a landscape including hills and valleys. The algorithm starts with a water level above the highest intensity followed by a continuous lowering of the level while uncovering more and more of the local maxima. In each step, the new uncovered data points are annotated by the label of adjacent labeled neighbors. Those data points that remain unlabeled are identified as a new peak and receive a new label. The highest data point among a set of new labeled positions defines the peak coordinate. Finally, the algorithm stops if all data points are labeled or the level drops below a defined threshold. An implementation of the of the WST for MCC/IMS data is provided by the IPHEX tool, see Section 2.8.2.

### Peak Model Estimation

At first, the PME method was designed to describe peaks with statistical mixture models of parametric distributions. Later it proved to be a new opportunity for high quality peak detection. The expectation maximization (EM) algorithm is the core of the algorithm and is used to optimize the parameters of a mixture model from a given set of starting values. The algorithm requires a given set of “seed” coordinates for each peak to be modeled. In general, any of the previously described peak detection methods is suitable to provide these initial “seeds”. However, this implies that the quality of the results strongly depend on the chosen seed-finding approach.

Utilizing the EM algorithm, each peak is described by a model function consisting of two shifted inverse Gaussian distribution and an additional peak volume parameter. Finally, the set of model functions plus a noise component describe the whole MCC/IMS measurement. According to the authors, this approach uses continuous functions providing real valued peak positions, which are more precise than the discrete results of the previous methods, but also introduces an additional level of variance and noise.

The PeaX software published by Kopczinski *et al.* provides the only implementation of this approach so far (128). It includes a seed-finding method described by Fong *et al.* (86), which is based on finding roots in the first derivatives of both spectra and chromatograms. For more details on PeaX tool, see Section 2.8.2.

## 2.2.4 Data Integration

Each peak detector produces a list of peaks for each measurement. However, the aim is to create a matrix of peaks and measurements that provides the intensity of the corresponding peak which can then be interpreted as a list of feature vectors utilized for

subsequent statistical analysis (following Sections). Therefore, two circumstances need to be corrected for:

1. False separate peaks in a single measurement that are too close to each other to be two different peaks in reality and should be merged to a single one instead;
2. A set of peaks that occur in different measurements at “slightly” different positions, which we will refer to as “peak clusters” in the remainder of this paper.

The peak merging method described by Boedecker *et al.* (37) can be utilized to resolved those issues. Initially, all peaks are sorted by descending intensity. Potential peaks  $P$  and  $Q$  with  $s(P) > s(Q)$  are merged (i.e. labelled  $P$ ) if the following conditions are satisfied:  $|t(P) - t(Q)| < 0.003$  and  $|r(P) - r(Q)| < 3.0 + r(P) \cdot 0.1$ . Note that this procedure only needs to be applied to the automated peak detection results. The VisualNow software package automatically applies the described algorithm directly on the resulting peak lists of a set of measurements.

Similarly to GC-MS, the final product is a data matrix containing observations in rows. The values in columns correspond to the relative amount of VOCs, i.e. their intensities. The label of each column (i.e. VOC1, VOC2 etc.) represents a specific MCC/IMS coordinate corresponding to a previously described peak cluster. Such general representation of the data allows one to apply various machine learning techniques.

## 2.3 Data Management

In 2007, Lesniak developed the first schema to organize IMS data and a database prototype (138). It categorizes the information into the following three different groups, “Patient”, “Biological Elements” and “Measurement” that are converted to an Oracle database schema. However, this preliminary study can only serve as a proof of concept, since it solely provided a manual insertion of data via the Oracle database interface. Moreover, the schema is fairly inflexible and not expandable as it is not able to store arbitrary entities, attributes and values, as well as relations between entities, which is beneficial to make the database adaptable to any kind of up-coming medical annotation data (108).

A more recent study presented the Advanced Breath Analysis platform (ABA-Cloud), also called ABA-Cloud (76). Its purpose is to document and store breath analysis studies to enable full automatic reproducibility. The studies are stored as Problem Solving Environments (PSE) encapsulating algorithms and data using learning platforms such as R (115), Octave (Eaton and Others) and Matlab (2). Each study is annotated and searched by predefined attributes or keywords. Working with uploaded data requires knowledge about the structure of the data and expert knowledge of the specific learning platform.

In parallel the Automatic Analysis Framework (AAF) was developed. It stores breathomics data in cloud based Taverna activities and enables the user to reanalyze the data automatically. The cloud based Taverna activities are provided for classification and defines a easily extendable Taverna workflow for the automatic analysis. Moreover the

usage of a modular evaluation procedure (k-fold cross validation) is provided to detect overfitting (144). The usage of PSEs and Taverna, enables both frameworks to store projects of arbitrary structure and content. However, this freedom leads to difficulties when automated analysis of combinations of different projects in terms of data, and or analysis and evaluation approaches and strategies are required.

Beyond the field of breath research, there are two areas overlapping with the proposed database framework.

**Metabolomics Databases** The first field covers the advanced processing and analysis of metabolomics and proteomics data, including software tools as MeltDB (160; 124) or the free software library OpenMS (204). MeltDB 2.0 is a web-based software platform responsible for storage, sharing, standardization, integration and analysis of metabolomics experiments (124). OpenMS is a software pipeline for automated analysis for mass spectrometric data. It is an open project and offers many additional analysis packages for special fields, such as proteomics or metabolomics (204).

**Clinical Databases** The second field deals with the problem of adequate and flexible data storage. Some of these solutions are commercial, for instance Oracle Clinical (Oracle Corporation, Redwood City, USA) and others are non-commercial like TrialDB (41) or SenseLab (159). These projects focus on the development of flexible clinical data repositories. While TrialDB focuses on clinical trials and outcome data, SenseLab provides various modules for multiple different areas such as Brain, Neuron or others.

## 2.4 Hypothesis Testing

Hypothesis testing is a technique of inferential statistics. Hypotheses are assumptions that can be tested on the bases of a set of measurable random variables. A statistical tests validates the so called null-hypothesis and controls the error (frequently called  $\alpha$ -error) of incorrectly rejecting it. Depending on the applied test, the so called p-value is calculated and compared to the  $\alpha$ -error. The null-hypothesis is rejected, if the p-value is smaller than the  $\alpha$ -error and proves the alternative hypothesis. However, accepting the null-hypothesis (p-value >  $\alpha$ -error) does not disprove the alternative hypothesis. In contrast, the methods of decision theory discussed later treat both (null and alternative) hypothesis and therefore equally (137).

Statistical tests can be categorized in different ways, depending on the type of hypothesis or the type of data, see Figure 2.4. Parametric tests, for instance can be applied to data sets following a known parameterized distribution, most frequently the normal distribution. The knowledge about the structure of the data enables more powerful test on the parameters of the distribution, for instance the mean or median. On the other hand, if the structure or distribution of the data is unknown, non-parametric tests are applied. Nominal random variables follow a non-parameterized nominal distribution and are therefore require a special type of non-parametric tests. Furthermore, the type of test depends on the number of compared groups or classes as well as whether the data is paired



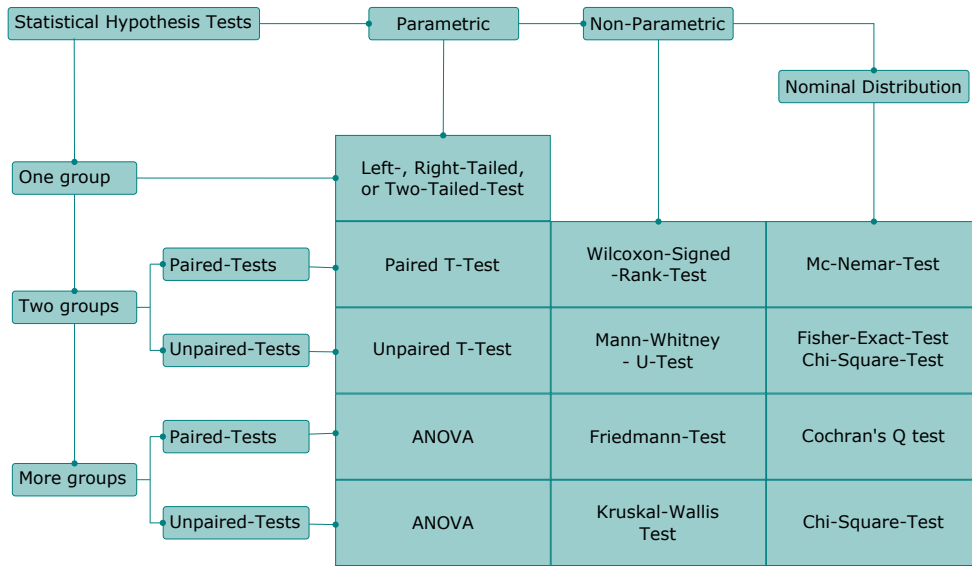


Figure 2.4: The Figure depicts the categorization of statistical hypothesis tests and examples (70). The "Nominal Distribution" column is a sub category of non-parametric tests and characterizes a group of tests suited for binomial or other nominal random variables. Further sub categories exist, like equal variances of the compared data sets, but are beyond the scope of this thesis.

or unpaired. Paired data often results from consecutive measurements of a set of objects, for example patients pre- and post-treatment. Unpaired tests are applied to evaluate the difference between groups of objects, for instance comparing diseases or treatments. On the bases of methodological reasons it is known that the intensities of molecules within the MCC/IMS chromatogram are not following a normal distribution (220). Hence, MCC/IMS breathomics studies focus on the application of non-parametric tests. Most frequently used are the Wilcoxon-signed-rank and Mann-Whitney-U test based on ranking the samples of the two groups. They verify whether the samples are drawn from the same distribution, more precisely, evaluating if one of two samples of independent observations tends to have smaller values than the others. Section 2.8 briefly describes several studies analyzing MCC/IMS data based on the Mann-Whitney U test.

### Misinterpretation of P-Values

In typical biomedical research, however, the sole use of p-values to show the significance of the finding is widespread. The p-value represents the likelihood of getting the observed value of the test statistic by chance, falsely rejecting the null-hypothesis. However, despite the clear definition, the inferential meaning of p-value is very often misinterpreted. For instance, it does not carry information about the magnitude of the differences and a non-significant p-value only indicates that the data is consistent with the null hypothesis. Recall, accepting the null-hypothesis ( $p\text{-value} > \alpha\text{-error}$ ) does not disprove the alternative hypothesis. More details on p-value misconceptions and explanations of the consequence of improper understanding and interpretation, was described by Goodman (92). Additionally, Malley et al. (151) provides information about the restrictions of the p-values usage in biomedical research. It clearly outlines that p-values have limited function as a quality measures. Specifically, the paper underlines the common misinterpretation of the p-value as a probability statement of the hypothesis being true. It does not give relevant information on physiological processes and might be even completely uninformative biologically.

### Multiple Testing Correction

Moreover, correcting over several hundred of tests by e.g. Bonferroni or False Discovery Rate (FDR) assumes that the separate p-values are independent and the relations between one VOC and next are set to zero. However, data produced in breathomics are frequently correlated as (often not considered) cascades of metabolic pathways and biochemical reactions connect to the measured VOCs. The significant p-value does not give the probability that repeating the experiment the same conclusion can be drawn. Consequently, Malley *et al.* suggest utilizing more sophisticated techniques and focusing on predicting the class (i.e. disease) of a given sample (i.e. patient) rather than on something that is merely significant. Obviously this is the case of machine learning techniques.

## 2.5 Multivariate Machine Learning

Breathomics, which typically comprises metabolomics studies of exhaled air, generates huge data sets where the number of compounds exceeds the number of samples. Moreover, co-linearity between measured variables exists. In many cases, breathomics data is, to some extent, sparse, i.e. a compound is present only in e.g. 20%-30% of the samples. These issues have to be taken into account while applying machine learning techniques (45). Multivariate statistical methods offer practical means of maximizing information recovery from complex breathomics data. Machine learning provides a plethora of methods to explore and understand complex data and thus obtain valuable information on biological changes. The analysis might start with data exploration and discovery. This part of the analysis is blind and unsupervised approach and thus gives an unbiased first view on the data, but might neglect useful information. Therefore, a typically approach continues with supervised analysis in which *a priori* knowledge of the data structure is utilized.

### 2.5.1 Exploratory Analysis

The term explorative analysis summarizes methods to evaluate main characteristics of a data set. Promoted by Tukey in the early sixties, it comprises measures like mean, median and quantiles as well as graphical analysis utilizing histograms, box or scatter plots, see Figure 2.5. Further, methods for variance analysis and multidimensional scaling, described in the following, are categorized as exploratory approaches.

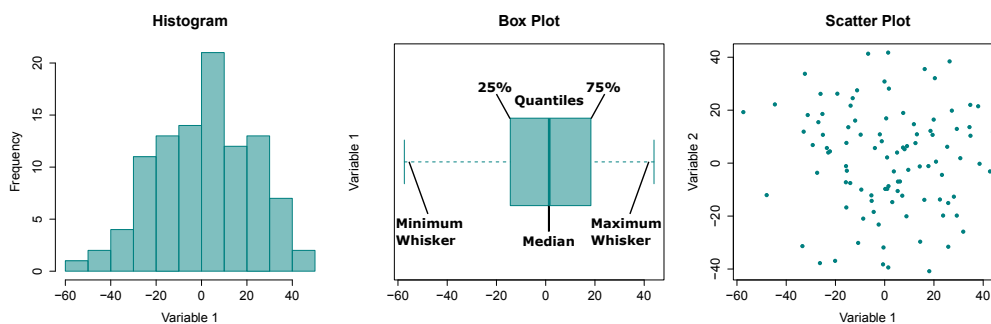


Figure 2.5: Examples for histogram, box and scatter plot. (1) The histogram shows the general distribution of the data and allows discover left or right shifts in the distribution. (2) The box plot depicts the quantiles and Whiskers of the distribution. The inner quantile range (IQR) is defined as the difference between the  $Q_1 = 25\%$  and  $Q_3 = 75\%$  quantiles.  $Q_2$  is equal to the median. The upper and lower Whisker are the minimum and maximum values of the data, except outliers are present. An outlier is a value that is smaller than  $Q_1 - 1.5IQR$  or larger than  $Q_3 + 1.5IQR$ . (3) The scatter plot depicts a relation between two variables.

### Principal Component Analysis

One of the most widely spread exploratory analysis technique is the principal component analysis (PCA). The aim is to select a small number of orthogonal principal components (PCs) that explain the majority of the variation in the data (199). Therefore, the first principal component chosen to be the direction of the largest possible variation in the data. Each following PC is selected as the most varying orthogonal component (102). Finally, the PCA results in two matrices known as scores and loadings, representing new coordinates for the samples and the weights of the linear combination of original variables to calculate the PCs, respectively. A score and a bi-plot plot can visualize the outcome of PCA. While the score plot presents the relations (i.e. similarity) between all samples in breathomics data, the bi-plot, depicts on the correlation between compounds and the difference in relative abundance between clusters of samples. Although the method is well suited to summarize the variation of high-dimensional data, it was not developed to find a pattern of variables that best distinguish classes of objects. In real world data, homogeneous groups of objects are might distributed along the directions of largest variance due to certain biases in the data and hence show in one of the first PCs. However, if the PCA does not show a class separation since the class information is not present in the high variance directions, does not mean that the data does not contain the necessary information (62).

In most breathomics studies, PCA is used for exploring variation and biases in the data. Clearly, PCA is a reasonable choice for such purposes. However, it is known that the outcome of PCA depends on the applied scaling and the potential presence of outliers (63). Moreover, the usage for classification is not suited and mostly leads to miss interpretation of the data. The reduction of multidimensional data within reason is suitable to remove noise variables lacking any variance, however, by removing low variance features, actual important information might be discarded.

### Multi-Dimensional Scaling

The visualization of high dimensional data is a challenging and complex task. Carotta integrates the so-called multi-dimensional scaling (MDS), a standard method for this purpose. It aims to find an embedding from the pairwise representation to a space of lower dimension, such that the distances are preserved (241). Given  $N$  different objects  $z$  in a high dimensional space  $p$ , the objects will be arranged in the low dimensional space  $p$  in such a way that the pairwise distances are most similar to original distances. Therefore, the objective is to minimize the squared distance of all pairwise distances, Equation (2.1) (132).

$$S(z_1, z_2, \dots, z_N) = \sum_{i \neq j} (||x_i - x_j|| - ||z_i - z_j||)^2 \quad (2.1)$$

The resulting 2-dimensional or 3-dimensional coordinates can now be visualized by a scatter plot. In contrast to PCA, MDS aims to preserve the pairwise distances between each of the two coordinates, influenced by all variables equally. Since these distances are the bases for the clustering, the MDA is a more reasonable choice for this purpose.

## 2.5.2 Unsupervised Statistical Learning

Unsupervised methods try to find hidden structures without incorporating external knowledge. Essentially popular unsupervised methods to analyze multivariate data are clustering algorithms (226). They identify groups (clusters) of data objects that are more similar to each other than to objects from other groups (101). These methods are based on similarity measures, for example determining multivariate distances such as Euclidean distance or Pearson correlation, see Section 2.5.2. Besides subgrouping, clustering is often used to reduce the amount of data. For instance, if the clusters are compact it may be sufficient to use only a part of the original data by removing redundancy and choosing representatives for each cluster (199). In other cases, the methods utilize to divide large data sets into reasonable packages of subsets that subsequently can be analyzed separately. The most common examples are hierarchical or k-means clustering, also referred to as prototype based clustering methods. Recently, they found various applications in breathomics studies (65; 203; 80; 52; 243).

More sophisticated clustering approaches are distribution-based (e.g. Gaussian Mixture Models (102)), density-based (130) or graph-based clustering methods (Spectral or Transitivity Clustering (146; 235)). A comprehensive study recently compared the performance of biomedical clustering methods indicated that they might be more suitable for certain data sets (236). Nevertheless, most of them require more in depth knowledge of the principles of the methodologies and the structure of the data, to be able to chose the right approach and correctly set the parameters. Thus, they have not been applied in breathomics before.

The following sections first introduce two common distance/similarity measures commonly used in the context of clustering. Subsequently, it focuses on two common clustering algorithms, namely hierarchical agglomerative clustering and transitivity clustering and discuss good practice and parameter. Finally, two common and adequate measures of clustering quality are described.

### Dissimilarity and Similarity Measures

(106): The pairwise relation of two data points is defined by a similarity or dissimilarity function. For a data set including  $N$  samples, the set of all pairwise relations can be calculated and stored into a  $N \times N$  matrix. It is generally called similarity or distance matrix. Clustering approaches each depend on either the similarity or dissimilarity matrix. However, there is an inverse relationship between similarity and dissimilarity and they can equally be converted into each other. A similarity matrix is converted into a dissimilarity matrix as follows: The entries of the new matrix are defined as  $d(x, y) = \max(|S|) - |s(x, y)|$ , where  $S$  is the matrix containing the original similarity and  $s(x, y)$  corresponds to the similarity of objects  $x$  and  $y$ . The similarity based on the dissimilarity is defined accordingly:  $s(x, y) = \max(|D|) - |d(x, y)|$ . A pairwise relation  $r$  can be symmetrical  $r(x, y) = r(y, x)$  or asymmetrical  $r(x, y) \neq r(y, x)$  which will subsequently result in a symmetrical or asymmetrical relation matrix. In this thesis we will focus on relations that are symmetrical, namely the Pearson and Spearman Correlation as well as the Euclidean distance.

**Pearson Correlation Coefficient** The Pearson correlation coefficient (154) is a measure of linear correlation. It is varying between  $-1$  and  $1$ , where  $-1$  is negative correlation,  $0$  is no correlation and  $1$  is positive correlation.

$$\text{cor}(x, y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2.2)$$

Hence, two features  $\{a, b\}$  with the correlation of  $\text{cor}(a, b) = x$  have an equally strong relation to each other than two features  $\{c, d\}$  with a correlation of  $\text{cor}(c, d) = -x$ . Depending its purpose, a clustering analysis can favor positive, negative or both relations. When used for clustering the general strength of the relation is of main interest and therefore focuses on the absolute value of the correlation.

**Spearman Correlation Coefficient** A non-parametric version of the Pearson product-moment correlation is the Spearman correlation coefficient. The corresponding value estimates how well one variable can be described as a monotonic function of another variable. Similar to the Pearson Correlation it varies between  $-1$  and  $1$ , where  $-1$  is negative correlation,  $0$  is no correlation and  $1$  is positive correlation. It is defined as the Pearson correlation coefficient between the ranks of variables (202). Again, we focus on the absolute value of the correlation.

**Euclidean Distance** The Euclidean distance (154) is the most commonly-used dissimilarity measure. It is defined by the following equation:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2.3)$$

The function is given by the Pythagorean theorem and is always greater than zero, besides the two points being equal.

### Hierarchical Agglomerative Clustering

Hierarchical clustering is one of the most widely-used clustering methods (101), that is based on the distances between the objects. There are two basic approaches: agglomerative and divisive, while this thesis focuses on the first. In contrast to the divisive “top down” approach, the hierarchical agglomerative clustering (HAC) algorithm starts by assigning every object to its own cluster. Subsequently, an iterative process merges the most similar (smallest distance) clusters. The result is a hierarchical structure base on a multivariate dissimilarity which can be represented as a tree, commonly called dendrogram. The choice of agglomeration or linkage method, defines the equations to recalculate the distances between two clusters, each containing a set of objects of different coordinates. Popular examples are the average- or complete-linkage specified as the average or the maximum of all pairwise dissimilarities of all objects between the two clusters, respectively. Various other linkage methods to recalculate the distances exist. The most

common ones and the corresponding equations are described below (102). Given two clusters  $X$  and  $Y$ :

**Average-linkage:** Uses the average distance of all objects between  $X$  and  $Y$ .

$$D = \frac{1}{|X||Y|} \sum_{x \in X, y \in Y} d(x, y)$$

**Complete-linkage:** Determines the maximum of all pairwise distances between the two clusters.

$$D = \max_{x \in X, y \in Y} d(x, y)$$

**Single-linkage:** Selection of the smallest pairwise distances, also known as “nearest neighbor clustering”.

$$D = \min_{x \in X, y \in Y} d(x, y)$$

**Ward-linkage:** Given  $\bar{X}, \bar{Y}$  the means of clusters  $X$  and  $Y$ . This distance minimizes the variance within the clusters. Therefore, it yields in clusters with equal size.

$$D = \frac{d(\bar{X}, \bar{Y})}{\frac{1}{|X|} + \frac{1}{|Y|}}$$

**Centroid-linkage:** This method uses the distance between the centroids of both clusters.

The choice of the linkage method largely influences the result of the clustering. For instance, the single linkage approach is prone to wrongly merging close clusters, a phenomenon called “chaining phenomenon” (116). A more stable result can be achieved using the average or Ward-linkage approaches. However, they are biased to find spherical group arrangements (184; 33). Therefore, a closer look on the resulting dendrogram is indispensable to verify the ultimate result.

### Transitivity Clustering

Transitivity clustering is based on the weighted transitive graph projection problem (234). A given similarity matrix is interpreted as a weighted similarity graph and split into a cost graph by removing edges with weights below a user-given threshold. Such a putatively intransitive cost graph  $G = (E, V)$  will be transformed into a transitive graph  $G'$  by adding and removing a minimal number of edges. In practice, the edge weights are taken into account, yielding a cost function for edge modifications that is to be minimized. In 2010, Wittkop *et al.* published an algorithm that tackles this NP-hard problem by combining exact and heuristic algorithms (235). The threshold influences the number of clusters, as the average similarity of objects within one cluster is above the threshold, while the average similarity of the object from different clusters is below the threshold. Consequently, a high threshold leads to many small clusters, while a low threshold has few, but bigger clusters. The Transitivity Clustering software also provides a hierarchical clustering mode.

### Application and Thresholds

Besides methodological delineation the main difference between the two approaches is the real-world interpretation of the threshold. In hierarchical clustering, it corresponds to the number of clusters. In contrast, in transitivity clustering, it corresponds to the similarity value  $S$ , for which the average similarity of all objects from different clusters is smaller than  $S$  (and the similarities between objects from the same cluster is higher than  $S$ , on average). The selection of the clustering method depends on the purpose of the study and the data sets at hand. Using hierarchical clustering usually appears beneficial if we may assume (or guess) a certain number of clusters. In data sets with few or no outliers, this might become problematic. If prior knowledge on a preferable similarity cutoff is available, transitivity clustering will be more appropriate. It is more robust to outliers, as it is independent of the number of clusters (*i.e.*, outliers would end up as singletons).

However, the main disadvantage of both methods is that it does not deliver information about the compounds which are responsible for the resulting clustering. A solution is for instance provided by methods like co-clustering also referred to as bi-clustering. This approach simultaneously clusters data in its rows and columns (samples and compounds) (44; 150).

### Quality Measures

Finally, clustering quality measures give evidence of how well the groups of objects are separated by the clustering. This can be assessed by either internal or external criteria. Internal criteria like the Silhouette value rely solely on the underlying data and favor clusterings with high inner and low between cluster similarity. In contrast, external indices for instance the F-Measure, compare the clustering result to a user-given gold standard, *i.e.*, the primary outcome variable.

**Silhouette Value** A prominent example for an internal quality measure is the silhouette value (184). It evaluates how well an object fits into the associated cluster depending on the paired dissimilarity to the objects within its cluster in contrast to the objects in all other clusters. It is defined as follows:

$$S(i) = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (2.4)$$

Here,  $a_i$  is defined as the average dissimilarity to all objects in the same cluster, while  $b_i$  is the dissimilarity to the so-called neighbor cluster, which is the cluster of the next lowest average dissimilarity to  $i$ . The average of all object silhouette values is called the overall silhouette value of a clustering. The value varies between one and minus one. If all elements are well clustered, the result will be one.

**F-measure** Let  $K$  be the gold standard defining a known grouping of the objects. The F-measure compares the clustering  $C$  to the gold standard, whereas  $t_{i,j}$  denotes the number of common elements of  $K_i, i \in \{1, \dots, m\}$  and  $C_j, j \in 1, \dots, n$ . It is defined as



follows:

$$F - measure(C, K) = \frac{1}{\sum_{i=1}^m |K_i|} \sum_{i=1}^m \left( |K_i| * \max_{i \leq j \leq n} \frac{2 * t_{i,j}}{|C_j| + |K_i|} \right) \quad (2.5)$$

The final F-measure among all clusters is varying between 0 and 1. While 0 corresponds to a poor overlap with the gold standard, 1 indicates a perfect match (168). This measure gives an impression of the clustering performance with respect to a user-defined gold standard. However, many biomedical data sets do not provide such a standard. In our case, though, we may utilize the outcome variables (disease annotation and/or the confounding factor annotations, respectively).

### 2.5.3 Supervised Learning

In contrast to unsupervised learning methods, supervised approaches aim is to infer a qualitative (classification) or quantitative (regression) outcome variable. Therefore, *a priori* knowledge for instance about diseases, treatments, other groupings within the data or a continuous measures is required.

In general, supervised learning methods aim to find a relation or pattern between a matrix of input variables or features (e.g. MCC/IMS peak intensities) and the outcome variable (e.g. a disease or a treatment group). All of these algorithms follow one or both of the two main goals:

- Optimal prediction of outcome variable
- Identification of important predictors

The goal is to find a pattern or set of rules within the predictor matrix that best predicts the outcome variable, not only on the data used for training, but on a separate data set not used for creating the model, the so called test data, see Section 2.7 for more details. However, often it is just as important to know which minimal set of predictors or also called features is most informative or predictive within the model. In practice it enables reducing the number of features needed and hence time and costs to acquire them. Nowadays, a wide range of linear and non-linear statistical learning approaches are available. In the following, I will introduce the ones utilized in my work and if applicable their capability for feature selection.

#### Naive Bayes

The most widely tested and straightforward linear method for statistical induction is known as the naive Bayesian classifier (134). In this method, each class is represented by a single probabilistic summary. To minimize the probability of error in the classification assignment, the state of action that maximizes the posterior probability is chosen each time. This is calculated by the Bayesian formula simply from the prior probabilities and the conditional densities. Despite its simplicity, the naive Bayes works quite well in practice and outperforms far more sophisticated techniques. The reason for this is that although bias can hurt the individual class density estimates, this might not influence the posterior probabilities as much, especially near the decision (102). The naive Bayes

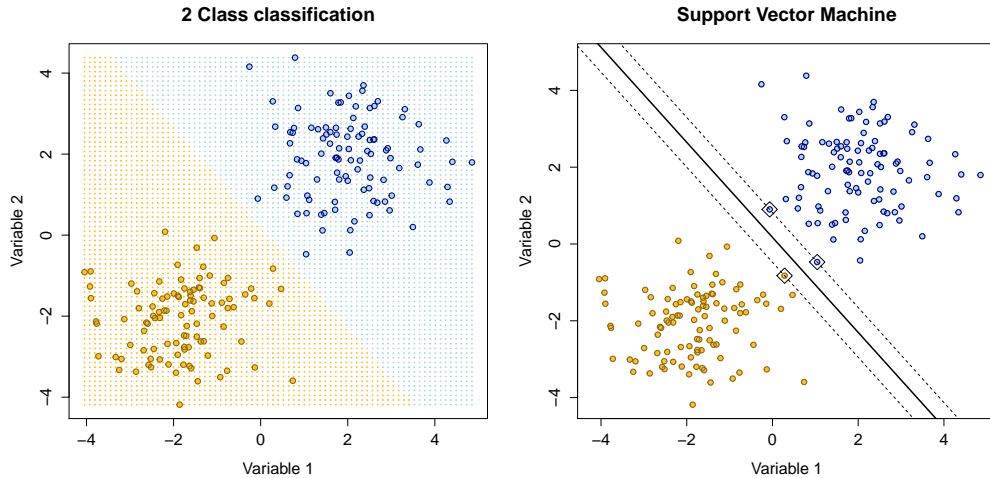


Figure 2.6: Example for a linear classification using support vector machine. The blue and yellow samples are separated by a SVM. The samples marked by a square are the so called support vectors.

method was implemented using the R package *NaiveBayes* (227), using the standard parameters with activated *usekernel* parameter (103).

### Support Vector Machine

The aim of a support vector machine (SVM) is to find a hyper plane that perfectly splits two classes while maximizing the thickness of the margins. This margin corresponds to the distance of the plane to the closest data point from either class. Nevertheless, in most real world examples, classes are overlapping. Therefore, the margin is maximized while penalizing the points that lie on the wrong side of the margin. In contrast to other linear methods, solely the points on the boundary or the wrong side of the margin support the split and, therefore, called "support vectors", see Figure 2.6

Although the SVM is only suitable for two class classification problems, it is a very powerful method. Due to the support vector approach the solution can be sparse since it is built on a limited number of samples, this reduces the variance (bias-variance-trade-off) without increasing the bias, and hence to some extent prevents the model from overfitting.

In practice, most implementation of the SVM provide multi-class classification by splitting the data set into a set of two class problems. The single SVM without the usage of a non-linear kernel belongs to the class of linear methods and therefore, also called "linear SVM". However, in 1992, Boser *et al.* suggested the application of the kernel trick as a solution to create non-linear SVM classifiers, for example by using the Gaussian radial basis function (39). See Section 2.5.3 for more details on kernels and kernel methods.

## Interpretation of Decision Trees

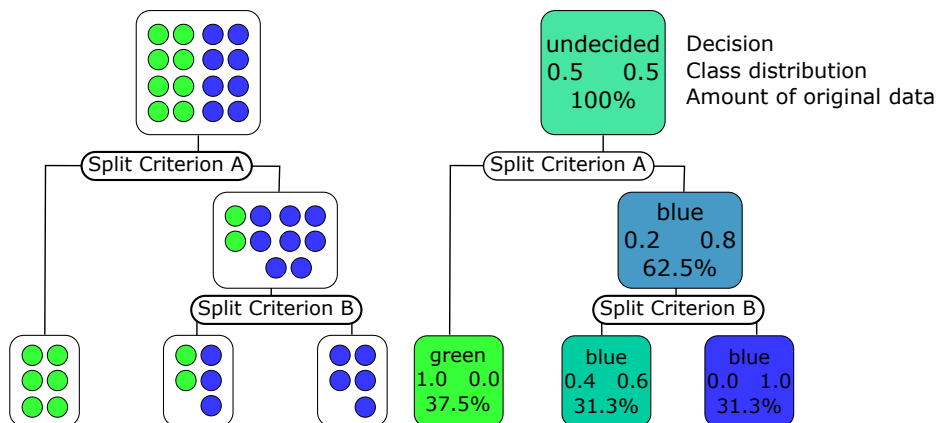


Figure 2.7: The figure on the left side depicts a DT and the separation of the data. The figure to the right shows the corresponding explanation of the key values, class distribution, the corresponding decision, and the percentage of the original data within this node.

### Decision Tree

The most straightforward non-linear methods are based on Decision Trees. An example is the classification and regression tree (CART) method proposed by Breiman (43). The goal of CART is to find exclusive regions in the data that contain homogeneous subsets of the data (i.e. defined classes). The outcome of the CART is presented as a binary tree. The tree is constructed by recursively dividing the samples from a parent node into two child nodes. Each node is described by a simple, logical rule based on one compound, see Figure 2.7. Note that, at each split, a different compound is used; nevertheless the same compound may appear more than once in a single tree. The splitting continues until the similarity of the samples within each node is the highest one or nodes contain the minimum number of samples (this has to be specified by the user). The splitting compounds intuitively serve as the most important set of features.

### Random Forest

The strong tendency of decision tree to overfitting results in a large variance between single trees on slightly different data sets. Therefore, the combination of ensemble trees reduces the overall variance of the models and thereby significantly increases the performance (199). The random forest (RF) method as proposed by Breiman (42) is a very popular example. RF builds a large collection of de-correlated trees each on a randomly selected subset of the original samples (called bootstrap aggregation (bagging)) and on randomly selected subsets of compounds (102). The error rate of the RF model depends on two factors: firstly the correlation between trees and secondly the strength

of the individual trees. Advanced ensemble tree approaches like RF seem simple and easy to understand. However, they are based on solid mathematical theories and hence outperform other methods on many data sets in terms of classification accuracy. An additional advantage is that they provide comprehensible importance estimations for the single features which allow for supervised dimension reduction and the direct abstraction of potential biomarker patterns (13; 231; 103). It is a non-parametric approach and therefore, does not assume any particular distribution in the data. In addition, the RF technique is more resistant to the different types of outliers and mislabeled samples and thus leads to generalizable models. The main disadvantage of RF is computational time on huge data sets.

**Feature Importance:** The random forest method provides two measures of importance, both dependent on the accuracy of the trees. The first, called Gini index, accumulates the improvement in the split-criterion, while growing the trees for each variable and corresponding splits. The second uses the left out samples (called out of the box samples, OOB) of each tree to measure the prediction strength of each variable; it is further referred to as OOB randomization. See Hastie *et al.* (2009) for more details (102).

### Kernel-Based Models

Another very common non-linear category of techniques are the so called kernel-based models. Kernel-based models require transformation of the data via specific functions called kernel. The key idea of kernel transformation is to map the non-linear problem in the original data into a higher-dimensional feature space, corresponding to a reproducing kernel Hilbert space (RKHS), in a manner that the problem becomes linear and thereby easily solvable. The kernel function calculates the inner products of the original compounds in the RKHS without the need of transforming the whole data set, which is called the "kernel trick" or kernel property. For more details, we refer to Hastie *et al.* (102). By mapping the original data of size ( $m \times p$  where  $m$  is the number of samples and  $p$  is the number of compounds) into the feature space, a kernel matrix of size  $m \times m$  is obtained. The kernel matrix has to be positive semi-definite and likely there are many kernel functions which fulfill this constraint. The simplest kernel function is the dot product of the data matrix. This is linear kernel and thus is considered as linear approach, see previous Section on Linear SVM.

A widely used non-linear kernel transformation is the radial basic function. This function has one parameter to tune, i.e. the width of the Gaussian function. The combination of SVM and the radial basis function is often called the radial SVM. In general, the kernel methods have a strong discrimination power, but the results are highly depending on the chosen kernel function. Furthermore, the combination with the support vector approach results in a more robust and less variable model, however, to complex kernel functions might still lead to overfitting of biased data selections (199).

The disadvantage of most kernel-based methods is the lag of information about the compounds importance, due to the kernel transformation. Hence, a direct interpretation of discriminatory compounds remains impossible. Nevertheless, a recent paper by Krooshof *et al.* (131) proposed a solution by applying a procedure based on non-linear bi-plot. Additionally, the general recursive stepwise forward feature selection or backwards feature

elimination approaches can be applied (98). The ranking reflects the importance of the compounds in the classification problem. However, the interpretation of highly non-linear classification problems remains a bottleneck. Section 2.8 will list a number of publications utilizing this techniques in MCC/IMS breath studies (198; 142; 16).

### Artificial Neural Networks

In 2013, Amato *et al.* (9) pictured the growing importance of artificial neural networks (ANN) in medical diagnostics. ANN is a two-stage regression or classification model which is typically represented as a layered network of neurons. The first level of an ANN contains an "input" layer including a node for each input variable. The last level is the the "output" layer, containing  $k$  output units, on for each class in a  $k$ -class classification problem. Between these levels a sets of novel features so called "hidden" layers are derived as linear combinations of the input variables. An activation function, which is usually chosen to be the sigmoid  $1/1 + e^{-v}$  within each hidden node leads to a non-linear model. The final "output" layer combines these features to model the target label. This creates a very powerful and flexible method to model non-linear problems. The interpretation of the effects of a compound in such a model however, is challenging. The most common "vanilla" neural net contains only one hidden layer and is called the single layer perceptron. Variants of this approach have been applied in various areas of breath analysis, for instance for analyzing sensor array results (191), mass spectrometry data (8).

### Performance Measures for Supervised Learning

Finally, all classification and regression results should be assessed by a set of key performance criteria. A performance measure gives evidence of how well the groups are predicted, or a continuous value is fitted by the model. The validation process of regression problems are primarily based on goodness of fit measures, analysis of the randomness of residuals and especially the performance on a separate test set. The most commonly used is the  $R$  squared or coefficient of determination ( $R^2$ ), which is 1 minus the fraction of the residual sum of squares and the total sum of squares.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.6)$$

Residual analysis allows for closer evaluation of the structure of the error. It includes graphical and analysis of the model structure using partial residual plots and assessment of hypothesis on the distribution of the model, auto-correlation and homoscedasticity (169). The classification model is assessed via a set of various different quality measures, starting with the so called confusion matrix, Table 2.1 and 2.2. It contains the counts of all true positives(TP), true negatives(TN), false positives(FP) and false negatives (FN) for a two class classification problem Table 2.1. Respectively, it contains the truly and the various forms of falsely classified samples for a multi-class classification Table 2.2. Table 2.3 lists some of the most important performance measures for classification and their equations for two and three class problems. The accuracy (ACC) is the most common measure simply assessing the percentage of correctly classified samples. However, in various applications, sensitivity (SEN) and specificity (SPE) are far more important and informative. For

Table 2.1: Confusion Matrix for two class classification

		Real Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 2.2: Confusion Matrix for three class classification:  $TP_X$  represents the number of samples correctly predicted as class X,  $F_{X \rightarrow Y}$  samples of label X, wrong classified as class Y, and  $TN_X$  accounts for the true negatives, the sum of samples correctly not predicted as class X, e.g.  $TN_A = TP_B + TP_C + F_{B \rightarrow C} + F_{C \rightarrow B}$ .

		Real Class			
		Class A	Class B	Class C	
Predicted Class	Class A	$TP_A$	$F_{B \rightarrow A}$	$F_{C \rightarrow A}$	$F_A = F_{B \rightarrow A} + F_{C \rightarrow A}$ $P_A = TP_A + F_A$
	Class B	$F_{A \rightarrow B}$	$TP_B$	$F_{C \rightarrow B}$	$F_B = F_{A \rightarrow B} + F_{C \rightarrow B}$ $P_B = TP_B + F_B$
	Class C	$F_{A \rightarrow C}$	$F_{B \rightarrow C}$	$TP_C$	$F_C = F_{A \rightarrow C} + F_{B \rightarrow C}$ $P_C = TP_C + F_C$
		$R_A$	$R_B$	$R_C$	

instance, in clinical diagnostics for a lethal disease, the sensitivity will have a higher weight than specificity. In contrast, when a treatment for a certain disease has severe side effects, the specificity is of greater importance. Additionally, positive predictive value (PPV) and negative predictive value (NPV) are very commonly used performance measures in medical science.

The so called receiver operating characteristic (ROC) curve depicts the balance of true positive rate (sensitivity) and false positive rate (1-specificity), see Figure 2.8. Therefore, the area under the ROC curve (AUC) serves as another very important measure of classification performance (78). Calculating the AUC for multi-class problems relies on separating the problem into multiple two class classification problems and weighting it by the class prevalence in the data:

$$AUC_{total} = \sum_{c_i \in C} AUC(c_i) \dot{p}(c_i) \quad (2.7)$$

Here  $AUC(c_i)$  is the area under the ROC curve for class  $c_i$ . The overall complexity of

Table 2.3: Selection of most important performance measures for the two and three class classification problem: ACC, SEN, SPE, PPV and NPV.

Performance	Classification Performance	
	Two - class	Three - class
ACC	$ACC = \frac{TP+TN}{P+N}$	$ACC = \frac{TA+TB+TC}{RA+RB+RC}$
SEN	$Sen(A) = \frac{TP}{(TP+FP)}$	$Sen(A) = \frac{TP_A}{(TP_A+FP_{A \rightarrow B}+FP_{A \rightarrow C})}$
SPE	$Spe(A) = \frac{TN}{(TN+FP)}$	$Spe(A) = \frac{TN_A}{(TN_A+FP_{B \rightarrow A}+FP_{C \rightarrow A})}$
PPV	$PPV = \frac{TP}{(TP+FP)}$	$PPV(A) = \frac{TP_A}{(P_A)}$
NPV	$NPV = \frac{TN}{(TN+FN)}$	

this approach is  $O(|C|n \log n)$ . More complex calculations for the multi-class AUC exist, but exceed the scope of this thesis. For an excellent introduction to ROC curves and AUC calculation as well as multi-class AUC the reader is referred to Fawcett *et al.* (78) and Hand *et al.* (100).

## 2.6 Longitudinal Analysis and Mixture Models

The combination of both, fixed and random effects in so-called mixed-effect models (MEMs) is useful for a wide variety of applications. However it is a particularly well suited technique to model longitudinal data (178; 133; 29). The MEMs are an extension of the more general linear model (LM), which only contain fixed effects, representing explanatory variables that are supposed to be not random, for instance the medication while modeling the disease progression.

In contrast, random effects or variance components assume a hierarchy of different populations within the data set. Consider the analysis of a group of persons undergoing a certain treatment, the treatment and the time are expected to be non random. However, each person has a specific "population" of measurements following a specific distribution, resulting in a random effect.

Various extensions of linear mixed models have been developed to better accommodate more complex data sets. In the frame of the proposed project, we will evaluate the capabilities of more advanced models such as non-linear extensions of MEMs (224), hierarchical MEMs (201), or multivariate MEMs (211). The latter is particularly suited to account for multiple variables evolving in parallel. Another category of methods that has been developed in the last decade are multi-variate longitudinal classification methods rely on, for instance, SVM. These are combined with various kernel methods like dynamic time

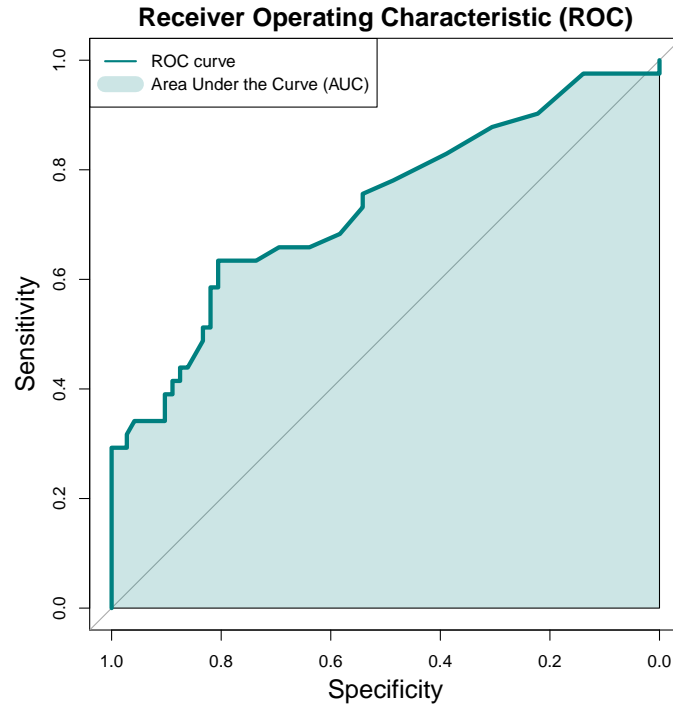


Figure 2.8: Receiver operating characteristics curve represents the relation between Sensitivity and 1-Specificity. The AUC is the area under this so called ROC curve. The shown diagram was plotted utilizing the R package *pROC* and the public R data set example *aSAH*.

warping kernels (167) or a hierarchical structure of multiple kernels (54; 242). Other techniques combine the concept of mixed effects least squares models and SVM (145). In contrast to the described supervised methods, a set of unsupervised approaches have been developed, for instance various clustering techniques (51; 136).

## 2.7 Validation and Permutation

In most clinical diagnostics studies the main aim is to get a comprehensive and robust description of the data and its predictive potential. Therefore, often multiple machine learning techniques are applied, evaluated and compared. However, the benefit of applying multiple techniques depends on the underlying problem. Within a set of linear and non-linear techniques, the more sophisticated non-linear approaches most likely deliver the highest accuracy (e.g. SVM, kernel-methods, RF) but might be too complex for the problem and prone to overfitting, or often lack interpretability. The goal is to compare approaches and determine the most suitable one for the given data and the corresponding implications (e.g. non-linear / linear). Several breathomics studies presented integrated



systems, including several machine learning methods, for VOC-based supervised classification into patient groups (175; 34).

As mentioned before, different techniques enable selecting the most informative attributes in the data. Their quality relies on the quality of the selected model. Therefore, an adequate validation is crucial to give certainty and robustness to the findings, in particularly to the differences between studied classes and the selected classifying compounds. The model performance and predictive ability can be assessed by the previously described quality measures, see Section 2.5.3.

### 2.7.1 Traditional Conservative Validation

Commonly, in supervised analysis, the model is constructed using a training set and subsequently verified with an independent test set. Moreover, an ideal data setting for statistical learning provides a sufficiently large number of samples, such that the data can be divided into training, validation and test set, also called three-way-split. The first subset is used to train a classification model with various parameters. During the next phase, each of the resulting models is "validated" on the second subset. Finally, the prediction power of the best performing model is assessed on the test set. The corresponding parameters can be used to build the final model. The test error or performance represents the most conservative but also most robust estimator for the real error. Popular algorithms for splitting the data are for example Kennard and Stone algorithm (123), Duplex technique (200) or random selection. Nevertheless, the ultimate manner of assessing prediction ability of a statistical model is to use a set of newly acquired samples coming from an independently sampled population. This ultimate validation should be generally favored.

### 2.7.2 Cross Validation

In the breathomics field, however, the sizes of data sets are often too small to apply the standard training-validation-test split. In those scenarios cross-validation (CV) is the most commonly applied approach. The easiest approach is the so called leave-one-out CV (LOO CV) in which the data set is divided into as many folds as the number of samples. During an iterative process each sample is excluded from the data once. The model is trained on the resulting set consisting of the remaining  $n-1$  samples ( $n$  equals the total number of samples) and tested on the excluded sample. Since each sample is excluded once, the accuracy is evaluated on the predictions of all samples. LOO CV is used when the number of available samples per class is low. However, the method has certain drawbacks, for instance, it requires  $n$  iterations of training and test and has a high variance if the prediction rule is unstable (73). A very common alternative is the  $k$ -fold CV, where  $k$  defines the number of subsets the data is divided into. Each subset is excluded once and used as a test set for a model trained on the remaining  $k-1$  subsets. If the number of samples is not sufficient for a three-way-split, it is advised to apply multiple cross validation runs, or a so called nested cross-validation (10). In contrast to the LOOCV, the  $k$ -fold CV is more conservative and less variable. However, to generate unbiased results it is crucial that all aspects of classifier training take place within the CV

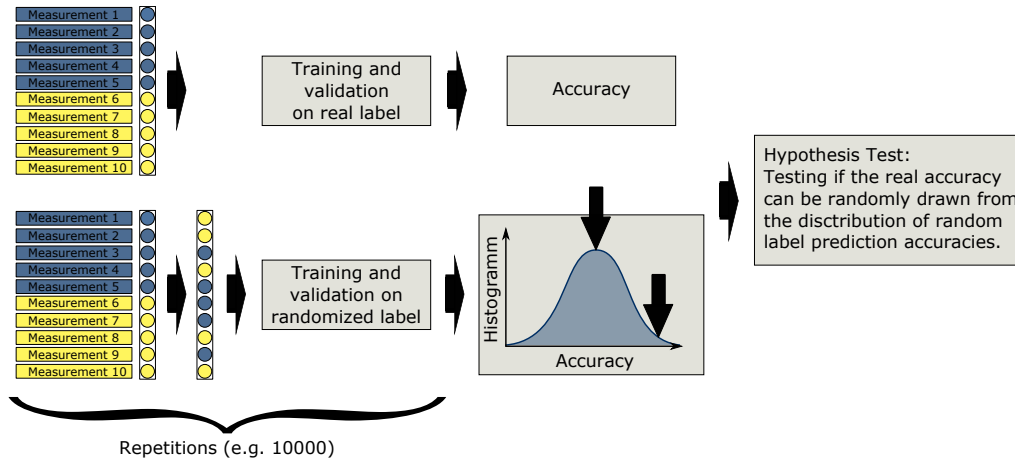


Figure 2.9: The Figure depicts the overview of the permutation tests procedure. The upper path shows the estimation of the real accuracy. The second path depicts the estimation of the distribution of random accuracies. Finally, it can be evaluated if a value equal to the real accuracy can occur by chance.

loop. This includes feature selection as well as classifier type and parameter optimization. Any violation of these rules can result in very biased estimates of the true error (212).

### 2.7.3 Randomization and Permutation Tests

Another approach to evaluate the robustness of a classification model is utilizing randomization and permutation tests. The key idea is to assess whether a real grouping of the data set is significantly better classified than a randomly assigned grouping (229; 91), see Figure 2.9. Therefore, at first the classification performance of the model trained on the real labels is evaluated. Subsequently, for the permutation test, real class labels are randomly permuted and a new classification model is built predicting these random classes and evaluating the random performance. A large number of repetitions of this step produces a distribution of random performances. It is expected that the mean random performance lies around 50%. However, many real world data sets allow complex statistical learning methods to overfit. The assumption is that the prediction of the real labels is in fact better than the random distribution. The Null-Hypothesis of the permutation test is that the real label is drawn out of the random distribution. A rejection of the test shows that the real prediction is significantly better than the random model (166).

## 2.8 Overview of Related Work in Breathomics

During the last decade, the advances and developments of various analytical high throughput and high resolution technologies have expanded the fields of application especially in breathomics. This enabled a large number of studies within the area of clinical breath

diagnostics. In the following I will present a selection of studies and software tools that are focusing on MCC/IMS and breathomics data analysis and therefore most relevant to this thesis.

### 2.8.1 Previous Breathomics Studies

**Statistical Tests and Correlation Analysis:** Many studies applied statistical hypothesis tests to differentiate various diseases or states traditionally from a set of healthy controls. While the first pilot studies were failing to account for the multiple testing problem, later conducted studies included Bonferroni or false discovery rate (FDR) corrections. The main focus of studies lies on the characterization of pulmonary diseases, differentiation between diseases or towards a healthy control group. For instance, chronic obstructive pulmonary disease (COPD), lung or bronchial cancer (31; 23; 231; 30). In other studies, simple statistics were applied to evaluate disease subtypes, for instance distinguishing COPD patients with and without alpha 1-antitrypsin deficiency (127).

Another study by Maddula *et al.* utilized correlation analysis of IMS data to detect relations of infectious agents in the airways (148).

**Unsupervised Learning and Principal Component Analysis:** Some breathomics studies utilize unsupervised learning techniques such as clustering, but the majority focuses on simpler methodology like PCA. For instance to evaluate the variation of metabolic breath profiles within different disease groups, like smoking related diseases such as COPD or lung cancer (232; 65).

However, more common and advanced are the statistical techniques utilized in breathomics studies involving traditional analytical techniques like ESI-MS or sensor based systems. Examples are the evaluation and discrimination of bacterial cultures of various strains (243; 56) or differentiation of patients with untreated pulmonary sarcoidosis from controls (69) applying PCA. More advanced studies use combinations of PCA, factor or clustering analysis to identify potential VOC biomarkers for lung cancer (52) or subphenotyping of COPD stages (80).

**Supervised Learning and Classification:** One of the first studies applying more advanced supervised learning methods on MCC/IMS breath data was conducted by Baumbach *et al.* in 2007. The software framework IMS2 was utilizing multiple advanced methods such as SVM, ANN and Naive Bayes classifier to predict lung cancer (16). Another pilot studies focussed on decision trees or probabilistic relational learning for medical diagnostics (85; 231).

However, in contrast to the relatively novel MCC/IMS technology, the data analysis of traditional analytical techniques in breathomics has proven the potential of breathomics data in combination with advanced statistical learning techniques. For instance, techniques like linear discriminant analysis were utilized to identify breath biomarkers for liver cirrhosis (60). The partial least squares discriminant analysis (PLS-DA) showed great potential in metabolomics in general (206) and was therefore applied to find early predictive signatures of asthma in children (199). Another popular method are artificial neural networks which have been applied to determine blood glucose levels (191) or identify patients with colorectal cancer (8).

## 2.8.2 State of the Art Software Packages

### MCC/IMS specific software

**VOCan** The commercial software package VOCan, is provided by B&S Analytik (Dortmund, Germany) the manufacturer of several MCC/IMS devices, such as, for instance, the BioScout, specialized for the analysis of human exhaled air. It is operating the MCC/IMS devices, handling the raw data flow and formatting it into the previously introduced raw data Excel files.

**VisualNow** The commercial software package VisualNow, is also developed by B&S Analytik. It is implemented in Java and is one of the state of the art software tools to visualize the whole MCC/IMS chromatogram as well as single IMS spectra in two- and three-dimensional plots. The acquired data of the MCC/IMS file includes a set of parameters describing the measurements', experimental setup, which are displayed in a separate area (35). Additionally, it provides all essential preprocessing methods like RIP-detailing, baseline correction, smoothing, as described previously, as well as the ability to manually and automatically pick peaks in the chromatogram based on algorithms developed by Bader in 2008 (12).

**IPHEX** The software package IPHEX (by A. Bunkowski, University Bielefeld, Germany) also supports the visualization of MCC/IMS chromatograms, including single spectra and total the ion current of the MCC as well as basic preprocessing. It further implements a peak detection based on the watershed transformation (49; 219) as described in Section 2.2. In contrast to the VisualNow, the IPHEX software tool provides a feature to visually compare a MCC/IMS measurement with a GC/MS measurement, by plotting the MCC and GC retention times next to each other. Unfortunately, due to the time dependent changes in temperature in the GC/MS the alignment between GC and MCC is not trivial. In a recent presentation, Sanders *et al.* proposed linear functions representing the relation, mainly focused on certain classes of compounds (189). However, a cross species mapping function would most likely require a non-linear function that includes more information about the molecule and its family (152).

**PeaX** PeaX is a command line tool for MCC/IMS raw data processing, developed at the Technical University of Dortmund. It focuses on the detection and modeling of peaks within the MCC/IMS raw data (128). It provides basic functionality in preprocessing, such as RIP-detailing, denosing, baseline correction and smoothing. However, its main purpose is the peak detection and peak modeling. Therefore, it provides several algorithms for finding potential peak candidates, clustering peak candidates and an expectation maximization algorithm to fit a mixed model for each peak.

### General Metabolomics, Breathomics and Clinical software

The selection of software products is restricted to the projects that are most relevant to this thesis.

**ABA-Cloud** The Advanced Breath Analysis platform documents and stores breath analysis studies to enable full automatic reproducibility. It also enables to submit keyword-based queries in order to search for conducted studies at collaborating research centers (76). See Section 2.3 for more details.

**AAF** Automatic Analysis Framework is a Taverna based project enabling the user to reanalyze existing data automatically (144). See Section 2.3 for more details.



# Chapter 3

## Data Sets

### 3.1 COPD and Lung Cancer Data

chronic obstructive pulmonary disease (COPD) is an umbrella term used to describe chronic lung diseases that cause a permanent blockage of airflow from the lungs, which is not fully reversible (WHO). The most prominent symptoms are breathlessness, a chronic cough, and excessive sputum production. Airways and lungs react to noxious particles or gases, like smoke from cigarettes or fuel, with an increased inflammatory response (GOL). However, despite the similarities, the disease is not just a simple "smoker's cough", but a leading cause of morbidity and mortality worldwide which is still widely under-diagnosed. The World Health Organization (WHO) reported it as one of the four most frequent causes of death. In the period between 2000 and 2011 alone, the disease caused 5.8% of all deaths worldwide (WHO). Estimates show that COPD will become the third leading cause by 2030.

Furthermore, Young *et al.* reported in 2009 that COPD is both a common and important independent risk factor for lung cancer (240). The National Cancer Institute (NCI) defined lung cancer as an "uncontrolled cell growth in lung tissue". It usually occurs in the cells lining the air passages mostly functioning as an air filter for dust and noxious particles (NCI). The small cell lung cancer and non-small cell lung cancer are two main subtypes (NCI). The survival rate of patients within five years is less than 20% depending on the stage of the lung cancer. Modern diagnosis is based on the microscopic visual analysis of the cells. However, despite the severity of the disease, the majority of bronchial carcinoma is still detected randomly during routine examinations.

The MCC/IMS breath profiles of COPD patients as well as healthy volunteers was utilized. The present data set was previously published by Westhoff *et al.* in 2011 (231) and kindly provided by our collaboration partners at the Lung Hospital in Hemer, Germany. The study comprised the exhaled air profiles of 42 COPD patients, 52 patients suffering from both, COPD and bronchial carcinoma (BC), as well as 35 healthy controls. The metabolomics analysis of the patients' breath was done using an ion mobility spectrometer coupled with a multi-capillary column, as described previously. An automated preprocessing by VisualNow and an expert driven manual peak evaluation resulted in a set of

120 volatile organic compounds. Each of the compounds present in at least three of the patients' measurements. All patients were recruited from cooperating German hospitals. This data set is utilized in multiple Chapters, to demonstrate the potential of modern computational tools in breathomics studies. In Chapter 4 it serves as an example data set that is stored in the MCC/IMS database. Additionally, the data is analyzed by state of the art supervised and unsupervised learning techniques in Chapters 5 and 8.

## 3.2 GC/MS - MCC/IMS Comparison Data

The following data set was specifically conducted to analyze and develop a methodological procedure to combine and compare GC/MS and MCC/IMS measurements. The state of the art manual comparison is tedious. Therefore, this data set serves as a prove of concept for the MIMA tool presented in Chapter 7. It requires the parallel analysis of the same mixture of chemicals using the two different technologies.

**Chemicals** To create such a data set, the following chemicals (purchased from Merck (Darmstadt, Germany)) were utilized in this experiment:

- Trans-anethole (98%)
- 2-heptanone (98%)
- 2-nonanone (99%)
- undecanal (97%)
- R-(+)-limonene (96%)
- R-(-)-carvone (99%)
- (-) menthol (99%)

**Preparation of MCC/IMS and GC/MS Reference Mixture** Each reference substance (10  $\mu$ L, menthol 10 mg) has been filled into a 2 mL reaction vial (CS-Chromatography Service GmbH, Langerwehe, Germany) with a screw cap including a gas permeable membrane. All 7 vials were placed together in a 250 mL Schott flask and incubated for 48 hours at room temperature. The flask was equipped with a two ports screw cap.

**MCC/IMS Measurement** Finally the samples were measured using a MCC/IMS (Type BioScout, B&S Analytik, Dortmund, Germany). For more details on the adjusted parameters see Table 1 in Maurer and Hauschild *et al.* 2014 (152). Ports of the Schott flask were used for sample collection and simultaneous synthetic air injection to avoid negative pressure. Synthetic air flow rates were adjusted to be equal to the sampling flow rate.



**GC/MS Sample Collection** The sampling is done by using TENAX tubes (GERSTEL, Mlheim, Germany). The tube is connected on one side to the sampling pump (Universal XR Pump Model PCXR8, SKC Inc., Pennsylvania, USA) and on the other to the Schott flask with the reference mixture. A flow of 150 mL/min for 6 min is adjusted to collect the sample. Ports of the Schott flask were used for sample collection and synthetic air injection. Synthetic air flow rates were equal to the sampling flow rate.

**GC/MS Measurement** The measurement of the TENAX tubes was done using a GC/MS (Agilent Technologies 6890N Network GC System/Agilent 5973 Network Mass Selective Detector) coupled with a cooling trap CIS (GERSTEL, Mlheim, Germany). The TENAX tubes are located in the thermo desorption auto sampler (TDSA2), which is connected to the thermo desorption system TDS 2 (GERSTEL). For more details on the parameter settings we refer to Table 2 in Maurer and Hauschild *et al.* 2014 (152).

### 3.3 Anonymous Data

To evaluate the quality of peak detection methods in Chapter 6, a comparison to the gold standard which equals to the manual peak picking by a domain experts is required. These specialists are familiar with the result of previous studies, as for instance COPD studies. In order to achieve an unbiased list of peaks, independent of previous results, this set of patients was recruited anonymously and based on a disease unknown to the person conducting the manual analysis.

The data set contains 69 MCC/IMS measurements: 39 from patients suffering from a certain disease and 30 healthy persons in a control group. Due to confidentiality reasons on sensitive patient data, this work has no permission to give more details on the disease. The single samples correspond to one of the two classes “healthy” or “not healthy”. This scenario therefore requires the implementation of a classifier distinguishing patients from controls, to evaluate whether the MCC/IMS is a valid diagnostic tool for this disease and subsequently to support a physician by proposing the most likely patients condition.

### 3.4 Artificial Data

The artificial data set is dedicated to demonstrate and clarify the capabilities of unsupervised learning methods in general. It consists of 16 samples and 12 metabolites associated with three metabolite groups. Each of these groups is related to one of the three predefined labels of the samples: (1) “health” (values: healthy (five), disease Subtype 1 (five), disease Subtype 2 (six)); (2) “smoking” (values: smoker (nine), non-smoker (seven)); (3) “nutrition” (values: apple juice (four), tea (four), orange juice (four), coffee (four)). The label “health” is our primary outcome variable, while “smoking” and “nutrition” shall be considered as potential confounding factors. Tables 3.1, 3.2 and 3.3 show the corresponding mean and standard deviations used to generate normal distributions from which the artificial dataset was sampled. Note that these simulated data are highly idealized. This serves the sole purpose of exemplifying and clarifying the use and functionality of unsupervised learning methods. In contrast to a real world data set, this artificial data

set that provides a known outcome, serves as a prove of concept for the Carotta tool presented in Chapter 8.

Metabolite	Healthy	Subtype 1	Subtype 2	Standard deviation
Metabolite 1	1	4	3	0.25
Metabolite 2	3	6	5	0.1
Metabolite 3	6	9	8	0.5
Metabolite 4	0	3	2	0.7

Table 3.1: Distribution of metabolite expression mean intensities and their variance (standard deviation) for artificial patients with class label “Health”.

Metabolite	Orange Juice	Apple Juice	Tee	Coffee	Standard deviation
Metabolite 1	0	2	4	6	1
Metabolite 2	2	4	6	8	1
Metabolite 3	5	7	9	11	0.5
Metabolite 4	-1	1	3	5	0.75

Table 3.2: Distribution of metabolite expression mean intensities and their variance (standard deviation) for artificial patients with class label “Nutrition”.

Metabolite	Non-Smoker	Smoker	Standard deviation
Metabolite 1	-1	5	1
Metabolite 2	1	7	1
Metabolite 3	2	8	2
Metabolite 4	-2	4	0.7

Table 3.3: Distribution of metabolite expression mean intensities and their variance (standard deviation) for artificial patients with class label “Smoking”.

### 3.5 Longitudinal Rat Breath Data

This data set is originating from a recent pilot study of Fink *et al.* (83) and was kindly provided by our collaboration partners at the Department of Anesthesiology, Intensive Care, and Pain Therapy, Saarland University Medical Center, Homburg (Saar), Germany. All rat experiments were conducted with approval from the Animal Care & Use Committee (Landesamt für Soziales, Gesundheit und Verbraucherschutz; Saarbrücken; Germany) and in accordance with the German Animal Welfare Act.

The exhaled air of 40 Male Sprague-Dawley rats (obtained from Charles River, Sulzfeld, Germany) was analysed. The rats were kept in the institutional animal facility under controlled conditions and fasting for a period of 12 hours. Afterwards, they were anesthetized, connected to the respirator and ventilated with highly purified synthetic air.

A measurement of the metabolite composition of their exhaled air was taken continuously every 20 Minutes using the MCC/IMS technique during the whole period of the experiment (maximal 12 hours). Subsequently, each rat underwent the surgical procedure according to the randomly assigned treatment group. A detailed description of the surgical procedures can be found in Fink *et al.* 2014 (83).

**CLI** Sepsis was induced using a standardized model of **Cecal Ligation and Incision** of the cecum as previously described (83).

**SHAM** Sham-operated rats were treated likewise but without ligation and incision.

Prior to connecting the rats to the MCC/IMS device and the surgical treatment, the background concentrations of volatile compounds from the respirator was measured for one hour and two to three measurements each. The combination of the respirator and corresponding rat measurements are further referred to as one experiment. A blood test at roughly 300 minutes, confirmed the induced/not-induced sepsis.

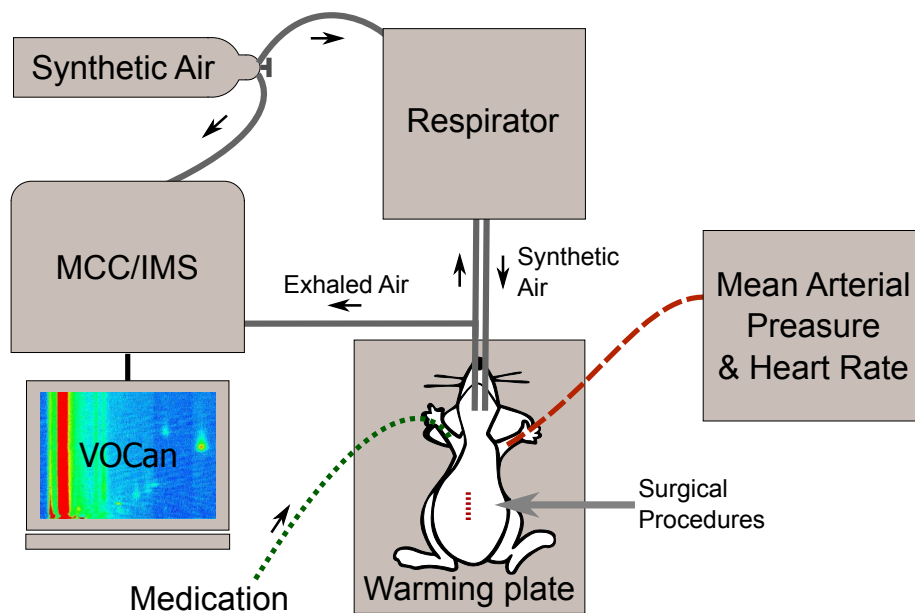


Figure 3.1: Shows a schematic of the operational procedure utilized for the sham operation or to induce sepsis.

See Fink *et al.* for more details (83). This data set is analyzed in Chapter 9. It is focusing on the evaluation of longitudinal data, such as introduced in this section.

### 3.6 Summary

The previously described data sets are utilized in one or several of the following methodological Chapters. Table 3.4 gives an overview of the data sets and in which Chapters they are used.

Table 3.4: This table depicts which Chapters utilize the previously described data sets.

		Chapters					
		Clinical Breath Data Management	Towards Robust Machine Learning in Breathomics	Evaluating Automated Peak Detection	Metabolite Identification with MIMA	Uncovering Hidden Structures with CAROTTA	Longitudinal Breath Analysis
Datasets	COPD and Lung Cancer	×	×			×	
	GC/MS - MCC/IMS Comparison Data				×		
	Anonymous Dataset			×			
	Artificial Data					×	
	Longitudinal Rat Breath Data						×

## Chapter 4



# Clinical Breath Data Management

*The IMSDB database framework was developed in a collaborative fashion during the course of the Master thesis of Till Schneider. The author of this PhD thesis was responsible for the conception and structure of the software project and advised the work. This includes most of the software engineering as well as the design of the database structure. The contribution of Till Schneider was the implementation of the project.*

The high-throughput character of MCC/IMS technology facilitates the application in large-scale health care screening studies. However, the computational methods to process MCC/IMS data are limited to a few preprocessing and analysis tools. While multiple issues have to be addressed to achieve a rapid analysis, one of the major issues is the efficient management of large amounts of MCC/IMS raw data as well as the increasingly complex information gathered in clinical studies.

Therefore, there is a need for a centralized and flexible data repository, which is capable of gathering metabolic data, like MCC/IMS chromatograms or GC/MS spectra, and heterogeneous biomedical data, e.g. patient records, in a well structured manner. Furthermore, the database application should support the storage of constantly evolving object-attribute-data. Additionally, the solution should provide the foundation for a statistical analysis and data mining. To achieve user accessibility and acceptance, we require a self-explanatory user interface, including automated data integration, data integrity validation, data backup and retrieval, as well as a basic data mining toolkit.

In summary, we need an intuitive software system for storing and managing breathomics and heterogeneous biomedical data that does not require any prior knowledge or technical

skills (195). This Chapter will first identify common requirements of a clinical breathomics database and evaluate to what degree they are fulfilled by state of the art software solutions. Subsequently, a detailed description of the structure and implementation of the IMSDB is given. This novel comprehensive database application and analysis platform combines metabolic maps with heterogeneous biomedical data in a well-structured manner. The IMSDB was previously published in (195), which is the main source of this Chapter.

**Objective:**

To build a flexible and intuitive system for breath data management and analysis, providing quick and easy access.

## 4.1 Requirements and State of the Art

A clinical breathomics database needs to combine the flexible management of medical data (diseases, medication, age, gender, etc.) and the efficient storage of analytical data (MCC/IMS, GC/MS) provided by physicians or biologists. Moreover, a set of major functionalities described in the following paragraphs have to be fulfilled.

**Flexibility and Heterogeneous Clinical Data** The database is required to store study-subject specific attributes, such as gender and age for human subjects or antibiotics resistances for bacterial samples, respectively. Other examples are important environmental factors such as nutrition, medication or diseases. These parameters can differ between various projects. Therefore, the system has to provide a flexible and generic data structure to store arbitrary information without changing the internal database scheme. The major drawback of most generic models is the performance loss under particular circumstances due to higher model complexity. In particular, more complex and attribute-centric queries perform worse (53). However, a proper balance between flexible and efficient modeling combined with performant database queries can alleviate these drawbacks.

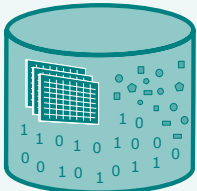
**Raw Data** To hold huge amounts of breathomics raw data, the developed system needs a fast and robust storage system. All further processing and analysis approaches, such as preprocessing and peak detection, rely on the ability to efficiently access the raw data. Additionally, experimental conditions, such as the temperature, carrier or drift gas flow, and settings of the electric field, are of particular interest. This information allows for the evaluation of the impact of various technical parameters and their optimization with respect to VOC based biomarker investigations.

**Metabolites** Another important issue includes the storage of extracted metabolomics or breathomics information. As previously described, the MCC/IMS chromatogram has three dimensions: (1) retention time (RT), (2) inverse ion mobility ( $1/K_0$ ) and (3) peak intensity. Metabolites can be interpreted as local maxima in the peak intensity. Under standardized conditions the coordinates RT and  $1/K_0$  are specific for a particular VOC.

The corresponding peaks are detected by a sequence of existing computational preprocessing algorithms for RIP-detailing, denoising, smoothing and peak detection. See Section 2.2.3 for more details. The provided metabolite information can be utilized by machine learning and statistical methods to identify potential biomarkers and to assess the quality of the various preprocessing techniques.

**Statistical Analysis** As a proof of concept, the system should provide simple statistical methods for classification and feature selection. A clear understanding of the detailed composition of the analyzed air is necessary to distinguish between disease specific biomarkers and compounds originating from confounding factors, such as nutrition or life-style. Therefore, the system needs to analyze the relation of peaks to an outcome variable selected from the available subject attributes.

**Usability** As described previously, the success of a software tool is highly dependent on its usability. An intuitive graphical user interface guides the user through the process of automated data integration, curation, backup and retrieval followed by data analysis and presentation.



### Requirements:

- ⇒ Flexible data handling of heterogeneous and evolving project attributes, such as clinical data and biological experiments.
- ⇒ Efficient storage of MCC/IMS raw data.
- ⇒ Storage of metabolite information in a reasonable way.
- ⇒ Functionality for simple statistical analysis and feature selection.
- ⇒ Intuitive and user friendly graphical user interface for data acquisition and storage.

## State of the Art

The first automated systems for IMS data analysis was published in 2007 by Baumbach *et al.* (16). In the same year, Lesniak developed the first database scheme to organize MCC/IMS data (138). However, the presented approaches do not fulfill the above requirements. The static database scheme of Lesniak is not able to store flexible entities, attributes and values, as well as relations between entities, which is essential to make the

database adaptable to any kind of biomedical data. The data analysis system of Baumbach *et al.* does not account for confounding factors, such as patient data.

A more recent study presented ABA-Cloud, a platform to document, store and re-analyse entire breath studies. The encapsulation into PSEs does not enforce a general common structure of breath data. Therefore, the integration and analysis of the data from different studies is tedious and error prone.

Beyond the field of breath research, there are two areas overlapping with the proposed database framework. The field of Omics-databases covers the advanced processing and analysis of metabolomics and proteomics data, including software tools such as MeltDB (160; 161) or the free software library OpenMS (204). The clinical-database field deals with the problem of adequate and flexible clinical data repositories described in Section 2.3. None of these projects are able to include both heterogeneous biomedical data, and metabolomic data. Moreover, they do not include data validation. Oracle Clinical and TrialDB for instance, make use of the entity-attribute-value (EAV) design (159) described later. Others use the more advanced EAV with classes and relationships (EAV/CR) approach like SenseLab (159). Those databases focus solely on the storage of heterogeneous biomedical data. However, they are not able to handle metabolomic data and neglect the validation of input data. See Section 2.3 for more details on the different systems.

For the IMSDB database framework, we aimed at combining the strengths of both research fields, to create a powerful solution that can handle metabolic data together with heterogeneous biomedical data.

## 4.2 Methods

In this Section, we first describe the data structure and information of the MCC/IMS data followed by the data model of the IMSDB framework and details of the implementation. Finally, I will introduce an unpublished extension of the IMSDB called ClinicalGUI, which focuses on user friendly data acquisition.

### 4.2.1 Structure of MCC/IMS Data

Commonly, MCC/IMS measurements, detected peaks and heterogeneous object data, are stored in a set of well-defined file types. The MCC/IMS related file types are generated by VisualNow, the firmware of the BioScout device (B&S Analytik, Dortmund, Germany), as described in Section 2.8.2 of the background Chapter. The most important file types are explained in the following.

**IMS Measurement File** The raw data is a comma separated text file that comprises experimental and technical information as well as the raw spectra measured by an MCC/IMS instrument. The file format is described in (214). In the following, we do not distinguish between MCC/IMS measurement files and MCC/IMS measurements, assuming that the meaning can be deduced from the context.



**PeakAn File** This file follows the Microsoft Excel xls format and contains information about a particular peak region retrieved from a set of MCC/IMS measurements. A peak region can be represented as a rectangle or ellipse defined by two corners describing the bounding rectangle MCC/IMS coordinates. Recall that each MCC/IMS coordinate is defined by RT and  $1/K_0$ . Ideally, all detected peaks within this defined peak region correspond to a particular molecular compound. The PeakAn file comprises the maximum signal intensity in the peak region and the corresponding MCC/IMS coordinate for each measurement.

**Object-Attribute-Table** A manually generated Microsoft Excel spreadsheet that provides heterogeneous attributes of objects (synonymous to patient) obtained during the examination at a specific time point. Each examination of an object is referred to as an object case. It is denoted by a row in the table. The columns comprise distinct attribute names and the data type of the attribute values. In order to identify an object case, an *ID* attribute corresponding to the *sample ID* in the IMS measurement is included in the first column. The name of the corresponding MCC/IMS measurement is given in column *imsfile*. A *class* attribute denotes the hospital or institution the object/patient belongs to. An example for an object-attribute-table is given in Table 4.1.

Table 4.1: This table illustrates the structure of the object-attribute-table format. The first two rows comprise identifying attribute names and the corresponding data types.

id	class	imsfile	attribute1	attribute2	...
string	string	string	boolean	date	...
304856	hospital_patient	file1.ims.csv	yes	01.01.2011	...
305082	hospital_patient	file2.ims.csv	no	01.01.2012	...
...	...	...	...	...	...

## 4.2.2 Structured Data Model

The data model can be divided into three main components, namely the MCC/IMS measurement, the generalized data structure and the peak component, which are described in detail in the following. The connected components model, the relations between MCC/IMS measurements, peaks and object cases are shown in Figure 4.1.

**1) Measurement component:** It contains entities related to the MCC/IMS measurements: *ImsMetaData*, *ImsRawData*, and *Alignment*, see Figure 4.1. The first represents experimental conditions including all technical parameters selected for the experiment and stored in the measurement file. *ImsRawData* comprises the MCC/IMS measurement file. Due to environmental and instrumental variations such as ambient temperature or pressure, MCC/IMS chromatograms can be slightly shifted. Therefore, the *Alignment* refers to parameters of a linear transformation for each measurement.

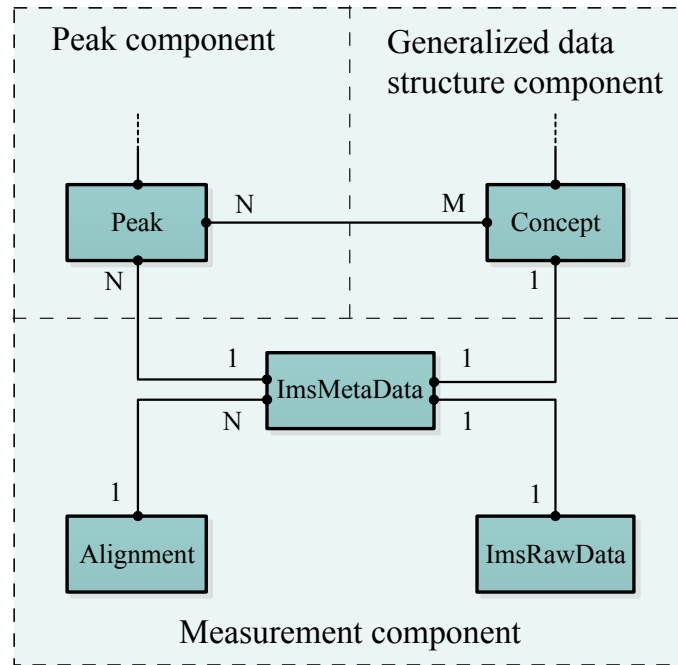


Figure 4.1: Overview of the three main database components and their relations. The central core entities of each component are linked directly. Auxiliary entities of generalized data structure and peak component can be found in the next paragraph. Associations and maximum cardinalities (one or many) are represented as solid lines labeled with characters 1, N, M.

**2) Generalized data structure component:** It models objects with arbitrary attributes and is also referred to as ontology component. In addition, the concept of relations allows for relationships between any two objects. See Baumbach *et al.* 2009 for related work (17; 18). The ontology component is similar to an entity-attribute-value (EAV) model for the organization of heterogeneous data, described in (159). Its main function in this work is to model sparse object data retrieved from an Object-Attribute-Table ensuring the flexibility of the system. New projects will require to store entirely different sets of attributes and parameters. In contrast to conservative database schemes, an EAV organization offers a general storage of attributes and values and does not require any changes in the database structure. Figure 4.2 shows the participating entities and relationships.

The central *Concept* entity represents an object, such as, for instance a patient case referring to a row of the Object-Attribute-Table (Table 4.1). Many-to-many associations between *Concept* and value entities (*ValueBoolean*, *ValueDate*, *ValueDouble*, *ValueInt*,

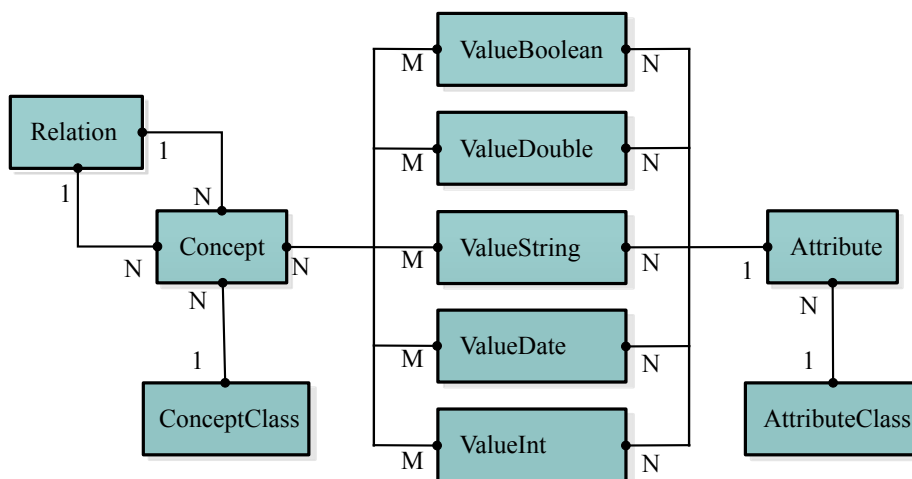


Figure 4.2: Overview of the generalized data structure (ontology) component. The central entity of this component *Concept* is associated to combinations of *Attribute* and value entities. In addition, a *Relation* is used to associate *Concept* entities to each other.

*ValueString*) represent the assignment of attribute values to the *Concept*. Each of these value entities, which include a value column of the corresponding type, is assigned to an *Attribute* entity by using a many-to-one association. An *Attribute* contains a unique *name* column. The application-logic ensures that the attribute-value-pairs are unique. The objects in *Attribute* and *Concept* can be further categorized with meta information (e.g. different institutions). They are associated to *AttributeClass* and *ConceptClass*, respectively. Finally, *Concepts* can be linked to each other by the entity *Relation*. This enables for instance the representation of the patient history that is modeled by the connection of multiple consecutive consultations. The relationship between an object such as a patient examination and the corresponding MCC/IMS measurement is represented by a one-to-one association between the entities *ImmsMetaData* and *Concept* (shown in Figure 4.1).

**3) Peak component** offers a semi-generic structure for peak data persistence. This comprises the compulsory parameters peak position and peak region. Moreover, it allows for exchangeable peak attributes such as the maximum peak intensity. This flexibility is required since modern automated peak detection methods like PME offer a variety of measures for molecule quantity such as the peak volume or a number of parameters fitting the peak model. Details of the peak component are shown in the Appendix Figure A.1. An association between *Peak* and *Concept*, as shown in Figure 4.1, allows the data model to support any kind of annotation. Chapter 7, for instance, presents a tool for automated peak identification utilizing either parallel GC/MS measurements or a MCC/IMS molecular database.

### 4.3 Implementation

The presented breathomics database relies on a PostgreSQL database management system (<http://www.postgresql.org>). The corresponding database application for integration, view and analysis is based on Java SE 6. The integrated data analysis functionality takes advantage of the free, comprehensive machine learning software package Weka (version 3.7.3) (99). The database model is fully incorporated in the Java application, which automatically generates the database scheme. Therefore, the Java code is annotated with meta data specifying the mapping between objects in Java and the relational database. The object-relational mapping (ORM) described in (15) deals with the automated persistence of Java classes to database tables. The free Java software package Hibernate (version 3.6) is utilized as an ORM tool (official website: <http://www.hibernate.org>). Figure A.1 in the Appendix illustrates the system architecture.

#### 4.3.1 Validation and Integration

The integration process is divided into three steps, where each step handles a specific kind of information: (1) MCC/IMS measurements, (2) peaks, (3) object/patient data. The IMSDB software comprises parsers, a comprehensive file and consistency validation logic, and business logic for the integration of data collections. The validation logic initially triggers the parsers to read the different files. Prior to the import process, potential violations of the specification such as missing entries, type errors and inconsistencies for cross-related file entries are listed and presented to the user. Multiple unique entry constraints such as the *PeakPosition* require an additional select query before inserting a new entry. Large datasets containing thousands of peaks would hence be unfeasible. Therefore, the so called chunk query peak persistence (CQPP) method was developed to select multiple entities at once (depending on the chunk size) instead of inefficiently executing one query for each new entry. The insertion of missing entities is then handled in a JDBC batch process. The analysis in Section 4.4.1 shows a significant increase in performance compared to the native peak persistence method, which includes a number of queries to insert only one peak entry with corresponding parameters.

#### 4.3.2 Data Retrieval and Reorganization into data sets

In a typical target data set, peak intensities are stored in a matrix, where peak regions correspond to columns and the corresponding MCC/IMS measurements to rows. Depending on the hypothesis, only measurements of particular sample objects should be included in the data set, i.e. hypothesis specific labels for each instance are required. Consider the following task: Differentiate all patients suffering from a certain disease (e.g. COPD,  $COPD=TRUE$  see Section 3) from a control group of all persons not suffering from this disease ( $COPD=FALSE$ ). Thus, all the respective instances are required to provide an associated value for the label  $COPD=\{TRUE, FALSE\}$ . To generate such data sets the data stored in EAV format is converted to a matrix of patients vs. attributes. This is done by applying a dynamic pivoting (cross-tabulation) strategy utilizing the PostgreSQL function *crossstab(text, text)* of the *tablefunc* module.

In a second step the peak information is inquired. Given the first partial data set of

objects, corresponding MCC/IMS measurements and the set of common peak regions for those measurements are extracted. A simplified query performing this task is shown in Listing 4.1. The query delivers only those peak regions common to each MCC/IMS measurement of the instances in the result set of patients.

Listing 4.1: Simplified hibernate query language (HQL) code for the retrieval of persistent *PeakRegion* entity identifiers in context with the generation of a target data set. Given a list of IMS measurement file names, the query returns only those *PeakRegions* which are associated to all (and not less) IMS measurements in the list. The semantic of the “group by - having” statement is explained in (122).

```
SELECT r.id
FROM
  ImsMetaData i join i.peaks p join p.peakRegions r
WHERE
  i.file in :fileNameList
GROUP BY r.id
HAVING count(i) = <fileNameList.size(>
```

### Data Set Analysis

In order to find relations between peaks and attributes such as pharmaceuticals or diseases, an automated analysis of a queried target data set is performed. The analysis incorporates an evaluation of the classification performance of a decision tree algorithm (WEKA, C4.5 decision tree), see Section 2.5.3 for details (182). Subsequently, the results are evaluated using ten-fold cross validation approach integrated in the WEKA library.

### Backup, Restore, and Versioning

To ensure data protection and to allow for recovery in the case of data loss, a backup and restore system was developed. The end-user can create backups of the current state of the IMS database and restore it to a previous state. This is ensured by triggering the PostgreSQL dump and restore command line tools in the graphical user interface. Integrating and restoring data may result in highly diverse changes. In order to track changes in the IMS database, it incorporates a custom version control system. Major changes, such as the integration of new MCC/IMS measurements or patient records, are logged and induce an update of the current database version.

## 4.4 Results and Discussion

The IMSDB application offers a graphical user interface (GUI) including several windows, tabs and dialogues to allow for an intuitive interaction between end-user and application. The following key features are currently supported by the GUI:

- Automated integration and validation of MCC/IMS data combined with object-attribute data.

- Presentation of database objects and their attributes, such as, for instance, patient records.
- Backup and restore platform with an integrated versioning.
- Dialogue-guided retrieval and automated decision tree analysis and validation of target data sets along with decision tree visualization.

#### 4.4.1 Database Upload

As explained in Section 4.3.1, the software provides a *CQPP* strategy, which reduces the overhead of individual operations concerning peak data upload coupled with batch inserts. Furthermore, the system is able to optionally store the raw data of MCC/IMS measurement files into the database. The running time of the previously described conservative and pivoting methods is compared using benchmarks datasets that reflect real world use-cases. Five data sets of different size, increasing in the number of measurements (25, 50, 75, 100, 125) as well as the number of peak entries (3-, 6-, 9-, 12-, 15-thousand), are uploaded to a clean database. The average MCC/IMS measurements file size is approximately 8MB. The benchmark is performed using the 64-bit database server PostgreSQL 9.0 (full logging activated) under 64-bit Windows 7 running on a Quad Core Intel Xeon CPU (2.4GHz) workstation with 24GB RAM and a SATA hard disk. Average running times of ten trials are reported in Figure 4.3.

#### 4.4.2 Decision Tree Classification and Performance Evaluation

As a proof of concept, the breathomics database application was evaluated using the COPD data set described in Section 3. The preprocessing of this data set resulted in 120 peak regions, which are evaluated for each MCC/IMS measurement. In this example, the goal is to find peaks that are related to the COPD annotation. The attributes of interest (*COPD=TRUE*, *COPD=FALSE:=["control"]*) are selected in the GUI. Subsequently, the software evaluates the data utilizing a ten-fold cross validation and the Weka machine learning package for decision trees (see Figure A.4 of the Appendix). The quality of the classification is validated by a set of performance measures shown in Table 4.4.2. See Section 2.5.3 for details. Finally, a decision tree is trained on the complete data set and the visualized in the GUI, see Figure 4.4.

### 4.5 Conclusion

The presented IMSDB is a comprehensive database and analysis platform capable of combining metabolic data, like MCC/IMS chromatograms, and heterogeneous biomedical data, e.g. patient records, into a centralized data structure. In contrast to the previous work of Lesniak (138), the ontology component enables the storage of constantly evolving object-attribute-data. In addition, the peak component ensures the fast storage and retrieval of metabolic data. As opposed to the open source database TrialDB (41), a database model was designed, that not only utilizes the EAV model but also features simple classes and relations. The combination of this design and the consistency-validation-

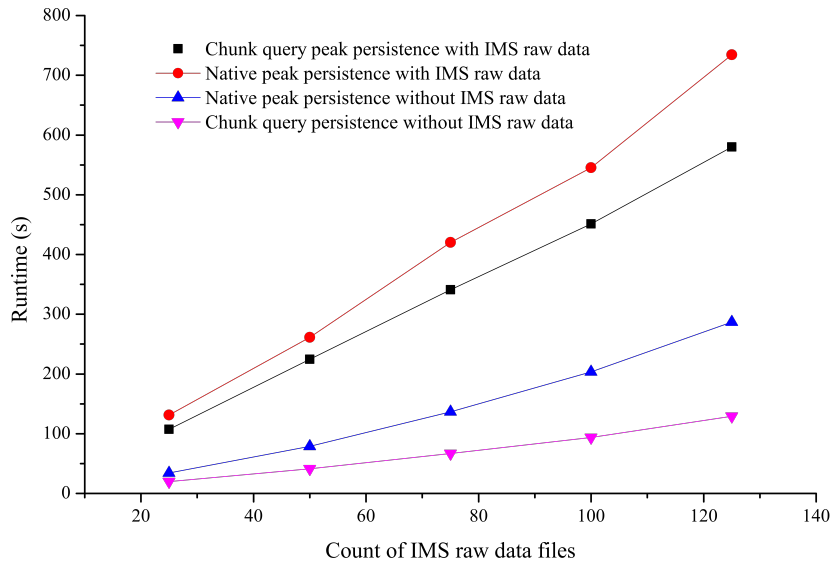


Figure 4.3: The plot depicts the average running times determined by ten repetitions of uploading data sets with an increasing number of IMS measurement files (25, 50, 75, 100, 125). Each raw file is associated with 120 peak regions. Four different uploading strategies are shown. The two methods CQPP (black line, pink line) and native peak persistence (red line, blue line) are thereby combined with the options for storing the IMS raw data measurement files into the database.

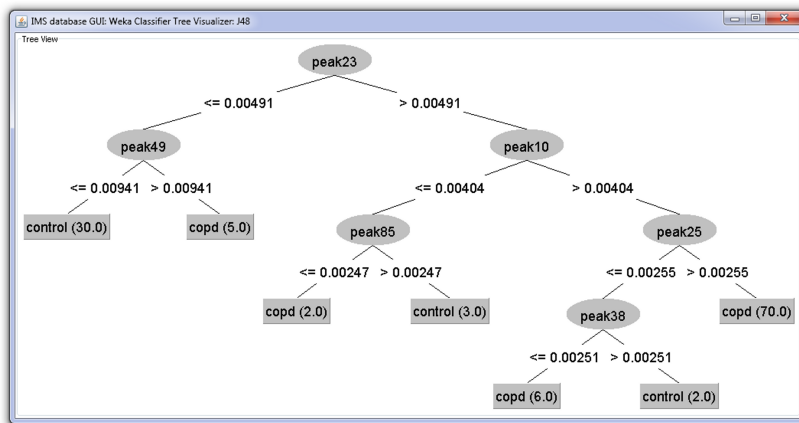


Figure 4.4: Illustration of the decision tree built by training on the data set that contains the combination of the labels (classes) COPD vs. Control.

and business-logic ensures the quality of the data, the relations and the object-attribute combinations. The extension of this design towards a full EAV/CR would decrease the

Table 4.2: Classification performance distinguishing COPD vs. control. The table on the left shows the confusion matrix and the table on the right the performance measures for the classification.

		Real	
		COPD	Control
Predicted	COPD	29	6
	Control	8	75

Measure	Performance
AUC	0.85
ACC	0.88
TPR	0.90
FPR	0.17

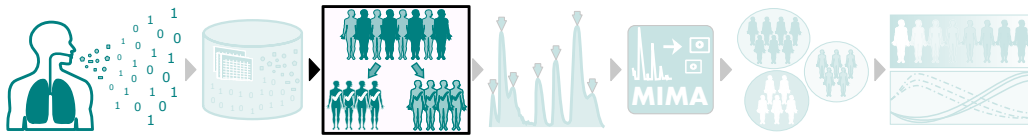
dependency on the application logic, but also decrease the performance of reading and writing in the system. In summary, the major contributions to the IMS biomarker community are:

- An intuitive software system for biologists, chemists, physicists and physicians that does not require any prior knowledge or technical skills.
- A general database model establishing the structure to combine heterogeneous biomedical and metabolic data for the purpose of research and diagnostics.
- Finally, this work provides an outlook on a potential biomarker discovery and validation platform by means of direct access to statistical learning methods.

The IMSDB is the combination of two powerful computational tools, an extensible breathomics database and a machine learning toolkit, accessible through an intuitive graphical user interface. This will accelerate and expand the opportunities of clinical diagnostic research in the near future. The work fills a gap that hindered efficient analysis of breath-based metabolomics in biomedical research. The IMSDB framework, the software package and an artificial test data set are publicly available at <http://imsdb.mpi-inf.mpg.de>.



## Chapter 5



# Towards Robust Machine Learning in Breathomics

*The statistical evaluation within this chapter was conceptualized and conducted by the author of this thesis. Prof. Dr. Jan Baumbach and Prof. Dr. Jörg Ingo Baumbach supported this project with their valuable insights on machine learning and breath analysis.*

The lack of sophisticated state of the art machine learning methods and robust evaluation, are the main reasons that prevent the successful application of breathomics in modern clinical diagnostics. This chapter evaluates the combination of MCC/IMS technology together with properly applied sophisticated machine learning techniques for (1) VOC-based feature selection and (2) supervised classification. In order to illustrate the potential of this combination, the previously described COPD data set serves as an example of an adequate analysis. To receive a broad overview of their potential for distinguishing the patient groups, several statistical learning methods have been integrated into the analysis. The content of this Chapter largely relies on a previously published study by Hauschild *et al.* in 2012.

### Objective:

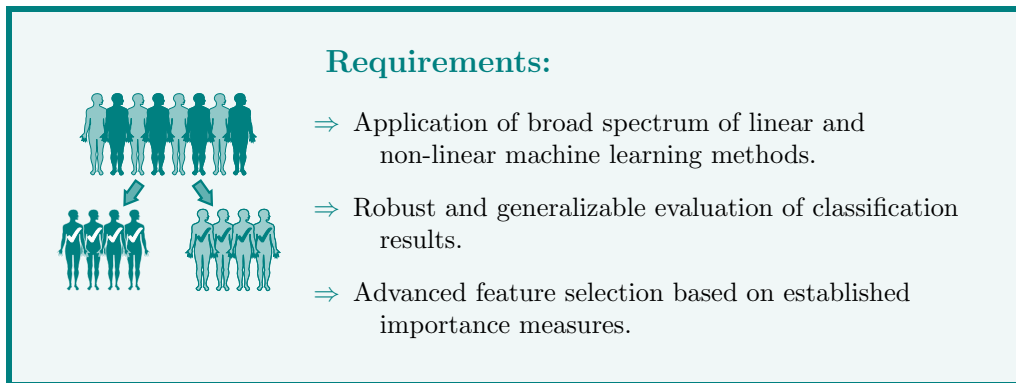
Proof of concept for the potential of sophisticated statistical learning methods for breathomics analysis.

## 5.1 Requirements and State of the Art

Typically, one of the major goals in breathomics is to identify the smallest subset of exhaled metabolites that can provide the most accurate and robust predictions regarding the clinical variable of interest.

Common statistical approaches in previous studies include basic methods for separability such as: hypothesis tests (e.g. Wilcoxon-Mann-Whitney test (WMW), (31; 79; 93; 197)), correlation analyses (148; 197; 129), and PCA (232; 114). Sporadically, more advanced learning methods demonstrating good predictability have been applied, such as decision trees (231), probabilistic relational learning (85), support vector machines (16; 142), or neural networks (8). Nevertheless, previous studies have mainly focused on separability rather than predictability, or otherwise fail to demonstrate the robustness of their prediction results. Thereby, these studies further fail to prove the reliability of the most important features selected. Moreover, the field has been lacking a comprehensive comparative analysis for a broad set of sophisticated methods in combination with MCC/IMS data to show their potential.

In contrast to previous studies, the focus of this work is a large-scale study of the potential in combining clinical MCC/IMS data and sophisticated statistical learning methods. In order to achieve a more robust estimate of the generalization error, results are evaluated by a cross validation scheme. Finally, a set consisting of the most informative features, based on established mathematical measures for feature importance are selected as potential biomarkers.



## 5.2 Methods

### 5.2.1 Data

The data set is previously described in chapter Materials as COPD data set, section 3.1. In total, the data comprises 119 volunteers: 35 healthy controls (HC) and 84 patients suffering from COPD, 54 out of the 84 COPD patients have also been diagnosed with bronchial carcinoma (BC). Unless stated otherwise, we use the following abbreviations as class labels:

**HC** = healthy controls (35)

**COPD** = patients suffering from chronic obstructive pulmonary disease (84)

**COPD+BC** = COPD patients also suffering from BC (54)

**COPD-BC** = COPD patients not suffering from BC (30)

### 5.2.2 Overview

First, the measurements are captured by the BioScout, an MCC/IMS device as previously described in Section 2.1, and the data set is saved on the attached computer (see Figure 5.1, step 1). Afterwards, the three-dimensional data files are preprocessed for noise reduction and VOC detection (Figure 5.1, step 2). Finally, a set of statistical learning techniques is applied and evaluated (Figure 5.1, steps 4 and 5) to estimate the classification accuracy. Additionally, some of these methods are able to identify the VOCs that are most relevant for the classification. To achieve the previously defined requirements with

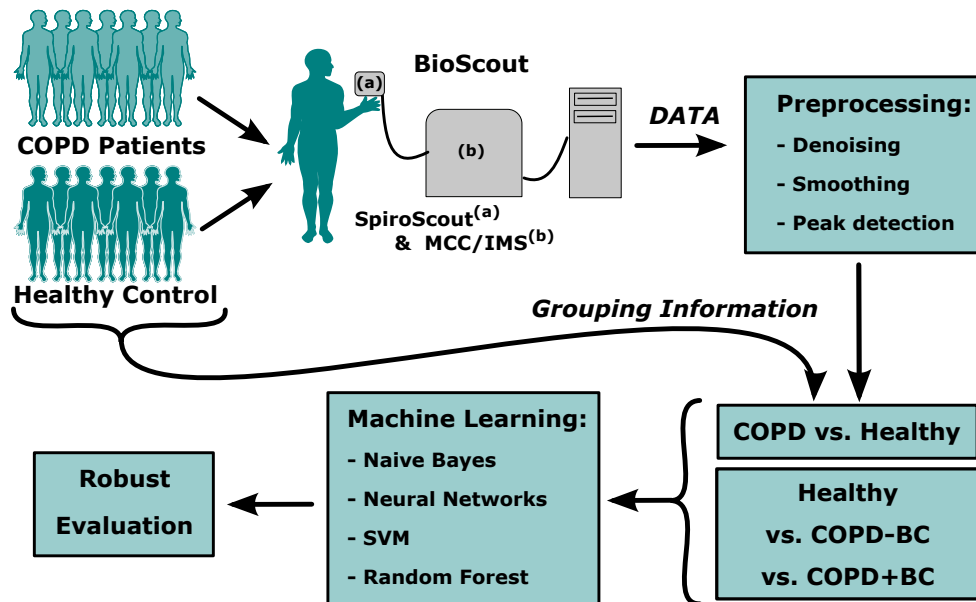


Figure 5.1: Overview of the integrated statistical approach to evaluate COPD-related metabolic MCC/IMS profiles. The numbering indicates the different steps of the approach.

respect to the described COPD data, the following two major questions are addressed (Figure 5.1, step 3):

1. Are advanced machine learning methods applied to breathomics data suitable to distinguish between healthy and COPD patients?

2. Is it possible approach the three class problem: healthy, COPD-BC, and COPD+CB, by combining breathomics data with the proposed set of tools?

### 5.2.3 Preprocessing

A MCC/IMS measurement contains 1 million intensities, 500 for the retention time axis and 2000 for the drift time axis. At first, general preprocessing methods for denoising, smoothing and peak detection, such as log-normal detailing and wavelet transformation must be applied, see Section 2.2 for details. Both preprocessing and peak detection is done with the aid of the VisualNow software (provided with BioScout by B&S Analytik), see Section 2.8.2. The software tool allows for both manual peak picking and execution of fully automatic peak finding algorithms. For this project, the data is preprocessed using the standard settings of VisualNow and the manual peak picking procedure. The domain expert determined 120 VOC positions. The final preprocessed data table consists of 119 measurements and the peak intensities for 120 compounds.

### 5.2.4 Methods

Several well established statistical supervised learning methods were selected, to obtain a broad overview of the potential of the data and the different classification techniques for clinical diagnostics. Thus, the project includes methods of varying complexity. On the one hand, long-established approaches were applied, such as decision tree, naive Bayes, and linear support vector machine (SVM). On the other hand, more recent and complex techniques were used: neural net, random forest, and radial SVM. The statistical analysis and feature selection was implemented in the R environment for statistical computing (Version 2.13.1) (1). A detailed description of the theoretical background of methods is provided in Section 2.5.3. A brief overview of the implementation and parameter settings is given in the following.

#### Naive Bayes

A widely used method for linear statistical modeling is known as the naive Bayes classifier (134). The naive Bayes method was employed using the R package *NaiveBayes* (227), with default parameters and the *usekernel* parameter set to “true”.

#### Decision tree

This method is a non-linear machine learning tool based on recursive partitioning. The decision tree implementation in the R package *rpart* (207) was used, with default parameters.

#### Artificial Neural Networks

A well established non-linear method is the artificial neural network. The basic model contains three layers: the input layer, the output layer, and at least one hidden layer. The complexity of the model increases by the number of hidden layers (102; 9). In our case, we set the number of hidden layers to 2. The weight decay and the maximal number

of iterations were kept at the standard settings 0 and 100, respectively. The R package used is called *nnet* (218).

### Random Forest

Random forest is based on the two strategies of bagging and bootstrapping, reducing the variance by building a forest of decorrelated decision trees (102). It provides two measures for feature importance: Gini index and mean decrease in accuracy. The random forest classification and feature selection were done using the *randomForest* R package, by Liaw and Wiener in 2002 (139), with standard parameters number of trees equal to 500, where each tree was grown to the maximum possible depth.

### Support Vector Machine

SVM is a widely used statistical learning method, especially in combination with a non-linear kernel like the Gaussian radial basis function. Both the linear as well as the radial SVMs were implemented with the *e1071* package (67). The *cost* and *tolerance* parameters of the linear SVM were set to 100 and 0.01 respectively, while the *cost* and *gamma* parameters of the radial SVM were fixed to 1000 and 0.1.

## 5.2.5 Statistical Analysis and Validation

After preprocessing and peak finding the previously described statistical learning methods are applied to solve two different classification tasks: COPD vs. HC and COPD-BC vs. COPD+BC vs. HC. However, the data set consists of a rather small number of samples, in our case 119. Thus, the evaluation using the ideal approach of training and test set as described in Section 2.7, would lead to relatively noisy estimates of the predictive performance. Therefore, cross validation is used to give a better estimate for the generalization error of the predictive model, see Section 2.7 on Validation for more details. Stratified cross validation was used to ensure that each subset covers instances from all classes, i.e. the proportions of the class labels are preserved within each subset. In this pilot study we are solely interested in getting a general idea of the suitability of MCC/IMS combined with statistical learning for clinical diagnostics. Therefore, we refrain from performing advanced parameter tuning of the learning methods. The R package *pROC* was used to compute various measures of model performance (183): sensitivity, specificity and the AUC, which is the area under the ROC curve. The class specific sensitivity and specificity of the three-class-problem are calculated with the equations described in Section 2.5.3.

**Feature Selection** Both random forests and linear support vector machines provide measures for feature importance. In order to investigate the feature importance of the set of 120 compounds, the Gini index of random forest and the coefficients fitted by the linear SVM model are evaluated for each cross validation (CV) run. Finally, the importance vectors are averaged over all ten CV models, for each of the features. As a proof of concept, the best ten features for each of the two methods will be discussed.

### 5.3 Results and Discussion

The first classification problem aims at distinguishing between COPD patients and the healthy control group. Table 5.1 shows the performance comparison of the results of the six different methods. The decision tree and naive Bayes, achieved an accuracy between 82% and 85% and an AUC of around 80%. The linear SVM performed slightly better with an AUC of 83% and an accuracy of around 87%. The more complex methods, i.e., neural net and radial SVM, gave an accuracy of 89% and AUCs of 86% and 87%, respectively. The best performing method classifying between COPD and HC was random forest, which had the best prediction accuracy of 94% as well as high values for AUC (92%), sensitivity (98%) and specificity (86%). As expected, methods with higher complexity performed better. They have a relatively low bias, which means they make less strong assumptions and fit the data more closely than the simpler methods. On the other hand, the basic methods performed surprisingly well in terms of AUC and accuracy. However, due to the unbalanced data set (COPD  $\approx$  70% vs. HC  $\approx$  30%), one has to take a closer look at the sensitivity and specificity. While the sensitivity was good (between 87% and 98%), the specificity of the more complex methods (80% to 85%) was in general higher. This indicates that for those methods, COPD patients are still accurately discovered, however, a larger percentage of healthy controls are falsely predicted to suffer from COPD. In clinical practice this increase in false positives could lead to false information for the patients and in the worst case to invasive follow-up examinations or harmful side effects of unnecessary medication.

Table 5.1: Results of the two-class-classification problem, evaluating the differences between COPD and the HC.

Method	AUC	Accuracy	Sensitivity	Specificity
Decision Tree	81	85	91	71
Linear SVM	83	87	92	74
Naive Bayes	79	82	87	71
Neural Net	86	89	93	80
Radial SVM	87	89	92	83
Random Forest	92	94	98	86

To the best of our knowledge, the presented work is the first comprehensive study about the performance of state of the art statistical learning tools for IMS-based metabolic profiling of COPD and bronchial carcinoma patients. A recent study presented by Westhoff et al. (2011) solely concentrated on rank sum tests and a decision tree trained on the VOC to classify COPD and HC. The best VOC 98, given by the rank sum test, was reported to result in a comparatively good training accuracy of 91%. However, the major issue of that study (also briefly discussed in (231)) is the focus on separability and training error as well as the lack of cross validation to avoid data overfitting.

**Receiver Operating Curves** The predicted classes of the supervised learning methods were used to evaluate the accuracy, area under the ROC curve, sensitivity and

specificity(103). In addition, some of the presented methods estimate the probabilities for each sample to belong to the one or the other class. Those results can be utilized to plot the so called receiver operating curve, see Section 2.5.3 for more details. Figure 5.2 shows the corresponding ROC curves for the given statistical learning methods.

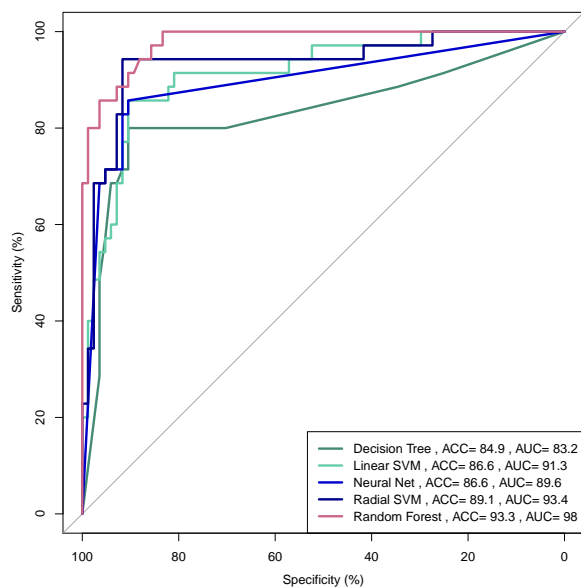


Figure 5.2: ROC curves of the statistical learning methods based on the estimated class probabilities.

### Feature Selection

In this classification setting, features directly corresponded to molecules in the human exhaled air. We are interested in finding those features that contribute most to the classification performance. Therefore, the coefficients of the linear SVM and the Gini index of the random forest were taken as a measure of importance. This enables to discover a linear as well as a non-linear feature pattern. Figure 5.3 shows the ten best features provided by the two methods linear SVM and random forest. The linear SVM feature importance depends on the coefficients of the variables according to their weight on the final linear model. The Gini index of the random forest is a summation of the decrease in node impurity over all trees. Tables 5.2 and 5.2 show the two resulting subsets of features.

Both ten best feature subsets did not overlap. This might result from the different underlying mathematical approaches, in one case a linear model, in the other a non-linear. Additionally, both subsets were compared to the peaks identified in the study of Westhoff *et al.* by using the rank sum test (231). Interestingly, five of the compounds found

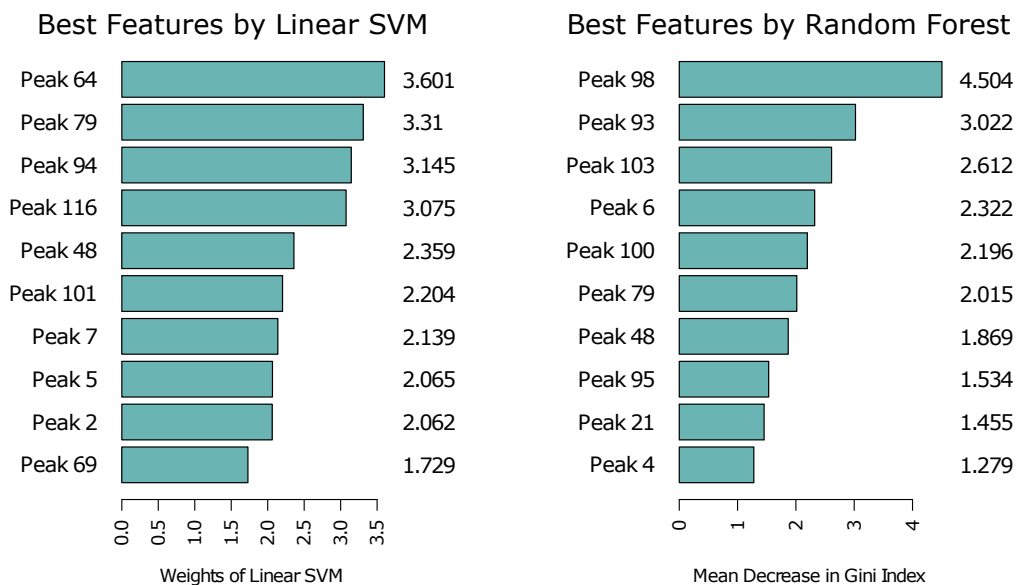


Figure 5.3: Barplot of the ten best features determined by the feature selection of the linear support vector machine and the random forest, on the COPD vs. HC classification. Depicted are the ten best features according to their weights generated by the linear SVM (left) and the Gini index provided by the random forest (right). The right Y-axis of each Figure lists the names of peaks/VOCs, while the left Y-axis states the importance.

Table 5.2: Comparison of the 10 best features selected by linear SVM to the features selected by random forest and the peaks identified by the study of Westhoff *et al.* 2011 using rank sum test (231). The  $\times$  sign indicates a match. The peaks are ordered by the rank of their weight in the linear SVM. Additionally, the coordinates in the MCC/IMS chromatogram, the inverse drift time ( $1/K_0$ ) and the retention time ( $RT$ ) is shown.

Peak Nr.	Linear SVM Selected Peaks									
	64	79	94	116	48	101	7	5	2	69
1/K <sub>0</sub>	1.119	0.844	0.514	0.55	0.693	0.932	0.563	0.544	0.535	0.904
RT	189.8	501.1	248.1	56.5	100.7	110.4	6.2	3.4	4.3	233.4
Linear SVM	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$
RF		$\times$			$\times$					
Westhoff, '11		$\times$			$\times$					

in this study have also been found to be important by random forest. On the one hand this proves the validity of the approach. However, it also indicates that further analysis of the compounds found by the variable selection performed is necessary, both for previously



mentioned peaks {48, 79, 98, 100, 103} and newly discovered peaks {93, 6, 48, 95, 21, 4}. Another study by Bessa *et al.* in 2011 analyzed a data set of 13 COPDs and 33 HCs,

Table 5.3: Comparison of the 10 best features selected by random forest to the features selected by linear SVM and the peaks identified by the study of Westhoff *et al.* 2011 using rank sum test (231). The peaks are ordered by the rank of the Gini index generated during the training of the random forest model. Additionally, the coordinates in the MCC/IMS chromatogram, the inverse drift time ( $1/K_0$ ) and the retention time ( $RT$ ) is shown.

Peak Nr.	Random Forest Selected Peaks									
	98	93	103	6	100	79	48	95	21	4
$1/K_0$	0.605	0.607	0.61	0.581	0.553	0.563	0.553	0.648	0.647	0.6
$RT$	22.3	16.9	18.9	78.5	49.6	20	17.9	17.6	21.9	11.9
Linear SVM						×	×			
RF	×	×	×	×	×	×	×	×	×	×
Westhoff, '11	×		×		×	×	×			

and reported a peak at position  $1/K_0 = 0.50$  and  $RT = 26$  to be the best discriminant for that classification scenario with 100% accuracy. However considering 120 VOCs and only 46 training and evaluation data points, does not allow for any save conclusions at this point (31).

### COPD+BC Classification

Table 5.4 depicts the classification results of the three-class problem. Prediction performance was low (accuracy  $\approx 70\%$ ) for each of the applied machine learning techniques, except random forest (accuracy  $\approx 79\%$ ). The AUC dropped by at least ten percent for all of the methods except naive Bayes, which may be due to its simplicity and robustness. Therefore, it remains unclear whether the information content in the data is high enough for distinguishing between the three groups of volunteers. In fact, all of the methods showed a very low sensitivity for the COPD class, which indicates that the differentiation between class COPD and COPD+BC is a difficult problem for all of the models. While the methods were still able to identify the HCs in a quite robust manner, most of the measurements of COPD patients were falsely predicted to suffer from both COPD and bronchial carcinoma, i.e. class COPD+BC. This was due to the fact that the number of BC patients was almost double the number of patients solely suffering from COPD. This is not surprising, since the cause of both diseases is highly dependent on the smoking behavior of the patients, and both are reported to be strongly related to each other. This is also supported by Young *et al.* (2009), where they identified COPD as a common and important independent risk factor for lung cancer. The prevalence of the different groups of COPD within lung cancer goes up to 60% (240). Consequently, we can assume that the probability for each of the COPD patients to get a bronchial carcinoma is comparatively high. Hence, we cannot exclude the possibility that an early stage bronchial carcinoma might have been undetected, particularly since most types of lung cancer are detected in late stages in the majority of cases.

Table 5.4: Results of the three-class-classification problem, evaluating the differences between COPD patients, COPD patients suffering from bronchial carcinoma, and the control group. The class specific sensitivity and specificity assessed for class COPD as well as COPD+BC, is based on the equations discussed in the methods section.

Method	AUC	Accuracy	COPD		COPD+BC	
			SEN	SPE	SEN	SPE
Decision Tree	70	60	23	82	69	65
Linear SVM	71	59	0	93	80	48
Naive Bayes	75	62	43	79	61	72
Neural Net	73	61	20	82	69	62
Radial SVM	73	62	0	91	78	57
Random Forest	79	67	6	99	85	55

Although there has been no study using modern machine learning methods on COPD MCC/IMS chromatograms, there were two studies for MCC/IMS data about BC patients. One study applied naive Bayes, multi layer perceptron, and SVM to a set of MCC/IMS chromatograms and achieved an outstanding performance (accuracy and AUC both 99%) (16). Despite the good results, the study has some limitations. At first, the prediction was done on a comparatively small sample and large feature set, where each chromatogram was separated by a grid, while each feature was calculated as the average intensity of the corresponding grid element. Secondly, the accuracy and AUC were evaluated on a single cross validation neglecting the possible variation between different selections of CV sets. Another study used relational probabilistic learning for the extraction of Markov logic network formulas and achieved a cross validation accuracy of up to 90% during classification (85). Unfortunately, a comparison with the results of the peak selection of those two methods is difficult, as the authors did not make the peak or grid positions publicly available. Besides, no further information on additional medical conditions of the patients, such as COPD, was known.

Although the evaluation of classification performance using repeated cross validation runs is an improvement to previous studies, we are still lacking more robust evaluations, such as permutation tests as presented in Chapter 6. Another problem for the analysis of COPD measurements is the medical COPD categorization itself. The patients were categorized with respect to the severity of their COPD disease, which is done according to a defined set of rules and tests. In addition, clinical practice and the subjective impression of the physician play an important role, as well as other factors that can influence the compounds in human breath, such as diet, smoking, and other secondary diseases, that have not necessarily been tested.

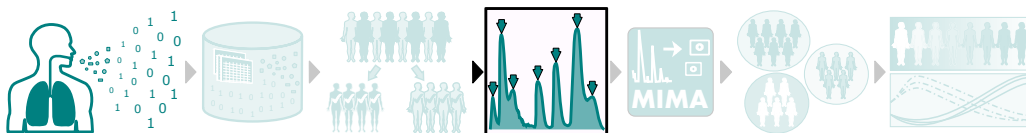
## 5.4 Conclusion

This study shows that ion mobility spectrometry data of human breath can generally be utilized for distinguishing between lung diseases such as COPD, if properly combined with

statistical learning methods. To demonstrate this, we evaluated sophisticated machine learning techniques on MCC/IMS chromatograms regarding their classification performance and ability to identify the most important molecular compounds. Therefore, the breath of 84 patients either suffering from COPD or both COPD and bronchial carcinoma was processed and compared with 35 healthy volunteers. The by far best test error (AUC 91%, ACC 94% for COPD vs. HC; AUC 79%, ACC 67% for COPD vs. COPD+BC vs. HC) were achieved with the random forest method. These results pinpoint a strong potential to separate healthy from COPD, but also suggest that a further examination of the differences between COPD and COPD+BC is needed. Linear SVM and random forest extracted 10 important VOCs, respectively. Moreover, five of these were confirmed to have discriminative power regarding COPD and healthy IMS chromatograms in other studies. Parameter tuning is indeed a valuable method to adjust the present statistical learning methods. However, it is only recommended on larger data sets and in combination with proper evaluation using permutation tests, for instance. In the future, we plan to determine whether these 20 molecules are biologically related to COPD and to eliminate those that are related to diet or other environmental influences. A further objective is to optimize the standardized learning methods in order to improve the prediction performance, on the one hand, and the identification of the smallest discriminating set of biomarkers, on the other hand. If such a set of biomarkers can be found it might lead to a tremendous progress in the field of COPD and cancer diagnostics. However, larger data sets of patients with pulmonary diseases and lung cancer are necessary.



## Chapter 6



# Evaluating Automated Peak Detection

*This chapter is based on a collaborative project with the Bioinformatics group at the technical university of Dortmund and funded through the Collaborative Research Center SFB 876 (SFB) and the International Max Planck Research School (IMPRS). The project partners Marianna D'Addario and Dominik Kopczynski contributed with the implementation and results of two automated peak detection methods: local maxima search and peak model estimation. The results of the manual peak picking were provided by B&S Analytik in Dortmund. The author of this thesis was primarily responsible for the study design, statistical analysis and evaluation of the results. Additionally the author contributed the results of the merged peak cluster localization and watershed transformation peak detection methods.*

Although sophisticated automated methods for peak finding exist, the majority of studies still rely on manual peak picking, which is the current gold standard approach. However, the daily increasing number of measurements, each with several dozens of potential peaks, exceeds the possibility of manual processing and increases the necessity of an automated data analysis and classification.

This chapter describes the evaluation of four state of the art approaches for automated IMS-based peak detection: LMS, WST with IPHEX, MPCL with VisualNow, and PME by PeaX. They are compared to the manual gold standard generated with the aid of a domain expert. To avoid a bias by previous knowledge of the expert on the disease specific peaks, the previously introduced anonymized breathomics data set of patients and controls is utilized for the comparison (see Section 3.3). In contrast to the previous intrinsic quality measures for peak detection, this study introduces two distinct qualitative criteria. First, established machine learning methods are utilized to systematically study the classification performance identifying patients on the bases of the detected peak lists

as features. The second criteria investigates the classification variance and robustness regarding perturbation and overfitting. This Chapter primarily relies on Hauschild *et al.* from 2013 (107).

**Objective:**

Comprehensive and robust evaluation of the power of available automated MCC/IMS peak detection methods in comparison to the manual gold standard.

**Outline** The next Section describes the state of the art in peak detection and further evaluates the requirements of a proper evaluation of these. The Methods Section 6.2 gives an overview about the preprocessing as well as the peak detection methods we evaluate. It further describes which machine learning and evaluation approaches we apply to assess the quality of the peak detection methods. Section 6.3 describes the comparison results and finally discusses these findings.

## 6.1 Requirements and State of the Art

One of the first steps in automatic high-throughput MCC/IMS data analysis is peak detection, as each so-called peak represents a specific analyte in the exhaled air (see Background Section 2.2.3). Such an automatic peak detection method should be as accurate as a human annotation but fast enough to cope with thousands of measurements. Several peak detection algorithms for MCC/IMS data have been proposed and described in the literature (180; 37; 49; 128). Previous to this study, these were generally evaluated using intrinsic quality measures, i.e., criteria that can be derived from the data, such as, for instance, the goodness of fit of the peak models to the initially measured data. However, for most projects that are searching for potential disease biomarkers it is important to optimize the set of detected peaks, rather than fit the shape of each peak mathematically as exact as possible. Therefore, the final objective is to find such an optimal peak set, such that the classification performance will allow us to distinguish between patients and controls in the best possible way. The strategy of this project is to evaluate four state of the art peak detection methods according to such data-independent criteria. Hence, these criteria reflect the performance and robustness of the resulting peak lists, regarding the final task: biomedical decision making.

A peak detection method transforms a raw data measurement (a matrix of values, see an example heat map representation in Figure 2.2) into a set of peak descriptors. These will be used as input features for two different classification algorithms, a linear support vector machine and a random forest model. Subsequently, each peak detection method will be evaluated by the corresponding qualitative criteria: the classification performances and the robustness of the models. Finally, an MCC/IMS data analysis expert, will manually select the peaks in the set of measurements. The resulting peak set as well as qualitative criteria will be compared to the results of the automated methods.

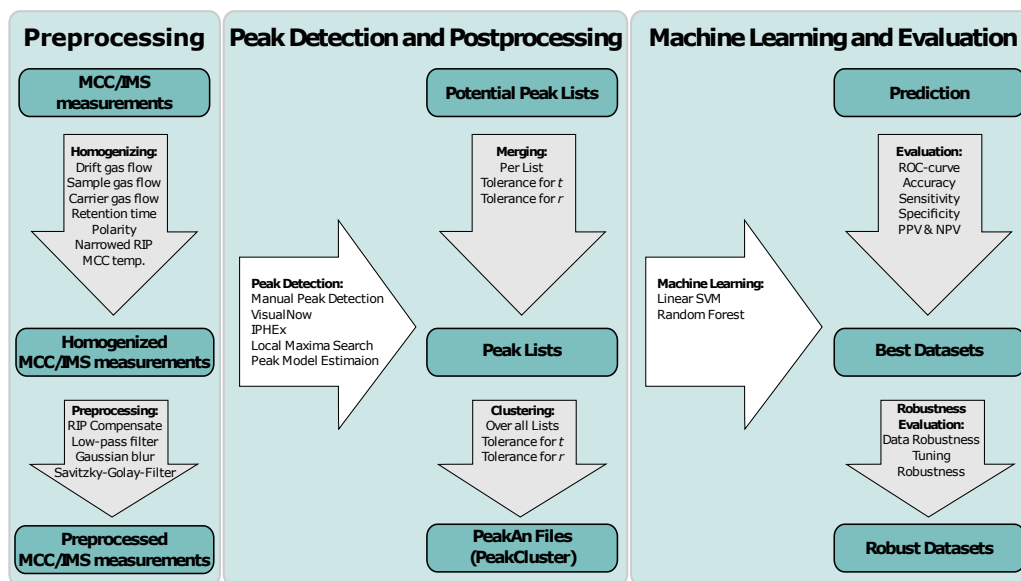
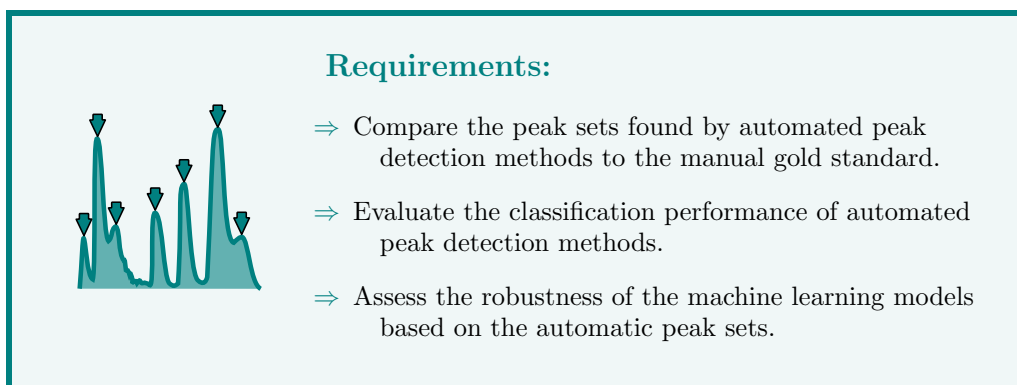


Figure 6.1: Shown is the structure of the evaluation pipeline. The three sections (1) preprocessing, (2) peak detection and postprocessing, and (3) machine learning and evaluation describe the details of the overall analysis. The figure is adapted from Hauschild *et al.* 2013 (104)



## 6.2 Methods

To evaluate the quality of five peak finding approaches, the previously introduced anonymized data set of human exhaled air is utilized, see Section 3.3 for more details. Figure 6.1 gives an overview of the following sections and the structure of the comparison process. At first, preprocessing approaches such as homogenization, filtering and individual preprocessing steps are described. In the second step the peak detection and merging approaches are applied to extract a feature matrix of measurements vs. metabolites. Subsequently, we

describe the evaluation scheme and the used machine learning methods.

## 6.2.1 Preprocessing

### Homogenizing and Filtering a Set of Measurements

To ensure comparability within the set of measurements that potentially originate from varying sources or hospitals, we define a list of rules and restrictions the measurements have to satisfy.

- Drift gas flow:  $100 \pm 5$  mL/min
- Sample gas flow:  $100 \pm 5$  mL/min
- Carrier gas flow:  $150 \pm 5$  mL/min
- MCC temperature:  $40 \pm 2$  °C
- Drift gas: the same value for all measurements in the set
- Polarity: the same value for all measurements in the set

This step determines the measurement with the smallest retention time range, the so-called cut-off time. Subsequently, in each MCC/IMS measurement, the set of measured spectra is reduced to those before the cut-off time. The aim of these rules is to ensure that all measurements in our set are generally comparable. Since most signals occurring at retention times smaller than 5 seconds originate from the MCC/IMS device itself they were discarded as well.

### Preprocessing MCC/IMS Raw Data

In order to reduce the noise and increase the quality of the raw measurements, the following four preprocessing steps are applied: The first step is the RIP detailing or RIP compensation filter, which subtracts the global median of each MCC spectrum from each data point. To further denoise the data, a low-pass and Gaussian filter as described in Section 2.2 are applied. They remove high frequencies and smooth the spectrum. Finally, a one-dimensional Savitzky-Golay filter smooths every single spectrum computing a weighted average across the drift time axis with a window size of 9 data points (192).

## 6.2.2 Peak Detection and Postprocessing

In order to evaluate the peak detection approaches, they are applied to the preprocessed MCC/IMS measurements. The current “gold standard” is the manual peak detection Visual Now done by a domain expert, see Section 2.2.3. This chapter will compare it to the four most common automated approaches:

**LMS** Automated local maxima search, Section 2.2.3

**WST** Automated peak detection via water shed transformation implemented in IPHEX, Section 2.2.3

**MPCL** Automated peak detection via merged peak cluster localization supported by VisualNow, Section 2.2.3



**PME** Peak model estimation approach is the focus of the PeaX tool, Section 2.2.3

All of them transform the given raw measurements into a list of potential peaks. A peak  $P$  is defined as a triple  $P = (r, t, s)$  containing the following parameters: retention time  $r$ , inverse reduced mobility  $t$  and signal intensity  $s$ . See Section 2.2.3 in the Background Chapter. In this chapter a set of multiple MCC/IMS measurements is given. Therefore, we extend this peak description by the measurement index  $i$  resulting in  $P = (i, r, t, s)$ . For convenience, we use the notations  $i(P) := i$ ,  $r(P) := r$ ,  $t(P) := t$ ,  $s(P) := s$  to define the projections on each component. Finally, we arrive at a list of peaks for each method and each measurement, i.e. each patient.

Subsequently, for each peak detector, the peak lists are merged to create a matrix of peaks and patients that provides the intensity of the associated peak. Implicitly, this matrix describes whether we observe a certain peak in a certain measurement (patient) or not, and if so with which intensity. It can further be interpreted as a list of feature vectors, which can be utilized for subsequent classification (next Section). Therefore, we need to (1) correct for peaks that are too close to each other to be two separate peaks in reality but should be merged to a single feature. We further need to (2) account for peaks that occur in different measurements at “slightly” different positions, henceforth referred to as “peak clusters”.

The peak merging method described by Boedecker *et al.* (37) is applied for this purpose. See Chapter Background, Section 2.2.4 for details. Note that this procedure is applied solely to the automated peak detection results. The manual peak detection with Visual-Now directly resulted in a list of peak clusters for each measurement.

### 6.2.3 Evaluation Methods

The result of the previous processing steps is a set of peak clusters for each peak detection tool and the intensity of the corresponding peak cluster for each measurement. Under the assumption that one peak cluster originates from one specific compound or metabolite, further bioanalytical techniques can determine the metabolite corresponding to each peak cluster. In modern biomarker research, we essentially seek to find correlations between those metabolites and a certain disease. This suggests a conditional relationship between the metabolic pathways of the disease and the produced indicator components. Therefore, it is desirable that a peak detection method is able to find all relevant peaks that trigger a good and robust classification performance. To compare the results of the different peak detection methods, two different steps are performed. First, the overlap of the peak lists and peak clusters of the compared methods are assessed. In a second Step the mentioned qualitative criteria are evaluated, based on the classification performance of the different detection methods.

#### Peak Position Comparison

Recall that a peak in a measurement is described as  $P = (i, t, r, s)$  (see Section 6.2.2), where  $i$  is the measurement index,  $t$  the inverse reduced mobility,  $r$  the retention time and  $s$  the intensity. The different peak detection methods result in different peak cluster

positions such that simply studying their intersection is infeasible because two peak clusters that refer to identical molecules would only be recognized as the same if the clusters had exactly the same coordinates. This is very unlikely. To overcome this problem, an adapted version of the merging conditions defined in Section 6.2.2 are created. Let  $V$ ,  $W$  be two peak lists generated by two different peak detection methods. Two peaks clusters  $P \in V$  and  $Q \in W$  of the same measurement  $i(P) = i(Q)$  are mapped (considered identical), if they fulfill the conditions defined in Section 6.2.2. The overlap of list  $V$  with list  $W$  is defined as the number of peaks in  $V$  that can be mapped to at least one peak in  $W$ . The resulting mapping count table is not symmetric, since each peak of list  $V$  can be mapped to more than one peak from list  $W$ .

### Machine Learning and Evaluation

The result of the postprocessing step is a feature matrix containing intensities of peak clusters that presumably resemble the abundances of the corresponding molecules for each of the patient measurements. These measurements or samples are each assigned by one of the two class labels:  $K$  for control or  $D$  for diseased.

Two distinct standardized machine learning methods are trained on this data to predict these class labels. The goal is to get an overview of the overall potential of the peak detection methods in combination with different classification strategies. On the one hand we chose a linear learning technique, namely the linear support vector machine. On the other hand the non-linear random forest method is selected to evaluate potential non-linear dependencies within the data. See Section 2.5.3 for a more detailed description of the methods.

**Evaluation:** In order to achieve a robust estimation of the classification quality, the evaluation embedded a ten-fold cross validation (CV) environment. In settings with comparably small data set sizes, splitting the data into training-, validation- and test-set leads to relatively noisy estimates of the predictive performance. Therefore, CV is used to give a more accurate estimate for the actual accuracy of the predictive model. To ensure that each subset covers the variety of both classes, the classes are balanced within each CV subset. Furthermore, the CV procedure was repeated 100 times using 100 different ten-fold cross validation sets. Thereby, the robustness of the different peak sets towards changes in the measurement set can be analyzed.

The classification performances are evaluated based on the feature matrices emerging from the five peak detectors by using different quality measures: (1) accuracy ACC, (2) the AUC, which is the area under the receiver operating characteristic curve (183), (3) the sensitivity, (4) the specificity, (5) the positive predictive value (PPV), and (6) the negative predictive value (NPV). Furthermore, we give mean and standard deviation of the AUC in boxplots.

Finally, we investigate the robustness and analyze, whether the feature sets and their model performance are susceptible to classification parameter tuning, using a single ten-fold cross validation set. We will pick the classifier that performs worse since the potential for improvement will be higher. The parameters are varied systematically. In addition, we randomize the class labels allowing us to judge the robustness to small parameter

changes on both: (1) the original class labels as well as (2) the randomized class labels (expected to lead to a decreased classification performance).

## 6.3 Results and Discussion

At first, the MCC/IMS measurements are preprocessed utilizing the previously described homogenization and quality improving preprocessing steps such as denoising and smoothing. The original data set contains 69 measurements of which two are discarded during the homogenization phase. Subsequently, the peak detection methods are applied and evaluated. Table 6.1 gives an overview on the results of the postprocessing.

Table 6.1: Number of peaks detected by all methods. Number of peak clusters after merging the peak lists (postprocessing).

Method	# Peaks	# Peak Clusters
Manual (VisualNow)	1661	41
LMS (PeaX)	1477	69
MPCL (VisualNow)	4292	88
WST (IPHEX)	5697	420
PME (PeaX)	1358	69

After merging the overlapping peaks of the peak lists, the MPCL (VisualNow) and WST (IPHEX) methods show the by far largest number of peaks, between 4,000 and 6,000. The manual peak picking, LMS as well as the PME (PeaX) methods find a similar amount of peaks, “only” about 1,500. The number of peak clusters is almost constant over all methods, it varies between 40 and 90. An exception is the IPHEX implementation of the WST peak picker, which finds 420 clusters. In theory, the underlying WST algorithm provides different parameters to tune. However, the IPHEX implementation does not support a change of these parameters and hence results in a huge number of peaks and peak clusters. For both VisualNow-based methods, manual as well as automated, the ratio between potential peaks and resulting peak clusters is comparably high. The reason lies within the VisualNow implementation. Once a potential peak was found in one measurement (out of the 67), it automatically assumes the presence of a peak at this position in all other 66 measurements (presumably with low intensities), even if no actual peak exists. This results in the observed high number of peaks that mainly correspond to noise. This, as demonstrated later, may lead to problems within the classification procedure. In contrast, the IPHEX, LMS and PME implementations only assign intensities to peak clusters for those measurements where a peak at the corresponding position was detected.

### 6.3.1 Peak Position Comparison

The comparison of the peak cluster lists between the different peak detection methods results in a matrix of overlaps, which is summarized in Table 6.2. The comparison of

Table 6.2: Overlap of the five peak detection methods. The overlap of the peak list  $A$  (row) and peak list  $B$  (column) is defined as the number of peaks in  $V$  that can be mapped to at least one peak in  $W$ . Note that the resulting mapping count table is not symmetric.

Method	Manual	LMS	MPCL	WST	PME	Software
Manual	1661	911	1522	1184	791	VisualNow
LMS	868	1477	1096	1074	1128	PeaX
MPCL	2667	2233	4292	2341	2082	VisualNow
WST	1112	1009	1157	5697	912	IPHEX
PME	737	1086	983	926	1358	PeaX

the peak cluster lists shows a large similarity of both peak lists created with VisualNow, i.e. most peaks found in the manual evaluation were also identified with the automatic MPCL peak detection of VisualNow. The IPHEX water level approach creates a huge set of peaks. Hence it finds many peaks detected by the other approaches as well ( $\approx 70\%$  on average). Nevertheless IPHEX appears to be less redundant than the automated VisualNow method since hardly any peaks from the other sets occur several times within the IPHEX set. LMS and the PME method overlap highly (in both directions,  $\approx 80\%$ ).

### 6.3.2 Evaluation via Statistical Learning

The evaluation of the classification performance is shown in Table 6.3 and Table 6.4. Table 6.3 presents the results of the linear support vector machine indicating that all methods perform almost equally well. The manual, LMS and MPCL peak detection methods perform worst, in terms of AUC as well as accuracy. The automatic peak detection in IPHEX shows a slightly better AUC and performs best in terms of 73% accuracy. The peak detection method that produced the most informative features for the linear method in terms of AUC  $\approx 82\%$  is the PME approach.

Table 6.3: Classification results of the linear support vector machine. The quality measures are the AUC, accuracy (ACC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

Method	AUC	ACC	Sensitivity	Specificity	PPV	NPV	Tool
Manual	77.4	70.9	69.7	72.4	75.7	65.9	VisualNow
LMS	77	67.8	70.6	64.4	71	64	PeaX
MPCL	76.6	68.3	66.8	70.1	73.4	63.1	VisualNow
WST	79.8	73	70.5	76	78.4	67.6	IPHEX
PME	82.2	72.2	77.2	66.1	73.7	70.1	PeaX

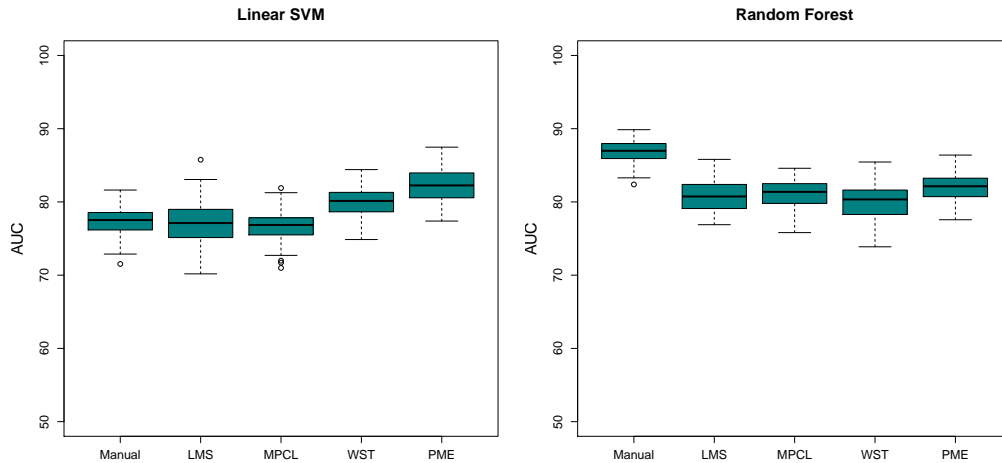


Figure 6.2: Boxplots of 100 runs of the ten-fold CV for the linear SVM and the random forest method.

Table 6.4 shows the classification results of the random forest method. Again, all methods vary little in their performance. The best set of features for this machine learning method was generated by the gold standard (Manual VisualNow). The manual detection shows an accuracy of  $\approx 76\%$  and an AUC of  $\approx 87\%$  and also outperforms all other peak detection methods in all of the quality indices. The peak model estimation performs slightly better in terms of AUC  $\approx 82\%$  and accuracy  $\approx 74\%$ , as well as most of the other measures.

Table 6.4: Classification results of the random forest. The quality measures are the AUC, accuracy (ACC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

Method	AUC	ACC	Sensitivity	Specificity	PPV	NPV	Tool
Manual	86.9	76.3	78.7	73.4	78.5	73.6	VisualNow
LMS	80.8	70.5	75	64.9	72.5	67.8	PeaX
MPCL	81.1	71.9	75.6	67.3	74.1	69.1	VisualNow
WST	80	68.9	72.8	64	71.4	65.6	IPHEX
PME	81.9	74.2	81.6	65	74.2	74.1	PeaX

### Data Robustness

Figure 6.2 shows boxplots of the list of AUCs generated by 100 runs of the ten-fold cross validation. The prediction results of the linear SVM with the manual and automated MPCL in VisualNow are the most stable, while the LME shows the highest variation.

The PME approach has a reasonable robustness and performs better than the simple methods in almost all runs. In comparison, the AUC-measured classification performance with random forest are most robust for the gold standard and the PME approach. The other automated methods introduce larger variations, in particular IPHEX.

### Tuning Robustness

Finally we investigate, if the feature sets and their model performance are susceptible to parameter tuning on both learning methods. Initially the analysis will focus on the worse performing classifier: the linear SVM. Therefore, we systematically vary the cost and tolerance parameters ( $\{0.1, 1.0, 100, 1000\}$  and  $\{0.01, 0.1, 1\}$ , respectively) and in a second run we randomize the class labels. The results of this analysis are shown in Figure 6.3, which plots the variance of the AUC for both the original labels (left) as well as the randomized labels (right). Figure 6.3 shows the results of the robustness analysis of random forest and linear SVM. Here the parameters of the random forest, number of trees ( $n_{tree} = \{100, 200, 300, 400, 500\}$ ) and number of features randomly sampled at each split ( $m_{try} = \{5, 10, 20, 30, 40\}$ ) are systematically varied. This parameter tuning was applied to the real data set as well as a data set containing randomized the class labels. The parameters tuned for the linear SVM are cost ( $cost = \{0.1, 1, 100, 1000\}$ ) and tolerance ( $cost = \{0.01, 0.1, 1\}$ ) parameter.

At first glance, Figure 6.3 indicates that the performance (AUC) of the manual and MPCL, as well as the IPHEX WST peak detection feature set can be heavily improved by tuning the parameter of the linear SVM classifiers. However, when considering the results for the randomized labels, these three tools seem to generate peak clusters that are prone to overfitting, most likely resulting from the high number of detected potential peaks. We would generally expect to observe a drastic drop in the classification quality for the randomized labels compared to the real labels, which is not clearly observed for all methods (overfitting), but LMS and PME. In addition to the comparably low susceptibility to overfitting, PME has a quite small variability in AUC indicating stable classification results. In contrast to the results of the linear SVM, the random forest tuning results show that this method is considerably more stable. Most data sets show considerable smaller potential for tuning of the random forest. Furthermore, for all data sets a drastic drop in classification quality for the randomized labels compared to the original labels can be observed. A classifier trained on a randomized class labels, is expected to result in random predictions. Therefore, the present drop is a sign for a reasonable predictor. One can generally say that random forest classification appears to be more robust in terms of overfitting on this data set.

## 6.4 Conclusion

To summarize, four different approaches for automated peak detection on medical MC-C/IMS measurements were compared: LMS, automated MPCL (VisualNow), IPHEX WST, and PME of PeaX in comparison to the gold standard, manual peak picking. In particular, their impact on the goodness of classical machine learning approaches that are used for separating patients into “healthy” and “not healthy” was investigated. This is

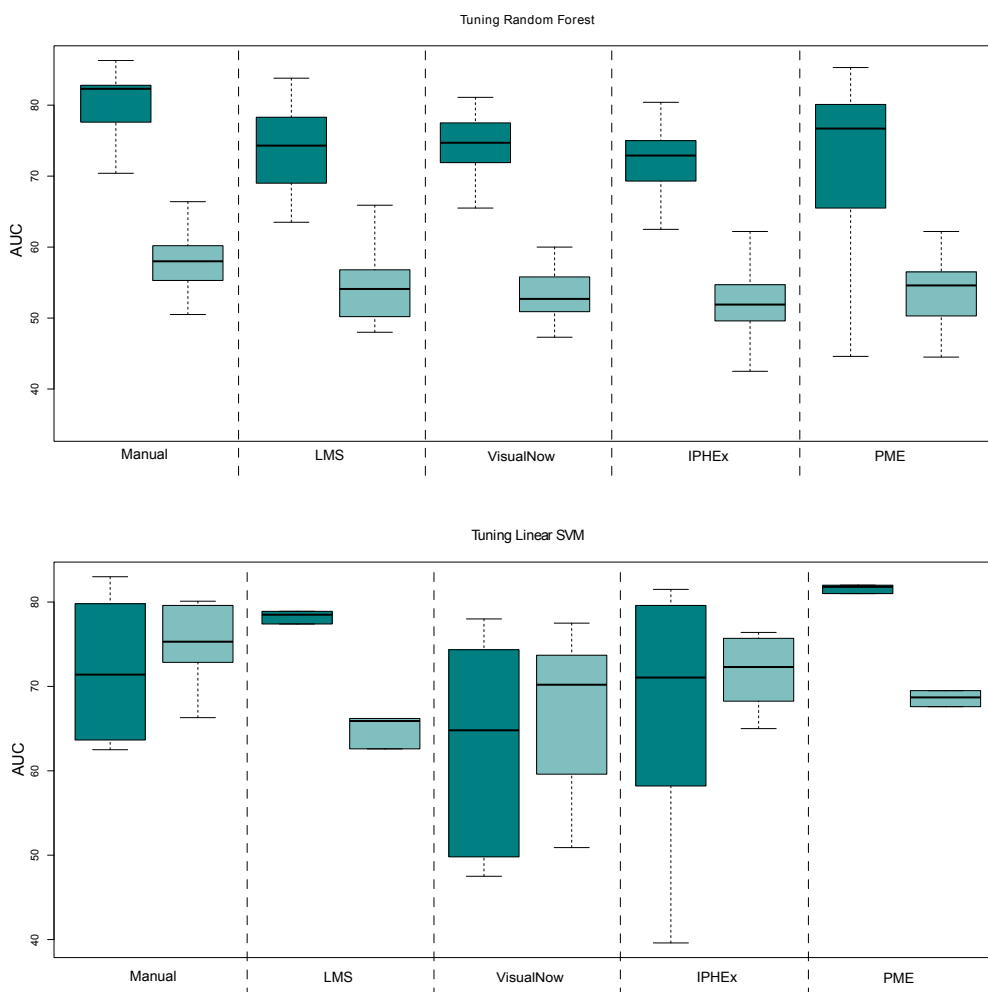


Figure 6.3: Boxplots illustrating the variation within the random forest and linear SVM tuning results on a single ten-fold cross validation set respectively. The dark cyan boxes show the results when tuning the original feature sets. The light cyan boxes show the results when tuning the randomly labeled feature sets.

crucial in current biomarker research since it hugely influences medical decision making. Our results indicate that the automated approaches can generally compete with manual peak picking protocols carried out by experts in the field, at least with regard to subsequent health status classification. Although the manual peak picking remains the gold standard, in medical studies yielding huge amounts of data one has to weigh the trade off between a slightly higher accuracy (manual) and a huge increase in processing speed (automatic). For example, a set of 100 MCC/IMS measurements is processed in less than

ten minutes, whereas an expert would need more than ten hours.

However, automated peak detection methods would process every kind of input, whereas a domain expert would immediately recognize erroneous data. Although all methods perform almost equally good, the quite recently developed peak model estimation approach slightly outperforms the other automated methods. PME is also robust against overfitting in all tested learning approaches.

We conclude that all current automatic peak picking methods provide generally good results. They are quite sensitive but might be improved in their specificity by optimizing the number of peaks they predict per measurement. This will, however, be difficult to implement while keeping the comparably high sensitivity rates.



## Chapter 7



# Metabolite Identification with MIMA

*This study was conducted in a collaborative fashion with Dr. Felix Maurer and Kathrin Eisinger of the microfluidics & clinical diagnostics department at the Korean Institute of Science and Technology (KIST). Kathrin Eisinger was responsible for the generation of experimental MCC/IMS and GC/MS data. Dr. Felix Maurer conducted the manual data analysis and comparison to the automatically mapped data. The author of this thesis developed the MIMA software package for automated metabolite identification based on parallel GC/MS measurements and the MCC/IMS substance database.*

The previous chapters focused on the detection of peaks within MCC/IMS measurements and the selection of those peaks suitable as potential biomarkers. The identification of such molecules is crucial to distinguish between those that originate from confounding factors or are directly related to a disease and can give further insights to molecular background. To address this challenging goal in complex samples such as human exhaled air, a reference database is necessary. Additionally, parallel measurements of GC/MS and MCC/IMS technology may improve the accuracy of the analyte identification. The MIMA (MS-IMS-Mapper) software tool was developed to automatically identify MCC/IMS peaks or generate a mapping between a MCC/IMS chromatogram and the corresponding GC/MS data. It was previously published in the shared first author paper by Hauschild and Maurer in 2014 (152), which is the main source of this Chapter. It demonstrates the power of MIMA by successfully identifying the analytes of a 7-component mixture.

**Objective:**

Develop a fast and easy computational method for assigning analyte names to yet un-assigned signals in MCC/IMS data

**Outline** In the following, the automated workflow for processing GC/MS and MCC/IMS measurements is introduced to analyse a same sample performed in parallel. After discussing the state of the art and requirements of such a system, the section will explain the experimental setting used to demonstrate MIMA's functionality in detail: An analyte mixture of 7 reference compounds will be analyzed by using MCC/IMS directly and additionally via TENAX tubes and thermo desorption GC/MS. The Section 7.2 will outline the workflow behind MIMA's automatic mapping procedure. A comparison between MIMA and other previous methods and tools will be presented in the discussion.

## 7.1 Requirements and State of the Art


The most straightforward identification process is to utilize an MCC/IMS reference database and manually match the entries to detected peaks. However, existing MCC/IMS reference databases are still in their infancy and do not currently identify all analytes.

As mentioned earlier, another common approach relies on supplementary GC/MS measurements of the same air sample. Decades of GC/MS research led to automated analysis software for both GC/MS preprocessing, peak detection as well as sophisticated identification. The mass spectrometric technology provides specific fragment profiles and therefore enables the identification of compounds using various commercial databases. The largest and most widely used is the NIST/EPA/NIH mass spectral library, which was developed by the NIST and contains several hundred thousand entries. In contrast, the MCC/IMS technology has developed more recently and, at time of publication, only preprocessing and peak detection methods have been automated.

The MCC/IMS peak identification via corresponding GC/MS data requires a parallel analysis of both, GC/MS and MCC/IMS data in many different manual steps. First the data is analyzed with the firmware of the GC/MS manufacturer and compared to the NIST-library (version 2.0, 2011)<sup>1</sup>. The first hits of the NIST-library search are subsequently screened by hand according to certain criteria for promising analytes. A typical exclusion criterion is whether a molecule is non-volatile and cannot occur in an exhaled air sample. Finally each of these analytes is searched for, within the MCC/IMS measurement, utilizing the MCC/IMS firmware and a given IMS reference database (MCCIMS-DB) file.

To overcome this time-consuming and tedious manual procedure of peak identification, the developed MIMA software tool has to fulfill the following requirements:

<sup>1</sup>see <http://webbook.nist.gov/chemistry>



**Requirements:**

- ⇒ An intuitive graphical user interface for the identification of VOCs detected by MCC/IMS.
- ⇒ Main Tasks:
  - I) Finding the closest known molecule in the given MCC/IMS database.
  - II) Mapping the MCC/IMS peaks to molecule peaks detected within a parallel GC/MS measurement.

## 7.2 Implementation

The suggested procedures rely on the previously developed methods for preprocessing and peak detection.

### 7.2.1 MCC/IMS - MCC/IMS Mapping

The first approach solely requires the MCCIMS-DB file. For each MCC/IMS peak, it searches for the closest fit within the given database file. We introduce four different definitions for closeness that the user can choose:

1. Minimal Euclidean distances between the coordinates.
2. Maximal overlap of the two defined peak regions.
3. Equivalent to option 1. and maximal overlap in case the distances are indecisive.
4. Equivalent to option 2. and minimal Euclidean distance in case the overlap is ambiguous.

### 7.2.2 MCC/IMS - GC/MS Mapping

The second approach relies on the parallel GC/MS measurements. Here, the whole process of NIST-library extraction and IMS reference database lookup can be organized in the following two main automation steps, see Figure 7.1.

**Step 1:** In the first step of the workflow, the OpenChrom software (version 0.8<sup>2</sup>, BMC Bioinformatics, 2010) is utilized to preprocess the GC/MS data (228). OpenChrom is an open source application for chromatography and mass spectrometry, which is capable of reading and analysing different kinds of MS data formats (e.g. GC/MS,

---

<sup>2</sup>see <https://www.openchrom.net/>

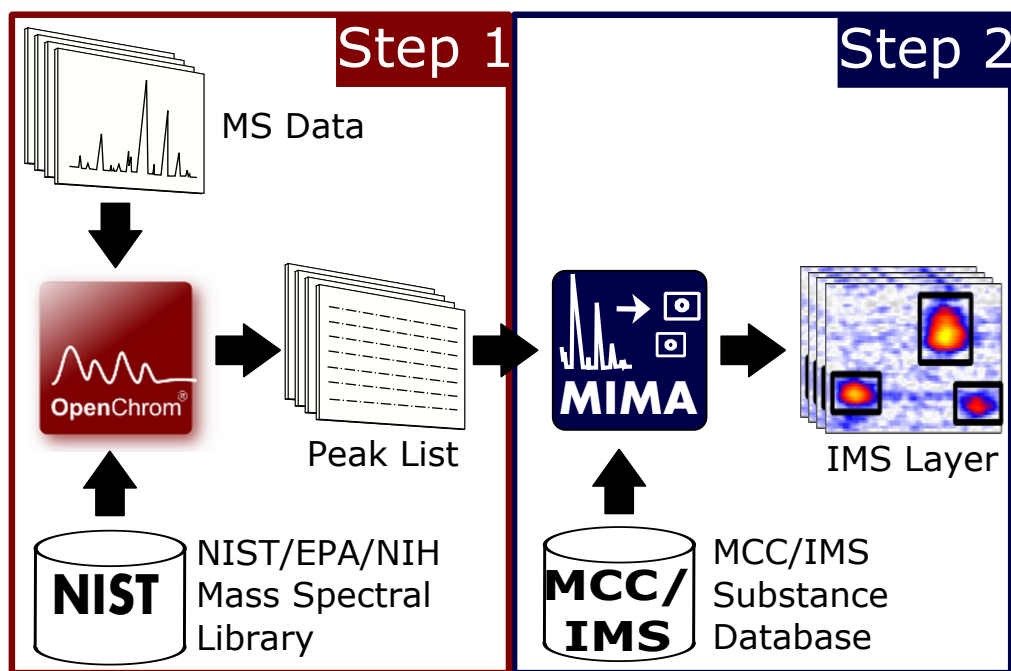


Figure 7.1: Schematic view of the step by step process

LC/MS, Py-GC/MS, HPLC/MS) (228). The software provides a batch functionality, to automatically process the chromatograms as follows:

- A) Automatic peak finding within the chromatogram. (In our example, a GC/MS measurement is utilized).
- B) The signature of each peak is matched against the NIST-library and annotated by a predefined number of most likely candidate molecules.

Finally, the procedure exports the list of MS peaks and a list of best matching NIST-library candidates for each peak.

**Step 2:** In the second step of the workflow, the exported files are processed by the MIMA software. For each MS peak in the list, the following steps are performed:

- A) The corresponding NIST-library entries and their chemical abstract service number (CAS-number) are extracted.
- B) Each CAS-number is matched to the entries within the IMS reference database file.
- C) The IMS references successfully mapped to a CAS number are added to the final IMS peak list.

Once all MS peaks are processed, the duplicate list entries are removed and the IMS peaks are saved to a new file, named IMS layer. An IMS layer mainly contains a list of analytes and the corresponding coordinates within the MCC/IMS-chromatogram (retention time RT and inverse reduced mobility  $1/K_0$ ), as well as the radii representing the typical peak expansion in both directions. This information is extracted from the MCC/IMS reference database file.

Note, the NIST-library contains non-volatile substance which might occur as first suggestion in the search. Therefore, the MIMA package supports several rules to take a set of NIST-library matches into account. The user can chose between three options (a) **All Analytes**, considering all proposed analytes with the GC/MS peak list; (b) **Best Matching Analyte**, selecting the first analyte in the GC/MS list that occurs in the MCCIMS-DB file and (c) **Best Analyte**, solely searching for the first NIST-library hit for each GC/MS peak in the MCCIMS-DB file. However, the matching to the MCC/IMS reference database extracts only appropriate (e.g. volatile) substances from the NIST-library selection.

Finally, the layer can be superposed on the IMS chromatogram, which corresponds to the processed mass spectrometric chromatogram.

## 7.3 Results

The reliability of the MIMA strategy was tested by parallel GC/MS and MCC/IMS measurements of a reference substance mixture containing 7 components. Subsequently an automated MIMA analysis of the GC/MS data was performed, resulting in the previously described IMS layer. This layer was overlayed on the MCC/IMS data that was measured in parallel, see Figure 7.2. In total 24 GC/MS substance candidates identified by the National Institute of Standards and Technology (NIST) database could be matched to the MCC/IMS reference database. By discarding candidates that do not show a peak in the MCC/IMS measurement, 17 signals in the IMS chromatogram could be identified using the GC/MS layer. The assigned signals are summarized in Table 7.3. In total, 12 of the 17 signals arise from the reference analytes as well as their dimers and trimers. Interestingly the analytes No. 5 and 14-17 were not part of the reference analyte mixture, but could, given their patterns known from previous studies, be clearly identified as decanal, n-nonan and heptanal. As component in many fragrances it may have its origin in the room air and entered the IMS during sampling. However, the appearance of all 7 used reference compounds in the automatically generated layer confirms the functionality of our software solution.

## 7.4 Discussion and Conclusion

The mapping of the GC/MS and MCC/IMS data was tested successfully by parallel measurements of a 7 components mixture of reference substances. The results demonstrate that the MIMA software tool extracts correct annotations for the MCC/IMS signals when processing the accompanying GC/MS measurement data with the aid of OpenChrom and

Table 7.1: Automatically identified signals

No.	CAS	compound
1	112-44-7	undecanal
2	104-46-1	anethol (trans-anethol)
3	6485-40-1	carvon (monomer)
4	6485-40-1	carvon (dimer)
5	112-31-2	decanal
6	2216-51-5	(-)-menthol (monomer)
7	2216-51-5	(-)-menthol (trimer)
8	821-55-6	2-nonanon (monomer)
9	821-55-6	2-nonanon (dimer)
10	5989-27-5	D-limonen (monomer)
11	5989-27-5	D-limonen (dimer)
12	110-43-0	2-heptanon (monomer)
13	110-43-0	2-heptanon (dimer)
14	111-84-2	n-nonan (monomer)
15	111-84-2	n-nonan (dimer)
16	111-71-7	heptanal (monomer)
17	111-71-7	heptanal (dimer)

NIST. Although the identification of all 7 reference compounds by the automatically generated layer demonstrated the functionality of the software, it needs to be emphasized that this shall not reflect a proper large-scale evaluation. However, it demonstrates the reduction in time consumption compared to similar manually performed data analysis steps which is MIMA's main aim.

Figure 7.3 shows the time expenditure evaluation of the automatic workflow compared to the manual procedure. While the MCC/IMS measurement is reasonably fast, the GC/MS measurement is equally time-consuming than the manual analysis (about 2h). However, the GC/MS procedure can be automated and run over night. The manual analysis of a standard human breath or air sample, comprising the GC/MS and MCC/IMS chromatogram evaluation (about 1h) as well as the combination of the results typically takes more than 2 hours, even for experienced users. The presented workflow with MIMA will reduce this time to about 10 to 15 minutes. Furthermore, all tools in the workflow are able to automatically process several measurements sequentially. An example workflow as well as step-by-step instructions are available at the MIMA website<sup>3</sup>.

In comparison to other methods our workflow improves the automatic peak identification in MCC/IMS chromatograms. Peak identification by alignment of MCC and GC retention times, for instance, needs a manual and time-consuming comparison of the chromatograms, which results in suggestions for the MCC/IMS signals. (117)

A previously presented software package for the comparison between MCC/IMS and GC/MS spectra has been published by Bunkowski *et al.* 2010 (48). They propose a mathematical function to convert MCC/IMS retention time in GC retention time and vice versa. Nevertheless, the major drawback of this software is the mathematical function,

<sup>3</sup><http://mima.mpi-inf.mpg.de>

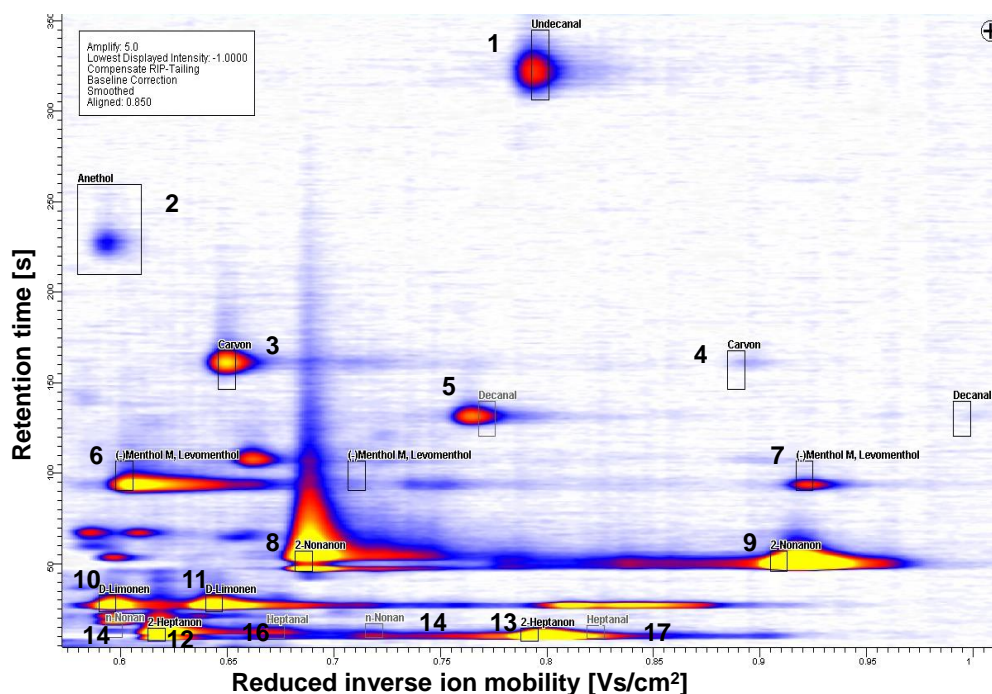


Figure 7.2: One MCC/IMS spectrum including the MIMA-generated layer (rectangles).

which was fitted on a small number of measured analytes, resulting in a reduced accuracy. Because of the relatively low number of analytes the curve could not be fitted exactly for all different analyte classes. Therefore, it remains a first approximation.

The main restriction of MIMA is the limited number of compounds in the existing IMS reference database (about 600 compounds) (37; 223), which are used in comparison with the NIST mass spectra database (212,961 compounds) to generate the layer. The fact that missing analytes cannot be displayed despite their presence in NIST necessitates the extension of the existing IMS reference databases. The current version of OpenChrom can provide a number of highest ranked candidate substances from the NIST suggestions. Given that the first suggestion is not in the IMS reference database (e.g. solid matter) the MIMA analysis can search for the next best NIST candidate to identify this peak. However, it is not able to choose the substance which fits best the IMS-measurement of interest. This remains a possible extension for future projects.

Still, this work presents the first software tool capable of automatic and thus effective mapping of MCC/IMS and GC/MS spectra by integrating two machinery-specific substance databases. This eases analyte annotation by using the MCC/IMS signals, which are stored in the MCC/IMS reference database file. Essentially, the proposed workflow reduces the workload for the spectra analysis from more than two hours to less than 15

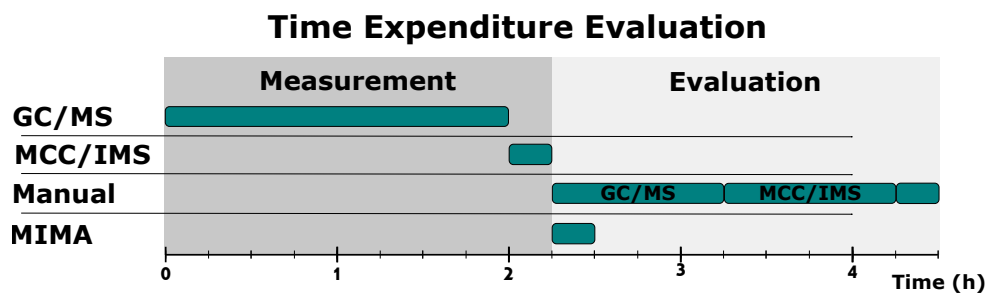


Figure 7.3: Time expenditure evaluation of the manual and automatic workflow, estimated for a standard human breath or room air sample containing 30 to 40 compounds performed by an experienced postdoctoral researcher in our lab.

minutes and enables a sequential processing of many large-scale experiments. Evaluation of the MIMA software on more complex samples will be carried out over the next years when more MCC/IMS reference annotations become available in the MCC/IMS databases.



## Chapter 8



# Uncovering Hidden Structures with CAROTTA

*The project was initiated during the course of the Bachelor thesis of Tobias Frisch that I supervised. Tobias Frisch developed a Java and R based software package including a data handling and visualization in Java, and machine learning in R. During the development of the Carotta framework, the author of this thesis refactored the structure of the data handling and processing towards a modular plug-in based system. The R functionality was replaced by Java packages. Finally, the author refactored the user interface and added an automatically generated parameter wizards.*

Carotta is a new cluster analysis framework dedicated to uncovering hidden substructures with the aid of sophisticated unsupervised statistical learning methods, as described in Section 1.3. This chapter examines the performance of transitivity clustering and hierarchical clustering with regard to two aspects. The first task is concerned with the identification of groups of patients with a similar VOC intensity pattern. The second task is concerned with the identification of groups of VOCs with a similar expression behavior across most patient breath samples. This analysis allows for the discovery of dependencies between metabolites, which has two advantages: On the one hand, the effect of potential confounding factors hindering disease classification can be eliminated. Examples for confounding factors can be behavioural or environmental factors, such as smoking, nutrition, or contaminations of the inhaled air. On the other hand, VOCs associated with disease subtypes or concomitant diseases can be identified.

Carotta is an open source software with an intuitive graphical user interface that can facilitate data handling, analysis and visualization. The back-end is designed in a modular fashion, which makes it easily expandable with additional features. Thus, new clustering methods and further statistical measures could be included in the future. Furthermore, neither extensive prior knowledge nor comprehensive technical skills are required for oper-

ating Carotta’s functionality. This functionality is demonstrated by means of one artificial data set. Moreover, to evaluate the power and applicability of the methodology, Carotta is applied to the real-world example data set on COPD described in Chapter 3 (106).

**Objective:**


Development of an intuitive graphical software framework for nested unsupervised learning on breathomics data.

## 8.1 Requirements and State of the Art

Several data analysis frameworks have been developed to process, visualize and analyze metabolomics data, particularly for GC/MS data. Some of them focus on the preprocessing of raw data, but include advanced methods for alignment, peak detection, and peak identification, e.g. mzMine (179). Others, such as the web application MeltDB, address issues concerning storage, sharing, and standardization of metabolomics data, as well as a binding to the R software package(115). This allows for the application of the whole wealth of statistical data analysis tools integrated in R (124). However, the use of R requires programming knowledge. More advanced services, such as XCMS Online (94) and MetaboAnalyst (238), offer advanced statistical analysis techniques. XCMS Online is optimized for LC/MS data and offers various parametric as well as non-parametric test statistics, as well as extended visualizations for meta-analysis (Venn diagrams, for instance). It further offers unsupervised learning techniques and visualization capabilities, mainly PCA and HAC. However, it lacks adequate measures for internal and external clustering quality, which are essential for evaluating the information content of the clusterings and for selecting reasonable clustering parameters/thresholds. The MetaboAnalyst web server also provides access to GC/MS data pre-processing, multivariate statistics, and PCA, but focuses mainly on supervised learning and time series analysis afterwards. It supports advanced learning methods, such as partial least squares, discriminant analysis, and random forests as well as an evaluation framework, with which cross-validation, permutation tests, and ROC curve analysis can be performed. However, it does not support features for systematically exploring the results of unsupervised data processing technologies.

The design of Carotta was encouraged by this lack of comprehensive and user-friendly software systems to fill the gap between the quickly emerging breathomics data sets and the challenges of current breath data analysis as described in Section 1.2. The aim of the Carotta project is a software application that provides easy access to advanced unsupervised learning techniques specifically designed for breath data analysis. It addresses two main goals, a user-friendly front-end as well as a flexible and modular back-end that allows for functional extensions. Since the target users are biomedical researchers, an access to these techniques without deeper knowledge of advanced computational approaches is necessary. This requires a step by step guidance through the different steps of unsupervised learning analysis, starting with the similarity function, clustering, cluster quality evaluation, filtering and dimensionality reduction. Moreover, to increase the understanding of the results and support in-depth investigation directly in the user interface, each

intermediate or final result should be presented to the user by an interactive visualization. Furthermore, a flexible plug-in system would allow future methods to be added in a straight-forward fashion. This includes all described methodology of the steps: similarity measure, clustering, clustering quality and dimension reduction.

The diagram shows three circular icons containing stylized human figures, arranged in a triangular pattern. To the right of these icons is the heading "Requirements:" followed by three bullet points, each starting with a right-pointing arrowhead (⇒).

**Requirements:**

- ⇒ User-friendly front-end supporting easy access to unsupervised learning methods
- ⇒ A flexible and easily extensible back-end.
- ⇒ Detailed visualization of data and results.

## 8.2 Structure and Implementation

The Carotta software framework provides interactive pipelines that can reveal hidden structures in metabolomics breath data. The typical Carotta workflow consists of six processing steps depicted in Figure 8.1. Each of the six steps will be described in detail in the following paragraphs.

**Step 1** In the first step of the Carotta workflow, the data sets are imported into the system and displayed. Prior to the import, the raw data is preprocessed by technology-specific preprocessing methods, such as baseline correction, denoising, as well as peak detection (for MCC/IMS and GC/MS), such that a data matrix, as shown in Figure 8.1, is generated. This can be done by corporate or open source software products such as VisualNow (B&S Analytics)(35) and OpenChrom(228). A typical data matrix contains the relative compound abundances for each sample measurement that correspond to objects of interest, such as patients. The matrix resulting from the preprocessing of a set of MCC/IMS measurements comprises the MCC/IMS peak intensities. This process is described in detail in Section 2.2 of the Background Chapter.

**Step 2** We define pairwise relation as a measure for the strength of an association between two objects that can either be a similarity or distance. The Carotta software supports three different measures: correlation coefficient (Pearson or Spearman), or Euclidean distance (154), that can be calculated on either study subjects (e.g., patients) or metabolites. The correlation coefficient corresponds to a similarity and the Euclidean distance to a dissimilarity matrix, see Section 2.5.2 for more details. These pairwise relations are stored in a matrix and depicted by a heat map. All further methods require either a similarity or a dissimilarity matrix; therefore, the dissimilarity matrix is converted into a similarity matrix, and *vice versa*, according to the requirements of the following method. This is done as follows: The conversion of a similarity matrix to a dissimilarity matrix is performed as follows. Let  $S$

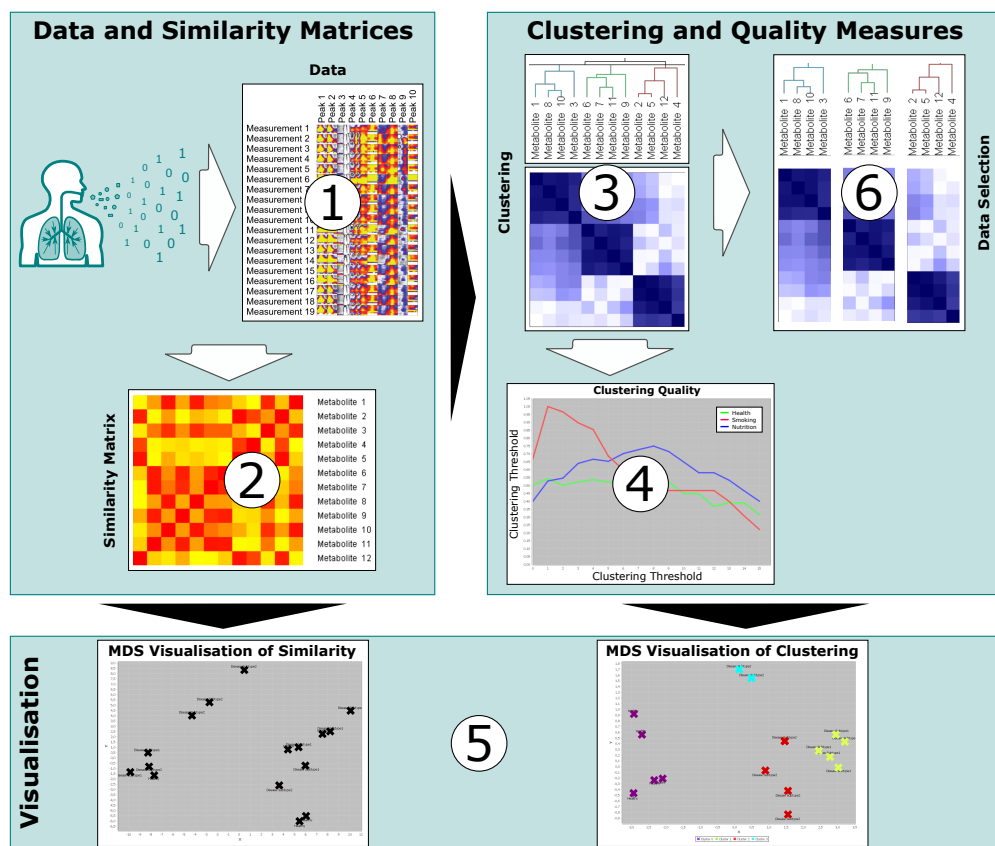


Figure 8.1: The Carotta pipeline consists of several steps: (1) importation of pre-processed data; (2) similarity calculation; (3) clustering; (4) clustering quality; (5) similarity or clustering visualization; (6) subset selection. Intermediate results can be inspected, optimized, and repeated with an arbitrary depth.

be the matrix containing the original similarity values, where  $s(x, y)$  corresponds to the similarity of object vectors  $x$  and  $y$ . Then  $\hat{d}(x, y) = \max(|S|) - |s(x, y)|$ , defines the dissimilarity of  $x$  and  $y$  constructing the dissimilarity matrix  $D$ . The conversion of the dissimilarity matrix to a similarity matrix is defined accordingly:  $\hat{s}(x, y) = \max(|D|) - |d(x, y)|$ , where  $D$  is the original dissimilarity matrix. Further, these pairwise relations can be visualized in a two-dimensional scatter plot by using MDS (241), such as depicted in Figure 8.1, Step 5. Details on the multidimensional scaling approach can be found in Section 2.5.1.

**Step 3** A clustering algorithm can be applied to the similarity or dissimilarity matrices calculated in Step 3. The system integrates two state of the art clustering algo-

rithms. The first algorithm is hierarchical agglomerative clustering (101), which is based on pairwise dissimilarities. The second algorithm is transitivity clustering (TC) (234), which is based on similarities. Details on the methodology of the clustering algorithms are given in Section 2.5.2. The result of a clustering algorithm is a list of clusterings, each corresponding to a certain threshold. The thresholds depend on the method and the selected parameters. For clarity, this Chapter will refer to the set of all clusterings as the clustering result and to each individual clustering solely as clustering.

**Step 4** The quality of these clusterings is evaluated and compared by means of two quality measures, the silhouette value and the F-measure.

**Step 5** At this step, several clusterings and clustering results may have been produced. The user can choose to visualize each individual clustering using the MDS coordinates of the underlying similarity. The resulting scatter plot is color-coded by cluster.

**Step 6** Finally, filtering methods can be utilized to select a subset of the data. For example, the user can choose a certain clustering and extract a representative for each cluster or select all objects in one cluster.

By repeating steps one to four on the selected subsets of the data, Carotta explores various layers of potentially hidden sub-structures. Especially the nested-clustering of metabolites and samples can reveal novel information; see Figure 8.2.

In short, Carotta can be used to split the set of metabolites into subsets (defined by clusters), which, in turn, can be used individually to inspect their association with the primary outcome variable, i.e., the disease. This allows for eliminating large sets of metabolites that correlate with potential confounding factors and not with the investigated disease. In this way, uninformative features are eliminated. Most notably, Carotta automatizes these steps and intuitively visualizes intermediate and final results.

### 8.2.1 Visualization of Data and Results

The graphical user interface (see Figure 8.3) is split into three basic regions: (1) the data and results area, showing a list of all generated results ordered in a tree-like structure; the categories correspond to the previously described processing steps (data, similarity, clustering results, clustering quality, visualization); (2) a details panel, reporting the parameter used for the calculation of the currently presented result; this also includes, for instance, general information on the data set (such as the minimum and maximum values); (3) the main result visualization panel displays the results of the different intermediate steps, as well as the final results. The following paragraphs describe the visualization of each of the steps introduced Section 8.2, in the graphical user interface.

**Data and similarity matrix.** Each data or similarity matrix as described by Step 1 and 2, is displayed as a heat map tagged by the corresponding metabolite names and sample label, on the columns and rows, respectively. Labels can be changed to arbitrary annotation details included in the original data matrix.

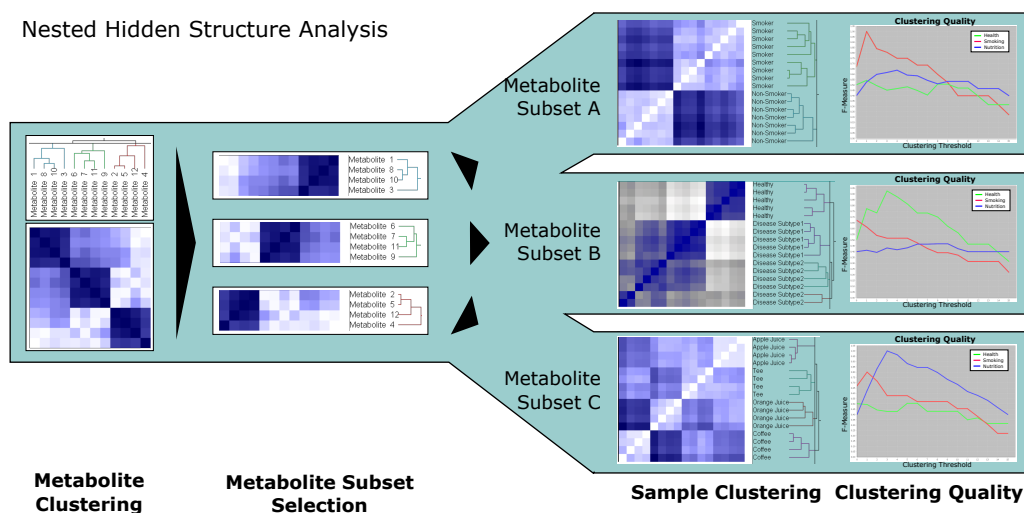


Figure 8.2: The subset selection allows for a **nested analysis** of the hidden structures in the data. Steps 2–4 from Figure 8.1 are repeated on a selected subset (or all subsets). In this example, the artificial data is first clustered by metabolites. In the next step, a threshold is selected that splits the data into three metabolite clusters and the data is separated accordingly. Subsequently, the data for each of these metabolites is used separately to access the distances between the samples (patients). These similarity matrices are further used for clustering the samples. Finally, an external quality measure is used to evaluate the association of each set of metabolites to the selected patient annotations (i.e., labels; here: *health*, *nutrition* and *smoking*). The F-measure plots show to what extent the metabolite clusters explain the different labels.

**Clustering.** The heat map of the underlying similarity matrix is displayed in the center of the visualization of the clustering results. The rows and columns are sorted by the corresponding clustering. For hierarchical clustering, results can be inspected interactively by selecting a clustering threshold. This is facilitated by sliders at each of the two dendrograms which can be altered by using the mouse. Leaf nodes correspond to clustered objects; inner nodes depict how the data set is split (top down) or merged (bottom up) during the clustering. For the transitivity clustering (*hierarchical = FALSE*), one can manually adjust the threshold through a bar on the right side. Depending on the selected cut, colors encode the resulting clusters. Note that this does not depend on the selected clustering method. The axis labels can be defined by the user.

**Cluster quality.** The quality of a clustering can be evaluated for varying cuts/thresholds and visualized by line plots, see Section 2.5.2 for details on the utilized quality measures. The F-measure is most appropriate to identify which threshold best reflects a given labeling or gold standard. Using Carotta, the F-measure can be calculated and visualized. In order to compare different class labels with respect to their consistency

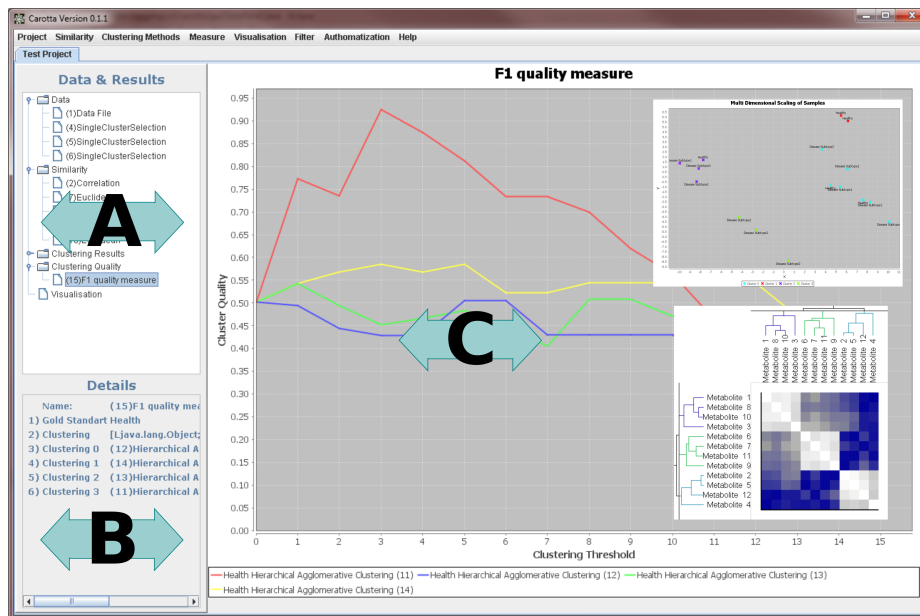


Figure 8.3: The graphical user interface is split into three basic regions: (A) the data and results area lists available (intermediate and final) results; (B) the details panel lists attributes of the data, such as mean or minimum values, or the history of previous processing steps; (C) the main result panel displays the visualization of the intermediate and final results.

with a specific clustering, several labels can be selected by the user and visualized in a single F-measure line plot. Equally, different clusterings can be selected and compared to a single label. However, if a data set does not provide a gold standard, a reasonable cut or threshold for the clustering can be identified by the internal quality measure: silhouette value.

**Multi dimensional scaling.** The visualization of the similarity of a set of objects is provided by a customizable scatter plot, based on coordinates determined by the MDS. Besides the custom-defined labeling, the depiction of a clustering result can be colored according to a chosen threshold. This representation can give the first indication of whether a clustering is “good”.

## 8.2.2 Modularity and Extendibility

Due to the modular structure of Carotta, new functionality can be integrated easily. Each of the previously described processing steps (similarity, clustering, cluster quality, and visualization) can be expanded by additional methods and implementations. Java reflections guarantee a comfortable plug-in system that does not require any further editing

of the previous code.

### 8.2.3 Import and Export

Convenient functions for exporting all intermediate and final results are included. The system can export all visualizations of Carotta, as described before. The user can choose between different resolutions of the resulting portable network graphics (PNG) image file. Carotta further supports exporting into an MS Excel file (e.g., a similarity matrix or the results of the quality measure).

### 8.2.4 Language and Packages

The Carotta software package and associated software libraries are purely Java-based. The source code is available at the project website and underlies the Apache License Version 2.0. More information on the technical aspects can be found in the Appendix and the following address: <http://carotta.compbio.sdu.dk/> (106). Carotta makes use of the following software packages:

- The TransClust package for transitivity clustering (235).
- The HAC package for hierarchical agglomerative clustering (190).
- JExcelApi 2.6.12 for parsing the excel sheet into the internal data structure (JEx).
- The JFreeChart 1.0.14 visualization (clustering quality, scatter plot of MDS) (90).
- JHeatChart 0.6 for creating the heat map (JHe).
- The MDSJ calculation of the multi-dimensional scaling (6).
- The Guava & Reflections & Javassist Google Core Libraries (gua) and the Javassist (57) are used for the reflections technology.
- log4j 2.0 for logging and debugging (3).

## 8.3 Results and Discussion

To demonstrate the abilities of the Carotta clustering framework, we analyze two data sets, described in Chapter 3. At first, we discuss the results of the artificial data set, followed by the real world COPD data set.

### 8.3.1 Results for Artificial Data Analysis

The artificial data set to demonstrate and clarify the capabilities of Carotta. It consists of 16 samples and 12 metabolites associated with three metabolite groups. Each of these groups is related to one of the three predefined labels of the samples: (1) *health*; (2) *smoking*; (3) *nutrition* while the label *health* is our primary outcome variable, the remaining



label *smoking* and *nutrition* represent potential confounding factors. See Section 3.4 for more details.

This data set is used to exemplify the power of Carotta. First, the Pearson correlation coefficient is used for calculating the similarities between all metabolite occurrences, and clustered them by both HAC and TransClust. Disregarding the clustering method, the silhouette value indicates, an optimum of three different metabolite clusters, as expected. As shown in Figure 8.2, the full data set is now split into three subsets, one for each cluster of correlating metabolites. Subsequently, for each cluster, as well as for the full data set, the Euclidean distance between all pairs of patient samples is computed for each cluster separately. The three resulting clusterings gained from the three metabolite subsets are compared against the clustering achieved with the full metabolite data. Therefore, the F-Measure is used to assess the overlap of the four different clusterings with the initially-designed class labels (*health* = green, *smoking*= red and *nutrition*=blue). Figure 8.4 shows the results of this clustering evaluation in terms of a line plot for each clustering and a colored line for each label. In this artificial example, the entire data set is heavily confounded by the influence of the *smoking*-related metabolites. Therefore, the clustering of the entire data set using all metabolites is heavily influenced by smoking related metabolites that dominate the distance measure for the patients, see red curve in Figure 8.4. Consequently, the clustering of the entire data set does not reflect any overlap with the gold standard label for health, depicted by the green curve. A detailed analysis of the F-Measure curves for the clustering on the three subsets of the data (A, B, C), reveals strong relations between the metabolite subsets and annotations: subset A and *smoking*, subset B and *health*, and subset C and *nutrition*. Thus, according to the F-Measure, the metabolites in subset A and C correspond to confounding factors. Solely the metabolites clustered in subset B contain information that relates to the gold standard label of the samples: *health*.

### 8.3.2 Results for COPD Data Analysis

COPD is an inflammatory lung disease characterized by a permanent blockage of airflow from the lungs, which is not fully reversible. Here, we study the exhalome of COPD patients using a data set from (231), described in detail in Chapter 3. It consists of metabolic maps from 42 COPD patients, 52 patients suffering from both COPD and bronchial carcinoma, as well as 35 healthy controls. The patients' breath was captured and analyzed using an MCC/IMS, as introduced in Section 2.1 of the background Chapter. This data set was evaluated utilizing Carotta, following the previously introduced standard workflow described in Section 8.2. At first, all 120 metabolites were clustered by HAC and the Pearson correlation (converted to dissimilarity, as explained above). Several thresholds were investigated, leading to a varying number and size of the clusters. This resulted in an optimal threshold at  $T = 40$ . Subsequently, the set of metabolites was split into 40 subsets, one for each cluster of correlating metabolites. All clusters with less than three compounds were excluded, yielding a total of 14 metabolite sets.

Finally, the hierarchical agglomerative clustering was performed on the correlation matrix (converted to the distance matrix, as previously described) of the patients for each of these metabolite sets. Carotta subsequently evaluates the overlap of the patient clusters with the three patient groups over varying clustering thresholds using the F-measure. Figure

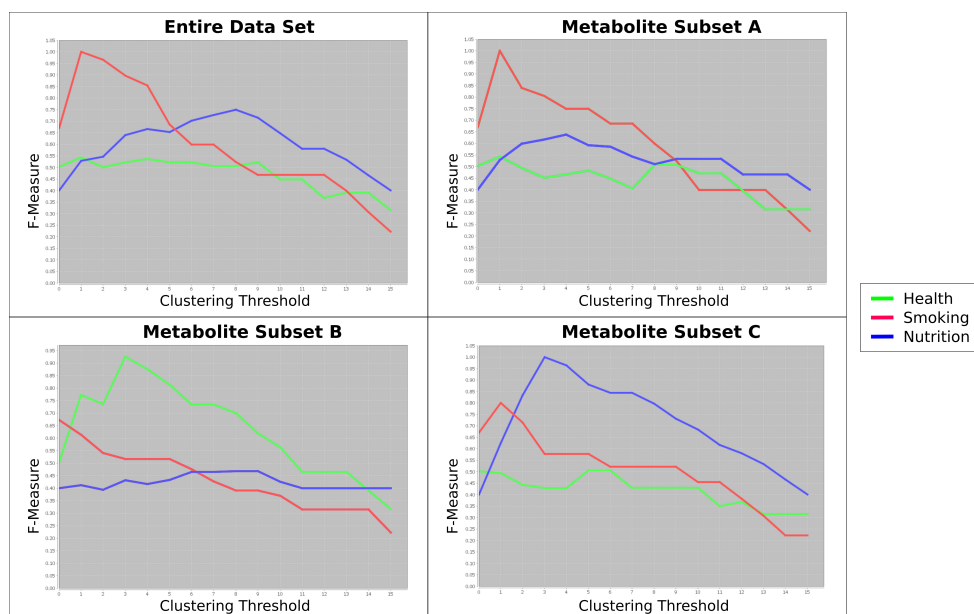


Figure 8.4: Result of the nested hidden structure analysis applied to the artificial data set. Each plot depicts the F-Measure evaluation of the clustering on one of the four different data sets: Entire Data Set, and each of the distinct Metabolite Subsets A-C. The subsets were previously determined by a clustering of the metabolites. The lines show the F-measure for the labels: *health* in green, *smoking* in red, and *nutrition* in blue, for every clustering threshold. The top left plot depicts the result for the entire data set, it shows a dominant effect of the confounder *smoking*, which overlays the main outcome variable *health*. The other three plots show the results for each cluster of correlated metabolites, separately. Each of the metabolite subsets clearly corresponds to one of the three labels: *health*, *smoking*, and *nutrition*.

8.6 plots the results for four of the 14 metabolite subsets, as well as the results when using the entire set of metabolites. For better visualization, only the five most interesting results are displayed in the Figure 8.6. Specifically, the F-measure is plotted for the entire metabolite set, the two metabolite sets with the highest F-measure and the two metabolite sets with the lowest F-measure. Each patient is annotated with one of three groups: *healthy*, *COPD*, or *COPD+BC*. Thus, the overlap of the clustering results at  $T \sim 2$ , corresponding to two splits of the data resulting in three clusters is investigated. One can see two metabolite clustering subsets, namely Subsets 1 and 14, that peak around three clusters. Both exceed the F-measure achieved using the full metabolite set.

The compounds within these clusters were compared manually (via their specific peak coordinates) to the results of previous COPD studies utilizing supervised learning methodology (103). Three of the compounds in Subset 1 were previously reported as potential biomarkers. This shows that the presented stepwise multi-dimensional clustering ap-

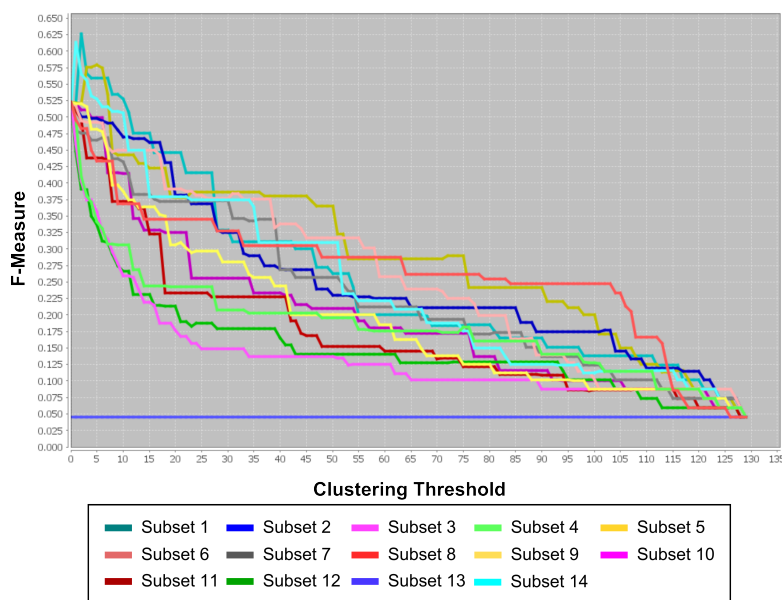


Figure 8.5: Comparison of the clustering results of the 14 metabolite clusters. The plot shows the F-measure for different clustering thresholds computed against the disease annotation (COPD, COPD with BC, and healthy). Since the annotation defines three groups of patients, the performance of the clustering results at a clustering threshold of 3 is of particular interest (x-axis).

proach points out putative COPD marker metabolites by using a purely unsupervised approach. In contrast, the metabolites in Subsets 3 and 4 show a rapid decrease in the F-measure for a growing number of clusters. The evaluation of the list of compounds within these clusters uncovered that these subsets contain the menthol trimer (Subset 3), as well as the menthol monomer and dimer (Subset 4) compounds, respectively. The occurrence of menthol in human exhaled air can be the result of various environmental and nutritional influences, for example tooth paste or candy. This exemplifies where Carotta is useful: when we expect yet uncharacterized confounders to exist, which have an effect on the metabolic patterns that we like to detect and exclude. The human exhaled air in particular can be influenced by various external factors, like nutrition and compounds in the environmental air. They do not need to be known *a priori*, however. The menthol example in human breath serves as a proof of concept here.

However, further analyses of the clustering results would be beneficial. It is of particular interest to focus on the investigation of the extent to which the elimination of putative confounding metabolites would improve the classification performance in a systematic statistical learning study. This clearly goes beyond the focus of this work and should be addressed in the future.

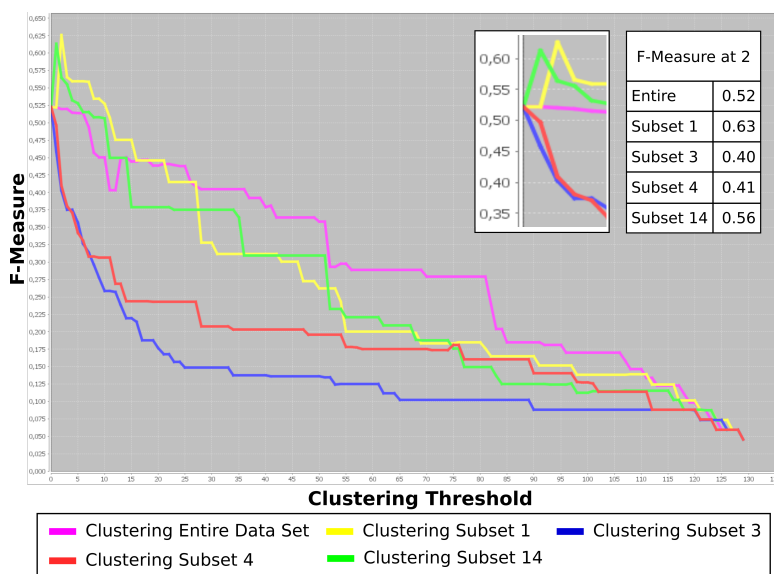


Figure 8.6: Comparison of the clustering results on the entire COPD data set, *i.e.*, using all metabolites and the four most interesting metabolite clusters, including two clusters showing the largest F-Measures and two clusters showing the smallest F-Measures. The plot shows the F-Measure for different clustering thresholds computed against the disease annotation (COPD with BC and healthy). The Y-axis corresponds to the clustering threshold, in this case the number of splits. Given three groups of patients in the annotation, we are particularly interested in the performance at  $T \sim 2$  (x-axis) resulting in three clusters. This is shown in more detail in the zoomed cutout, as well as the table of F-measure values at this position. Two subsets of metabolites overlap with the patients' disease annotation better than the clusterings based on the entire metabolite set. The two other metabolite subsets result in reduced F-measures, indicating a relation to confounding factors, in this case menthol.

### 8.3.3 Comparison to Existing Software

A number of other data analysis frameworks for metabolomics data processing are available. However, most of them focus on a single task, such as preprocessing, storage and distribution, or statistical analysis. Examples for this are mzMine (179) or MeltDB (115). Other services, such as XCMS Online (94) and MetaboAnalyst (238) provide advanced learning techniques for metabolomics data. Similar to Carotta, XCMS Online offers unsupervised learning techniques and visualization capabilities. However, in contrast to Carotta, it does not provide means for systematically exploring adequate measures for internal and external clustering quality, which are essential for evaluating the information content of the clusterings and for selecting reasonable clustering parameters/thresholds. The MetaboAnalyst web server is designed for GC/MS data processing and supervised learning. Nevertheless, the system is lacking features for the systematic analysis of clustering results. In contrast, the focus of Carotta lies in the *de novo* detection of confounding

factors. For instance, Carotta can be used for performing an unsupervised analysis of breath data sets which detangles potential biomarkers and confounders.

Existing methods for such multi-dimensional clustering, such as bi-clustering or co-clustering (44; 205), do not provide graphical frameworks for systematically exploring the parameter space. With Carotta, however, a sequence of various clustering combinations of metabolites and samples can easily be used for investigating all results visually and systematically while using different validity measures.

It needs to be emphasized that the main focus of Carotta is breath data analysis, yet its utility is neither limited to MCC-IMS data nor to breath gas profiling. Applications in transcriptomics (gene expression data) or related omics fields are generally possible. This thesis focuses on breath data only. This kind of data is infested with yet undiscovered confounders that emerge from the environment, nutrition or ambient air. In addition to the high prevalence of systematic confounders in breath data, it might also be prone to various types of technological noise. An extensive analysis of their effects is needed and introduces interesting new research questions.

Unlike all other tools, with the exception of MetaboAnalyst, Carotta allows for direct processing of a metabolomics peak matrix (regardless of the utilized technology). MetaboAnalyst, however, does not support systematic clustering exploration. The MCC/IMS community has established a number of standard procedures for pre-processing, and a set of integrated tools has been developed in the past; see (199; 61; 37). Like other frameworks, Carotta does not yet include all of these standard procedures for pre-processing. However, its flexible plugin architecture can be used for implementing preprocessing procedures, among other novel features, such as additional statistical learning approaches.

## 8.4 Conclusion

In this chapter the Carotta software for *de novo* detection of confounding factors and disease sub-types is presented. It is open source and comes with an intuitive graphical user interface for unsupervised breathomics data analysis and visualization. The flexible back-end design can be easily extended with plugins, such as new clustering methods and statistics. It intuitively guides the user through four steps: (1) (dis-)similarity matrix computation; (2) clustering; (3) clustering evaluation; and (4) visualization and interpretation of the results. This process requires neither extensive knowledge or advanced technical skills to operate and is therefore suitable for non-technically trained personnel. By means of an artificial data set, the functionality and applicability of the Carotta software framework for revealing hidden structures and confounding factors was demonstrated. Additionally, the power of Carotta in practice was exemplarily shown by the reanalysis of the real-world data set on COPD. It demonstrated how Carotta is useful in finding potential informative metabolite clusters containing substances also supported by previous studies. Most notably, it identified confounder metabolites (e.g., menthol), which are related to nutrition and to the environment rather than to the primary outcome variable (disease annotation, *i.e.*, COPD and lung cancer). The Carotta software framework offers easy access to extensive clustering analysis for non-technical personal working in the area of breathomics. It is publicly available at <http://carotta.compbio.sdu.dk> (106).



## Chapter 9



# Longitudinal Breath Analysis

*This chapter is based on a collaborative project with the Department of Anesthesiology, Intensive Care, and Pain Therapy, Saarland University Medical Center, Homburg (Saar), Germany and the Division of Biostatistics, University of Southern California (USC), Los Angeles, United States. The collaboration partners in Homburg were responsible for sample preparation, generation of the longitudinal experimental data set and assisted with the interpretation of the results. Prof. Dr. Sandy Eckel from the USC supervised the project and gave valuable insights into longitudinal statistical analysis techniques. The author of this thesis designed and implemented the pipeline for longitudinal analysis of breath data and conducted the statistical evaluation of the example data set.*

The majority of current breathomics research is restricted to investigations on cross-sectional snapshots of a specific disease. However, most diseases traverse various stages during their development, or emerge towards different subtypes depending on influences such as genetics, environment or medication. Common examples are for instance the evolution of genetic variants in cancer (164; 95), COPD stages (88), sepsis progression (176) or lateral sclerosis (239). Thus an analysis that focuses on a single time point of the patients' metabolite pattern may overlook potential biomarkers. These so-called longitudinal biomarkers show a relative change over time, which can be much more informative than a single observation. A recent study of Langley *et al.* suggests that this might be the case for sepsis in primates as well (135). Sepsis is the leading cause of death among critically ill patients in USA and Europe and the number of incidences and sepsis-related deaths is increasing (153; 194). It is caused by an uncontrolled, systemic, inflammatory response to bacterial, viral or fungal infection. However, we lack a rapid and accurate identification method for sepsis and its causative microorganisms, which would be essential to increase the chances of survival. Today, using the current diagnostic gold standard, at least 24 hours are needed to get preliminary information on the causal organism, using cultures of blood and other body fluids or tissues.

**Objective:**

Development of a comprehensive framework for longitudinal breath analysis using the specific example of sepsis in rats.

## 9.1 Requirements and State of the Art

More recent approaches to diagnose sepsis use real-time polymerase chain reaction or calorimetry to obtain a more rapid diagnosis of bacteria or fungi based sepsis. However, the causative microorganism can be identified in the blood in only 20-40% of patients with severe sepsis (193). Large false positive and false negative rates are common. Other metabolomics or proteomics-based studies proposed more than 170 different compounds as potential biomarkers, but so far none of them have sufficient specificity or sensitivity to be routinely employed in clinical practice (176). To address the unmet clinical need for early and reliable diagnosis of sepsis, rapid and cost-efficient diagnostic tools are needed (193).

**MCC/IMS a Diagnostic Alternative?** As previously discussed, the MCC/IMS approach has several key advantages compared to other analytical methods for measuring metabolites in exhaled breath. Some of these advantages are crucial for a diagnostic tool for sepsis. For example, MCC/IMS can be used to rapidly collect and analyze a sample and is a robust and easy to handle instrument in a clinical setting (24; 20).

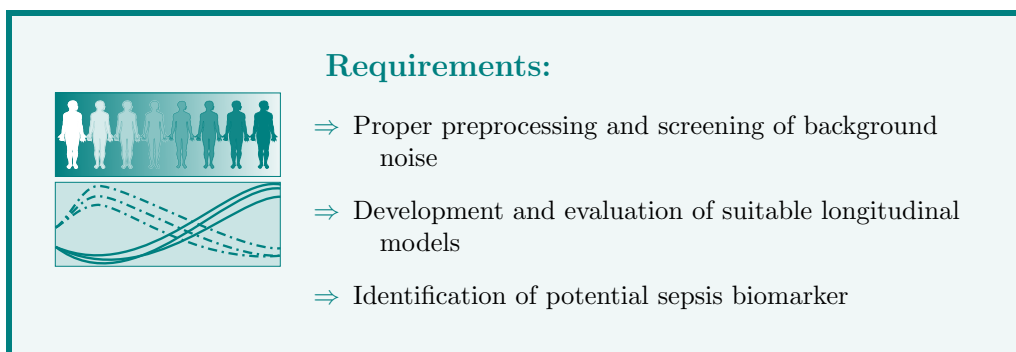
Methodological developments over the last decade have led to frequent use of MCC/IMS for medical and biological applications, for instance evaluating pulmonary diseases, lung cancer and airway infections (84). Of particular interest for the study of the causal agents of sepsis, previous studies have shown the potential of MCC/IMS technology to identify and distinguish between bacterial species (71; 147; 181; 118). Furthermore, a recent study of Guamn *et al.* successfully utilized GC/MS and IMS to analyze the breath of rats with systemic inflammation.

These results encouraged a pilot study by Fink *et al.* to investigate whether systemic inflammation or sepsis influence the metabolite composition of the breath in rats. The recently published results identify several compounds that significantly differentiate rats with sepsis from control rats. (83). See Section 3.5 in Chapter Materials for more details. Although, the study collected repeated breath samples over the entire timecourse of the infection, the recently published analysis used only cross-sectional data, which might lead to non-robust results, false positives or false negatives.

In contrast, this project aims to analyze the time dependent behavior of breath metabolites by adapting established methods for longitudinal data analysis to breathomics. Among others, a common problem in studies of clinical diagnostics are confounding factors resulting for instance from instrumental background. Therefore, in order to focus on the real informative metabolites a screening approach is needed to sort out the background noise. Once a suitable longitudinal model is created, the aim is to identify possible markers for sepsis. In contrast to previous approaches, the identification will focus on the



time course of various volatile organic metabolites within the rat breath, during sepsis progression. Furthermore, a better knowledge of the changes in the metabolic pattern of the progressing disease could give more detailed insights of their biological background, or indicate states that require a specific medication.



## 9.2 Methods

This study utilizes the longitudinal rat breath data introduced by Fink *et al.* and described in Section 3.5. It consists of two treatment groups: One group underwent an operation that induced sepsis (CLI) and the control group underwent a sham operation (SHAM).

Figure 9.1 shows the four general analysis steps in this project:

- A** At first, the rat breath is captured and analyzed by the MCC/IMS device, and the resulting MCC/IMS chromatograms are preprocessed by a well established set of methods to extract the compound signals.
- B** In the next step a well established screening is utilized to eliminate compounds that belong to the background noise of the ventilation system.
- C** In the model selection phase, the structure of the data is analyzed by fitting a separate longitudinal model for each compound. Subsequently, the model optimizing the overall quality with respect to the data is selected.
- D** Finally we evaluate the model results and investigate which compounds show the best potential for the discrimination of rats with induced sepsis. False discovery rate correction is used to adjust for the multiple comparison problem.

### 9.2.1 Preprocessing

**RIP detailing and baseline correction** As described earlier, each MCC/IMS chromatogram contains a characteristic structure called the reactant ion peak (RIP). In order to remove this structure, this workflow applies one of the most widely used techniques. It subtracts the vector of median values (median spectrum) from each

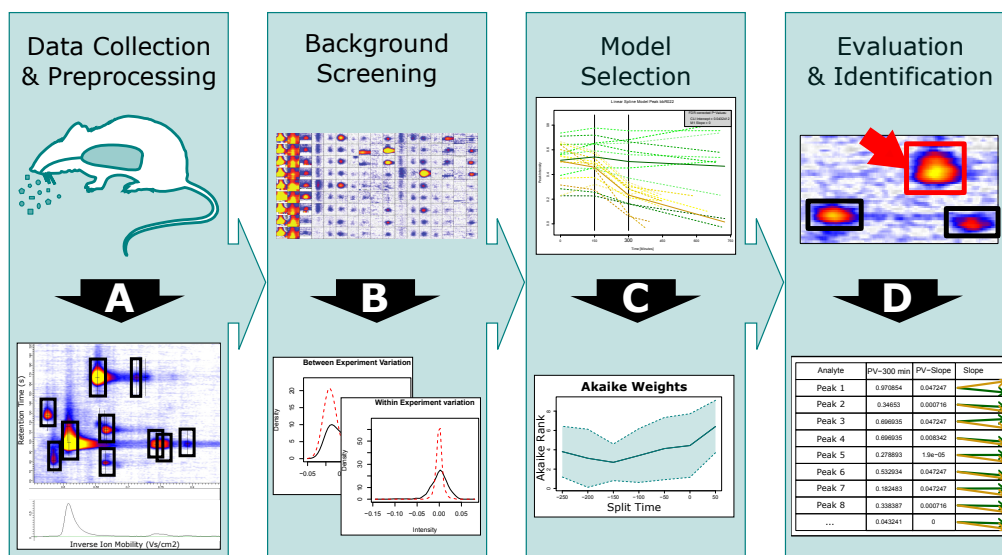


Figure 9.1: Workflow for longitudinal breath data analysis framework. The four steps contain (A) MCC/IMS analysis of rat breath and subsequent preprocessing of MCC/IMS chromatograms; (B) a mature screening for background noise; (C) model selection phase evaluating the quality of longitudinal models; (D) model evaluation and identification of most discriminating biomarkers.

IMS spectrum in the data matrix (108).

**Denoising and Smoothing** To improve the signal-to-noise ratio in the data, the median filter is applied to remove noise from the data. Finally, a Savitzky - Golay filter and a Gaussian filter are both applied to smoothen the data (199; 108).

**Peak Picking** In the last step of data preprocessing, the peaks corresponding to molecules within the samples are detected. Therefore, we utilize a semiautomated approach to find a set of peaks. At first, the merged peak cluster localization which is provided by the commercial software package VisualNow (B&S Analytik, Dortmund, Germany) is applied. Finally the list of peaks is manually curated by a domain expert. See Section 2.2.3 for more details (199; 108; 107).

## 9.2.2 Background Screening

Background noise signals originating from experimental settings or random events can hugely influence the signal pattern of the breath. In order to investigate the influence of this contamination on the rat breath, MCC/IMS measurements of the respirator and the ventilation system are taken. The data set contains two to three measurements, taken every 20 minutes over the period starting one hour prior to the connection of the rats to

the respirator system. This data enables the detailed evaluation and differentiation of the volatile compounds originating primarily from the ventilator system as compared to those from the rat metabolism. We hypothesize that all compounds of interest show a larger variation in the breath of the rats as compared to the background noise. Therefore, we developed a screening approach which excludes peaks that vary more in the background distribution of the respirator samples than in the distribution of the rat breath samples. To analyze this variation, two simple linear regression models are estimated for each compound (one for the measurements of the compound intensity from the background, respirator samples (S) and one for the measurements of the compound intensity during the rat breath sample (T)), as shown in Equation 9.1. In this Equation, compound intensity is indexed by  $i$ , which denotes the rat identification number, and by  $j$ , which indexes the repeated samples over time.

Each set of measurements is thus grouped by the corresponding rat experiment.

$$\begin{aligned} S_{i,j} &= u_i^S + \epsilon_{i,j}^S \\ T_{i,j} &= u_i^T + \epsilon_{i,j}^T \end{aligned} \quad (9.1)$$

Here,  $S_{i,j}$  and  $T_{i,j}$  indicate the respirator and rat intensities for a certain time point, respectively and the individual experiment means for respirator and rat are modeled by  $u_i^S$  and  $u_i^T$ . Finally,  $\epsilon_{i,j}^S$  and  $\epsilon_{i,j}^T$  correspond to the residuals of the model. This model structure enables the comparison of the within-rat variance (variance in  $\epsilon_{i,j}^S$ ) to the within-ventilator variance (variance in  $\epsilon_{i,j}^T$ ) as well as the between rat variance (variance in  $\mu_i^T$ ) to the between ventilator variance (variance in  $\mu_i^S$ ). The first comparison elucidates the variation over time within the same experiment within each rat compared to the variation over time in the corresponding ventilator measurements. The second evaluates the variation in the mean values of the rats compared to the variation of the ventilator means. Therefore, we compared each pair of variances using the Brown-Forsythe test. The Brown-Forsythe test is a statistical test for the equality of group variances based on performing an ANOVA on a transformation of the response variable (46). In contrast to Levene's test, the Brown-Forsythe test utilizes the median instead of the mean to get a more robust result. The analysis is based on the two following assumptions:

1. All compounds of interest show a larger variance either within or between rats compared to the background noise.
2. Components that wash out in the respiratory system of the rat are not of interest for our study.

To exclude background noise VOCs from the set of selected analytes, the one-sided version of the Brown-Forsythe test implemented in the R package `car` (87) is utilized, see Equation 9.2. Let  $z_{ij} = |y_{ij} - \tilde{y}_j|$  where  $\tilde{y}_j$  is the median of group  $j$  and  $n_j$  the number of observations in that group. We further define  $N$  as the number of observations and  $p$  as the number of groups. Thus the F statistic of the the Brown-Forsythe test is the ANOVA

on  $z_{ij}$ , while  $\tilde{z}_{.j}$  and  $\tilde{z}_{..}$  are the group and overall means of  $z_{ij}$ .

$$F = \frac{(N - p)}{(p - 1)} \frac{\sum_{j=1}^p n_j (\tilde{z}_{.j} - \tilde{z}_{..})^2}{\sum_{j=1}^p \sum_{i=1}^{n_j} (z_{ij} - \tilde{z}_{.j})^2} \quad (9.2)$$

All components showing a p-value below 0.005 for both of the tests are included into the final set of components.

### 9.2.3 Model Selection

In order to appropriately model the variation of the single rats as well as the variation within groups, a linear mixed-effects model was applied. The combination of both fixed effect and random effect models is a state of the art technique for modeling longitudinal data, see Section 2.6 for more details (178; 133). In our approach, we are expecting the influence of the treatment groups on the measurements to be non-random, modeled by fixed-effects. In contrast, the individual rats are treated as random effects, based on the assumption that their intercepts and slopes follow a random normal distribution centered by their treatment group mean. However, the probability distribution for multiple measurements of a single rat have the same structure. This allows the explicit modeling and analysis of between- and within-individual variation. The fundamental framework of the linear mixed model is shown in Equation 9.3.

$$\begin{aligned} Y_{i,j} &= (\beta_0 + \beta_1 \text{CLI} + u_{0,i}) \\ &\quad + (\beta_2 + \beta_3 * \text{CLI} + u_{1,i}) * t_j + \epsilon_{i,j} \\ Y_{i,j} &= \text{Intensity for rat } i \text{ at time } j \\ \text{CLI} &= \text{Indicator for class CLI} \\ u_{0,i} &= \text{Rat intercept \& } u_{1,i} = \text{Rat slope} \\ \epsilon_{i,j} &= \text{Residual error} \\ t_j &= \text{time } j \end{aligned} \quad (9.3)$$

**Alignment of Time Points** For comparability, the time points of the treatment (either control="SHAM" or sepsis="CLI") in the rats are aligned. All rats tested positive for their particular treatment at about 300 minutes after their treatment. For practical reasons, this time point of interest is shifted to zero and denoted by  $T_0$  in the following. All other time points are adjusted accordingly.

**Model Selection** Frequently, the driving force of a septic condition is an underlying bacterial infection in the organism. Therefore, in this study the sepsis is induced by intestinal bacteria. We assume that bacterial growth strongly influences the observed time course of the septic condition. This growth typically follows four different phases, (1) lag

phase, (2) exponential, (3) stationary, and (4) death phase (244). Thus, we derive the assumption that these disease stages will be reflected in the metabolite patterns over time. The death phase however is an exception and not relevant for our analysis, since the rats will likely reach septic shock before entering this phase. In order to adjust the metabolites model according to this knowledge, we refer to a recent study of Sekse *et al.* 2012. They proposed a linear spline model that contains two knots producing three linear intervals to directly model bacterial growth (196). Additionally, more advanced non-linear models like polynomial, or cubic splines have been proposed. Nevertheless, the amount of data for this first longitudinal attempt is limited. It is therefore appropriate to restrict the complexity of the model and follow the approach proposed by Sekse *et al.* and apply a two knot linear spline model.

In this study, we focus on the difference of the treatment groups at our time point of interest ( $T_0$ ) and the behavior of the molecular species resulting in this divergence. Therefore, the second knot point is fixed at  $T_{K2} = T_0$ . However, the position of the first knot point  $T_{K1}$  has to be determined empirically using the data. In order to find the optimal time point, a set of spline models with varying knot positions ( $T_{K1} = 50, 100, 150, 200$  or  $250$  minutes) is constructed for each compound. For comparison, the simple linear model as well as the single spline (knot at  $T_0$ ) are added. All in all we are fitting 7 models for each compound.

**Akaike Weights** To evaluate the optimal knot point position, the Akaike weights are calculated for the set of 7 models of each component. The Akaike weights are straightforward to compute from the raw Akaike information criterion (AIC) and provide an intuitive interpretation as the probabilities of the given model as compared to the best model in an AIC sense (222). Essentially, the relative likelihood  $L$  of the model  $i$  is estimated using the shifted AIC values and normalized so that  $\sum w_i(\text{AIC}) = 1$ , see Equation 9.4.

$$\begin{aligned} \Delta_i(\text{AIC}) &= \text{AIC}_i - \min(\text{AIC}) \\ w_i(\text{AIC}) &= \frac{e^{(-0.5*\Delta_i(\text{AIC}))}}{\sum_{k=1}^K e^{(-0.5*\Delta_k(\text{AIC}))}} \end{aligned} \quad (9.4)$$

The sum of all Akaike weights for each knot position results in the overall quality, while the maximum sum value indicates the optimal position for the second knot point denoted by  $T_{K1}$ .

**Linear Mixed Effects Spline Model** The overall hypothesis is that the two treatment groups CLI and SHAM will show different intensities in a given compound at the time point of interest or have different time-course patterns in the compound. Therefore, the intercepts at the time point  $T_0$  and the slope of the preceding spline section have to be assessed and compared using hypothesis tests. Additionally, a re-parameterization of the time variable in the spline model is necessary. The times are shifted and scaled such that

$T_{K2} = T_0 = 0$  and  $|T_{K1} - T_{K2}| = 1$ .

$$\begin{aligned}
 T < k_1 & \begin{cases} t_1 & = 1 - k_1 \\ t_2 & = k_1 \\ t_3 & = 0 \end{cases} \\
 k_1 \leq T < k_2 & \begin{cases} t_1 & = 0 \\ t_2 & = t \\ t_3 & = 0 \end{cases} \\
 k_2 \leq T & \begin{cases} t_1 & = 0 \\ t_2 & = k_2 \\ t_3 & = t - k_2 \end{cases}
 \end{aligned} \tag{9.5}$$

Thus, the spline segments are modeled according to Equation 9.5 (89). Subsequently, each compound is modeled by the linear mixed effect spline model, utilizing the previously determined knot points according to the given Equation 9.6.

$$\begin{aligned}
 Y_{i,j} &= (\beta_0 + \beta_1 \text{CLI} + u_{0,i}) \\
 &+ (\beta_2 + \beta_3 * \text{CLI}) * (t_{1j} + t_{2j} + t_{3j}) \\
 &+ u_{1,i} * t_j + \epsilon_{i,j}
 \end{aligned} \tag{9.6}$$

## 9.2.4 Evaluation and Identification

Finally, we evaluate the hypothesis that the SHAM and the CLI treatment groups are varying in the slope of their second spline segment or consequential at the point of interest  $T_0$ . The conditional t-tests are based on the maximum likelihood estimates of the variances. It is testing the marginal significance of the fixed effect coefficient (slope and intercept) with the other fixed effects in the model. To account for the problem of multiple testing a post hoc method called false discovery rate (FDR) correction is applied to adjust the p-values. The FDR method proposed by Benjamini and Hochberg is less conservative compared to the Bonferroni approach, thus has improved statistical power (26). The FDR correction is applied separately for each hypothesis (intercept and slopes). This is reasonable since intercept and slope and are dependent variables (58).

## 9.2.5 Implementation

The MCC/IMS raw data was preprocessed using the commercial VisualNow software package. The statistical modeling and evaluation of the longitudinal models is implemented in R (115) utilizing the following additional packages:

*nlme* for modeling linear mixed effect spline models (177);

*car* to performe the one-sided Brown-Forsythe test (87);

*sm* to create the density figures (40);

## 9.3 Results

The pre-processing of the raw MCC/IMS data as described before, resulted in a set of 100 peaks of potentially interesting compounds.

### 9.3.1 Peak Screening

In order to remove the peaks related to the background noise of the ventilation system, the between and within experiment variation was analyzed separately. As an example, Figure 9.2 shows the comparison of the distribution of the rat and respiratory group means and residuals of the fitted model for an example peak, respectively.

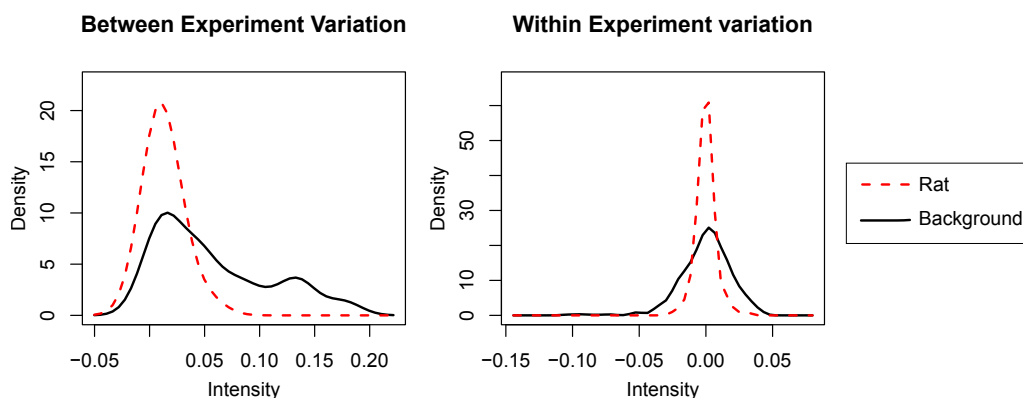


Figure 9.2: The first plot shows the distribution of the rat and respiratory experiment means. The second plot shows the distribution of the rat and respiratory residuals. The variances of both the experiment means and residual values of the described linear rat model vs. a simple linear respirator model are compared based on the Brown-Forsythe test.

For each of the compounds, the Brown-Forsythe test is performed, on both sets of these intercept and residual values, comparing rat vs. respirator variation. This background screening evaluation of the distributions of the respirator and rat measurements identified 43 peaks that showed a significantly larger variation of peak intensities within the background compared to the rat measurements. It follows that the set of compounds is reduced from 100 to 57. These 57 compounds are selected for further processing.

### 9.3.2 Model Selection

To account for the nonlinear bacterial growth we follow the proposed two knot linear spline model of Sekse *et al.* 2012. We extend this approach by combining it with a random mixed effects model. While  $T_{K2}$  is fixed to our point of interest, we evaluate the potential

of 6 different spline models based on varying positions for  $T_{K1} = 50, 100, 150, 200, 250$  minute as well as the single knot spline and compare these to a simple linear model. The sum of the Akaike weights is used to assess the model quality over all peaks, shown in Table 9.3.2. The highest value indicates the best fitting model. The results show that the optimal position for the first knot point is at  $T_{K1} = 150$ . Thus, this model is used for all further computations.

Model	Sum of Akaike Weights
Linear	1.82
Single spline	5.16
250	8.27
200	7.46
150	13.43
100	12.65
50	8.22

### 9.3.3 Model Evaluation

Finally the resulting model was evaluated on all 57 relevant compounds. The FDR corrected hypothesis tests show significant difference in the slope at 300 minutes for only two of the components. However, twenty components showed a significant difference in the slope of the treatment groups, see Table 9.3. Figure 9.4 gives an example for the spline model of one of the two peaks that shows significance in both tests.

## 9.4 Discussion

We detected 100 VOCs in total in the exhaled air of the examined rats. A comparison of the longitudinal intensity progression showed that 20 of these volatiles had statistical significant differences between septic rats and control rats regarding the intercept or the slope of the spline model (Figure 9.3). These volatile compounds were identified as ketones, hydrocarbons, alcohols and aldehydes, which is consistent with earlier findings in the breath of rats (97) and mice (215). The publication by Fink *et al.* that introduced the data used for this study, focused on the comparison of peak intensities in two hourly increments (83). They revealed two compounds with statistically significant differences in septic rats compared to control animals. In accordance with the results of that previous study we found the analytes acetone and 3-pentanone with a significant reduction in septic animals. This decrease in ketone bodies is most probably not due to low blood sugar levels (165), but due to the increased utilization of ketone bodies by the peripheral tissue which is not compensated by liver ketogenesis (165). In addition to these two components, the analysis revealed 18 VOCs which differ between the two groups. These include three further ketones, namely 3-hydroxy-2-butanone, 2-octanone and 2, 3-heptanedione,



Analyte	P-Value $T_0$	P-Value Slope	Slope	Analyte-Class
Ethylbenzene	0.97	0.047		Hydrocarbons
2-Heptanol	0.35	0.00072		Alcohols
3-Octanol	0.7	0.047		Alcohols
3-Hydroxy-2-Butanone	0.7	0.0083		Ketones
2-Phenylethanol	0.28	1.9e-05		Alcohols
2/3-Heptanedione	0.53	0.047		Diketones
gamma-Terpinene	0.18	0.047		Hydrocarbons
5-Methyl-3-Heptanone	0.34	0.00072		Ketones
2-Pentanone	0.043	0		Ketones
2-Hexanone	0.37	0.00074		Ketones
1-Pentanol	0.41	0.00093		Alcohols
2-Propanol	0.18	0.0041		Alcohols
Butanal	0.33	0.032		Aldehydes
Acetone	0.043	0.0015		Ketones
1-Propanol	0.18	4.6e-05		Alcohols
2-Octanone	0.28	0.00011		Ketones
2-Hexanol	0.55	0.0086		Alcohols
Butanal	0.37	3.1e-05		Aldehydes
3-Hydroxy-2-Butanone	0.81	0.00072		Ketones
n-Nonane	0.18	2e-06		Hydrocarbons

Figure 9.3: Table of components that show a significant treatment difference in either the intercept or the slope of the spline model. The given p-values are corrected for multiple comparison using the false discovery rate correction (26). The arrows indicate the direction and strength of the change in the analyte intensity (CLI in yellow and SHAM in green). The last column indicates the analyte class of the component.

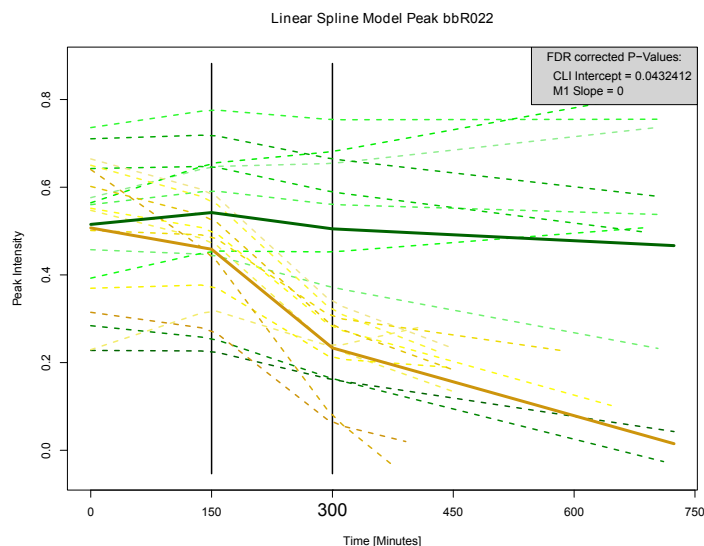


Figure 9.4: As an example, this plot shows the fitted two knot spline model of the compound bbR022. The vertical lines represent the spline knot positions. The dashed lines represent the predicted intensities from the model for each individual rat, while the solid bold lines represent the average intensity for rats in each group. Both rats and class curves are colored in green and yellow according to their grouping, SHAM and CLI respectively.

with a significant increase in septic rats after 5h. 3-hydroxy-2-butanone was found to be strongly released by *Staphylococcus aureus* (82) a facultative pathogen bacteria. 2-octanone and 2,3-heptanedione were detected in the headspace of *Enterobacter cloacae* respectively *Proteus mirabilis* cultures (118) which are common parts of the gut flora. Another VOC which may be associated with intestinal bacteria strains is 1-pentanol. It is produced by various bacteria but prominently by *Escherichia coli* (38) and strongly increased in CLI animals after 5 hours. It has been detected in the headspace of *Escherichia coli* infected blood cultures (7). A hypothesis to explain the late increase of those bacterial associated VOCs would be the required bacterial reproduction time. Whether these VOCs are bacterial products or otherwise related to the sepsis remains unclear. Another three molecules, 1-propanol, butanal and 2-hexanone show significant differences between septic rats and the control. Those have already been reported as significantly different between lipopolysaccharide treated rats and control animals by Fink *et al.* 2014. Treatment with lipopolysaccharide (LPS), which is an outer membrane part of gram-negative bacteria, is mimicking a bacterial infection and causes an endotoxemic shock as in microbial sepsis but is lacking the functional microorganisms. Therefore it is likely that septic and endotoxemic shock share a common pattern in the breath. Summarizing the above, it should be noted that it is unclear which of these VOCs might be useful for diagnostics to indicate a septic state, especially because sepsis may progress differently in humans than

in rodents. The presented framework was able to detect a new pattern of compounds which differ between septic and healthy animals, however, certain limitations remain. A major issue is the small number of samples that limits the generalizability of the study. This restricts the design to linear splines instead of more complex models, such as, for instance, cubic or polynomial splines. Furthermore, the transferability of the findings resulting from a rodan study to humans is a subject of frequent discussions. Therefore, additional work is needed to determine which exhaled VOCs are of diagnostic benefit in humans.

## 9.5 Conclusion

Despite its limitations, our study provides an important extension of prior work on sepsis progression and suggests future directions for clinical studies on exhaled air and computational breath analysis. Moving the point of view from single time points to sepsis progression and the analysis from hypothesis tests to accurate time dependent models, enabled us to find 18 additional potential biomarkers. We discovered septic-typical phenomena like a decrease of ketone bodies and the occurrence of compounds found in facultative pathogen bacteria or species that are common parts of the gut flora, during the course of the sepsis. Nevertheless, future work might require new and larger clinical studies, for instance, the combination of several biomarkers (176) and more complex non-linear models.

**Breath analysis for clinical diagnostics** Several general lessons can be drawn from this work regarding the future of breath analysis studies and clinical diagnostics utilizing exhaled air as an assessment:

- Cross-sectional time point analysis might not always be robust enough to perturbations and individual patient variation.
- A thorough screening approach for background noise was established, which is generalizable to future longitudinal studies.
- Longitudinal analysis of the time course of volatile organic compounds in the rat breath introduces a novel prediction model for sepsis progression.
- Longitudinal analysis might give better insights on the biological background of sepsis progression in the future.

This study shows the potential of longitudinal studies to be beneficial for various kinds of diseases, where single time point analysis is too noisy. It can be concluded that the MCC/IMS, suitable for online monitoring, in combination with advanced modeling techniques like mixed effect models, present an innovative tool to potentially complement conventional diagnostic techniques.



## Chapter 10

# Discussion and Conclusion

The innovations in clinical diagnostics during the last century made a huge step from initial conservative anamnesis to the analysis of diseases not only on cellular but molecular level. Many of these techniques rely on body fluids or tissue samples and are, therefore, often invasive and painful. This is further exacerbated by the fact that some techniques show low accuracy or are too slow for fast progressing lethal diseases (84). These facts drive the search for non-invasive and fast alternative diagnostic tools, for instance based on the analysis of metabolites within the human exhaled air. The novel field of breathomics emerged. The potential of well established analytical technologies such as MCC/IMS or GC/MS for clinical diagnostics is beyond dispute. Variations of these methods allow the discovery of a broad spectrum of components potentially related to specific diseases. Prior to this project, studies mainly relied on manual and/or non - robust processing and statistical analysis. The computational methods were mostly limited to automated preprocessing approaches. In combination with small sample sizes, this led to little concordance between studies, especially in terms of robust biomarker selection(173). In contrast, the past decades showed that bioinformatics methods allow for tremendous advances in various areas of biology and medicine, such as genetics (121), proteomics (162), or metabolomics (81). The main focus of this thesis is to bring the benefits of these advanced bioinformatics methods to the evolving field of breathomics and to drive the transition towards more automatization and robustness of computational analysis in breathomics research. The following sections will discuss, how these six studies address the major challenges described in the introduction.

### 10.1 Data Accumulation and Heterogeneity

Handling the challenge of increasingly large amounts of metabolomics and increasingly complex clinical data, requires an intelligent way of structuring and storing data. The key factors are the fast storage and retrieval, as well as the management of constantly evolving heterogeneous study and instrumental variables in a comprehensive centralized data repository. The presented IMSDB package combines two powerful computational tools, namely an extensible database and a machine learning toolkit. Both are fully

accessible through an intuitive graphical user interface (195). This database combines the advantages of a clinical open source data base like TrialDB (41) with a IMS metabolomics database schema such as developed by Lesniak (138). Additionally, this expands the opportunities of the diagnostic research in terms of confounder elimination, environmental factor evaluation and the organization of large scale clinical studies. The integrated statistical learning and feature selection process utilizes simple decision tree as proof of concept and ensures general extendibility by other data mining packages, for instance Weka (99). Further, the accessibility of the postgres database via Hibernate or ODBC allows the advanced user to attach various statistical tools such as the RapidMiner (126) toolkit or R (115). However, the following paragraph will discuss a number of limitations of this software package that have to be addressed. So far, the IMSDB solely relies on company specific MCC/IMS data format from VisualNow (B&S Analytik, Dortmund, Germany), as well as the predefined format of the patient table for clinical data. In recent years, other companies developed products for MCC/IMS analysis. Although they are mainly focusing on industrial process monitoring or drug detection and solely provide restricted access to the raw data, a general database for MCC/IMS research should support data formats of a broad spectrum of manufacturers. Incorporating data of analytical techniques utilized for breathomics such as GC/MS would complement the software by guiding it towards a more general clinical breathomics database. Along with growing sample and metabolite numbers, more adequate and advanced learning strategies for prediction and feature selection are required. These challenges on the way towards a more robust statistical analysis and evaluation and how they have been addressed in this thesis are discussed in the following section.

## 10.2 Robust and Generalizable Statistical Analysis and Evaluation

Prior to this thesis, the majority of studies focused on single variable statistical inference methods like hypothesis tests, or utilized variance analysis methods like PCA to find a relation to the label of interest (147; 230; 31; 23; 231). These methods entirely neglect the possibility of multi-compound relations, or rely on the assumption that the direction of the largest variance contains the most valuable information. In practice, this assumption is true in some, but not all cases. However, if no relation is found, this by no means implies that there is no information in the data. Moreover, confounding variables, such as technological biases, often generate large variances and conceal the important information.

Therefore, a handful of studies introduced more advanced statistical learning techniques (16; 85; 231). Their main goal was optimizing classification performance. However, these studies showed a few shortcomings: (1) insufficient validation, (2) only single methods evaluated, (3) not accounting for overfitting. In contrast, several studies within this thesis, demonstrated a thorough analysis and evaluation of sophisticated statistical learning methods. In 2012 we conducted a broad analysis of various linear and non-linear machine learning methods, evaluated via a ten fold cross validation schema, shown in Chapter 5. This demonstrated the general advantage of more sophisticated non-linear methods ap-

plied to MCC/IMS data over previous studies. Subsequently, the advantage of this robust approach was confirmed and extended while evaluating the peak detection methods, see Chapter 6. It demonstrated that, the selected cross validation set has a large influence on the classification performance on small data sets. This indicates that the evaluation of the performance variance on a number of random cross validation sets is crucial to get a deeper understanding of the real performance. It enabled a more robust evaluation and estimation of the real prediction accuracy. However, it has been reported, that repeated cross validation depending on number of samples and number of repetitions might lead to an overestimation of the performance variance (210). The same study further showed that MCC/IMS data sets influenced by the chosen peak detector are prone to overfitting if the machine learning methods are heavily tuned. Therefore, we can conclude that future MCC/IMS studies should ideally use permutation tests as shown in Chapter 6 to assess the extent of overfitting and reliability of the achieved performance.

In summary, future studies and especially automated analysis systems should provide and recommend such an evaluation framework in general. Additionally, similar to the study of Horsch *et al.*, the presented results indicate that non-linear ensemble tree methods and in particular random forest are favorable on MCC/IMS data sets. However, due to time constraints, several critical statistical issues haven't been addressed in this work. For instance, unbalanced data sets, as present in many breathomics studies (232; 103), have not been addressed by proper methodologies such as bootstrapping. Additionally, sophisticated supervised learning methods have been applied but not customized for MCC/IMS data.

**Robust and Generalizable Biomarker Candidate Detection** The second major challenge, besides an accurate prediction, is the selection of robust biomarker candidates. Previous studies focused on single VOC relations or unsuitable methods such as PCA. Although, one or more of the principal components might show discriminative power, the estimation of the individual influence of each molecule is non-trivial. A single study performed by Finthammer *et al.* established a interpretable model of equations utilizing the peak intensities (85). However, similar to PCA, the assessment of the importance of single VOC's remains difficult. Therefore, an analysis of the importance of each single molecules within a MCC/IMS data set utilizing more complex multi-compound approaches has been missing. This thesis, addressed this issue on several levels. The results in Chapter 5, present a set of most important features via random forest and linear SVM, which have been confirmed by comparisons with previous studies. The procedure of averaging the importance measures over a set of cross validation runs, resulted further in a more robust estimation of single feature importance. Another interesting approach is shown in Chapter 9. Here, features are not solely evaluated by a single time point, but their change over the time course of developing diseases. The results suggest that the time course of a component may be more informative than a single time point. More studies should investigate the evolution of metabolites in human exhaled air. However, the main limitation of this study is the focus on single features rather than the combination of peaks. Another issue that was not addressed in this thesis is the problem of correlated features. Tolosi *et al.* (208) showed that correlation in features can distort the importance values reported by various learning methods. The importance of information

that is similarly encoded by a large cluster of correlated features, will be largely underestimated.

In conclusion, although some weak points exist, this study introduced some of the most relevant feature selection approaches to computational breathomics. Further, it described a novel sequence of methods to discover longitudinal breath features.

### 10.3 Peak Detection Evaluation

A fully automated data analysis requires the users' trust in reliable results of each step of the processing pipeline. Although standard preprocessing procedures for quality enhancement are widely accepted in the community, automated methods for peak detection were neglected. Therefore, we conducted a broad analysis on the performance of automated MCC/IMS peak detection, as described in Chapter 6. In summary, all available software tools for MCC/IMS peak detection were evaluated by their ability to provide a valuable set of peaks optimizing the classification performance of the underlying data set. A comparison of the results to the manual selection shows a similar or only slightly decreased prediction accuracy of the approaches. Hence, for the first time we proved that these methods can be integrated into an automated analysis system, without the loss of important information for a classification task. However, the study was limited to a single data set and the methods were not optimized for their parameters. A recent study by collaborators within the SFB in Dortmund targeted this issue. They cross-compared the results of several preprocessing, peak detection, peak merging, and statistical learning techniques on three different data sets. Their results accredit our findings: (1) automated peak detection can compete with manual; (2) random forest showed the best overall results; (3) different peak detection approaches show less influence than the applied machine learning methods. Nevertheless, a few factors are still underrepresented in the results: The investigations did not evaluate the one-to-one measurement comparison of peak findings and do not account for overlaying peaks.

This work laid the foundation for further studies such as Horsch *et al.* (113). We showed that in order to gain the trust of the breathomics community in automated technologies, an adequate assessment of their quality is necessary. Therefore, this work set the foundation for more automatization in MCC/IMS data preprocessing and analysis.

### 10.4 Unknown Metabolites

Although the identity of an important metabolite that is key to the prediction of a disease is not essential for its performance as a biomarker, it might provide proof and reliability for its value as a predictor. Moreover, it encourages the research of the metabolic pathways of the disease that can ensure the direct relation of marker and disease. MIMA provides the first automatization of such an identification utilizing a mapping between MCC/IMS and the NIST-library search results of the corresponding GC/MS measurement. A major limitation of the method, however, is the small number of VOCs in the MCC/IMS compound database. Previous approaches, therefore, focused on the direct mapping of the GC and MCC peaks (48). Unfortunately, there is no linear nor con-



tinuously increasing relation between GC and MCC retention times. Depending on the temperature program in the GC column, components might switch positions influenced by their specific gas pressure (66). Other studies suggest that the corresponding relation between GC and MCC retention times depends on their specific molecular family (189). Nevertheless, the presented software tool provides a further step towards an automated pipeline as previously described.

A more difficult task is the identification of molecules present in low quantities. In particular this is the case when molecules can not be detected by GC/MS or can not be found in the MCC/IMS database. In these situations, the unsupervised learning methods such as provided by the Carotta software may aid the search for a possible solutions. For instance, a clustering of metabolites might yield a set of neighboring molecules that have been identified before. These similar components might allow to draw conclusions on the identity of the VOC of interest. This can either be decided for example upon common pathways or similar molecular properties.

## 10.5 Background Noise and Confounding Factors

The metabolomic composition of the breath and air in general is very vulnerable to perturbations such as environmental changes, nutrition, life style or instrumental background. Therefore, it is crucial to extract these influences before the data set is processed by the previously described learning and biomarker detection methods (156). The longitudinal study of Chapter 9 introduced a preprocessing step for screening volatile molecules originating from the instrumentation for ventilation of the rats. Previous studies have shown that certain components such as plastic tubes or other parts within the ventilating device produce a particular background noise (237). To separate this noise from the particles of interest, the variance of the respirator measurements are compared with the rat measurements, to reject molecules that are more active and variable in the background measurements. However, this study provided specific information about potential sources of confounding factors. If such information is missing, other approaches are required. Some very recent studies introduced unsupervised learning methods to analyze breath (155; 47; 80). Nevertheless, they lack in-depth evaluations using internal or external quality measures such as the Silhouette value or F-measure. Moreover, none of the existing studies emerged with a software or bioinformatics toolbox addressing the community's need for automatic unsupervised processing of breathomics data. This is provided by the Carotta software package presented in Chapter 8. The unsupervised methodology of Carotta can support the identification of metabolite subgroups that are related to confounding influences such as tooth paste. Thereby future studies can utilize Carotta to identify and eliminate disturbing influences of confounding factors. Nevertheless, the connection to a database of subject or patient attributes like the IMSDB as well as a proper identification (MIMA) of the component is required to explicitly identify certain molecules as confounding variables. Additionally, other promising approaches for multi-dimensional clustering exist, such as bi-clustering or co-clustering (see, e.g., (44; 205)). These might provide a more thorough assessment of the various influences of confounding and non-confounding factors on the metabolite pattern of the breath. In summary, both studies provide a solid example of confounding factor analysis and screening. They

further highlight the importance of such screenings for future studies.

## 10.6 Heterogeneous Diseases and Disease Stages

The existence of disease heterogeneity and stages is rarely dealt with, but are widely discussed in the breathomics community. Many pulmonary diseases are categorized by certain clinical symptoms but might originate from entirely different sources. One of the most prominent examples is COPD (chronic obstructive pulmonary disease), which describes a sum of diseases permanently affecting the lung function (88). The most recently published Carotta cluster analysis framework is dedicated to uncover, such hidden substructures or diseases subgroups. The combination of several similarity measures and sophisticated unsupervised statistical learning methods allow for a thorough analysis and discovery of potential subgroups. To ensure the accuracy of these clusterings, certain quality measures can assess the reliability of the results. Carotta presents the first breathomics dedicated software tool for automated disease subtyping and metabolite grouping. The study is one of the first applications of unsupervised learning in the area of breathomics. The main contribution of an easily accessible tool like Carotta is to further promote the application of powerful approaches like clustering in this field.

However, general unsupervised learning techniques such as clustering do not entirely solve the problem of evolving disease stages. Therefore, time dependent analysis of rat sepsis progression and its influence in the breath was initiated. The unpublished longitudinal analysis of this breath data represents the first of its kind. The effects of time on each single rat as well as the two treatment groups were modeled. Although, mixed effect models are common for longitudinal analysis, it is novel to evaluate bacterial driven sepsis via a spline structure. This particular model facilitates the observation of the rat metabolism reacting upon the bacterial intoxication. Admittedly, the present model does only evaluate a linear relations towards the grouping variable and treats each molecule separately. Additionally, the sample size makes the study prone to miss interpretation and does not allow a robust estimate. Nevertheless, this study serves as a proof of concept that the integration and analysis of time dependent data, especially in terms of disease progression, shows great promise. Time dependent data analysis further, provides a general pipeline as a starting point for future experiments and studies and shows current drawbacks, weaknesses as possible targets for improvement.

## 10.7 Usability, Maintainability and Re-Usability

In order to develop tools that serve the needs of inexperienced computer users or non computer science researchers, such as usability, maintainability and re-usability have to be met. The presented software tools: IMSDB, MiMa and Carotta aim to fulfill these requirements. The common programming language utilized for the provided tools is Java. This offers the advantage of easy portability to various platforms, for instance Windows or Linux. The usage of the packages requires the installation of current version of the Java Runtime Environment and postgres (IMSDB). Although, the installation is not very difficult, it might require the assistance of a more advanced computer user. All

packages provide an intuitive graphical user interfaces promoting data handling, analysis and visualization, in particular for non computer experts (106) Especially the intuitive menus and wizards guarantee an easy handling of the data import and result export. The visualization of results, for instance the decision tree in the IMSDB or the heat maps, dendogramms and multi-dimensional scaling plots in Carotta, further enhance the benefit of the packages.

All systems support extendibility, via a well defined structure and consequent use of standard interfaces. The IMSDB for example is easily expandable by integrating further WEKA libraries, or a direct connection of other machine learning tools via ODBC. The MiMa software provides simple extension of further data formats such as GC or MCC/IMS component lists. Finally, Carotta offers a comprehensive plug-in system for additional methods for all the described steps in the framework. This includes novel similarity definitions, clustering approaches, quality measures as well as methods for dimension reduction. All developed software packages are publicly available under the following addresses.

**IMSDB** ⇒ <http://imsdb.mpi-inf.mpg.de>

**Carotta** ⇒ <http://carotta.sdu.dk>

**MIMA** ⇒ <http://mima.mpi-inf.mpg.de>

This ensures a full re-usability of the tools as well as the underlying Java code.



# Chapter 11

## Conclusion



Figure 11.1: Schema of the steps in the automated pipeline: (1) Clinical IMS Database (2) Peak Detection Evaluation; (3) Metabolite Identification; (4) Supervised Learning; (5) Unsupervised Learning; (6) Longitudinal Analysis.

The vision of breathomics is a novel non-invasive tool for clinical diagnostics and medical research. This was further enhanced by the latest developments in high-performance methods optimized for measuring human exhaled air. While these devices for breathomics exist, we are facing the typical biomarker research barrier that related technologies (GWAS, NGS, microarrays, etc.) have seen before and successfully addressed with dedicated bioinformatics programs (72). The primary objective of my PhD project is to overcome the described obstacles in “computational breathomics” and to move from biomarker discovery to validation, from separability to predictability. Therefore, established technologies from computer science and machine learning needed to be adapted and integrated into breathomics research. Table 11.1 shows which of the briefly depicted study objectives addressed the previously described challenges. A centralized data structure can be the starting point for all the following analysis steps starting from preprocessing and ending with the proposal of biomarkers. The identification of the metabolites adds value to the results of potential confounders and biomarkers. The establishment of more advanced supervised, unsupervised and longitudinal learning methods will provide a large benefit in future clinical studies. Each of the introduced projects represents a small piece of the big picture of an automated processing and analysis system. The combination of these pieces sets the basis for an automated pipeline as depicted in Figure 11.1. Such comprehensive computational systems providing automated analysis from raw data to prediction would move the field forward towards more automatization, standardization

Table 11.1: This table depicts which projects addressed the previously described challenges.

		Challenges						
		Data Accumulation and Heterogeneity	Manual Peak Picking	Unknown Metabolites	Robust Statistics and Biomarkers	Background and Confounding Factors	Heterogeneous Diseases and Disease Stages	Usability, Maintainability, and Re-usability
Objectives	IMS Database	✓			✓			✓
	Evaluating Peak Detection		✓		✓			
	Metabolite Identification			✓				✓
	Robust Machine Learning				✓			
	Unsupervised Learning					✓	✓	✓
	Longitudinal Breath Analysis				✓		✓	

and generalization. In summary, this thesis presented pioneering contributions on several levels of MCC/IMS based data processing that have a major impact in the field of computational breathomics.

# Chapter 12

## Outlook

Despite the achievements of this work, some challenges to further improve automated computational breath analysis remain. The following paragraphs will outline the most critical issues.

**Generalized Platform** As of now, the projects described in this thesis are treated separately and addressed in independent software tools or scripts. Consequently, each program requires installation, operation and additional import and export of data. This circumstance is time consuming, confusing and not user friendly. Therefore, the most important future goal is the combination of the described projects into a sequential pipeline and to integrate it into a single user friendly system. Moreover, such a system could integrate the previously developed preprocessing methodology, such as denoising or peak detection, to further improve usability. Like in Carotta, an easy extensibility with novel pre- and postprocessing as well as methods for statistical analysis, enables a straight forward reanalysis of previously collected data sets with the goal to gain new insights.

**Centralized Web Service** In practice, several professions such as physicians, chemists, biologists, physics and statistical computer scientists are collaborating on one project. Therefore, an additional advantage will be to design the system as a centralized web service, which will enable easy but secure access. Updates and changes to the data will be instantly available to all participants. Moreover, every project member could add new methodology or combine these to a pipeline of processes and directly share this or the results with project members. User and access management will allow handling many projects with various rights for different sets of persons. Easy initiation of collaborations by simple changes of rights or, for instance, merging of projects.

**Data Expansion** Smolinska *et al.* proposed the fusion of data originating from several analytical technologies in order to improve predictive power (199). Therefore, the integration of, for instance, GC/MS measurements and the proposed data fusion methods (199) will allow for advanced cross platform analysis. For further advances, this data fusion might not only contain breathomics but also be extended to other

metabolomics or other omics data. In addition, the integration of GC/MS data in combination with the MIMA tool and NIST-library will provide automated peak identification. This knowledge allows for the connection to systems biology tools like KeyPathwayMiner (5) and would reveal more detailed information about the underlying molecular mechanisms of the disease.

In addition to a centralized storage and analysis framework, an integrated system would benefit from the following general improvements.

**Data Management** The IMSDB currently relies on a powerful PostgreSQL database management system, which requires extra installation and handling by a more advanced computer user. However, PostgreSQL could be substituted by Java file based databases such as the H2 database engine (4) or HyperSQL database (96). This would eliminate the required expert steps and ensure an easy handling by non computer expert users.

**Method and Parameter Evaluation** The evaluation approach described in Chapter 6 and by Horsch *et al.* assesses a broad spectrum of preprocessing, peak detection, merging and machine learning methods. However, the proposed generalized platform in combination with additional analytical data and statistical methods, will further allow for fully automated evaluation of all of the processing steps. In particular, this enables the comparison of various analytical technologies as well as other experimental parameters.

**Peak Identification** MIMA provides an easy and fast identification of MCC/IMS metabolites. Its main limitation is the dependence on the IMS component library. Since the analytical assessment of metabolite coordinates is comparably expensive, the expansion of this library is fairly limited. Different studies, however, have indicated the possibilities of MCC to GC mapping, in particular within a group of similar structured molecules (189). Hence, there is great potential to build models for similar molecules from the database. The key would be to define the physical and chemical properties that are most important to characterize the molecules. Consequently, these properties could be utilized as features to build a general retention time and/or mobility prediction model. Moreover, these models and the feature space spanned by the most important properties could yield information for the optimal selection of the next metabolite to be characterized. A set of unknown molecules that explores a void area in this space, would optimize the information gain and enlarge the database most efficiently.

**Robust and Generalizable Statistical Learning** This thesis presented two projects that introduced robust statistical analysis of MCC/IMS data by means of two examples, COPD data and peak evaluation. The analysis was implemented utilizing the statistical platform R. However, R requires certain expert knowledge in statistical programming. In contrast to this, a future system for supervised learning would be integrated into the previously described platform and allow for user friendly handling similar to Carotta. Besides various supervised learning methods, such a system would integrate various evaluation approaches. The most important examples are cross validation, variance or confidence intervals estimation as well as permutation



tests. Another issue are unbalanced data sets, as present in many breathomics studies (232; 103). The field of machine learning offers several methods to handle over representation of certain labels. Popular examples are bootstrapping and artificial balancing of the data sets by over or under-sampling. A general advanced analysis framework would provide and propose this functionality to the user, if required.

**Unsupervised Learning** Another future goal is the fusion of such a system with the Carotta framework in order to combine the functionality of supervised and unsupervised learning methods. For instance, a common application is the clustering of features in order to reduce redundancy and correlation in the data set. As previously explained, these are frequent sources for overfitting or distorted importance values (208).

**Longitudinal Analysis** In order to improve the modeling of longitudinal breath data, two main developments are inevitable. At first, we need to move from single towards multiple component models enabling interactions between several molecules. In addition to linear or linear-spline models, non-linear models, such as polynomials, need to be considered. However, more complex models require a larger number of experiments to achieve robust results. Although very few studies are dealing with longitudinal analysis of breath so far, this is expected to change in the near future.

The evaluated analytical technologies in combination with modern bioinformatics approaches present a promising tool for future clinical diagnostics. However, a lack of large screening studies and automated computational analysis hinders a translation to the world outside laboratories. Clearly, the presented PhD thesis is the first step towards such an automated analysis system. The described possible future optimizations will further enable this automatization. Finally, the combination with larger clinical studies and standardized as well as real world optimized devices will pave the way towards clinical practice. Consequently, a close collaboration between the research fields, starting with physics and medicine to statistical bioinformatics, will play a key role in this development.



# Bibliography

- [GOL] Global initiative for chronic obstructive lung disease. global strategy for diagnosis, management, and prevention of COPD: update 2014.
- [gua] Google core libraries, URL: <https://code.google.com/p/guava-libraries/wiki/Release16>.
- [JEx] Java excel API - a java API to read, write, and modify excel spreadsheets, URL:<http://jexcelapi.sourceforge.net/>.
- [JHe] JHeatChart - java library for generating heat map charts, URL:<http://www.javaheatmap.com/>.
- [NCI] National cancer institute, URL: [www.cancer.gov](http://www.cancer.gov).
- [WHO] World health organization, URL:<http://www.who.int/en/>.
- [1] (2011). *R: A language and environment for statistical computing*.
- [2] (2012). *MATLAB and Statistics Toolbox Release 2012b*. The MathWorks, Inc., Natick, Massachusetts, United States.
- [3] (2014). log4j 2.0, apache software foundation, URL:<http://logging.apache.org/log4j/2.x/>.
- [4] (2016). H2 database engine.
- [5] Alcaraz, N., Pauling, J., Batra, R., Barbosa, E., Junge, A., Christensen, A. G., Azevedo, V., Ditzel, H. J., and Baumbach, J. (2014). KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape. *BMC Systems Biology*, 8:99.
- [6] Algorithmics-Group (2009). MDSJ: Java library for multidimensional scaling (version 0.2), available at <http://www.inf.uni-konstanz.de/algo/software/mdsj/>.
- [7] Allardyce, R. A., Langford, V. S., Hill, A. L., and Murdoch, D. R. (2006). Detection of volatile metabolites produced by bacterial growth in blood culture media by selected ion flow tube mass spectrometry (SIFT-MS). *Journal of Microbiological Methods*, 65(2):361–365.

- [8] Altomare, D. F., Lena, M. D., Porcelli, F., Trizio, L., Travaglio, E., Tutino, M., Dragonieri, S., Memeo, V., and Gennaro, G. D. (2013). Exhaled volatile organic compounds identify patients with colorectal cancer. *British Journal of Surgery*, 100(1):144–151.
- [9] Amato, F., Lopez, A., Pena-Mendez, E. M., Vanhara, P., Hampl, A., and Havel, J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, 11(2):47–58.
- [10] Anderssen, E., Dyrstad, K., Westad, F., and Martens, H. (2006). Reducing over-optimism in variable selection by cross-model validation. *Chemometrics and Intelligent Laboratory Systems*, 84(1-2):69–74.
- [11] Armenta, S., Alcalá, M., and Blanco, M. (2011). A review of recent, unconventional applications of ion mobility spectrometry (IMS). *Analytica Chimica Acta*, 703(2):114–123.
- [12] Bader, S. (2008). *Identification and Quantification of Peaks in Spectrometric Data*. Thesis.
- [13] Bader, S., Urfer, W., and Baumbach, J. I. (2007). Reduction of ion mobility spectrometry data by clustering characteristic peak structures. *Journal of Chemometrics*, 20(3-4):128–135.
- [14] Basanta, M., Koimtzis, T., Singh, D., Wilson, I., and Thomas, C. L. P. (2007). An adaptive breath sampler for use with human subjects with an impaired respiratory function. *Analyst*, 132(2):153–163.
- [15] Bauer, C. and King, G. (2006). *Java Persistence with Hibernate*. Manning Publications Co., Greenwich, CT, USA.
- [16] Baumbach, J., Bunkowski, A., Lange, S., Oberwahrenbrock, T., Kleinbölting, N., Rahmann, S., and Baumbach, J. I. (2007). IMS2 - an integrated medical software system for early lung cancer detection using ion mobility spectrometry data of human breath. *Journal of Integrative Bioinformatics*, 4(3):75.
- [17] Baumbach, J., Rahmann, S., and Tauch, A. (2009a). Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms. *BMC Systems Biology*, 3:8.
- [18] Baumbach, J., Tauch, A., and Rahmann, S. (2009b). Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks. *Brief Bioinform*, 10(1):75–83.
- [19] Baumbach, J. I. (2006). Process analysis using ion mobility spectrometry. *Analytical and Bioanalytical Chemistry*, 384(5):1059–1070.
- [20] Baumbach, J. I. (2009a). Ion mobility spectrometry coupled with Multi-Capillary columns for metabolic profiling of human breath. *J. Breath Res. FIELD Full Journal Title:Journal of Breath Research*, 3:1–16.

- [21] Baumbach, J. I. (2009b). Ion mobility spectrometry coupled with multi-capillary columns for metabolic profiling of human breath. *Journal of Breath Research*, 3(3):1–16.
- [22] Baumbach, J. I. and Eiceman, G. A. (1999). Ion mobility spectrometry: Arriving on site and moving beyond a low profile. *Applied Spectroscopy*, 53(9):338A–355A.
- [23] Baumbach, J. I., Maddula, S., Sommerwerck, U., Besa, V., Kurth, I., Bödeker, B., Teschler, H., Freitag, L., and Darwiche, K. (2011). Significant different volatile biomarker during bronchoscopic ion mobility spectrometry investigation of patients suffering lung carcinoma. *International Journal for Ion Mobility Spectrometry*, 14(4):159–166.
- [24] Baumbach, J. I. and Westhoff, M. (2006). Ion mobility spectrometry to detect lung cancer and airway infections. *Spectroscopy Europe*, 18(6):22–27.
- [25] Beauchamp, J., Kirsch, F., and Buettner, A. (2010). Real-time breath gas analysis for pharmacokinetics: monitoring exhaled breath by on-line proton-transfer-reaction mass spectrometry after ingestion of eucalyptol-containing capsules. *Journal of Breath Research*, 4(Copyright (C) 2010 American Chemical Society (ACS). All Rights Reserved.). CAPLUS AN 2010:699470(Journal; Online Computer File).
- [26] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.
- [27] Berger, D. (1999a). A brief history of medical diagnosis and the birth of the clinical laboratory. part 1–Ancient times through the 19th century. *MLO: medical laboratory observer*, 31(7).
- [28] Berger, D. (1999b). A brief history of medical diagnosis and the birth of the clinical laboratory. part 2–Laboratory science and professional certification in the 20th century. *MLO: medical laboratory observer*, 31(8).
- [29] Bernal-Rusiel, J. L., Greve, D. N., Reuter, M., Fischl, B., and Sabuncu, M. R. (2013). Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models. *Neuroimage*, 66:249–260.
- [30] Besa, V., Teschler, H., Kurth, I., Khan, A. M. M., Zarogoulidis, P., Baumbach, J. I. I., Sommerwerck, U., Freitag, L., and Darwiche, K. (2015). Exhaled volatile organic compounds discriminate patients with chronic obstructive pulmonary disease from healthy subjects. *International journal of chronic obstructive pulmonary disease*, 10:399–406.
- [31] Bessa, V., Darwiche, K., Teschler, H., Sommerwerck, U., Rabis, T., Baumbach, J. I., and Freitag, L. (2011). Detection of volatile organic compounds (VOCs) in exhaled breath of patients with chronic obstructive pulmonary disease (COPD) by ion mobility spectrometry. *International Journal for Ion Mobility Spectrometry*, 14(Copyright (C) 2011 American Chemical Society (ACS). All Rights Reserved.):7–13. CAPLUS AN 2011:453257(Journal; Online Computer File).

- [32] Bijland, L. R., Bomers, M. K., and Smulders, Y. M. (2013). Smelling the diagnosis: a review on the use of scent in diagnosing disease. *The Netherlands journal of medicine*, 71(6):300–307.
- [33] Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, 83(3):377.
- [34] Blatt, R., Bonarini, A., and Matteucci, M. (2010). *Pattern Classification Techniques for Lung Cancer Diagnosis by an Electronic Nose*. Springer.
- [35] Bödecker, B., Vautz, W., and Baumbach, J. I. (2008). Visualisation of MCC/IMS - data. *International Journal for Ion Mobility Spectrometry*, 11(1):77–82.
- [36] Bödecker, B., Davies, A. N., Maddula, S., and Baumbach, J. I. (2010). Biomarker validation - room air variation during human breath investigations. *International Journal for Ion Mobility Spectrometry*, 13(3-4):177–184.
- [37] Bödecker, B., Vautz, W., and Baumbach, J. I. (2008). Peak finding and referencing in MCC/IMS-data. *International Journal for Ion Mobility Spectrometry*, 11(1):83–87.
- [38] Bos, L. D., Sterk, P. J., and Schultz, M. J. (2013). Volatile metabolites of pathogens: a systematic review. *PLoS Pathog.*, 9(5).
- [39] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.
- [40] Bowman, A. W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The kernel approach with S-Plus illustrations: The kernel approach with S-Plus illustrations*. OUP Oxford.
- [41] Brandt, C. A., Gadagkar, R., Rodriguez, C., and Nadkarni, P. M. (2004). Managing complex change in clinical study metadata. *Journal of the American Medical Informatics Association*, 11(5):380–391.
- [42] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- [43] Breiman, L., Friedman, J., Stone, C., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC, New York.
- [44] Bro, R., Papalexakis, E. E., Acar, E., and Sidiropoulos, N. D. (2012). Coclustering-a useful tool for chemometrics. *Journal of Chemometrics*, 26(6):256–263.
- [45] Broadhurst, D. I. and Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2(4):171–196.
- [46] Brown, M. B. and Forsythe, A. B. (1974). Robust tests for equality of variances. *Journal of the American Statistical Association*, 69:364–367.
- [47] Broza, Y. Y., Zuri, L., and Haick, H. (2014). Combined volatolomics for monitoring of human body chemistry. *Scientific Reports (Nature Publishing Group)*, 4:4611.

- [48] Bunkowski, A. (2010). Software tool for coupling chromatographic total ion current dependencies of GC/MSD and MCC/IMS. *International Journal for Ion Mobility Spectrometry*, 13(3-4).
- [49] Bunkowski, A. (2011). *MCC-IMS data analysis using automated spectra processing and explorative visualization methods*. Thesis, University Bielefeld: Bielefeld, Germany.
- [50] Buszewski, B., Keszy, M., Ligor, T., and Amann, A. (2007). Human exhaled air analytics: Biomarkers of diseases. *Biomedical Chromatography*, 21(6):553–566.
- [51] Carraro, S., Rezzi, S., Reniero, F., Heberger, K., Giordano, G., Zanconato, S., Guillo, C., and Baraldi, E. (2007). Metabolomics applied to exhaled breath condensate in childhood asthma. *American Journal of Respiratory and Critical Care Medicine*, 175(10):986–990.
- [52] Chatterjee, S., Castro, M., and Feller, J. F. (2013). An e-nose made of carbon nanotube based quantum resistive sensors for the detection of eighteen polar/nonpolar VOC biomarkers of lung cancer. *Journal of Materials Chemistry B*, 1(36):4563–4575.
- [53] Chen, R. S., Nadkarni, P., Marengo, L., Levin, F., Erdos, J., and Miller, P. L. (2000). Exploring performance issues for a clinical database organized using an entity-attribute-value representation. *Journal of the American Medical Informatics Association : JAMIA*, 7(5):475–487.
- [54] Chen, Z.-Y., Fan, Z.-P., and Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research*, 223(2):461–472.
- [55] Cheng, Z. J., Warwick, G., Yates, D. H., and Thomas, P. S. (2009). An electronic nose in the discrimination of breath from smokers and non-smokers: a model for toxin exposure. *Journal of Breath Research*, 3(3). Toxicology.
- [56] Cheung, W., Xu, Y., Thomas, C. L. P., and Goodacre, R. (2009). Discrimination of bacteria using pyrolysis-gas chromatography-differential mobility spectrometry (Py-GC-DMS) and chemometrics. *Analyst*, 134(3):557–563.
- [57] Chiba, S. (2014). Javassist (java programming assistant), URL:<http://www.csg.ci.i.u-tokyo.ac.jp/~chiba/javassist/>.
- [ChromaTOF] ChromaTOF, L. <http://www.leco.com>.
- [58] Conneely, K. N. and Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *American Journal of Human Genetics*, 81(6):1158–1168.
- [59] Cornu, J.-N. N., Cancel-Tassin, G., Ondet, V., Girardet, C., and Cussenot, O. (2011). Olfactory detection of prostate cancer by dogs sniffing urine: a step forward in early diagnosis. *European urology*, 59(2):197–201.

- [60] Dadamio, J., Van den Velde, S., Laleman, W., Van Hee, P., Coucke, W., Nevens, F., and Quiryneen, M. (2012). Breath biomarkers of liver cirrhosis. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences*, 905:17–22.
- [61] D’Addario, M., Kopczynski, D., Baumbach, J. I., and Rahmann, S. (2014). A modular computational framework for automated peak extraction from ion mobility spectra. *BMC Bioinformatics*, 15:25.
- [62] Daszykowski, M. (2007). From projection pursuit to other unsupervised chemometric techniques. *Journal of Chemometrics*, 21.
- [63] Daszykowski, M., Kaczmarek, K., Heyden, Y. V., and Walczak, B. (2007). Robust statistics in data analysis - a review basic concepts. *Chemometrics and Intelligent Laboratory Systems*, 85(2):203–219.
- [64] de Hoffmann, E. and Stroobant, V. (2007). *Mass Spectrometry: Principles and Applications, 3rd Edition*. Wiley-Interscience.
- [65] de Laurentiis, G., Paris, D., Melck, D., Montuschi, P., Maniscalco, M., Bianco, A., Sofia, M., and Motta, A. (2013). Separating Smoking-Related diseases using NMR-based metabolomics of exhaled breath condensate. *Journal of Proteome Research*, 12(3):1502–1511.
- [66] Dettmer-Wilde, K. and Werner, I. (2014). *Practical Gas Chromatography*. Springer.
- [67] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2011). *e1071: Misc Functions of the Department of Statistics (e1071)*. TU Wien.
- [68] Dragonieri, S., Annema, J. T., Schot, R., van der Schee, M. P. C., Spanevello, A., Carratu, P., Resta, O., Rabe, K. F., and Sterk, P. J. (2009). An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD. *Lung Cancer*, 64(2):166–170. Dragonieri, Silvano Annema, Jouke T. Schot, Robert van der Schee, Marc P. C. Spanevello, Antonio Carratu, Pierluigi Resta, Onofrio Rabe, Klaus F. Sterk, Peter J. Elsevier ireland ltd Clare.
- [69] Dragonieri, S., Brinkman, P., Mouw, E., Zwinderman, A. H., Carratu, P., Resta, O., Sterk, P. J., and Jonkers, R. E. (2013). An electronic nose discriminates exhaled breath of patients with untreated pulmonary sarcoidosis from controls. *Respiratory Medicine*, 107(7):1073–1078.
- [70] du Prel, J.-B., Röhrig, B., Hommel, G., and Blettner, M. (2010). Auswahl statistischer testverfahren. *Deutsches rzteblatt International*, 107(19):343–348.
- [71] Dwivedi, P., Puzon, G., Tam, M., Langlais, D., Jackson, S., Kaplan, K., Siems, W. F., Schultz, A. J., Xun, L., Woods, A., and Hill, H. H. (2010). Metabolic profiling of *Escherichia coli* by ion mobility-mass spectrometry with MALDI ion source. *Journal of Mass Spectrometry*, 45(12):1383–1393.
- [Eaton and Others] Eaton, J. W. and Others. GNU octave.



- [72] Eckel, S. P., Baumbach, J., and Hauschild, A.-C. (2014). On the importance of statistics in breath analysis - hope or curse? *Journal of Breath Research*, 8(1):012001.
- [73] Efron, B. and Tibshirani, R. (1997). Improvements on Cross-Validation: The 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.
- [74] Ehmann, R., Boedeker, E., Friedrich, U., Sagert, J., Dippon, J., Friedel, G., and Walles, T. (2011). Canine scent detection in the diagnosis of lung cancer: Revisiting a puzzling phenomenon. *European Respiratory Journal*, pages erj00517–2011+.
- [75] Eiceman, G. A. and Karpas, Z. (2005). *Ion Mobility Spectrometry*, volume 1. CRC Press, Taylor & Francis, Boca Raton, 2 edition.
- [76] Elsayed, I., Ludescher, T., King, J., Ager, C., Trosin, M., Senocak, U., Brezany, P., Feilhauer, T., and Amann, A. (2013). ABA-cloud: support for collaborative breath research. *Journal of Breath Research*, 7(2):026007.
- [77] et al Van, B. J. J., Berkel, J. J. B. N. V., Dallinga, J. W., Möller, G. M., Godschalk, R. W. L., Moonen, E. J., Wouters, E. F. M., and Schooten, F. J. V. (2010). A profile of volatile organic compounds in breath discriminates COPD patients from controls. *Respiratory medicine*, 104(4):557–563.
- [78] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.
- [79] Fay, M. P. and Proschan, M. A. (2010). Wilcoxon-Mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39.
- [80] Fens, N., van Rossum, A. G. J., Zanen, P., van Ginneken, B., van Klaveren, R. J., Zwinderman, A. H., and Sterk, P. J. (2013). Subphenotypes of Mild-to-Moderate COPD by factor and cluster analysis of pulmonary function, CT imaging and breathomics in a Population-Based survey. *Copd-Journal of Chronic Obstructive Pulmonary Disease*, 10(3):277–285.
- [81] Fernie, A. R., Trethewey, R. N., Krotzky, A. J., and Willmitzer, L. (2004). Metabolite profiling: from diagnostics to systems biology. *Nature Reviews Molecular Cell Biology*, 5(9):763–769.
- [82] Filipiak, W., Sponring, A., Baur, M. M., Filipiak, A., Ager, C., Wiesenhofer, H., Nagl, M., Troppmair, J., and Amann, A. (2012). Molecular analysis of volatile metabolites released specifically by *Staphylococcus aureus* and *Pseudomonas aeruginosa*. *BMC Microbiology*, 12:113.
- [83] Fink, T., Baumbach, J. I., and Kreuer, S. (2014a). Ion mobility spectrometry in breath research. *Journal of Breath Research*, 8(2):027104.
- [84] Fink, T., Wolf, A., Maurer, F., Albrecht, F. W., Heim, N., Wolf, B., Hauschild, A. C., Bödeker, B., Baumbach, J. I., Volk, T., Sessler, D. I., and Kreuer, S. (2014b). Volatile organic compounds during inflammation and sepsis in rats: A potential breath test using ion-mobility spectrometry. *Anesthesiology*.

- [85] Finthammer, M., Beierle, C., Fisseler, J., Kern-Isberner, G., Möller, B., and Baumbach, J. I. (2010). Probabilistic relational learning for medical diagnosis based on ion mobility spectrometry. *International Journal for Ion Mobility Spectrometry*, 13(2):83–92.
- [86] Fong, S. S., Rearden, P., Kanchagar, C., Sasseti, C., Trevejo, J., and Brereton, R. G. (2011). Automated peak detection and matching algorithm for gas Chromatography-Differential mobility spectrometry. *Analytical Chemistry*, 83(5):1537–1546.
- [87] Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage.
- [88] Garcia-Aymerich, J., Gómez, F. P., Benet, M., Farrero, E., Basagaña, X., Gayete, A., Paré, C., Freixa, X., Ferrer, J., Ferrer, A., Roca, J., Gáldiz, J. B., Sauleda, J., Monsó, E., Gea, J., Barberà, J. A., Agustí, A., and Antó, J. M. (2011). Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax*, 66(5):430–437.
- [89] Gauderman, W. J., Vora, H., Mc Connell, R., Berhane, K., Gilliland, F., Thomas, D., Lurmann, F., Avol, E., Kunzli, N., Jerrett, M., and Peters, J. (2007). Effect of exposure to traffic on lung development from 10 to 18 years of age: a cohort study. *Lancet*, 369(9561):571–577.
- [90] Gilbert, D. and Morgner, T. (2005). JFreeChart, URL:<http://www.jfree.org/jfreechart/index.html>.
- [91] Good, P. (2013). *Permutation Tests: A Practical Guide to Resampling Methods for Testing*.
- [92] Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Semin Hematol*, 45(3):135–40.
- [93] Gordon, S. M., Szidon, J. P., Krotoszynski, B. K., Gibbons, R. D., and O’Neill, H. J. (1985). Volatile organic compounds in exhaled air from patients with lung cancer. *Clinical Chemistry*, 31(8):1278–1282.
- [94] Gowda, H., Ivanisevic, J., Johnson, C. H., Kurczy, M. E., Benton, H. P., Rinehart, D., Nguyen, T., Ray, J., Kuehl, J., Arevalo, B., Westenskow, P. D., Wang, J., Arkin, A. P., Deutschbauer, A. M., Patti, G. J., and Siuzdak, G. (2014). Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Analytical Chemistry*, 86(14):6931–6939.
- [95] Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381):306–313.
- [96] Group, T. H. D. (2016). HyperSQL DataBase.
- [97] Guaman, A. V., Carreras, A., Calvo, D., Agudo, I., Navajas, D., Pardo, A., Marco, S., and Farre, R. (2012). Rapid detection of sepsis in rats through volatile organic compounds in breath. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, 881-882(Copyright (C) 2012 U.S. National Library of Medicine.):76–82.

- [98] Guyon, I., Weston, J., and Barnhill, S. (2002). Gene selection for cancer classification using support vector machine. *Machine Learning*, 46:389–422.
- [99] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.
- [100] Hand, D. and Till, R. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. 45(2):171–186.
- [101] Hastie, T., Tibshirani, R., and Friedman (2001). *The elements of statistical learning*, volume 1. Springer New York.
- [102] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- [103] Hauschild, A. C., Baumbach, J. I., and Baumbach, J. (2012a). Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification. *Genetics and molecular research*, 11(3):2733–2744.
- [104] Hauschild, A. C., Baumbach, J. I., and Baumbach, J. (2013a). Paving the way for automated clinical breath analysis and biomarker detection. Highlight paper presentation at the German Conference on Bioinformatics, Göttingen, Germany.
- [105] Hauschild, A. C., Baumbach, J. I., and Baumbach, J. (2014). Novel developments in computational clinical breath analysis and biomarker detection. Highlight Talk at 13th European Conference on Computational Biology (ECCB).
- [106] Hauschild, A.-C., Frisch, T., Baumbach, J. I., and Baumbach, J. (2015). Carotta: Revealing hidden confounder markers in metabolic breath profiles. *Metabolites*, 5(2):344–363. Accessed: 2015-05-31.
- [107] Hauschild, A. C., Kopczynski, D., D’Addario, M., Baumbach, J. I., Rahmann, S., and Baumbach, J. (2013b). Peak detection method evaluation for ion mobility spectrometry by using machine learning approaches. *Metabolites*, 3(2):277–293.
- [108] Hauschild, A. C., Schneider, T., Pauling, J., Rupp, K., Jang, M., Baumbach, J. I., and Baumbach, J. (2012b). Computational methods for metabolomic data analysis of ion mobility spectrometry data - reviewing the state of the art. *Metabolites*, 2(4):733–755.
- [109] Herbig, J., Mueller, M., Schallhart, S., Titzmann, T., Graus, M., and Hansel, A. (2009). On-line breath analysis with PTR-TOF. *J. Breath Res. FIELD Full Journal Title: Journal of Breath Research*, 3(2). 9 Biochemical Methods Ionimed Analytik GmbH, Innsbruck, Austria. Journal; Online Computer File 1752-7155 written in English.
- [110] Hill, H. H., Siems, W. F., Stlouis, R. H., and McMinn, D. G. (1990). Ion mobility spectrometry. *Analytical Chemistry*, 62(23):A1201–A1209. ISI Document Delivery No.: EK803 Times Cited: 158 Cited Reference Count: 61.

- [111] Ho, T. J., Kuo, C. H., Wang, S. Y., Chen, G. Y., and Tseng, Y. F. J. (2013). True ion pick (TIPick): a denoising and peak picking algorithm to extract ion signals from liquid chromatography/mass spectrometry data. *Journal of Mass Spectrometry*, 48(2):234–242.
- [112] Hoffmann, N., Keck, M., Neuweiger, H., Wilhelm, M., Hogy, P., Niehaus, K., and Stoye, J. (2012). Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets. *BMC bioinformatics*, 13.
- [113] Horsch, S., Kopczynski, D., Baumbach, J. I., Rahnenführer, J., and Rahmann, S. (2015). From raw ion mobility measurements to disease classification: a comparison of analysis processes. Technical report, PeerJ PrePrints.
- [114] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24.
- [115] Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314.
- [116] Jain, A. K. and Maheswari, S. (2013). Survey of recent clustering techniques in data mining. *Journal of Current Computer Science and Technology*, 3.
- [117] Jünger, M., Bödeker, B., and Baumbach, J. I. (2010). Peak assignment in multi-capillary column/ion mobility spectrometry using comparative studies with gas chromatography/mass spectrometry for VOC analysis. *Analytical and bioanalytical chemistry*, 396(1):471–482.
- [118] Jünger, M., Vautz, W., Kuhns, M., Hofmann, L., Ulbricht, S., Baumbach, J. I., Quintel, M., and Perl, T. (2012). Ion mobility spectrometry for microbial volatile organic compounds: A new identification tool for human pathogenic bacteria. *Applied Microbiology and Biotechnology*, 93:2603–14.
- [119] Kanu, A. B., Hampikian, G., Brandt, S. D., and Hill, H. H. (2010). Ribonucleotide and ribonucleoside determination by ambient pressure ion mobility spectrometry (IMS). *Analytica Chimica Acta*, 658(1):91–97.
- [120] Karpas, Z., Guaman, A. V., Calvo, D., Pardo, A., and Marco, S. (2012). The potential of ion mobility spectrometry (IMS) for detection of 2,4,6-trichloroanisole (2,4,6-TCA) in wine. *Talanta*, 93(Copyright (C) 2012 American Chemical Society (ACS). All Rights Reserved.):200–205.
- [121] Katsanis, S. H. and Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. *Nature Reviews Genetics*, 14(6):415–426.
- [122] Kemper, A. and Eickler, A. (2006). *Datenbanksysteme - Eine Einführung*. Oldenbourg, München, Germany, 6th edition.
- [123] Kennard, R. W. and Stone, L. (1969). Computer aided design of experiments. *Technometrics*, 11:137–148.

- [124] Kessler, N., Neuweger, H., Bonte, A., Langenkamper, G., Niehaus, K., Nattkemper, T. W., and Goesmann, A. (2013). MeltDB 2.0-advances of the metabolomics software system. *Bioinformatics*, 29(19):2452–2459.
- [125] King, J., Kupferthaler, A., Frauscher, B., Hackner, H., Unterkofler, K., Teschl, G., Hinterhuber, H., Amann, A., and Högl, B. (2012). Measurement of endogenous acetone and isoprene in exhaled breath during sleep. *Physiological Measurement*, 33(3):413–428.
- [126] Klinkenberg, R. (2013). Introduction to data mining and RapidMiner. In *Rapid-Miner: Data Mining Use Cases and Business Analytics Applications*, pages 1–2. Chapman and Hall/CRC.
- [127] Koczulla, R., Hattesoehl, A., Schmid, S., Bödeker, B., Maddula, S., and Baumbach, J. (2011). MCC/IMS as potential noninvasive technique in the diagnosis of patients with COPD with and without alpha 1-antitrypsin deficiency. *International Journal for Ion Mobility Spectrometry*, 14(4):177–185.
- [128] Kopczynski, D., Baumbach, J. I., and Rahmann, S. (2012). Peak Modeling for Ion Mobility Spectrometry Measurements. In *Proceedings of 20th European Signal Processing Conference*.
- [129] Kreuder, A. E., Buchinger, H., Kreuer, S., Volk, T., Maddula, S., and Baumbach, J. I. (2011). Characterization of propofol in human breath of patients undergoing anesthesia. *International Journal for Ion Mobility Spectrometry*, 14(4):167–175.
- [130] Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A. (2011). Density-based clustering. *WIREs Data Mining Knowl Discov*, 1(3):231–240.
- [131] Krooshof, P. W., Ustun, B., Postma, G. J., and Buydens, L. M. (2010). Visualization and recovery of the (bio)chemical interesting variables in data analysis with support vector machine classification. *Analytical Chemistry*, 82(16):7000–7.
- [132] Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.
- [133] Laird, N. M. and Ware, J. H. (1982). Random-Effects models for longitudinal data. *Biometrics*, 38:963–974.
- [134] Langley, P. and Sage, S. (1994). Induction of selective bayesian classifiers. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*.
- [135] Langley, R. J., Tipper, J. L., Bruse, S., Baron, R. M., Tsalik, E. L., Huntley, J., Rogers, A. J., Jaramillo, R. J., O'Donnell, D., Mega, W. M., Keaton, M., Kensicki, E., Gazourian, L., Fredenburgh, L. E., Massaro, A. F., Otero, R. M., Fowler, V. G., Rivers, E. P., Woods, C. W., Kingsmore, S. F., Sopori, M. L., Perrella, M. A., Choi, A. M., and Harrod, K. S. (2014). Integrative "omic" analysis of experimental bacteremia identifies a metabolic signature that distinguishes human sepsis from systemic inflammatory response syndromes. *American Journal of Respiratory and Critical Care Medicine*, 190(4):445–455.

- [136] Lee, S., Zipunnikov, V., Shiee, N., Crainiceanu, C., Caffo, B. S., and Pham, D. L. (2013). Clustering of high dimensional longitudinal imaging data. In *Proceedings of the 2013 International Workshop on Pattern Recognition in Neuroimaging, PRNI '13*, pages 33–36, Washington, DC, USA. IEEE Computer Society.
- [137] Lehmann (1986). *Testing statistical hypotheses*, volume 150. Wiley New York et al, second edition edition.
- [138] Lesniak, T. (2007). Entwurf, erprobung und bewertung eines informationsschemas fuer untersuchungen von metaboliten. Diploma thesis, University of Dortmund, Dortmund, Germany.
- [139] Liaw, A. and Wiener, M. (2002). Classification and regression by random forest. Technical Report 2:1822, R News.
- [140] Ligor, M., Ligor, T., Bajtarevic, A., Ager, C., Pienz, M., Klieber, M., Denz, H., Fiegl, M., Hilbe, W., Weiss, W., Lukas, P., Jamnig, H., Hackl, M., Buszewski, B., Miekisch, W., Schubert, J., and Amann, A. (2009). Determination of volatile organic compounds in exhaled breath of patients with lung cancer using solid phase microextraction and gas chromatography mass spectrometry. *Clinical Chemistry and Laboratory Medicine*, 47(5):550–560. Times Cited: 4.
- [141] Ligor, T., Ligor, M., Amann, A., Ager, C., Bachler, M., Dzien, A., and Buszewski, B. (2008). The analysis of healthy volunteers' exhaled breath by the use of solid-phase microextraction and GC-MS. *J. Breath Res. FIELD Full Journal Title:Journal of Breath Research*, 2(4). 9 Biochemical Methods Department of Anaesthesiology and Critical Care Medicine, Innsbruck Medical University, Innsbruck, Austria. Journal 1752-7155 written in English.
- [142] Liua, M., Wang, M., Wang, J., and Li, D. (2013). Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and chinese vinegar. *Sensors and Actuators B: Chemical*, 133.
- [143] Lommen, A. (2009). MetAlign: Interface-Driven, versatile metabolomics tool for hyphenated Full-Scan mass spectrometry data preprocessing. *Analytical Chemistry*, 81(8):3079–3086.
- [144] Ludescher, T., Feilhauer, T., Amann, A., and Brezany, P. (2013). Towards a high productivity automatic analysis framework for classification: An initial study. In Perner, P., editor, *Advances in Data Mining. Applications and Theoretical Aspects*, volume 7987 of *Lecture Notes in Computer Science*, pages 25–39. Springer Berlin Heidelberg.
- [145] Luts, J., Molenberghs, G., Verbeke, G., Huffel, S. V., and Suykens, J. A. K. (2012). A mixed effects least squares support vector machine model for classification of longitudinal data. *Computational Statistics & Data Analysis*, 56(3):611–628.
- [146] Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4).

- [147] Maddula, S., Blank, L., Schmid, A., and Baumbach, J. I. (2009). Detection of volatile metabolites of *escherichia coli* by multi capillary column coupled ion mobility spectrometry. *Analytical and Bioanalytical Chemistry*, 394(3):791–800.
- [148] Maddula, S., Rabis, T., Sommerwerck, U., Anhenn, O., Darwiche, K., Freitag, L., Teschler, H., and Baumbach, J. (2011). Correlation analysis on data sets to detect infectious agents in the airways by ion mobility spectrometry of exhaled breath. *International Journal for Ion Mobility Spectrometry*, 14(4):197–206.
- [149] Maddula, S., Rupp, K., and Baumbach, J. I. (2012). Recommendation for an upgrade to the standard format in order to cross-link the GC/MSD and the MCC/IMS data. *International Journal for Ion Mobility Spectrometry*, 15(2).
- [150] Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, 1(1):24–45.
- [151] Malley, J. D., Dasgupta, A., and Moore, J. H. (2013). The limits of p-values for biological data mining. *Biodata Mining*, 6.
- [152] Maurer, F., Hauschild, A. C., Eisinger, K., Baumbach, J., Mayor, and Baumbach, J. I. (2014). MIMA a software for analyte identification in MCC/IMS chromatograms by mapping accompanying GC/MS measurements. *International Journal for Ion Mobility Spectrometry*, 17(2):95–101.
- [153] Mayr, F. B., Yende, S., and Angus, D. C. (2014). Epidemiology of severe sepsis. *Virulence*, 5(1):4–11.
- [154] Merkl, R. and Waack, S. (2009). *Bioinformatik Interaktiv*.
- [155] Meyer, N., Dallinga, J. W., Nuss, S., Moonen, E., van Berkel, J., Akdis, C., van Schooten, F., and Menz, G. (2014). Defining adult asthma endotypes by clinical features and patterns of volatile organic compounds in exhaled air. *Respiratory Research*, 15(1):136.
- [156] Miekisch, W., Herbig, J., and Schubert, J. K. (2012). Data interpretation in breath biomarker research: pitfalls and directions this work was presented at the breath analysis summit 2011 (11-14 september 2011, parma, italy). *Journal of breath research*, 6(3):036007.
- [157] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940. ACM.
- [158] Mieth, M., Schubert, J. K., Groger, T., Sabel, B., Kischkel, S., Fuchs, P., Hein, D., Zimmermann, R., and Miekisch, W. (2010). Automated needle trap heart-cut GC/MS and needle trap comprehensive two-dimensional GC/TOF-MS for breath gas analysis in the clinical environment. *Analytical chemistry*, 82(6):2541–51.

- [159] Nadkarni, P. M., Marengo, L., Chen, R., Skoufos, E., Shepherd, G., and Miller, P. (1999). Organization of heterogeneous scientific data using the EAV/CR representation. *Journal of the American Medical Informatics Association*, 6(6):478–493.
- [160] Neuweger, H., Albaum, S. P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J., and Goesmann, A. (2008). MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, 24(23):2726–2732.
- [161] Neuweger, H., Baumbach, J., Albaum, S., Bekel, T., Dondrup, M., Hüser, A. T., Kalinowski, J., Oehm, S., Pühler, A., Rahmann, S., Weile, J., and Goesmann, A. (2007). CoryneCenter - an online resource for the integrated analysis of corynebacterial genome and transcriptome data. *BMC Systems Biology*, 1:55.
- [162] Ng, E. W., Wong, M. Y., and Poon, T. C. (2014). Advances in MALDI mass spectrometry in clinical diagnostic applications. *Topics in current chemistry*, 336:139–175.
- [163] Nixon, M. and Aguado, A. S. (2008). *Feature Extraction & Image Processing, Second Edition*. Academic Press, 2nd edition.
- [164] Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28.
- [165] Ohtoshi, M., Jikko, A., Asano, M., Uchida, K., Ozawa, K., and Tobe, T. (1984). Ketogenesis during sepsis in relation to hepatic energy metabolism. *Research in Experimental Medicine (Berl)*, 184(4):209–219.
- [166] Ojala, M. and Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863.
- [167] Orsenigo, C. and Vercellis, C. (2010). Time series gene expression data classification via l 1-norm temporal SVM. In Dijkstra, T., Tsivtsivadze, E., Marchiori, E., and Heskes, T., editors, *Pattern Recognition in Bioinformatics*, volume 6282 of *Lecture Notes in Computer Science*, pages 264–274. Springer Berlin Heidelberg.
- [168] Paccanaro, A., Casbon, J. A., and Saqi, M. A. S. (2006). Spectral clustering of protein sequences. *Nucleic Acids Research*, 34(5):1571–1580.
- [169] Paul (1981). Efficient computing of regression diagnostics. *The American Statistician*, 35(4):234–242.
- [170] Pereira, J., Porto-Figueira, P., Cavaco, C., Taunk, K., Rapole, S., Dhakne, R., Nagarajaram, H., and Camara, J. S. (2015). Breath analysis as a potential and non-invasive frontier in disease diagnosis: an overview. *Metabolites*, 5(1):3–55.
- [171] Perl, T., Bödeker, B., Jünger, M., Nolte, J., and Vautz, W. (2010). Alignment of retention time obtained from multicapillary column gas chromatography used for VOC analysis with ion mobility spectrometry. *Analytical and Bioanalytical Chemistry*, 397(6):2385–2394.



- [172] Perl, T., Carstens, E., Hirn, A., Quintel, M., Vautz, W., Nolte, J., and Jünger, M. (2009). Determination of serum propofol concentrations by breath analysis using ion mobility spectrometry. *British Journal of Anaesthesia*, 103(6):822–827.
- [173] Perl, T., Vautz, W., and Kunze-Szikzay, N. (2016). Atemgasdiagnostik - was geht und was geht (noch) nicht? IMS-Anwendertreffen 2016.
- [174] Peters, S., van Velzen, E., and Janssen, H. G. (2009). Parameter selection for peak alignment in chromatographic sample profiling: objective quality indicators and use of control samples. *Analytical and Bioanalytical Chemistry*, 394(5):1273–1281.
- [175] Phillips, C. O., Syed, Y., Parthalain, N. M., Zwiggelaar, R., Claypole, T. C., and Lewis, K. E. (2012). Machine learning methods on exhaled volatile organic compounds for distinguishing COPD patients from healthy controls. *Journal of Breath Research*, 6(3):036003.
- [176] Pierrakos, C. and Vincent, J. L. (2010). Sepsis biomarkers: a review. *Critical Care*, 14(1).
- [177] Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-111.
- [178] Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.
- [179] Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *Bmc Bioinformatics*, 11.
- [180] Purkhart, R., Hillmann, A., Graupner, R., and Becher, G. (2012). Detection of characteristic clusters in IMS-spectrograms of exhaled air polluted with environmental contaminants. *International Journal for Ion Mobility Spectrometry*, pages 1–6.
- [181] Purkhart, R., Kohler, H., Liebler-Tenorio, E., Meyer, M., Becher, G., Kikowatz, A., and Reinhold, P. (2011). Chronic intestinal Mycobacteria infection: discrimination via VOC analysis in exhaled breath and headspace of feces using differential ion mobility spectrometry. *Journal of Breath Research*, 5(2):027103.
- [182] Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- [183] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., and Müller (2011). pROC: an open-source package for r and s+ to analyze and compare ROC curves. *BMC Bioinformatics*.
- [184] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65.

- [185] Rupp, K. (2012). Breath analysis using MCC/IMS and GC/MSD. *Technical report for Collaborative Research Center SFB 876 Providing Information by Resource-Constrained Data Analysis*, page 88.
- [186] Ruzsanyi, V. (2013). Ion mobility spectrometry for pharmacokinetic studies—exemplary application. *Journal of breath research*, 7(4).
- [187] Ruzsanyi, V., Baumbach, J. I., Sielemann, S., Litterst, P., Westhoff, M., and Freitag, L. (2005a). Detection of human metabolites using multi-capillary columns coupled to ion mobility spectrometers. *Journal of Chromatography A*, 1084(1-2):145–151.
- [188] Ruzsanyi, V., Baumbach, J. I., Sielemann, S., Litterst, P., Westhoff, M., and Freitag, L. (2005b). Detection of human metabolites using multi-capillary columns coupled to ion mobility spectrometers. *Journal of Chromatography A*, 1084(1-2):145–151.
- [189] Sanders, D. and Wortelmann, T. (2016). Datenbank software zur Substanz-Identifizierung in GCxIMS messungen. 6. IMS Anwendertreffen, Hannover, Germany, March 2016.
- [190] Sape-Research-Group (2014). Hac - a java class library for hierarchical agglomerative clustering, URL:<http://sape.inf.usi.ch/hac/>.
- [191] Saraoglu, H. M. and Kocan, M. (2010). Determination of blood glucose Level-Based breath analysis by a quartz crystal microbalance sensor array. *Sensors Journal, IEEE*, 10(1):104–109.
- [192] Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.
- [193] Schaub, N., Frei, R., and Muller, C. (2011). Addressing unmet clinical needs in the early diagnosis of sepsis. *Swiss Med Wkly*, 141.
- [194] Schmidt, L. (2016). Blutvergiftung - eine der häufigsten todesursachen. *Frankfurter Allgemeine*.
- [195] Schneider, T., Hauschild, A. C., Baumbach, J. I., and Baumbach, J. (2013). An integrative clinical database and diagnostics platform for biomarker identification and analysis in ion mobility spectra of human exhaled air. *Journal of Integrative Bioinformatics*, 10:733–755.
- [196] Sekse, C., Bohlin, J., Skjerve, E., and Vegarud, G. E. (2012). Growth comparison of several *Escherichia coli* strains exposed to various concentrations of lactoferrin using linear spline regression. *Microbial Informatics and Experimentation*, 2:5.
- [197] Simenhoff, M. L., Burke, J. F., Saukkonen, J. J., Ordinario, A. T., and Doty, R. (1977). Biochemical profile of uremic breath. *The New England Journal of Medicine*, 297:132–135.

- [198] Smolinska, A., Blanchet, L., Coulier, L., Ampt, K. A. M., Luider, T., Hintzen, R. Q., Wijmenga, S. S., and Buydens, L. M. C. (2012). Interpretation and visualization of Non-Linear data fusion in kernel space: Study on metabolomic characterization of progression of multiple sclerosis. *PLoS ONE*, 7(6).
- [199] Smolinska, A., Hauschild, A. C., Fijten, R., Dallinga, J., Baumbach, J., and van Schooten, F. (2014). Current breathomics ? a review on data pre-processing techniques and machine learning in metabolomics breath analysis. *Journal of Breath Research*, 8(2):027105.
- [200] Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics*, 19(415):415–428.
- [201] Snijders, T. (2014). Multilevel analysis. In Lovric, M., editor, *International Encyclopedia of Statistical Science*, pages 879–882. Springer Berlin Heidelberg.
- [202] Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1).
- [203] Statheropoulos, M., Mikedi, K., Agapiou, A., Georgiadou, A., and Karma, S. (2006). Discriminant analysis of volatile organic compounds data related to a new location method of entrapped people in collapsed buildings of an earthquake. *Analytica Chimica Acta*, 566(2):207–216.
- [204] Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., and Kohlbacher, O. (2008). OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics*, 9(1):163.
- [205] Sun, P., Speicher, N. K., Röttger, R., Guo, J., and Baumbach, J. (2014). Bi-Force: large-scale bicluster editing and its application to gene expression data biclustering. *Nucleic Acids Research*, 42(9).
- [206] Szymanska, E., Saccenti, E., Smilde, A. K., and Westerhuis, J. A. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*, 8(1):S3–S16.
- [207] Therneau, T. M. and Atkinson, B. (2011). rpart: Recursive partitioning. r package version 3.1-50. R port by Brian Ripley.
- [208] Toloşi, L. and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994.
- [209] Trygg, J., Gabrielsson, J., and Lundstedt, T. (2009). *Background Estimation, Denoising, and Preprocessing*. Comprehensive chemometrics. Elsevier, Elsevier.
- [210] Vanwinckelen, G. and Blockeel, H. (2012). On estimating model accuracy with repeated cross-validation. In *BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning*, pages 39–44.

- [211] Vardhan, A., Prastawa, M., Sadeghi, N., Vachet, C., Piven, J., and Gerig, G. (2015). Joint longitudinal modeling of brain appearance in multimodal MRI for the characterization of early brain developmental processes. In Durrleman, S., Fletcher, T., Gerig, G., Niethammer, M., and Pennec, X., editors, *Spatio-temporal Image Analysis for Longitudinal and Time-Series Image Data*, volume 8682 of *Lecture Notes in Computer Science*, pages 49–63. Springer International Publishing.
- [212] Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91.
- [213] Vautz, W., Baumbach, J. I., and Jung, J. (2006a). Beer fermentation control using ion mobility spectrometry - results of a pilot study. *Journal of the Institute of Brewing*, 112(2):157–164.
- [214] Vautz, W., Bödeker, B., Bader, S., and Baumbach, J. I. (2008). Recommendation of a standard format for data sets from GC/IMS with sensor-controlled sampling. *International Journal for Ion Mobility Spectrometry*, 11(1-4):71–76.
- [215] Vautz, W., Nolte, J., Bufe, A., Baumbach, J. I., and Peters, M. (2010). Analyses of mouse breath with ion mobility spectrometry: a feasibility study. *Journal of Applied Physiology*, 108(3):697–704.
- [216] Vautz, W., Nolte, J., Fobbe, R., and Baumbach, J. I. (2009). Breath analysis-performance and potential of ion mobility spectrometry. *J. Breath Res. FIELD Full Journal Title:Journal of Breath Research*, 3.
- [217] Vautz, W., Zimmermann, D., Hartmann, M., Baumbach, J. I., Nolte, J., and Jung, J. (2006b). Ion mobility spectrometry for food quality and safety. *Food Additives and Contaminants*, 23(11):1064–1073.
- [218] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth edition*. Springer.
- [219] Vincent, L. (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE transactions on pattern analysis and machine intelligence*, 13(6):583–598.
- [220] Vogtland, D. and Baumbach, J. I. (2009). Breit-Wigner-function and IMS-signals. *International Journal for Ion Mobility Spectrometry*, 12(3):109–114.
- [221] Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4(3):274–290.
- [222] Wagenmakers, E. J. and Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review*, 11(1):192–196.
- [223] Wang, M., Vandermaar, A. J., and Srivastava, K. D. (2002). Review of condition assessment of power transformers in service. *Electrical Insulation Magazine*, pages 12–26.

- [224] Wang, W. L. and Lin, T. I. (2014). Multivariate t nonlinear mixed-effects models for multi-outcome longitudinal data with missing values. *Stat Med*, 33(17):3029–3046.
- [225] Watson, J. T. and Sparkman, O. D. (2007). *Introduction to Mass Spectrometry: Instrumentation, Applications, and Strategies for Data Interpretation*. John Wiley & Sons Ltd., west sussex, England.
- [226] Webb, A. (2002). *Statistical Pattern Recognition*. John Wiley & Sons Ltd.
- [227] Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). *klaR Analyzing German Business Cycles. Data Analysis and Decision Support*. Springer.
- [228] Wenig, P. and Odermatt, J. (2010). OpenChrom: A cross-platform open source software for the mass spectrometric analysis of chromaographic data. *BMC Bioinformatics*, 11(405).
- [229] Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J. J., van Duijnhoven, J. P. M., and van Dorsten, F. A. (2008). Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81–89.
- [230] Westhoff, M., Litterst, P., Freitag, L., and Baumbach, J. I. (2007). Ion mobility spectrometry in the diagnosis of sarcoidosis: Results of a feasibility study. *Journal of Physiology and Pharmacology*, 58:739–751.
- [231] Westhoff, M., Litterst, P., Maddula, S., Bödecker, B., and Baumbach, J. I. (2011). Statistical and bioinformatical methods to differentiate chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control by breath analysis using ion mobility spectrometry. *International Journal for Ion Mobility Spectrometry*, 11(4):139–149.
- [232] Westhoff, M., Litterst, P., Maddula, S., Bödecker, B., Rahmann, S., Davies, A. N., and Baumbach, J. I. (2010). Differentiation of chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control group by breath analysis using ion mobility spectrometry. *International Journal for Ion Mobility Spectrometry*, 13(3-4):131–139.
- [233] Whittle, C. L., Fakharzadeh, S., Eades, J., and Preti, G. (2007). Human breath odors and their use in diagnosis. *Annals of the New York Academy of Sciences*, 1098:252–266.
- [234] Wittkop, T. (2010). Clustering biological data by unraveling hidden transitive substructures.
- [235] Wittkop, T., Emig, D., Lange, S. J., Rahmann, S., Albrecht, M., Morris, J. H., Böcker, S., Stoye, J., and Baumbach, J. (2010). Partitioning biological data with transitivity clustering. *Nature Methods*, 7(6):419–420.
- [236] Wiwie, C., Baumbach, J., and Röttger, R. (2015). Comparing the performance of biomedical clustering methods. *Nature Methods*, 12(11):1033–1038.

- [237] Wolf, A., Baumbach, J. I., Kleber, A., Maurer, F., Maddula, S., Favrod, P., Jang, M., Fink, T., Volk, T., and Kreuer, S. (2014). Multi-capillary column-ion mobility spectrometer (MCC-IMS) breath analysis in ventilated rats: a model with the feasibility of long-term measurements. *Journal of Breath Research*, 8(1):016006.
- [238] Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., and Wishart, D. S. (2012). MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Research*, 40(Web Server issue):W127–133.
- [239] Yamanaka, K., Chun, S. J., Boillee, S., Fujimori-Tonou, N., Yamashita, H., Gutmann, D. H., Takahashi, R., Misawa, H., and Cleveland, D. W. (2008). Astrocytes as determinants of disease progression in inherited amyotrophic lateral sclerosis. *Nature Neuroscience*, 11(3):251–253.
- [240] Young, R. P., Hopkins, R. J., Christmas, T., Black, P. N., Metcalf, P., and Gamble, G. D. (2009). COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. *European Respiratory Journal*, 34(2):380–386.
- [241] Zerzucha, P. and Walczak, B. (2012). Concept of (dis)similarity in data analysis. *TrAC Trends in Analytical Chemistry*, 38(0):116–128.
- [242] Zhang, D. and Shen, D. (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PLoS ONE*, 7(3).
- [243] Zhu, J. J., Bean, H. D., Kuo, Y. M., and Hill, J. E. (2010). Fast detection of volatile organic compounds from bacterial cultures by secondary electrospray Ionization-Mass spectrometry. *Journal of Clinical Microbiology*, 48(12):4426–4431.
- [244] Zwietering, M. H., Jongenburger, I., Rombouts, F. M., and van 't Riet, K. (1990). Modeling of the bacterial growth curve. *Applied and Environmental Microbiology*, 56(6):1875–1881.

# Acronyms

- AAF** Automatic Analysis Framework. 35
- ABA-Cloud** Advanced Breath Analysis platform. 35
- ACC** accuracy. 49, 50, 94, 181
- AIC** Akaike information criterion. 129
- ANN** artificial neural networks. 49, 55
- AUC** area under the ROC curve. 50, 81, 94, 96, 97, 98
- BC** bronchial barcinoma. 59, 118, 178
- CART** classification and regression tree. 47
- CI** chemical ionization. 30
- COPD** chronic obstructive pulmonary disease. 59, 72, 74, 77, 78, 79, 81, 82, 83, 85, 86, 109, 117, 118, 117, 118, 121, 176, 178, 181, 182
- CQPP** chunk query peak persistence. 72, 74, 176
- CV** cross validation. 81, 97, 177
- DT** decision tree. 46, 176
- EAV** entity-attribute-value. 67
- EI** electron ionization. 30
- EM** expectation maximization. 34
- FDR** false discovery rate. 130, 132
- GC/MS** gas chromatography/mass spectrometry. 15, 17, 27, 28, 30, 56, 60, 61, 65, 66, 71, 110, 111, 119, 124

- GIM** growing interval merging. 32
- GUI** graphical user interface. 73, 74
- HAC** hierarchical agglomerative clustering. 42, 110, 116, 117
- IMPRS** International Max Planck Research School. 89
- IMS** ion mobility spectrometer. 27, 28, 124, 137, 175
- KIST** Korean Institute of Science and Technology. 101
- LMS** local maxima search. 33, 89, 95, 96, 98
- MCC** multi capillary column. 27, 28, 92, 175
- MCC/IMS** multi capillary column coupled with an ion mobility spectrometer. 15, 18, 27, 28, 29, 30, 32, 33, 34, 35, 36, 45, 48, 54, 55, 56, 59, 60, 61, 62, 63, 66, 67, 68, 69, 70, 71, 72, 73, 74, 77, 78, 79, 80, 83, 86, 92, 95, 111, 117, 121, 124, 125, 126, 175, 176, 181
- MDS** multi-dimensional scaling. 40, 111, 115, 116
- MPCL** merged peak cluster localization. 32, 33, 89, 95, 96, 98
- MS** mass spectrometer. 30
- NCI** National Cancer Institute. 59
- NIST** National Institute of Standards and Technology. 105, 107
- NIST-library** NIST/EPA/NIH mass spectral library, which was developed by the NIST. 18, 30, 102, 103, 104, 105, 140, 147
- NPV** negative predictive value. 49, 50, 94, 181
- ORM** object-relational mapping. 71
- PCA** principal component analysis. 39, 40, 77, 110, 138, 139
- PCs** principal components. 39
- PME** peak model estimation. 32, 34, 71, 89, 95, 96, 98
- PPV** positive predictive value. 49, 50, 94, 181
- PSE** Problem Solving Environments. 35, 67
- PTR/MS** proton transfer reaction mass spectrometry. 15, 17, 27



- PTR/TOF/MS** proton transfer reaction time of flight mass spectrometry. 17
- RF** random forest. 47
- RIP** reactant ion peak. 30, 32, 56, 175
- RKHS** reproducing kernel Hilbert space. 48
- ROC** receiver operating characteristic. 50, 81, 110, 171
- SEN** sensitivity. 49, 50, 181
- SFB** Collaborative Research Center SFB 876. 89, 140
- SPE** specificity. 49, 50, 181
- SPME-GC/MS** solid phase micro extraction/gas chromatography coupled with mass spectrometry. 27
- SVM** support vector machine. 46, 48, 51, 52, 55, 80, 81, 98, 175
- ToF** time-of-flight. 30
- VOC** volatile organic compound. 27, 66, 77, 79, 80, 82, 83, 86, 109, 132, 140
- WBMPD** wavelet-based multiscale peak detection. 32
- WHO** World Health Organization. 59
- WMW** Wilcoxon-Mann-Whitney test. 77
- WST** water shed transformation. 32, 34, 89, 95, 96, 98



# List of Figures

1.1	Key objectives of this thesis: (1) Clinical IMS Database; (2) Peak Detection Evaluation; (3) Metabolite Identification; (4) Supervised Learning; (5) Unsupervised Learning; (6) Longitudinal Analysis. . . . .	21
2.1	Sample and data flow: BioScout device; corresponding $CO_2$ profile; resulting MCC/IMS chromatogram. The $CO_2$ profile allows for precise analysis of air originating from certain parts of the respiratory system. For instance, solely end tidal air. Pictures of the BioScout device taken from the B&S Analytik website, <a href="http://www.bs-analytik.de/">http://www.bs-analytik.de/</a> . . . . .	28
2.2	An example of MCC/IMS chromatogram. The X-axis corresponds to the reduced inverse ion mobility $1/K_0$ ( $Vs/cm^2$ ) and it is proportional to the drift time (IMS), while the Y-axis corresponds to the retention time $r$ (MCC) which is proportional to the substance's affinity for the stationary phase. The colors reflect the signal height: the yellow color for the highest signal and the white color the lowest [white < blue < purple < red < yellow].	29
2.3	Example of a processing strategy of MCC/IMS data involving RIP-detailing (Step 1) and denoising and baseline correction (Step 2), peak picking (Step 3). . . . .	31
2.4	The Figure depicts the categorization of statistical hypothesis tests and examples (70). The "Nominal Distribution" column is a sub category of non-parametric tests and characterizes a group of tests suited for binomial or other nominal random variables. Further sub categories exist, like equal variances of the compared data sets, but are beyond the scope of this thesis.	37
2.5	Examples for histogram, box and scatter plot. (1) The histogram shows the general distribution of the data and allows discover left or right shifts in the distribution. (2) The box plot depicts the quantiles and Whiskers of the distribution. The inner quantile range (IQR) is defined as the difference between the $Q_1 = 25\%$ and $Q_3 = 75\%$ quantiles. $Q_2$ is equal to the median. The upper and lower Whisker are the minimum and maximum values of the data, except outliers are present. An outlier is a value that is smaller than $Q_1 - 1.5IQR$ or larger than $Q_3 + 1.5IQR$ . (3) The scatter plot depicts a relation between two variables. . . . .	39

2.6	Example for a linear classification using support vector machine. The blue and yellow samples are separated by a SVM. The samples marked by a square are the so called support vectors. . . . .	46
2.7	The figure on the left side depicts a DT and the separation of the data. The figure to the right shows the corresponding explanation of the key values, class distribution, the corresponding decision, and the percentage of the original data within this node. . . . .	47
2.8	Receiver operating characteristics curve represents the relation between Sensitivity and 1-Specificity. The AUC is the area under this so called ROC curve. The shown diagram was plotted utilizing the R package <i>pROC</i> and the public R data set example <i>aSAH</i> . . . . .	52
2.9	The Figure depicts the overview of the permutation tests procedure. The upper path shows the estimation of the real accuracy. The second path depicts the estimation of the distribution of random accuracies. Finally, it can be evaluated if a value equal to the real accuracy can occur by chance. . . . .	54
3.1	Shows a schematic of the operational procedure utilized for the sham operation or to induce sepsis. . . . .	63
4.1	Overview of the three main database components and their relations. The central core entities of each component are linked directly. Auxiliary entities of generalized data structure and peak component can be found in the next paragraph. Associations and maximum cardinalities (one or many) are represented as solid lines labeled with characters 1, N, M. . . . .	70
4.2	Overview of the generalized data structure (ontology) component. The central entity of this component <i>Concept</i> is associated to combinations of <i>Attribute</i> and value entities. In addition, a <i>Relation</i> is used to associate <i>Concept</i> entities to each other. . . . .	71
4.3	The plot depicts the average running times determined by ten repetitions of uploading data sets with an increasing number of IMS measurement files (25, 50, 75, 100, 125). Each raw file is associated with 120 peak regions. Four different uploading strategies are shown. The two methods CQPP (black line, pink line) and native peak persistence (red line, blue line) are thereby combined with the options for storing the IMS raw data measurement files into the database. . . . .	75
4.4	Illustration of the decision tree built by training on the data set that contains the combination of the labels (classes) <i>COPD</i> vs. <i>Control</i> . . . . .	75
5.1	Overview of the integrated statistical approach to evaluate COPD-related metabolic MCC/IMS profiles. The numbering indicates the different steps of the approach. . . . .	79
5.2	ROC curves of the statistical learning methods based on the estimated class probabilities. . . . .	83

5.3	Barplot of the ten best features determined by the feature selection of the linear support vector machine and the random forest, on the COPD vs. HC classification. Depicted are the ten best features according to their weights generated by the linear SVM (left) and the Gini index provided by the random forest (right). The right Y-axis of each Figure lists the names of peaks/VOCs, while the left Y-axis states the importance. . . . .	84
6.1	Shown is the structure of the evaluation pipeline. The three sections (1) preprocessing, (2) peak detection and postprocessing, and (3) machine learning and evaluation describe the details of the overall analysis. The figure is adapted from Hauschild <i>et al.</i> 2013 (104) . . . . .	91
6.2	Boxplots of 100 runs of the ten-fold CV for the linear SVM and the random forest method. . . . .	97
6.3	Boxplots illustrating the variation within the random forest and linear SVM tuning results on a single ten-fold cross validation set respectively. The dark cyan boxes show the results when tuning the original feature sets. The light cyan boxes show the results when tuning the randomly labeled feature sets. . . . .	99
7.1	Schematic view of the step by step process . . . . .	104
7.2	One MCC/IMS spectrum including the MIMA-generated layer (rectangles). 107	
7.3	Time expenditure evaluation of the manual and automatic workflow, estimated for a standard human breath or room air sample containing 30 to 40 compounds performed by an experienced postdoctoral researcher in our lab. . . . .	108
8.1	The Carotta pipeline consists of several steps: (1) importation of pre-processed data; (2) similarity calculation; (3) clustering; (4) clustering quality; (5) similarity or clustering visualization; (6) subset selection. Intermediate results can be inspected, optimized, and repeated with an arbitrary depth. . . . .	112
8.2	The subset selection allows for a <b>nested analysis</b> of the hidden structures in the data. Steps 2–4 from Figure 8.1 are repeated on a selected subset (or all subsets). In this example, the artificial data is first clustered by metabolites. In the next step, a threshold is selected that splits the data into three metabolite clusters and the data is separated accordingly. Subsequently, the data for each of these metabolites is used separately to access the distances between the samples (patients). These similarity matrices are further used for clustering the samples. Finally, an external quality measure is used to evaluate the association of each set of metabolites to the selected patient annotations (i.e., labels; here: <i>health</i> , <i>nutrition</i> and <i>smoking</i> ). The F-measure plots show to what extent the metabolite clusters explain the different labels. . . . .	114

- 8.3 The graphical user interface is split into three basic regions: (**A**) the data and results area lists available (intermediate and final) results; (**B**) the details panel lists attributes of the data, such as mean or minimum values, or the history of previous processing steps; (**C**) the main result panel displays the visualization of the intermediate and final results. . . . . 115
- 8.4 Result of the nested hidden structure analysis applied to the artificial data set. Each plot depicts the F-Measure evaluation of the clustering on one of the four different data sets: Entire Data Set, and each of the distinct Metabolite Subsets A-C. The subsets were previously determined by a clustering of the metabolites. The lines show the F-measure for the labels: *health* in green, *smoking* in red, and *nutrition* in blue, for every clustering threshold. The top left plot depicts the result for the entire data set, it shows a dominant effect of the confounder *smoking*, which overlays the main outcome variable *health*. The other three plots show the results for each cluster of correlated metabolites, separately. Each of the metabolite subsets clearly corresponds to one of the three labels: *health*, *smoking*, and *nutrition*. . . . . 118
- 8.5 Comparison of the clustering results of the 14 metabolite clusters. The plot shows the F-measure for different clustering thresholds computed against the disease annotation (COPD, COPD with BC, and healthy). Since the annotation defines three groups of patients, the performance of the clustering results at a clustering threshold of 3 is of particular interest (x-axis). 119
- 8.6 Comparison of the clustering results on the entire COPD data set, *i.e.*, using all metabolites and the four most interesting metabolite clusters, including two clusters showing the largest F-Measures and two clusters showing the smallest F-Measures. The plot shows the F-measure for different clustering thresholds computed against the disease annotation (COPD with BC and healthy). The Y-axis corresponds to the clustering threshold, in this case the number of splits. Given three groups of patients in the annotation, we are particularly interested in the performance at  $T \sim 2$  (x-axis) resulting in three clusters. This is shown in more detail in the zoomed cutout, as well as the table of F-measure values at this position. Two subsets of metabolites overlap with the patients' disease annotation better than the clusterings based on the entire metabolite set. The two other metabolite subsets result in reduced F-measures, indicating a relation to confounding factors, in this case menthol. . . . . 120
- 9.1 Workflow for longitudinal breath data analysis framework. The four steps contain (A) MCC/IMS analysis of rat breath and subsequent preprocessing of MCC/IMS chromatograms; (B) a mature screening for background noise; (C) model selection phase evaluating the quality of longitudinal models; (D) model evaluation and identification of most discriminating biomarkers. . . . . 126

- 9.2 The first plot shows the distribution of the rat and respiratory experiment means. The second plot shows the distribution of the rat and respiratory residuals. The variances of both the experiment means and residual values of the described linear rat model vs. a simple linear respirator model are compared based on the Brown-Forsythe test. . . . . 131
- 9.3 Table of components that show a significant treatment difference in either the intercept or the slope of the spline model. The given p-values are corrected for multiple comparison using the false discovery rate correction (26). The arrows indicate the direction and strength of the change in the analyte intensity (CLI in yellow and SHAM in green). The last column indicates the analyte class of the component. . . . . 133
- 9.4 As an example, this plot shows the fitted two knot spline model of the compound bbR022. The vertical lines represent the spline knot positions. The dashed lines represent the predicted intensities from the model for each individual rat, while the solid bold lines represent the average intensity for rats in each group. Both rats and class curves are colored in green and yellow according to their grouping, SHAM and CLI respectively. . . . . 134
- 11.1 Schema of the steps in the automated pipeline: (1) Clinical IMS Database (2) Peak Detection Evaluation; (3) Metabolite Identification; (4) Supervised Learning; (5) Unsupervised Learning; (6) Longitudinal Analysis. . . 145
- A.1 Overview of the peak component which models IMS peaks. The central *Peak* belongs to a particular *ImsMetaData* entity and is associated to a *PeakValueDouble* that is, in turn, bound to a *PeakAttribute* for the representation of an intensity, volume or any other peak parameter. A *Peak* is further associated to a unique *PeakPosition* comprising retention time and inverse mobility coordinates. Furthermore, a *Peak* belongs to multiple *PeakRegions* which correspond to a rectangle in the IMS coordinate system. Multiple *Peak* entities (e.g. from different IMS chromatograms) are associated to a particular *PeakRegion* retrieved from the corresponding PeakAn file. . . . . 185

- A.2 Overview of the layered system architecture based on a recommendation presented in (15). The communication between objects (denoted by arrows) of the layers starts at the presentation layer and ends at the persistence layer. This from top-to-bottom approach makes each layer dependent on the next lower layer, whereas the lower layer does not depend on the top layers. The presentation layer includes a graphical user interface (relying on Java Swing) where commands according to specific concerns can be triggered. In this work, the commands are modeled as java classes implementing a method to perform the specific concern which is part of the business logic. The business layer corresponds to the business logic and encapsulates all command classes and additional data containers which are not directly related to database tables. Persistent entity classes and their modifiers (data access objects), however, are directly related to database tables and therefore included in the persistence layer. The Hibernate interfaces (query, transaction, annotation, session) in this context are used in the persistence layer for database specific operations such as SQL query execution, transaction demarcation and JDBC connection pooling, where the latter is handled by integration of the free software c3p0 (<http://sourceforge.net/projects/c3p0/>, accessed october 2012) into the Hibernate framework. . . . . 186
- A.3 This snapshot of the graphical user interface shows a tree which includes the patient data of a project. Leafs of the tree correspond to particular patient cases. When selecting a leaf, the stored attributes of this patient concept are shown on the right side. In order to protect anonymity, relevant identifying characteristics like identifiers were altered in this example. . . 187
- A.4 This snapshot of the database application presented in this work illustrates the decision tree classification performance, which is retrieved by means of a ten-fold cross-validation, for a target data set comprising the labels (classes) *COPD* vs. *control*. . . . . 188



# List of Tables

2.1	Confusion Matrix for two class classification . . . . .	50
2.2	Confusion Matrix for three class classification: $TP_X$ represents the number of samples correctly predicted as class X, $F_{X \rightarrow Y}$ samples of label X, wrong classified as class Y, and $TN_X$ accounts for the true negatives, the sum of samples correctly not predicted as class X, e.g. $TN_A = TP_B + TP_C + F_{B \rightarrow C} + F_{C \rightarrow B}$ . . . . .	50
2.3	Selection of most important performance measures for the two and three class classification problem: ACC, SEN, SPE, PPV and NPV. . . . .	51
3.1	Distribution of metabolite expression mean intensities and their variance (standard deviation) for artificial patients with class label “Health”. . . . .	62
3.2	Distribution of metabolite expression mean intensities and their variance (standard deviation) for artificial patients with class label “Nutrition”. . . . .	62
3.3	Distribution of metabolite expression mean intensities and their variance (standard deviation) for artificial patients with class label “Smoking”. . . . .	62
3.4	This table depicts which Chapters utilize the previously described data sets. . . . .	64
4.1	This table illustrates the structure of the object-attribute-table format. The first two rows comprise identifying attribute names and the corresponding data types. . . . .	69
4.2	Classification performance distinguishing COPD vs. control. The table on the left shows the confusion matrix and the table on the right the performance measures for the classification. . . . .	76
5.1	Results of the two-class-classification problem, evaluating the differences between COPD and the HC. . . . .	82
5.2	Comparison of the 10 best features selected by linear SVM to the features selected by random forest and the peaks identified by the study of Westhoff <i>et al.</i> 2011 using rank sum test (231). The $\times$ sign indicates a match. The peaks are ordered by the rank of their weight in the linear SVM. Additionally, the coordinates in the MCC/IMS chromatogram, the inverse drift time ( $1/K_0$ ) and the retention time ( $RT$ ) is shown. . . . .	84

5.3	Comparison of the 10 best features selected by random forest to the features selected by linear SVM and the peaks identified by the study of Westhoff <i>et al.</i> 2011 using rank sum test (231). The peaks are ordered by the rank of the Gini index generated during the training of the random forest model. Additionally, the coordinates in the MCC/IMS chromatogram, the inverse drift time ( $1/K_0$ ) and the retention time ( $RT$ ) is shown. . . . .	85
5.4	Results of the three-class-classification problem, evaluating the differences between COPD patients, COPD patients suffering from bronchial carcinoma, and the control group. The class specific sensitivity and specificity assessed for class COPD as well as COPD+BC, is based on the equations discussed in the methods section. . . . .	86
6.1	Number of peaks detected by all methods. Number of peak clusters after merging the peak lists (postprocessing). . . . .	95
6.2	Overlap of the five peak detection methods. The overlap of the peak list $A$ (row) and peak list $B$ (column) is defined as the number of peaks in $V$ that can be mapped to at least one peak in $W$ . Note that the resulting mapping count table is not symmetric. . . . .	96
6.3	Linear SVM . . . . .	96
6.4	Classification results of the random forest. The quality measures are the AUC, accuracy (ACC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). . . . .	97
7.1	Automatically identified signals . . . . .	106
11.1	This table depicts which projects addressed the previously described challenges. . . . .	146

# Appendices



## Appendix A

### IMSDB

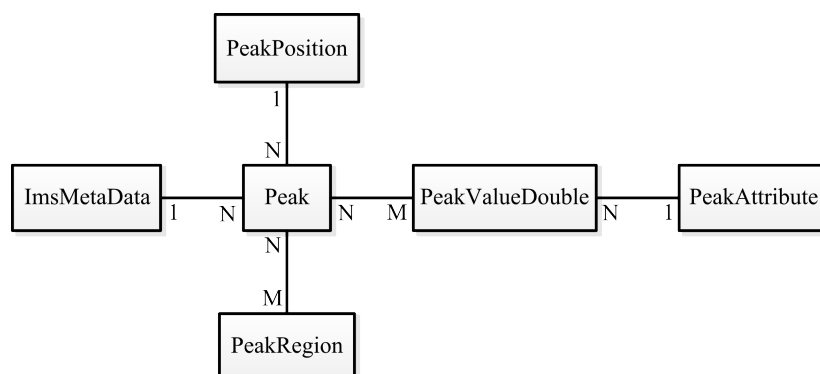


Figure A.1: Overview of the peak component which models IMS peaks. The central *Peak* belongs to a particular *ImMetaData* entity and is associated to a *PeakValueDouble* that is, in turn, bound to a *PeakAttribute* for the representation of an intensity, volume or any other peak parameter. A *Peak* is further associated to a unique *PeakPosition* comprising retention time and inverse mobility coordinates. Furthermore, a *Peak* belongs to multiple *PeakRegions* which correspond to a rectangle in the IMS coordinate system. Multiple *Peak* entities (e.g. from different IMS chromatograms) are associated to a particular *PeakRegion* retrieved from the corresponding PeakAn file.

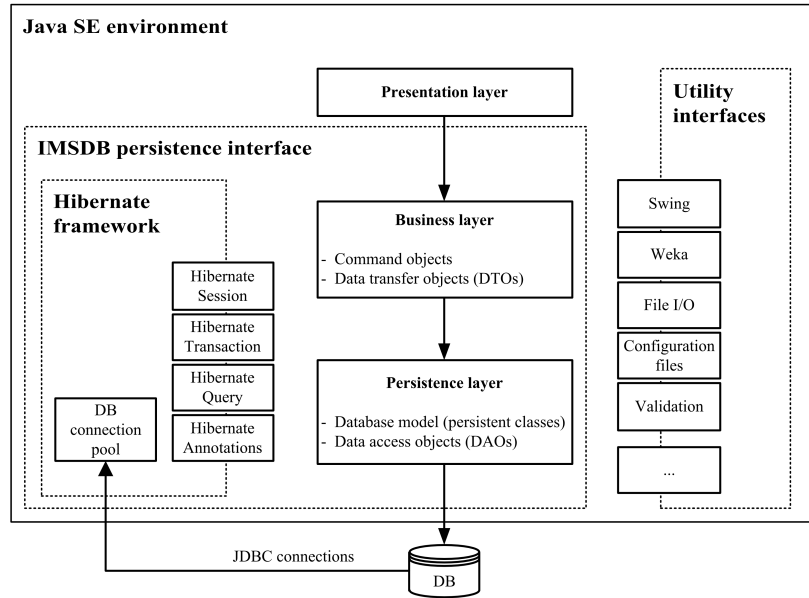


Figure A.2: Overview of the layered system architecture based on a recommendation presented in (15). The communication between objects (denoted by arrows) of the layers starts at the presentation layer and ends at the persistence layer. This from top-to-bottom approach makes each layer dependent on the next lower layer, whereas the lower layer does not depend on the top layers. The presentation layer includes a graphical user interface (relying on Java Swing) where commands according to specific concerns can be triggered. In this work, the commands are modeled as java classes implementing a method to perform the specific concern which is part of the business logic. The business layer corresponds to the business logic and encapsulates all command classes and additional data containers which are not directly related to database tables. Persistent entity classes and their modifiers (data access objects), however, are directly related to database tables and therefore included in the persistence layer. The Hibernate interfaces (query, transaction, annotation, session) in this context are used in the persistence layer for database specific operations such as SQL query execution, transaction demarcation and JDBC connection pooling, where the latter is handled by integration of the free software c3p0 (<http://sourceforge.net/projects/c3p0/>, accessed october 2012) into the Hibernate framework.

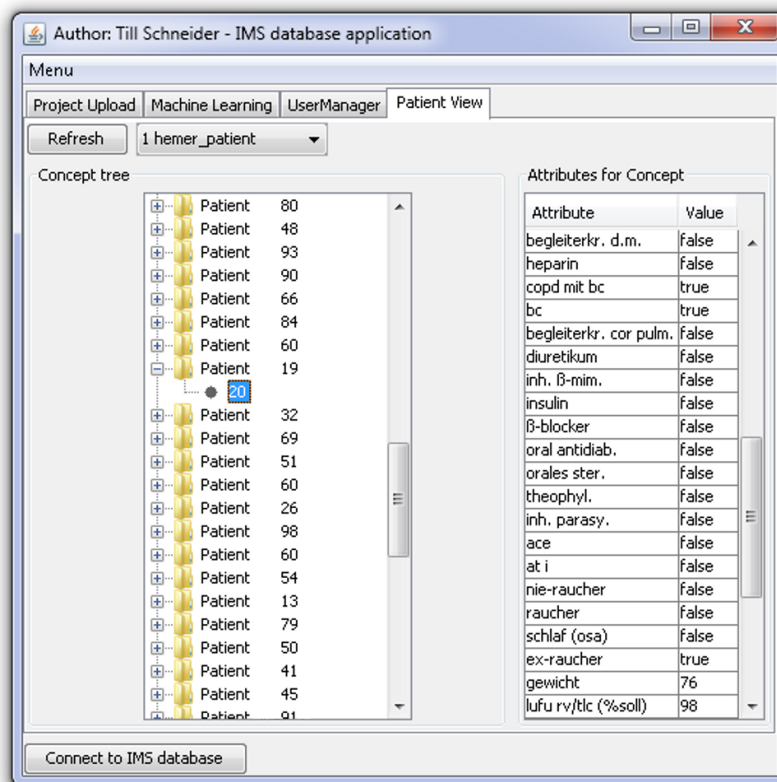


Figure A.3: This snapshot of the graphical user interface shows a tree which includes the patient data of a project. Leafs of the tree correspond to particular patient cases. When selecting a leaf, the stored attributes of this patient concept are shown on the right side. In order to protect anonymity, relevant identifying characteristics like identifiers were altered in this example.

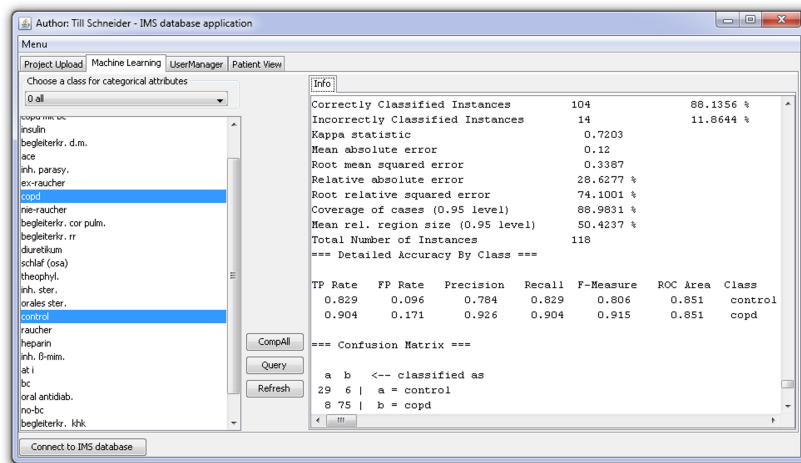


Figure A.4: This snapshot of the database application presented in this work illustrates the decision tree classification performance, which is retrieved by means of a ten-fold cross-validation, for a target data set comprising the labels (classes) *COPD* vs. *control*.