MCC-IMS data analysis using automated spectra processing and explorative visualisation methods

Zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften an der Technischen Fakultät der Universität Bielefeld vorgelegte Dissertation

von

Alexander Bunkowski

15. August 2011

Alexander Bunkowski
Adalbert-Stifter-Str. 4
33613 Bielefeld
abunkows@cebitec.uni-bielefeld.de

Supervisors: Prof. Dr. Jens Stoye
PD. Dr. Jörg Ingo Baumbach

Printed on acid-free paper.

Contents

1. Motivation and overview **7**

2. Background **11**
2.1. Metabolomic analysis 11
2.1.1. Breath and headspace analysis 12
2.2. Ion mobility spectrometer 13
2.2.1. Experimental setup 15
2.2.2. MCC-IMS Data 16
2.3. Data analysis procedures and related work 17
2.3.1. BBImsAnalyse Software 18
2.3.2. VisualNow 19

3. Requirements **21**
3.1. Spectra and image analysis 21
3.2. Project and result management 22
3.3. Required data formats and concepts 23
3.3.1. Measurement files 23
3.3.2. Peak lists 24
3.3.3. Area annotations 24
3.3.4. Heatmap images 24
3.3.5. Meta information 25
3.4. Data analysis strategies 26
3.5. Summary 28

4. Methods **31**
4.1. Spectra pre-processing 31
4.1.1. Normalisation and alignment 32

4.1.2. Baseline correction and RIP compensation 34
4.1.3. Filtering 35
4.2. Peak detection 36
4.3. Data analysis 38
4.3.1. Defining intensity thresholds 39
4.3.2. Detecting optimal intensity thresholds 40
4.4. Visualisation and exploration 42
4.4.1. Heatmap visualisation 43
4.4.2. Peak intensity visualisation 44
4.5. Summary 45

5. IPHEx - IMS Peaklist and Heatmap Explorer 47
5.1. Concept 48
5.2. User interface and primary components 49
5.2.1. Drag and drop interface 51
5.2.2. *Project browser* 52
5.2.3. *Heatmap explorer* 53
5.2.4. *Chart viewer* 56
5.3. Specialised analysis tools 57
5.3.1. Correlation viewer 57
5.3.2. Principal component analysis 58
5.3.3. GC/MS comparison 58
5.3.4. Spot table 59
5.3.5. Reference subtraction 60
5.3.6. Polar glyph visualisation 62
5.4. Workflow and general usage of IPHEx 63
5.4.1. Selecting files and starting the *project browser* . . . 63
5.4.2. Defining meta information 64
5.4.3. Starting the *heatmap explorer* 64
5.4.4. Generating peaklists 64
5.4.5. Setting up analyte information and retention time alignment 64
5.4.6. Top level analysis 66
5.5. Summary 66

6. Application examples 69
6.1. MCC/IMS signals in human breath related to sarcoidosis-results of a feasibility study using an automated peak finding procedure 69
6.1.1. Data 70
6.1.2. Analysis and visualisations 70
6.1.3. Results and discussion 73
6.2. Ion mobility spectrometry of human pathologic bacteria . 75
6.2.1. Background and objectives 75

6.2.2. Analysis and visualisation 75
6.2.3. Results and discussion 77
6.3. Large scale time series investigation 78
6.3.1. Background and objectives 78
6.3.2. Analyis and visualisation 78
6.3.3. Results and discussion 79

7. Discussion and conclusion **83**

8. Outlook **87**

A. APPENDIX **91**

Summary

Ion Mobility Spectrometry (IMS) is a method to characterise chemical substances on the basis of velocity of gas-phase ions in an electrical field. The data resulting from an IMS measurement is a number of spectra sorted by retention time. Each spectrum contains a series of values and each value represents the amount of ionised molecules at one specific drift time. Recent advantages in the field of Ion Mobility Spectrometry lead to an highly increased amount of data per measurement as well as measurements per experiment. Due to the usage of a Multi-capillary Chromatographic Column (MCC) as pre-separation technique the task of analysing and interpreting the resulting data completely changed and now includes a pseudo coloured image in addition to the classic spectra data. Analysing and comparing a high number of these images and their corresponding spectra is almost impossible and extremely time consuming with the methods used so far.

Different methods for spectra processing, data analysis, visualisation and project management were developed and combined in one software called 'IMS Peaklist & Heatmap Explorer' (IPHEx) to challenge this task. IPHEx is the first software system supporting the analysis, management, and visualisation of large amounts of MCC-IMS measurements in parallel. It is currently used for the investigation of metabolomic experiments with a focus on the analysis of exhaled air, headspace samples of cell and bacteria cultures, as well as general screening of ambient air. While the main methods of IPHEx are designed to process three dimensional data obtained from different MCC-IMS devices, it also handles GC-MS based data for comparison and substance identification as well as several other information obtained from flat and Excel files. It became the standard analysis platform at the *Leibniz-*

Institut für Analytische Wissenschaften ISAS e.V. for MCC-IMS data and showed its potential during the examination of experiments performed in cooperation with the *Lungenklinik Hemer - Zentrum für Pneumologie und Thoraxchirurgie*, the *University Göttingen - Department of Anesthesiology, Emergency and Intensive Care Medicine*, the *Charité - Universitätsmedizin Berlin* and several others. It is also used for experimental purposes at the *Korea Institute of Science and Technology, Saarbrücken* and the *B&S Analytik GmbH, Dortmund*.

The application to many different experiments and tasks demonstrates that the requirements have successfully been addressed and the software and therefore the underlying methods and concepts are suited to analyse large amounts of IMS data in an efficient way. With IPHEx, a complete analysis environment exists, which offers a solution for all analysis, management, and visualisation tasks which are necessary to perform a comprehensive investigation of large amounts of MCC-IMS data.

CHAPTER 1

Motivation and overview

Ion Mobility Spectrometry (IMS) is a method to characterise chemical substances on the basis of velocity of gas-phase ions in an electrical field.

The first developments in the field of ion-molecule detectors and ion-mobility were done by the U.S. Army during the 1960s to detect nerve agents. With a type of ion mobility cutoff filter, ions over a certain size which are typical for nerve agents could be distinguished from smaller ions, commonly formed in clean air. An overview of the different contracts and patents which were set up during this time is given in the book *Ion Mobility Spectrometry* of Eiceman and Karpas [1].

When the first civil research programs in Ion Mobility Spectrometry for chemical analyses started in 1970 at the University of Waterloo in Canada, a single spectrum per measurement could be obtained in about five minutes [2]. Further major developments were performed by Baim and Hill in 1982, allowing among other advantages, the usage of gas chromatographic methods to pre-separate a sample and thus the recording of several spectra per measurement [3].

In 1971, the first modern breath tests were performed by Linus Pauling by detecting volatile organic compounds (VOC) in cryogenically concentrated samples of human breath. Assays using gas chromatography/mass spectrometry have since identified more than 3000 different VOCs in human breath [4].

The here presented work was performed at the *ISAS Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.* which started research in the field of Ion Mobility Spectrometry in 1991. Until 1996 no pre-

separation technique was used and thus only single spectra recorded and investigated. With the first usage of a Multi-capillary Chromatographic Column (MCC) to perform a pre-separation, complex mixtures containing many different chemical compounds could be analysed. Several hundred spectra per measurement were recorded from now on and as a result of that, new methods were demanded to handle, analyse, and visualise this greatly increased amount of data. About one second is needed to record a single spectrum, and depending on the aim of an analysis about five hundred spectra are recorded per sample, resulting in a record time below ten minutes per measurement. This low time enables to record many measurements per project, which further increases the amount of data to handle. After an initial introduction of the IUPAC-Format for IMS spectra [5] this data was stored in an specialised data format for GC/IMS applications [6, 7].

To cover these requirements, different methods for spectra processing, data analysis and visualisation were developed. All of these were combined in one Software called 'IMS Peaklist & Heatmap Explorer' (IPHEx) to enable a fast and coherent analysis of IMS data with all its associated information.

The following chapters describe the methods and concepts which are necessary to enable the analysis of MCC/IMS data as well as the resulting software system and the application of it. Chapter 2 begins with an overview of the scientific background of IMS and its applications. It provides information about the origin of the investigated samples, describes the used techniques and devices and gives an overview about the data and the general analysis procedure.

Chapter 3 analyses the requirements which have to be addressed to enable a successful MCC/IM data analysis and leads to a number of different data structures, processing approaches, and graphical user interfaces which have to be established.

The structure of Chapter 4 follows the order of necessary methods to enable the data analysis from spectra pre-processing to peak detection and data analysis, concluding with methods for visualisation and exploration.

Chapter 5 describes the *IMS Peaklist and Heatmap Explorer* IPHEx, which was built to apply the developed methods to IMS data and visualise, manage and compare their results. It starts with a description of the different data structures that are created and used by the software and a detailed description of the user interface and its different compounds. It also includes a description of the general workflow of IPHEx and shows how the different parts of the software are typically applied to perform an IMS data analysis with IPHEx.

In Chapter 6 examples of metabolomic experiments and their analysis using the IPHEx software are described. Using the IPHEx software system, a detailed and fast analysis of large data sets and an integration of various existing information is possible, which is necessary to analyse these kind of metabolomic experiments.

A discussion of the System is provided in Chapter 7 followed by an outlook in Chapter 8.

CHAPTER 2

Background

This chapter introduces the scientific background of IMS and its applications. The different terms, areas and instruments which are needed to understand the basics of this thesis are pointed out. The first part describes the general field in which this work is settled and provides information about the origin of the investigated samples. It is followed by a description of the used techniques and devices and concludes with a specification of the generated datasets and an overview of the data analysis procedure.

2.1. Metabolomic analysis

Metabolites are substances which are created and/or used by the metabolism of an organism. The study of chemical processes involving metabolites provides a view inside the biochemical status of a cell or an organism. All metabolites of an organism are termed metabolome. The qualitative and quantitative analysis of all metabolites, present at one specific time in an organism, a tissue or a cell, is termed metabolomics [8]. It can give insight into the functions of unknown genes, cell systems or organisms, with respect to the response to external stimuli. By comparing metabolite occurrences and their levels in two or more sets of samples, significant differences may be identified. For example comparing metabolites from subjects suffering from a disease and healthy subjects can help to develop new diagnoses and treatments.

These comprehensive analyses are based on modern analytical and computational approaches and can be divided into two general steps. At first, the biological samples have to be extracted, eventually derivatisised and measured with a proper analytical method. The second step is the identification of the measured analytes with the use of existing databases and the comparison of different samples using computational approaches. For global metabolic profiling, the most popular analytical methods used today are based on mass spectrometry, especially in combination with gas or liquid chromatographic systems [9, 10].

In this work, volatile organic compounds are analysed to retrieve information about metabolites by screening either the exhaled air of humans (breath analysis), or the gas space above a biological sample (headspace analysis).

2.1.1. Breath and headspace analysis

Blood and urine analysis are standard techniques in medical examinations for various tasks and deliver information about the health status of a patient. During the last years several studies showed that exhaled air is a carrier of information about the metabolic state, including information about infections, medications, and diseases [11, 12, 13]. In the majority of present studies, a non invasive method for early diagnosis or therapy monitoring should be developed by identifying disease-specific biomarkers in the breath of patients.

Headspace analysis is the analysis of all chemical compounds present in the gas space above a sample. A solid, liquid or gas sample is recorded and used for the determination of volatile organic compounds. This technique has evolved over many years and gained worldwide acceptance for example to analyse blood and urine as well as residual solvents in pharmaceutical products.

The primary aim of these analyses in this context is to support and extend the field of breath diagnostics with laboratory studies of biomolecules at comparatively low pressure. For example the identification of volatile organic compounds which are caused by clinically important bacterial species (see 6.2) under controlled laboratory conditions can help to determine an infection with the respective bacterium of the human lung. Furthermore the metabolic profile of lung cancer cell lines can be compared to those from normal lung cells to support the task of diagnosing lung cancer effectively. Other applications, for example the comparison of the metabolome of cells grown on different medium, or under different experimental conditions are possible.

Existing analytical methods used to perform breath and headspace analysis based on mass spectrometry are gas chromatography mass spectrometry (GC-MS) [14, 11], proton transfer reaction mass spectrometry (PTR-MS) [15, 16], and selected ion flow tube mass spectrometry (SIFT-MS) [17, 18]. Beside these mass spectrometry based approaches there also exist different types of sensors [19, 20, 21, 22, 23], electronic noses [24, 25], and ion mobility spectrometry which was used in this work and is described in detail in the following section.

The IMS technology is very well suited to perform breath analysis because of several different facts:

- Humid air can be analysed directly, which is a major problem for many other analytical approaches. The detection limit goes down to the level of picogram per litre.
- A relatively low technical expenditure is necessary for the construction of an IMS, which directly results in lower costs of the device compared to the mass spectrometry based methods.
- The small size, low weight and power consumption as well as the short time needed to acquire a measurement are further practical advantages of the technology.

Due to these facts, an installation and usage of the IMS device inside a hospital and thus a direct sampling of the breath is possible. This negates the influences of adsorbents or tedlar bags on the sample which are otherwise needed to transport a breath sample to a laboratory. Even ambulant examinations and usage at general practitioners are possible and will become more likely once the technology develops further.

2.2. Ion mobility spectrometer

The term ion mobility spectrometer refers to the operating principle of the device. The gas phase ion mobilities of analytes are measured by ionising the gas and moving them under ambient pressure with a weak electrical field towards a Faraday plate after passing an electrical ion shutter (see Figure 2.1).

The measurement starts when the sample is brought into the ionisation and reaction region. The different molecules are ionised using a ^{63}Ni β-radiation source. While the ion shutter is closed the probe leaves the reaction region through the gas outlet. During a short opening stage of the shutter a small amount of the ionised probe is brought into the drift region. Under the influence of an electrical field the ions

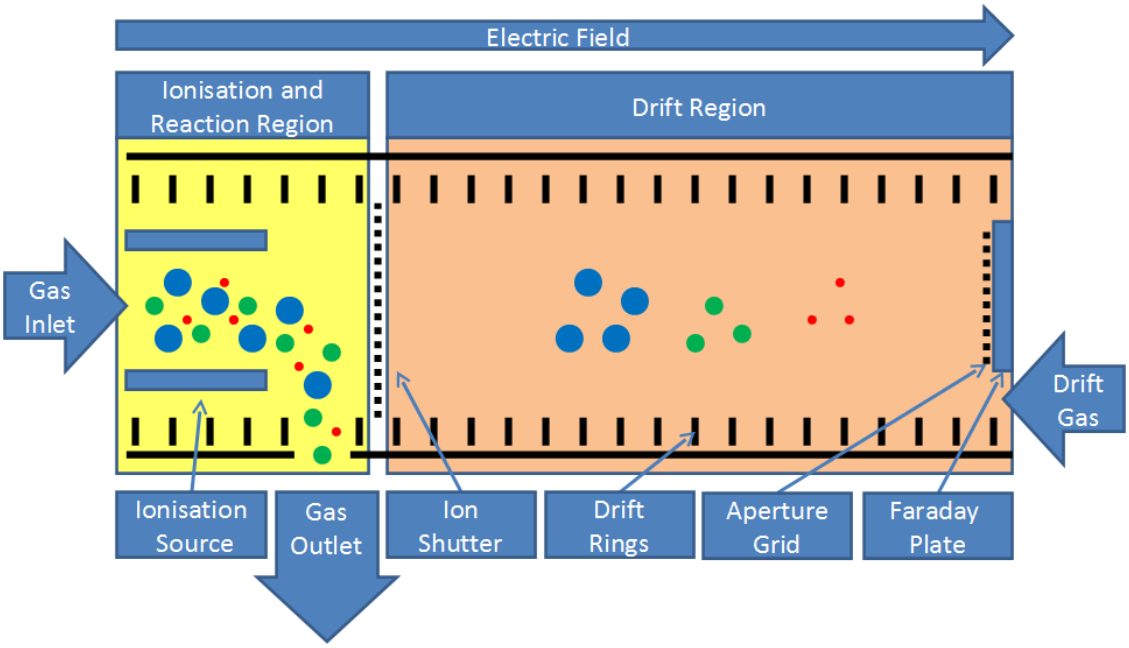


Figure 2.1.: IMS with closed ion shutter and a probe in the drift region. The sample enters the device trough the gas inlet on the left side and moves towards the faraday plate at the right side.

drift towards the Faraday plate at the end of the drift region and collide with drift gas molecules coming from the opposite direction. Due to this, the ions are decelerated in a varying manner depending on their structure and reach a steady drift velocity. In the ideal case, if no chemical reactions take place, the ions separate into different clusters and reach the Faraday plate. There, the ions are neutralised upon collision, causing a current flow of 10 to 1000 pA, which is amplified and converted into voltage. This time depended current is measured during an interval of half the shutter opening time and is called the ion mobility spectrum. The drift velocity (v_d, in units of cm/sec) normalised to the strength of the electric field (E, in units of V/cm) is termed the ion mobility:

$$v_d = KE \tag{2.1}$$

The proportionality coefficient, K, is called the mobility coefficient of the ion in units of $cm^2V^{-1}sec^{-1}$. Calculated mobility coefficients are usually normalised to the reduced ion mobility (k_0) as shown in Equation 2.2 with the values P_0 = 760 Torr and T_0 = 273 Kelvin while P and T are the respective measured values inside the drift tube. T is temperature in Kelvin and P is pressure in torr of the gas atmosphere through which the ions move.

$$k_0 = K\frac{T_0}{T}\frac{P}{P_0} \tag{2.2}$$

Reactant ions are formed continuously by the radioactive ionisation source and are brought into the drift region due to the electrical field. When no analyte molecules are available, the reactant ions move towards the Faraday plate and form a spectrum which contains the Reactant Ion Peak (RIP) [1].

2.2.1. Experimental setup

The detection of metabolites in complex mixtures like headspaces of cell and bacteria cultures as well as human breath results in a high number of detected analytes.

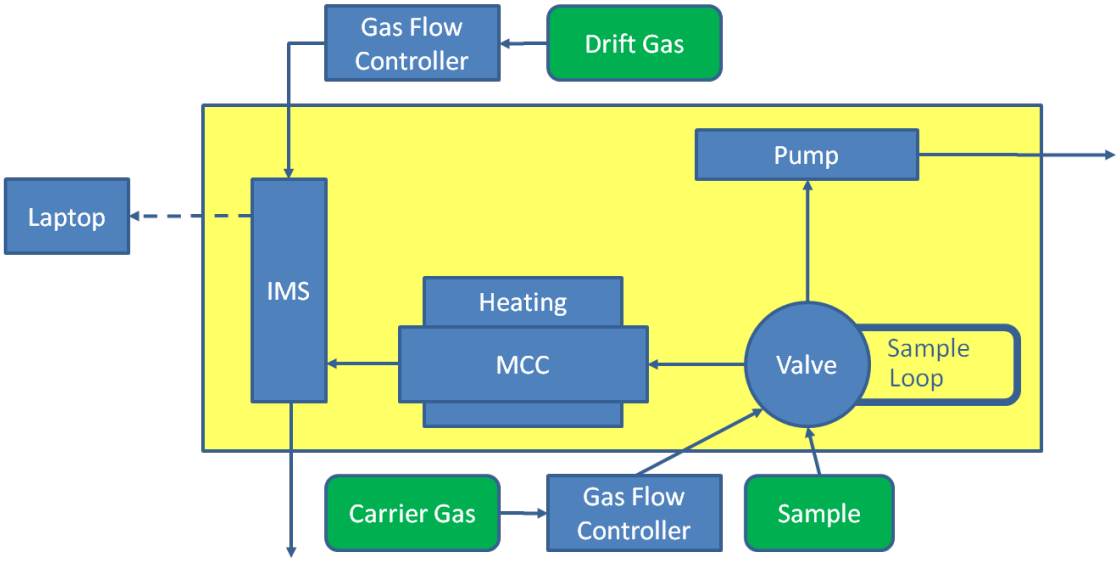


Figure 2.2.: Complete experimental setup of an IMS coupled to a multi-capillary chromatographic column and a sample loop to allow the recording of headspace and breath sample under controlled conditions.

To avoid overlapping of the signals and to improve the separation and identification capabilities, a multi-capillary chromatographic column (MCC) is used for pre-separation. To keep the speed of the pre-separation constant under varying conditions, a heating unit is used to stabilise the temperature at a constant value. Furthermore, a sample loop is used to buffer the sample and transport it into the MCC at a constant pressure, using a drift gas (Figure 2.2). The integration of the MCC increases the recorded data from one spectrum per sample

to a set of several hundred spectra. Thus a third parameter called retention time is added to each signal, in addition to the $1/k_0$ and signal intensity. In general, an analysis with this experimental MCC-IMS setup consist of a number of measurement sequences recorded under different experimental conditions or taken from different persons. A sequence consist of an instrumental blank, a sample of the room air or the medium, and the measurement of the exhaled air or headspace sample. The instrumental blank is taken to exclude general contamination of the device and ensure that no substances are left from previous samples. A room air or medium sample is analysed to identify substances which occur in the ambient air or medium and thus are taken to account afterwards when analysing the sample. The third and most important record is the probe sample, carrying information about the volatile organic compounds of the target. The comparison of these measurements to identify analytes and determine similarities and differences between classes of measurements is the general aim of an experiment.

2.2.2. MCC-IMS Data

The data resulting from an MCC-IMS measurement is a file with a number of spectra sorted by retention time. Each spectrum contains a series of values and each value represents the amount of ionised molecules at one specific drift time.

Each spectrum consist of a number of single scans which are averaged to one spectrum (Figure 2.3). The time needed for one single scan is 100 milliseconds, typically 10 scans are used to build one spectrum which leads to an acquisition speed of one spectrum per second. This is necessary to improve the noise to signal ratio and thus separate weak signals from noise.

Typically a measurement consist of 500 single spectra, where each of them contains 2000 values. This leads to a set of one million datapoints, where each datapoint consist of the three parameters retention time, drift time and intensity. Each spectrum contains a dominant signal, called the Reactant Ion Peak which is caused by the ionisation source of an IMS and further peaks which contain information about the measured substances of the sample.

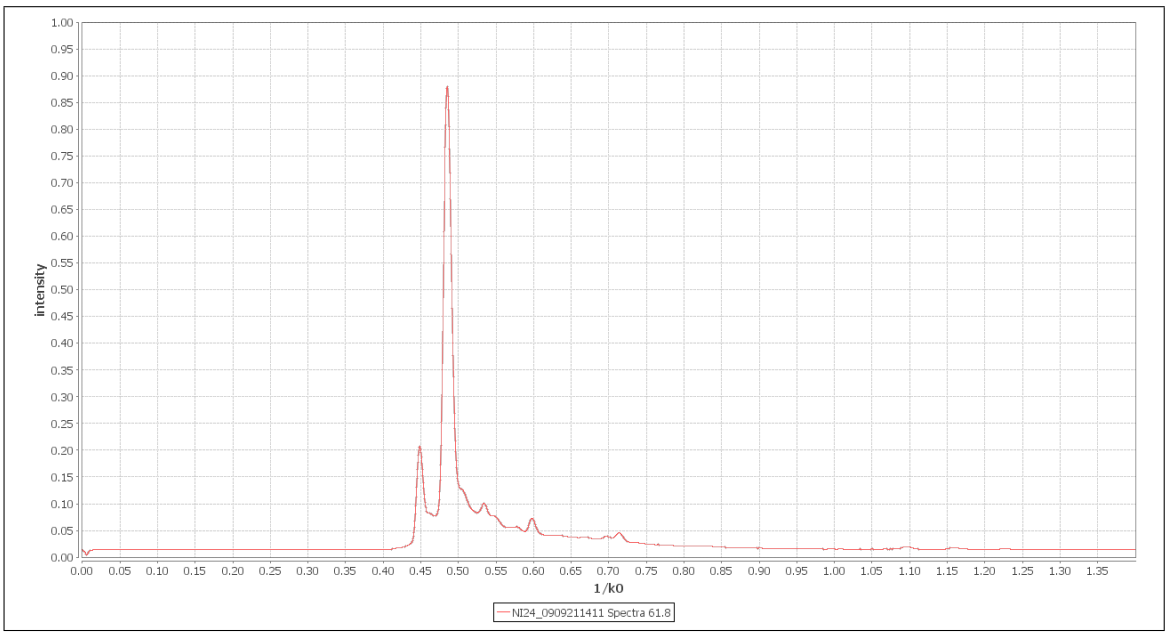


Figure 2.3.: One IMS spectrum, averaged over 10 single scans. The x-axis shows $1/k_0$, while the y-axis is ued to display the signal intensity.

2.3. Data analysis procedures and related work

In 1991, when the first IMS was used at the ISAS, no MCC was used for pre-separation and thus only one spectrum per measurement was recorded [26]. The resulting spectrum, containing about 2000 data-points, was investigated with common available software applications like Microsoft Excel and Origin (www.originlab.com). Due to the low complexity, no dedicated software was needed. With the application of the MCC in 2001 [27] the amount of data highly increased, since multiple spectra per measurement were recorded and new strategies to analyse the datasets were developed. A first step was to convert the set of single spectra into a matrix and visualise it with the help of Origin as a pseudo colorised map. With a combination of these maps and selected single spectra the analysis was done completely manually. A second factor which increased the amount of data was a change of the experimental focus in 2002 from single compound analysis [28] to complex applications like process analytics [29], emission of surfaces [30], cell and bacteria cultures [31] and the analysis of exhaled air. Due to the complexity and aims of those experiments, more measurements per experiment were recorded and compared with respect to external stimuli. Analysing and comparing a high number of these

images and their corresponding spectra is almost impossible with the methods used so far.

The analysis of MCC-IMS data can not be performed with available spectrometric software for MS with chromatographic columns, since the structural differences are too big. This is mainly because GC-MS/LC-MS analysis is based on investigating two dimensional total ion count spectra and using the mass decomposition spectra to identify compounds at specific locations, while the MCC-IMS data needs to be analysed based on a three dimensional, image like spectrum. On the other hand, existing image analysis tools are hard to apply because they lose the information of the underlying spectra.

Several people were acquired as diploma, master, and PhD students to investigate those problems. The first one was Sabine Bader who worked on the identification and quantification of peaks in spectrometric data [32, 33]. After her work, several diploma and master theses were set up by the ISAS in cooperation with Dortmund University and Bielefeld University to find solutions for the data analysis process. As a result of this process, one software called *BBImsAnalyse* was developed and will be described in the following section.

2.3.1. BBImsAnalyse Software

The BBImsAnalyse Software was developed by Bertram Bödeker during his diploma thesis [34] at the ISAS and was the first software explicitly designed for MCC-IMS data analysis based on a heatmap representation. It offers preprocessing functions a well as the possibility to setup a basic annotation layer for regions. It is also possible to compare different measurements based on layers with an integrated tool and export resulting values to an excel file for further analysis.

Visualisation

Within the central window of 2.4, the whole chromatogram from the data file is shown as two-dimensional-plot, where the signal height is colour-coded.

A single peak can be marked by a mouse-click with a cross line.

Thus, the single spectrum at the selected retention time becomes available in lower window and one single chromatogram at the selected ion mobility in the window on the right side.

Also in the lower window the parameter of the data point at the current mouse position like $1/k_0$, k0, and the intensity are visible. Analogous, in the right window the retention time at the position of the cross line is shown.

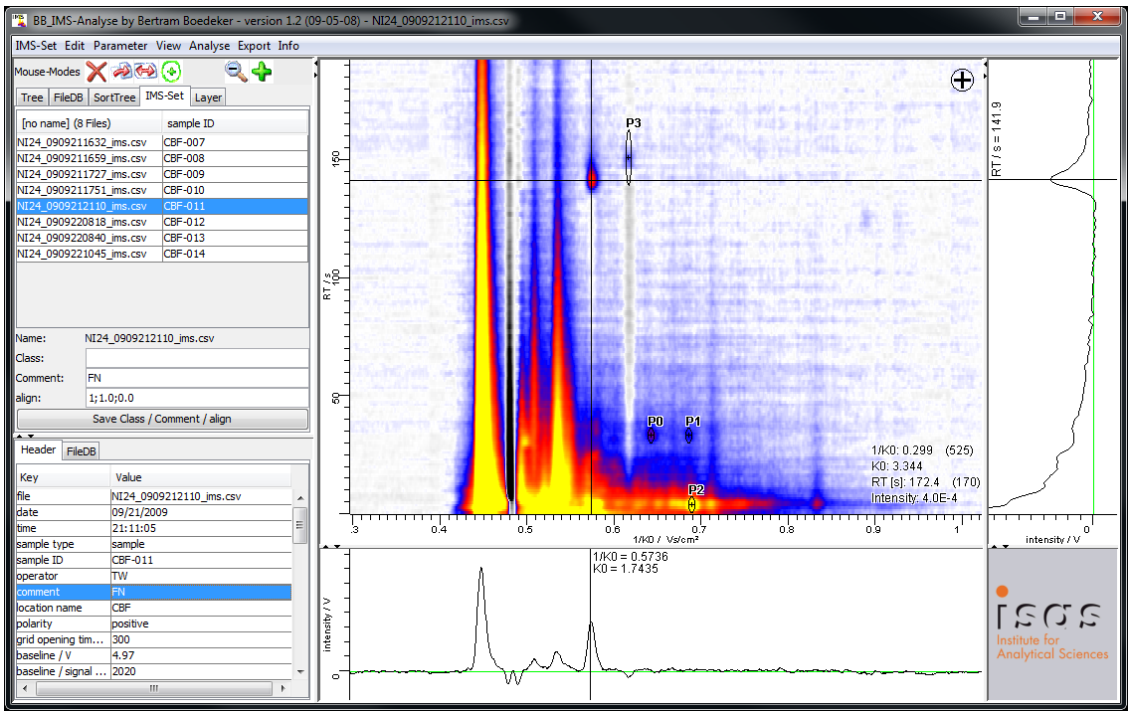


Figure 2.4.: Screenshot of the BBImsAnalyse software user interface. A heatmap in the centre visualises one measurement, the bottom and right parts show representation of single spectra which were selected.

To compare different peaks, it is necessary to mark the peak position in the central window manually one after the other [35].

Peak comparison

The Software enables the selection of an area in the heatmap directly and the search of the same area in further measurements. The findings can be visualised as shown in Figure 2.5 together with some peak characteristics. In this case eucalyptol as an analyte frequently found in human breath samples was used to illustrate the application of the peak comparison procedure [36].

2.3.2. VisualNow

The BBImsAnalyse Software was further developed, improved, used, and renamed to VisualNow by the *B&S Analytik GmbH, Dortmund* [37, 38]. Those improvements include the possibility to analyse and compare an increased amount of measurements and apply statistical values as well as Box-and-Wisker-Plot visualisation to the results.

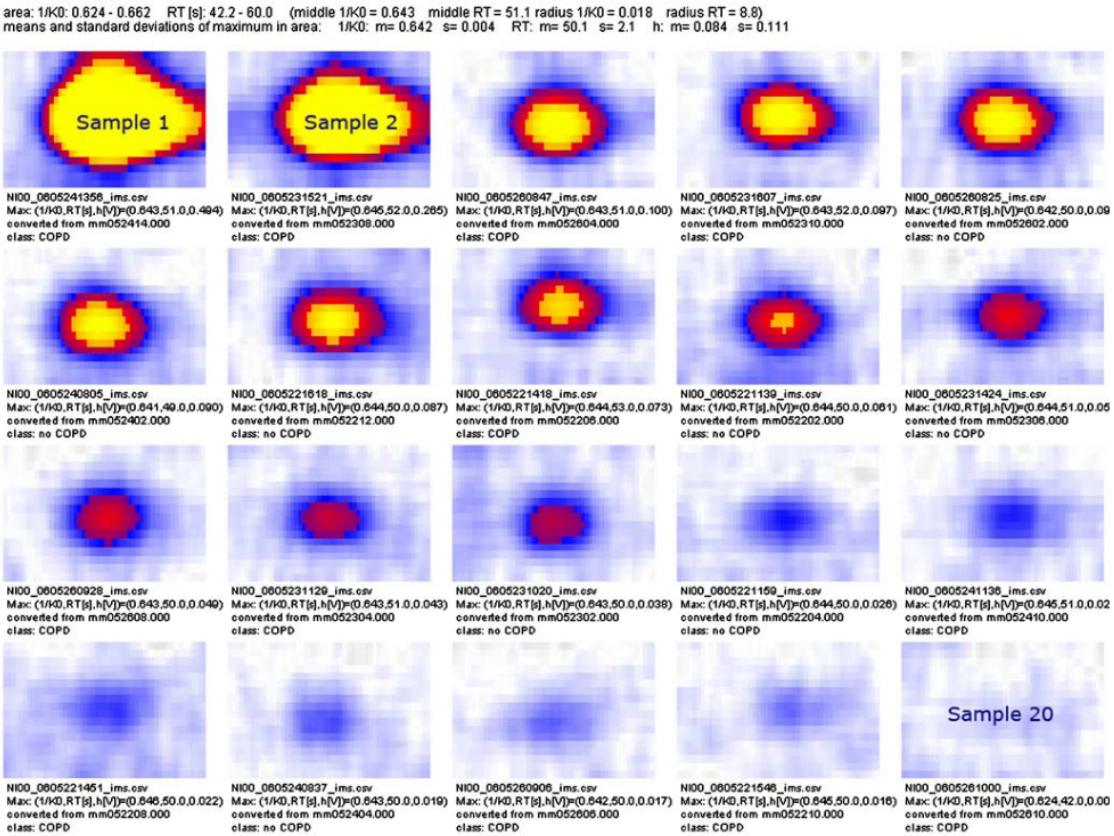


Figure 2.5.: Peak comparison of a marked area in 20 different samples of human breath ordered by signal intensity from the upper left (highest value) to lower right (no relevant signal intensity) [36].

CHAPTER 3

Requirements

The previous chapter showed that current developments in the application of the MCC-IMS technology make an analysis with existing software tools complicated and time consuming. With the change from a single spectrum per measurement to several hundreds and a drastically increasing number of measurements per experiment, new concepts for data handling and analysis need to be discovered.

Two main requirements can be formulated, regarding the current developments and problems.

- A method needs to be found to enable a combination of spectra based and image based analysis features in one environment.
- A high number of measurements and analysis results need to be stored and maintained in a project oriented way.

To achieve both goals, they need to be analysed in detail as each of them consists of several tasks to solve and require the design of different data structures, processing approaches, and graphical user interfaces.

3.1. Spectra and image analysis

The data recorded by an MCC-IMS is always effected by a certain amount of noise. To improve the visualisation and all subsequent analysis, filtering methods need to be applied to reduce these negative

effects. Since there exist several different MCC-IMS devices, and small differences of device parameter and ambient effects can influence the measurement, it is necessary to ensure that the data can be compared to each other. Therefore different alignment and baseline correction methods should be applied. The reactant ion peak (RIP) which is caused by the ionisation source of an IMS needs to be regarded in detail, because it dominates every spectrum and can overlay important signals, which causes problems in further analysis and visualisation. The key information every IMS measurement contains is the information about the type and amount of substances in the sample. Therefore functionality to annotate, identify and quantify peaks needs to be provided. In the ideal case, the data matrix is transformed into a list of all contained substances which forms the fundament for all subsequent analysis. To enable this and provide a general analysis environment for MCC-IMS measurements, visualisation methods are required. Since the data contains more image like information than other spectrometric methods but also contains more value based and axis depended data than standard image processing tasks, a combined analysis needs to be done based on both, the generated image and the individual spectra. A visual analysis environment needs to provide an overview as well as functionality to zoom and filter, and provide details-on-demand [39].

3.2. Project and result management

All measurements are typically taken in an experimental context, meaning there exist several measurements that are related to a specific analysis goal and thus need to be compared. Therefore, a list of all measurements which belong to one experiment should be provided by a project managing interface. All measurement files contain certain parameters and labels which were assigned during the sample recording and should be provided within this project manager as well. Additionally there often exists external information which is related to the measurements. In case of headspace samples this can be information about the medium, type and number of cells or age for example. When analysing exhaled air, there typically exists information about the respective person, like type of disease, gender, age and values that originate from blood or similar analysis. This additional information is important for further analysis and can be divided into two general types. The first one to group the measurements into different classes to compare those classes against each other, and the other one to include additional parameters for the analysis. Such in-

formation is termed meta information and needs to be displayed in the project manager. The information about the type and amount of substances in each sample should be included in the project manager view as well. Therefore a procedure needs to be developed to select specific substances or regions of measurements and provide methods to sort, filter and compare them.

Combining all this information in one view would allow a structured analysis of a large set of measurements. Since the amount of data to display can become large, techniques to select relevant information and hide currently unnecessary data need to be included. For those experiments with a high number of measurements it is important to avoid the need of displaying every single measurement and analysing it by hand and one by one. Only a few measurements should be viewed as representatives of their respective classes or even one for the whole experiment. Methods should be provided to identify and filter peculiar measurements and highlight them for further investigation. Furthermore peaks should be detected automatically because labelling them by hand can be extremely time consuming if not impossible in large projects. Nevertheless a software system should provide a method to perform a manual validation and editing of the peaks to correct possible errors and handle unexpected cases. The last point which should be regarded to reduce the time needed to analyse projects with a high number of measurements is that in many cases for similar projects there exists a subset of necessary analysis and annotation tasks which are identical. Therefore, functionality should be provided to transfer information between projects and compare them.

3.3. Required data formats and concepts

The analysis of the requirements lead to a number of necessary data formats and concepts which need to be handled and provided.

3.3.1. Measurement files

The first and most obvious necessary data format is a representation of the measurement file itself. The file consists of a header which contains many information about the device and measurement parameters as well as some information about the recorded sample which can be created while recording it. Furthermore it contains a list of all recorded spectra with information about retention time, drift time and spectra number. Each of those spectra is a list of measured values. The format of this file is a simple character separated values file

whose structure was defined by the ISAS. A data object needs to be created which enables access to the header information and all spectra together with the retention time and drift time coordinates.

3.3.2. Peak lists

A second data structure is needed to determine and compare the type and amount of substances in each sample. This structure called peak list needs to cover information about all peaks that occur in one measurement and is represented as a list of peaks. The parameter to describe one peak are peak position, consisting of retention time position and drift time position, the area of all elements belonging to a peak, the volume and the maximal intensity of the peak. As this information is very important for the analysis and an exchange with other software tools may be wanted, a simple exchange format is needed. Therefore peak lists should be saved in a tab separated text file where each line represents one peak and thus contains all necessary parameters separated by tabs. A header as first line is needed to define the type of parameter in each column.

3.3.3. Area annotations

The identification of a peak should be treated separately from the peak lists, since the parameter of a detected peak can not change, but the assignment of a substance name to a peak can change or become more precisely once an analysis progresses or the databases for identifications grow. The position parameter of identified substances typically contains a small range of drift and retention times where they can occur. Therefore an annotation type needs to be defined which contains the name of the substance, describes the position of a substance in form of a retention time index and a drift time index, as well as tolerance values for both indices. Existing annotations need to be read from external locations and a mechanism to define and save such annotations to use them in further projects should be established.

3.3.4. Heatmap images

A visual representation of the data is necessary to explore and work with the measurements, visualise peak lists, and define annotations. A suitable concept for this task is the generation of a heatmap image. A heatmap is a graphical representation of data that offers a possibility to visualise three parameters of a large number of data objects in

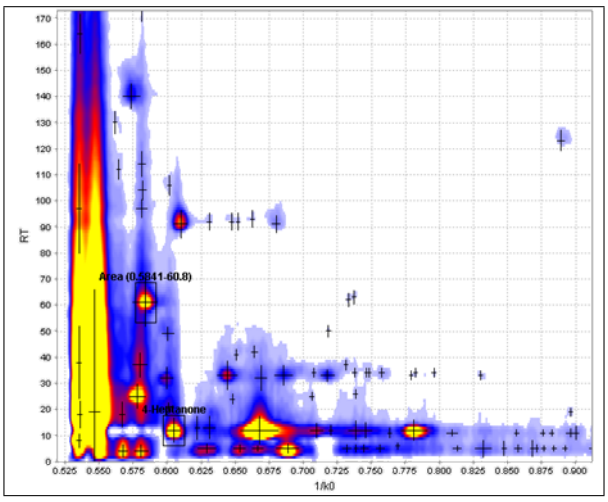


Figure 3.1.: Example of a heatmap created from a MCC-IMS measurement with labeled peaks and annotations.

one image. Using this method, two parameter of an object are used to determine the position of a pixel inside the image while the third parameter is encoded by colour. The term *heat* is used because typically low values are represented by blue colour and implicate a low temperature while higher values are represented by red and yellow colours which are often associated with high temperatures. Using the retention time and drift time indices of all values of a measurement as coordinates and the intensity values to encode the colour enables the visualisation of a whole measurement as one image. Furthermore peak coordinates can be drawn on this image and annotations to identify and compare peaks can be defined. An example of a heatmap with peak coordinates and annotations is shown in Figure 3.1.

3.3.5. Meta information

Meta information is additional available information for measurements and typically provided in a table format which was defined by the lab personnel. It contains information about persons whose exhaled air was measured or about type and properties of headspace samples. One line of such a table contains a sample ID or filename of the associated measurement as an identifier and a number of additional entries which reflect classes, diseases, cell type and number, or parameters from former blood and urine analysis. This additional information is called meta information and needs to be integrated into an analysis environment. Functionality to extend this information or create it in case it is inexistent needs to be provided.

3.4. Data analysis strategies

Once these requirements are fulfilled and the required data formats are designed, strategies need to be found to perform a detailed analysis of the resulting data.

	class	value	substance a	substance b
measurement 1	a	34	0.26	0.56
measurement 2	a	73	0.32	0.74
measurement 3	b	45	0.65	0.54
measurement 4	b	83	0.19	0.57
measurement 5	c	24	0.18	0.75

Table 3.1.: One simple table structure which includes basic information to analyse a project.

The general goal of an analysis is to compare the level of one or more substances in a set of measurements. When all required data is integrated into a project environment, a table view should be provided with one row per measurement and one column per relevant data type (Figure 3.1). The relevant data types are those that assign values or classes to measurements which originate from the measurement files, the meta information, and the quantified peaks which were selected via an area annotation. Depending on the type and aim of the experiment, different analysis and visualisation techniques are necessary to gain insight into the data and possibly reveal dependencies and specific characteristics.

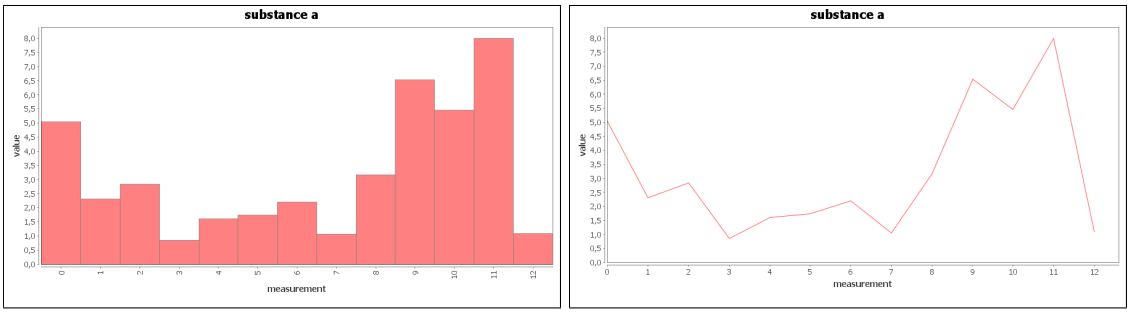


Figure 3.2.: Two possible visualisations of a series of thirteen substance values taken from different measurements.

The easiest case would be an investigation of one single substance in a set of measurements and can be visualised in a diagram where the x-axis is used to display the measurements and the y-axis is used to display the amount of a substance or value. A chart which uses lines

or bars seems to be a viable representation of each series (Figure 3.2). Special attention needs to be given to the order of the measurements on the x-axis. If the experiment contains some kind of time dependent analysis, the order can be given by the time, otherwise an order needs to be defined.

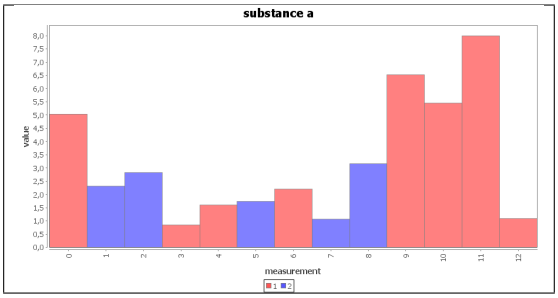


Figure 3.3.: A bar chart visualisation of a series of thirteen substance values, using two different colours to distinguish assigned classes

Many experiments aim at comparing the level of substances in different classes of measurements. This case can be covered by performing a bar-chart visualisation and assigning the same colour to all bars which originate from same classes (Figure 3.3).

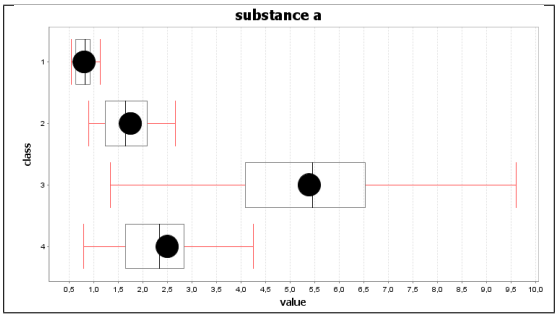


Figure 3.4.: A box plot to visualise the intensitiy distribution of one substance in four different classes.

When the amount of measurements, classes or substances to compare exceed a certain number, bar and line charts are no longer applicable due to the limited number of simultaneously displayable series. For this task, box plots seem to be a feasible solution (Figure 3.4). A high amount of substances and classes can be visualised in parallel this way, while the number of peak intensities is unlimited since they are grouped to boxes.

Another analysis goal is a determination of dependencies between substance intensities or between series which are available as meta

information and substances. A low number of these can be analysed by plotting them all into one diagram in the previously described way. But again, when the number of series exceeds a feasible number, other techniques are required. A correlation analysis can be performed using the correlation coefficient, to determine the similarity of two series. This can be visualised, for example, by assigning all series to both axes of a diagram and visualise the correlation coefficient at their intersections.

An existing standard analysis and visualisation technique which is often performed in this kind of experiments is the principal component analysis. Especially when different classes are assigned to measurements and a lot of substances should be regarded, it can be used to investigate differences between the classes.

To verify the results of an analysis it is important to provide a method to display the origin of the regarded values. Every value which was visualised with the above described charts and diagrams needs to be linked to the underlying peak and it is position inside the heatmap image of a measurement. This allows to verify if one or more particular values which are important to corroborate an analysis result were recorded correctly and are not influenced or caused by noise, misalignments or device errors. Therefore the heatmap image, project environment and diagrams need to be closely connected to each other to allow a tracking of the intensity values back to peak structures.

3.5. Summary

The detailed analysis of the requirements to perform a MCC-IMS data analysis can be summarised to a final list.

- Methods to perform filtering, alignment, peak detection, and visualisation for general processing of IMS data need to be defined and implemented.
- Explorative analysis of heatmaps with functionality to annotate substances is a key feature to gain inside into MCC/IMS data and needs to be designed and integrated carefully.
- Project oriented management of measurements and meta information with integrated information about quantified peaks and identified substances needs to be provided.
- Analysis of a project needs to be designed based on comparison and investigation of quantified peaks.

- Parallel work on different projects is needed to transfer knowledge from existing projects to new ones and minimise redundant work.

To cover the requirements and handle the necessary data formats and concepts, methods to process and analyse MCC-IMS data were designed and are shown in the following chapter and a software system was created which is described in Chapter 5.

CHAPTER 4

Methods

Based on the requirement analysis in the previous chapter four different types of methods have to be addressed to enable the analysis of IMS data. The structure of this chapter follows the order of necessary methods for this task from spectra pre-processing to peak detection and data analysis, concluding with methods for visualisation and exploration.

All figures and visualisations in this chapter were created with the IPHEx software which is described in Chapter 5 based on original datasets taken from the experiments shown in Chapter 6.

4.1. Spectra pre-processing

The intensity matrix as described in Section 2.2.2 is influenced by different negative effects which need to be compensated when analysing measurements. Pre-processing of IMS data is divided into three different types. Normalisation and alignment methods are crucial to compare different measurements to each other and verify that identical substances from different measurements have comparable drift- and retention times as well as comparable intensity values. Baseline correction and RIP compensation methods are needed to separate signals from the background and enable a interpretable visualisation. Filtering methods are necessary to denoise the data for peak detection methods and increase the quality of heatmap visualisations.

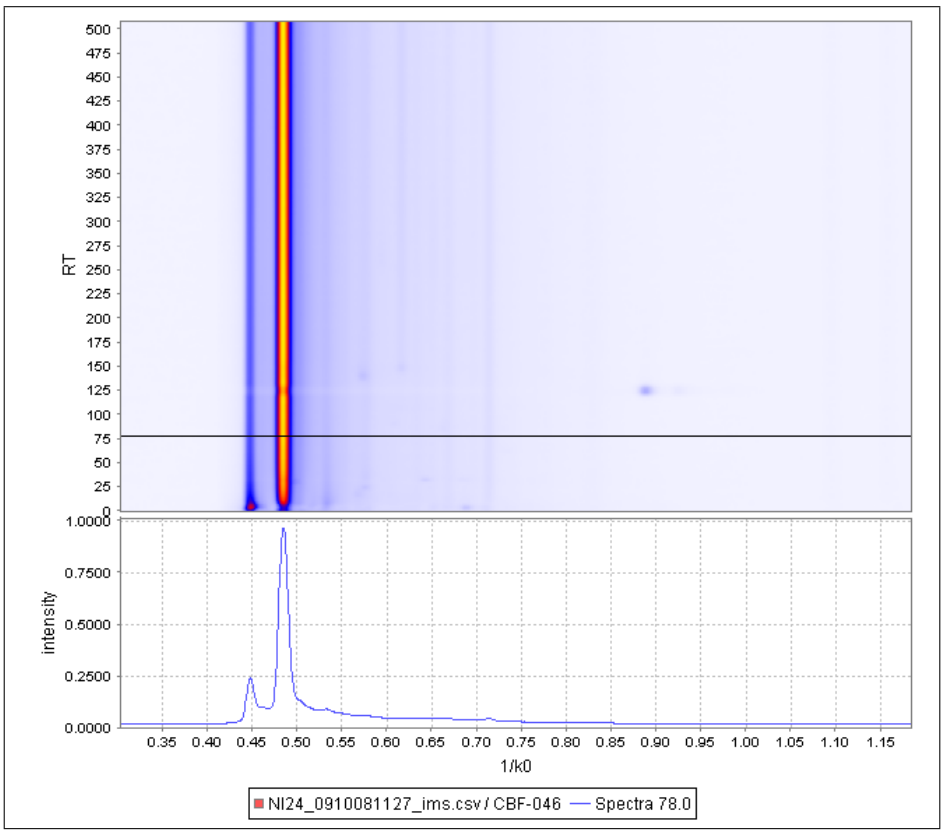


Figure 4.1.: Heatmap with detail view of a selected single spectrum. The heatmap consists of 500 single spectra which are pseudo coloured using a heatmap paint scale.

4.1.1. Normalisation and alignment

To align spectrometric data, sampling points have to be found which refer to identical substances and occur in every measurement. Alignment methods for IMS technology benefit from the fact that the RIP (see Section 2.2) occurs in every measurement and that its intensity refers to the strength of the ionisation energy of an IMS.

$1/k_0$ alignment

The first step in the alignment process is transforming the $1/k_0$ axis by moving the spectra until the RIP $1/k_0$ value matches the default position of 0.485. For this procedure, termed $1/k_0$ *alignment*, the position of the RIP needs to be determined exactly. A spectrum subset needs to be selected which ensures that the maximum of the RIP is not influenced by analytes or disturbing effects. The probability for analytes to occur decreases over the retention time, because most of

the molecules of a volatile gas sample have a small collision cross section and pass the MCC in a short amount of time. Therefore, the last 100 spectra of a measurement are used for the subset. The RIP is detected by finding the maximum of every spectrum of the subset and putting its $1/k_0$ value into a sorted list. Afterwards the median of this list delivers the $1/k_0$ position of the RIP. Due to the robustness of this procedure this can be done fully automatically without additional user input or parameter settings.

Intensity normalisation

The next step is a normalisation of the intensity axis to verify the comparability between measurements. This is important especially when comparing data recorded by different devices. Therefore, the absolute values are transformed to values relative to the maximal available ionisation energy. Since the maximum of the RIP reflects the maximal available ionisation energy of one specific IMS, the maximal intensity of the subset formed for the $1/k_0$ alignment is chosen for normalisation.

Retention time alignment

The chromatographic column is responsible for the pre-separation along the retention time axis and divides the sample into a number of single spectra. The time needed for analytes to pass the column depends on several factors such as temperature, pressure, and condition of the inner surface. Although much effort is put into keeping these factors stable, small variations cause analytes to occur at slightly different retention time points. These so called *retention time shifts* need to be corrected to allow exact comparison between measurements and analyte identification. The RIP can not be used as a reference here since it occurs in every spectrum. Furthermore no reference analyte is brought into the probe and no analyte could be found which occurs constant in every measurement. But experiments showed that there are experiment specific analytes which can be manually chosen to align the retention time axis. For example in cell culture probes (see Section 2.1.1) there are analytes which are caused by the medium and such caused by the cells. Since the aim of these experiments is to compare different cell cultures on the same medium or, vice versa, the same cell culture on different media, analytes $ca_1, \ldots, ca_n$ can be found that occur in every measurement. A retention time target $RTT(ca)$ for at least one of such constantly occurring analytes has to be chosen to find a retention time alignment factor (RTF) for each measurement.

This is done by dividing the target position of an analyte $RTT(ca)$ by the actual position of an analyte in the measurement $RTA(ca)$,

$$RTF = \frac{RTT(ca)}{RTA(ca)} \tag{4.1}$$

.

After scaling the retention time of every spectrum of a measurement with this factor, the retention time axis is called *aligned* and can be used for comparison and identification. If more constantly occurring analytes with differing retention times are available, this process can also be done more than once on specific retention time intervals, but experiments showed (see Chapter 6.3) that the correction using one analyte is sufficient for further analysis and using more analytes can lead to false deformations.

4.1.2. Baseline correction and RIP compensation

While the fact that the RIP occurs in every spectrum is a great advantage for $1/k_0$ alignment and intensity normalisation, it is also obstructive for further analysis and visualisation. It dominates every measurement but contains nearly no relevant information about the analytes of a sample.

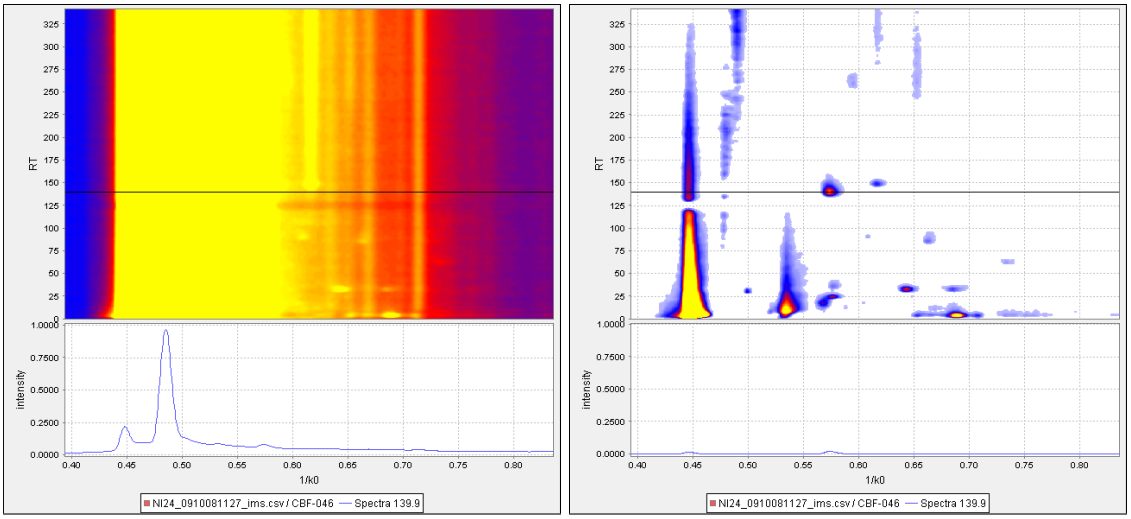


Figure 4.2.: Influence of the RIP compensation filter on an IMS heatmap. Left: Raw measurement; Right: Compensated RIP

One major problem is the so called *RIP-tailing*, the part of every spectrum from 0.50 to 0.80 $1/k_0$ where the RIP increases the baseline of a whole spectrum with a diminishing value. This causes a difficult separation of peaks from noise and problems when altering the colour

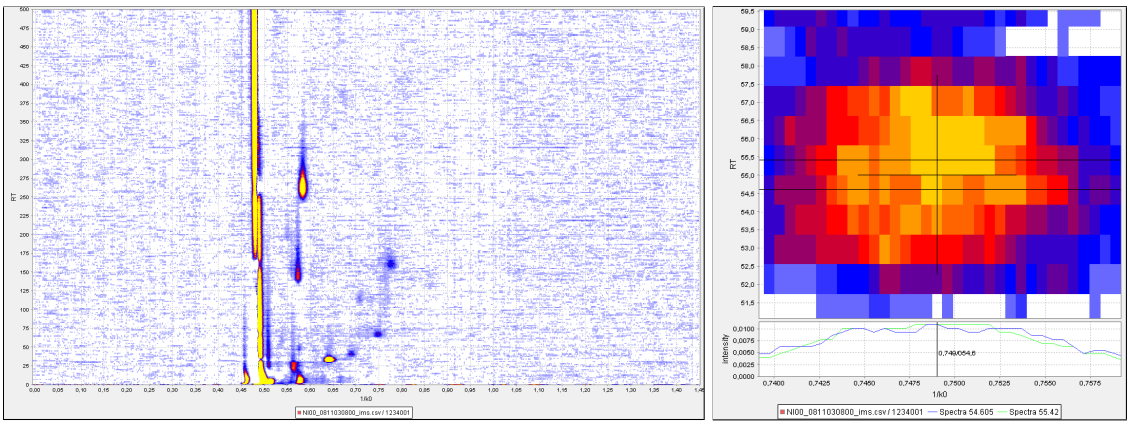


Figure 4.3.: Two heatmap images of a measurement with compensated RIP *without* applying any filter methods. The left part shows a complete measurement while the right part shows a detailed zoomed view of one single peak.

space of a heatmap for better peak visualisation as shown in the left part of Figure 4.2. To compensate for this effect, the RIP needs to be determined and subtracted from every spectrum. Therefore, all values with the same $1/k_0$ are brought into an ascendingly sorted list and the 25% quantile for every $1/k_0$ is selected. These values form the *low quantile spectrum* and represent the part of a measurement which all spectra have in common. The quantiles are chosen because the theoretically necessary minimum of each $1/k_0$ is zero in most cases due to noise and thus unusable for this task. Subtracting the low quantile spectrum removes the RIP together with its tailing and also corrects the baseline of every spectrum.

4.1.3. Filtering

Different filter operations are necessary to remove noise from the measurement and thus improve the quality of the data for further processing steps and visualisation. To determine the quality of filter operations and to verify that the position and quantification of analytes is not altered, control criteria are chosen. The first criterion is the area from 0.1 to 0.4 $1/k_0$ where no analytes can occur. The second criterion is the position and intensity of peaks in the measurement. An optimal filter reduces the noise of the first area while keeping the values of the peaks constant.

Two different filters showed an improvement of the data under these conditions and were therefore composed as a default filter pipeline for IMS data. The filter pipeline first uses a median filter and afterwards a Gaussian filter to remove noise from the data. An example of the

performance of this filter pipeline is given here, the detailed examples to show the result of the single filters are available in the Appendix.

The median filter approach was chosen to eliminate single noise fragments from the data. Basically it replaces every value of the measurement with the median of the neighbour values. As Figure 4.3 shows, the number of values of a measurement which do not form peak structures and thus contain no information needed for further analysis is high.

$$\hat{f}(x,y) = \underset{(s,t)\in S_{xy}}{median}\{g(s,t)\}. \tag{4.2}$$

Since the time needed to process and visualise measurements depends on the number of datapoints to handle, reducing them without losing information is advantageous. In the pipeline the filter uses a five times five area as neighbour values, resulting in a list of 25 values to determine the median. The median filter is applied first to eliminate single isolated values which otherwise would be blurred and thus turned into several values by following filters.

The Gaussian filter was applied to smooth the remaining values and preprocess them for the peak detection and visualisation methods. It is termed Gaussian filter because the resulting impulse response is a Gaussian function.

$$G(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4.3}$$

Where *Sigma* is the standard deviation of the Gaussian distribution.

$$s(x,y) = \sum_{(i,j)\in M} f(x+i, y+j)h(i,j) \tag{4.4}$$

Equation 4.3 can be used to create a gaussian filter mask G and Equation 4.4 shows how to convolute a given mask h with the original heatmap image f.

In the filter pipeline it is used with the smallest possible two dimensional mask size three times three, to prevent a distortion of the later determined peak intensities.

The effect of the filter pipeline on MCC-IMS data is shown in Figures 4.3 and 4.4.

4.2. Peak detection

Peak detection is the process of generating a list of detected peaks out of the intensity matrix of every measurement. This is needed for two main reasons. Firstly it is much faster to process a list of peaks than to

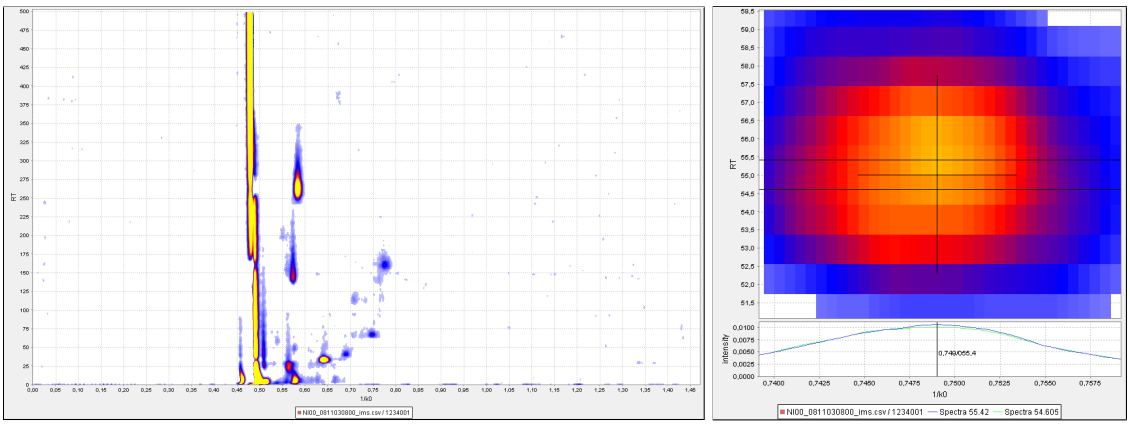


Figure 4.4.: Two heatmap images of a measurement with compensated RIP after applying the filter pipeline. The left part shows a complete measurement while the right part shows a detailed zoomed view of one single peak.

process the complete data matrix of an MCC-IMS measurement. Secondly, the alignment of the axis is not accurate down to the level of single data points. The position of the maxima of identical analytes can vary through different measurements. Although these variations are usually low, it eliminates the possibility of direct datapoint comparison through different measurements.

A suitable method to detect peaks in this kind of data is the *water shed transformation*, which was successfully applied by Wegner et al. for spot detection in two-dimensional gel electrophoresis images [40]. They illustrate the idea of this transformation by thinking of the data as a topological permeable map which is submerged into an imaginary water basin. In case of MCC-IMS data the map has to be flipped that the peaks touch the water surface first. The water that enters the map forms a pool for every elevation and with the rising water level more pools emerge which are called *regions*. The borders where two regions touch each other are termed *water sheds*. This concept is implemented by putting all values of a measurement into a list, descendingly sorted by their intensities. In addition a segmentation matrix of the same size is initialised. For each item the adjacent fields of the segmentation matrix are checked. One of three actions is performed afterwards, depending on the number of regions to which these adjacent fields belong:

a) If none of the adjacent fields are defined inside the segmentation matrix then a new region is created. An identifier for this region is defined at the respective position and the point is added to this region.

b) If one of the adjacent fields already belongs to this region, then the current point is added to a region and marked in the segmentation matrix.

c) If adjacent fields of an entry belong to different regions, the coordinates are marked as water shed in the segmentation matrix.

The result of this method is a list of regions which are surrounded by water sheds. The maximum of each region is used to determine the maximum of a peak and all members of the region are used to determine the volume.

After the peak detection all detected and quantified analytes of one measurement are available as a list of peaks. The $1/k_0$ and RT parameters of a peak are termed *peak position* and the peak maximum is termed *peak intensity*.

4.3. Data analysis

After the successful peak detection, areas need to be defined to compare peaks from different measurements. Those areas consist of a name, a RT and a $1/k_0$ parameter to determine its position, and two values to define the tolerance values for each parameter. Those values can either be taken from a list of known substances which were already determined in previous MCC-IMS measurements or by setting up a new area with a user interface.

Subsequently a dataset can be regarded as a table with one line for each measurement and one column for each substance as proposed in Figure 3.1 in the Requirements chapter. The main IMS data analysis questions which occur in the biological, chemical and medical context are then settled in the field of correlation and classification.

Correlation analysis is always needed when additional experimental parameters are known and dependencies between these parameters and occurring substances or dependencies between multiple substances should be detected.

Classification analysis enables the partition of measurements into different groups. This can be done by unsupervised techniques to detect similarities between measurements without any prior knowledge or supervised when the groups are already known.

For these kind of data and questions a vast number of methods is available. Commonly used techniques like *support vector machines* and *artificial neural networks* are able to perform supervised classification techniques, but the explanation of the results is difficult to understand, since they contain a black box model. This means that

there is no information available how important a specific analyte is for the classification result or which analytes are characteristic for a class. Since this information is necessary or even the aim of many IMS experiments, especially in the breath diagnostics field, analysis methods which contain any kind of black box are not suited to challenge this task. Unsupervised techniques like clustering methods are as well difficult to apply because the similarity between the different measurements is usually high. Many regions of the measurements are often similar and substances which are important for a specific experiment need to be determined. Therefore, approaches to analyse IMS data based on peak intensities while retaining the information about the peak characteristics and their related classes are described in the following.

4.3.1. Defining intensity thresholds

The basic criteria for a data analysis of spectrometric data are the detected and quantified peaks. One key question in the IMS data analysis context is if there exist one or more peaks whose intensities correspond to a previously defined class of measurements. Such relation between classes and intensities can be used to predict classes of so far unknown datasets, and gives information which peaks and thus analytes are possibly produced or caused by measurements of one class. This question can also be formulated as the search for an intensity threshold which allows the separation of a dataset into the previously defined classes.

The ideal case would be that one peak occurs on exactly the same position in every measurement of one class and never in any other. Since the position and intensity of peaks can be influenced by many different factors like pressure, temperature, air flow, and assigned classes can be inaccurate the chance that such peaks exist and can be found is low.

A more probable case is shown in Figure 4.5 where a relation between peak and class is visible although no clear separation is possible. Tolerance values have to be defined to determine the maximal allowed difference between the identified position of an analyte and the detected peak positions. To enable the partitioning into two classes, an intensity threshold has to be defined. All measurements with a specific peak equal or above this threshold are regarded as members of the first class, all other as members of the second class. An intensity threshold is considered as appropriate when the number of wrongly assigned classes through this procedure, compared to the previously defined classes, is low.

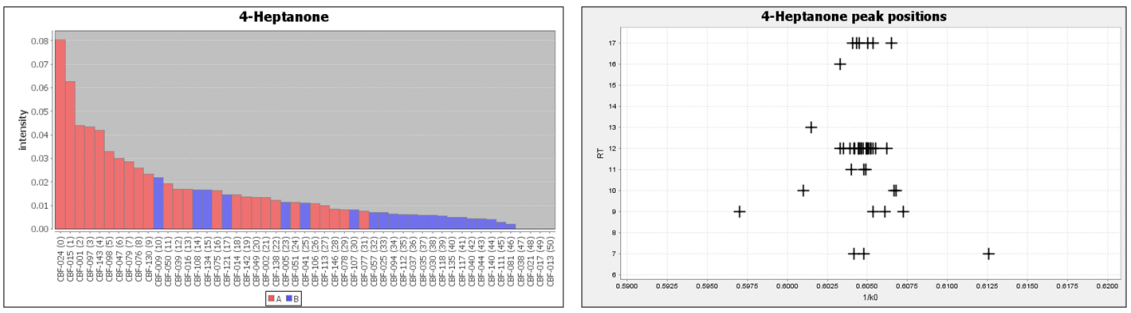


Figure 4.5.: The left part shows the peak intensities of 4-Heptanone in two different classes. A relation between classes and peak intensity seems probable. The right part shows the distribution of the peak positions. While most of the peaks are very close to the proposed position of 4-Heptanone in the centre, some differ within defined tolerance boundaries

As an experiment typically consist of a high number of measurements and detected peaks, a manual analysis is time consuming in the best case and impossible in the worst case. An optimisation of the tolerance values can increase the chance of detecting intensity thresholds and increase the quality of an analysis. These findings lead to the development of an automatic procedure to rate peaks based on their relation to a class and determine their optimal boundaries.

4.3.2. Detecting optimal intensity thresholds

To detect appropriate intensity thresholds, the high number of possible peaks and thresholds need to be rated by an automatic process. Therefore a scoring procedure was developed to determine the quality of all possible thresholds.

To determine if two peaks from different measurements belong to an identical analyte, two initial tolerance values are defined, one for retention time and one for $1/k_0$. For the retention time tolerance, 10% of the peak's retention time plus a fix value of 5 seconds is chosen. The percentaged value fits the tolerance to flow factors which can vary between different pre-separation columns while the fix value covers position shifts caused by noise. Since the $1/k_0$ values of peaks are known to be very constant, the tolerance has to cover only small shifts caused by noise, leading to a tolerance value of 0.015 $1/k_0$.

All detected peaks from all measurements are inserted into a list, sorted by their retention time. For every peak in the list, a subset is created which contains all peaks with a retention time difference smaller than or equal to the tolerance. This is done by first moving down in the list starting at the peak's original RT until the tolerance or

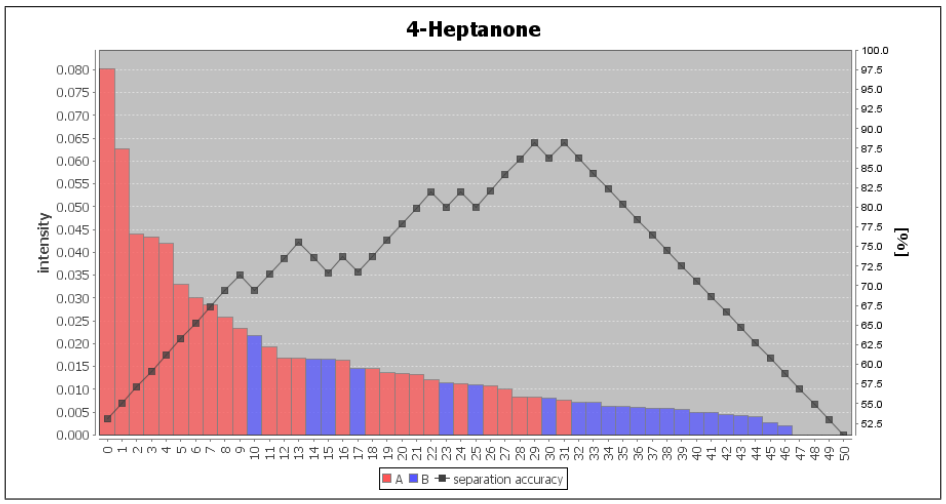


Figure 4.6.: Peak intensities of 4-Heptanone displayed as bars and labeld with previously known classes A and B. The intensity is shown on the left axis. The separation accuracy graph shows the percentaged number of correctly assigned classes when assigning all measurements containing peaks with an intensity equal to or higher than this value to class A and all others to class B. The right axis shows this percentage.

the end of the list is reached, afterwards moving up in the same manner. From this subset, all elements are removed where $1/k_0$ distance to the peak exceeds the $1/k_0$ tolerance. Peaks from this subset which belong to an identical measurement are a strong evidence that the tolerance values are too high and should be altered to prevent this effect. The peak which is closest to the starting peak is kept while the others are moved to a duplication list to create optimised tolerance values.

Using this method, a subset of peaks is assigned to every detected peak. To rate them, the peaks of every subset are sorted in descending order by their intensity. Now a score is assigned to every possible intensity threshold by counting the number of correct threshold based class assignments compared to the known classes to find appropriate intensity thresholds. Figure 4.6 shows an example where one of two possible optimal scores is achieved when setting the threshold to an intensity of 0.0083 resulting in separation between samples nr. 30 and nr. 31. Separating at this point assigns 44 of the 51 measurements to the correct class which equates approximately 86.3 percent. This score gives information about the quality when assigning all measurements that contain this peak with an intensity higher or equal to a first class and below this intensity to a second class.

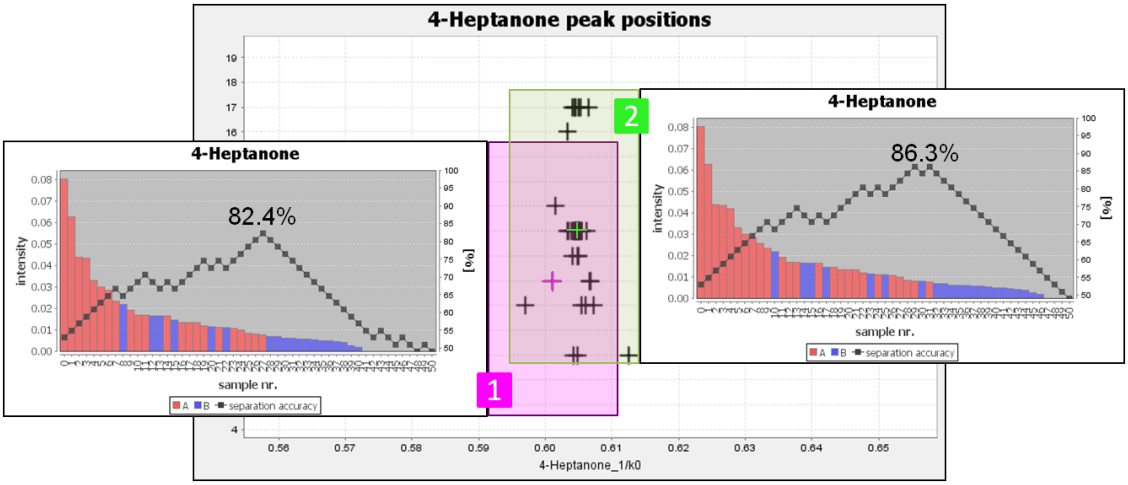


Figure 4.7.: Automatic centering effect of the scoring procedure. The black crosses mark the positions of the peaks. The purple cross is the center used to form the subset shown in the left diagram, the green cross is the center for the subset shown in the right diagram. The maximal achieved score is written in each diagram, showing that the subset formed by the peak in case 2 is a better separation criterion than the subset formed by the peak in case 1.

Sorting the list of all peaks descending by their assigned accuracy score orders them by their performance when using them as a separation criteria. Comparing every peak to every other peak inside the tolerance boundaries leads to multiple entries which have a very similar position and thus refer to the same analyte. Although they have a very similar position, some of them differ in their assigned score which can be seen in Figure 4.7. Peaks which have a higher distance to the centre of their possible group typically achieve a lower accuracy score. This is because a peak in a measurement that does not fit into the tolerance boundaries is considered as not present and marked with an intensity value of zero. Under the assumption that these excluded peaks are approximately equally distributed in the two classes, this results in an automatic detection of an optimal score centre.

4.4. Visualisation and exploration

A visual representation of the data which displays all information about an experiment and allows detailed examination of different areas is one of the most important parts of an IMS analysis. The pre-processed and aligned spectra need to be displayed together with detected peaks

and identified analytes and defined tolerance boundaries. This is necessary for verification of quality and results of the measurements as well as quality of the performed pre-processing, alignment, peak detection an intensity thresholding methods. Interactive visualisation also allows to identify important analytes and areas of the measurement as well as general interpretations.

Since all analysis of IMS data are based on parallel examinations of many measurements, techniques for parallel visualisations are necessary. The key informations every IMS measurement contains are the concentrations of the analytes. These are described through the parameters of the detected peaks. Visualising and comparing a high number of these parameters gives information about the characteristics of analytes and thus enables interpretation and understanding of the data.

4.4.1. Heatmap visualisation

The normalised spectra are composed to an intensity matrix and displayed as a heatmap. A heatmap is a visualisation technique where three parameters of an arbitrary number of data points can be displayed in parallel. Two parameters describe the position of each pixel on a two dimensional map, the third parameter is encoded by the colour of the pixel.

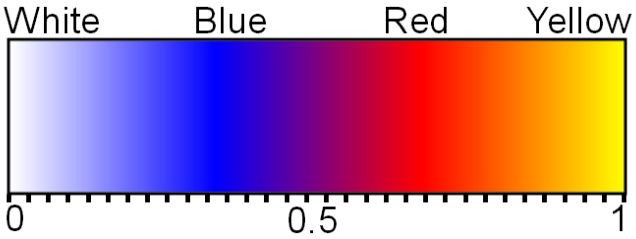


Figure 4.8.: Colour space used to generate the heatmap visualisation.

The position of each IMS data point on this map is determined by its $1/k_0$ value in horizontal position along the X-axis and its RT value in vertical position along the Y-axis. The intensity is mapped to a colour space shown in Figure 4.8. Intensity values which are lower or equal zero after the pre-processing are excluded from the visualisation to reduce the time needed for generation and thus enable a fast browsing through a high number of measurements. The emerging free areas in the heatmap are use to paint grid lines for precise determination of $1/k_0$ and RT values of analytes on the axis. The colour space is

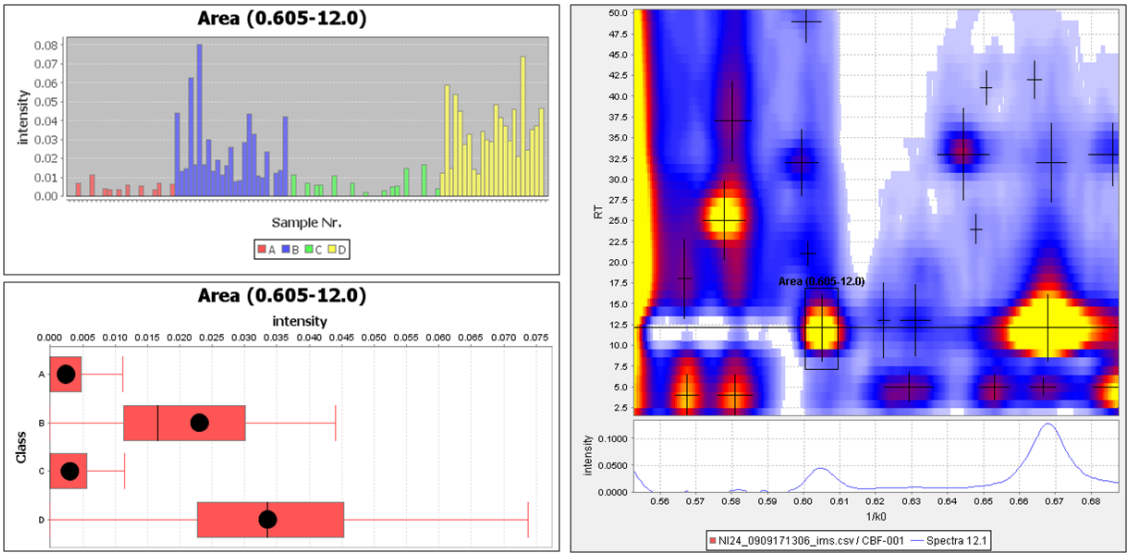


Figure 4.9.: Visualisation of the intensity distribution of one single peak. The bar chart and the boxplot on the left show the distribution of the selected analyte on the right in 4 different classes. The intensity of the selected analyte is higher in classes B and D than in classes A and C. The lower part of the right picture shows one selected spectrum, marked with a black line in the heatmap.

designed to map values between zero and one to a colour range starting with white going on to blue, red, and ending with yellow for high intensities.

4.4.2. Peak intensity visualisation

After a successful peak detection and the definition of tolerance boundaries, all intensity values of peaks which satisfy the constraints can be visualised in different ways. Since many experiments aim at the comparison of intensities in different classes, visualisation methods should support class based analysis.

Bar-charts and box-plots are easy to interpret types of visualisations and enable the determination of peak intensity distributions in different classes as shown in Figure 4.9. The classes can be encoded by colour in bar-charts and used to form subsets for the box-plot visualisation. Both methods have advantages and disadvantages depending on the number of measurements an experiment contains and the aim of the analysis. Bar-charts show one value for every single peak intensity, which is perfect for a detailed analysis, but no longer displayable once the number of measurements exceeds a certain threshold. Box-plots

abstract from the single peak intensities but enable the visualisation of a high number of them. Furthermore they show statistical information like median and mean as well as ranges and possibly outliers.

4.5. Summary

This chapter described methods for spectra pre-processing, peak detection, data analysis, visualisation and exploration which where developed for IMS data. With this set of techniques and methods the fundament exists to construct a software which enables the analysis of IMS data.

CHAPTER 5

IPHEx - IMS Peaklist and Heatmap Explorer

This chapter describes the *IMS Peaklist and Heatmap Explorer* IPHEx, which was built to apply the developed methods to IMS data and visualise, manage and compare their results. In contrast to existing software systems it enables an analysis based on the parallel exploration and visualisation of many measurements without the need of regarding every single measurement in detail. The results of an analysis are typically tables and charts giving information about peak intensities, especially their distribution in different previously defined classes or over a time period.

The software was developed in Java, using the NetBeans integrated development environment and makes strong use of the JFreeChart java library to display professional quality charts.

This chapter starts with a review of the different data structures that are created and used by the software to organise and analyse an experiment. After that, a detailed description of the user interface and its different compounds is given. The last section of this chapter is about the general workflow of IPHEx and shows how the different parts of the software are typically applied to perform an IMS data analysis with IPHEx.

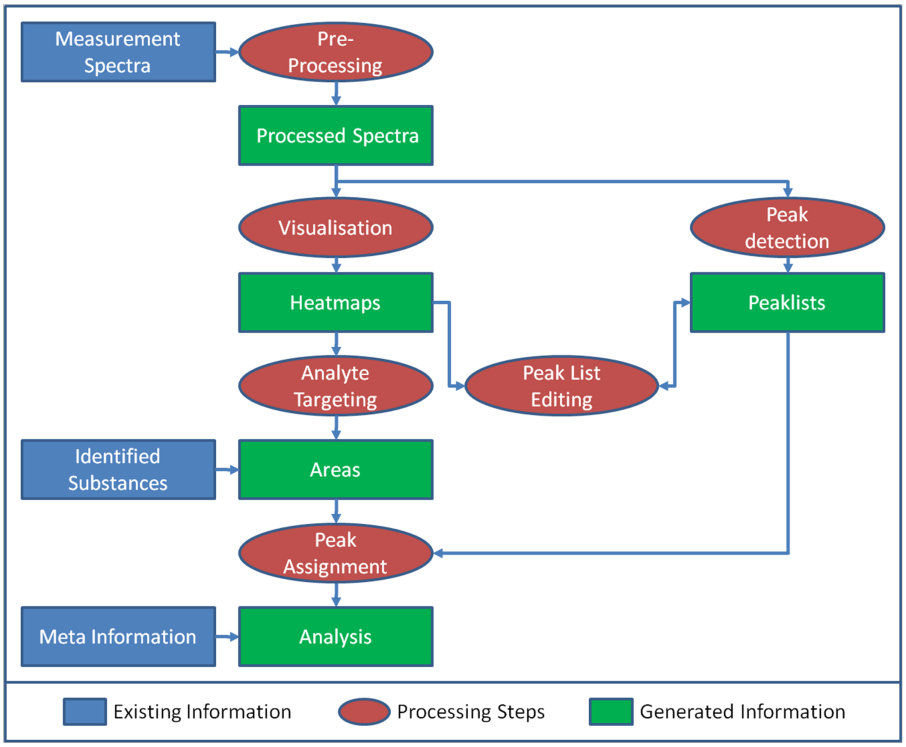


Figure 5.1.: Schema of the different data types and processing steps.

5.1. Concept

The IPHEx software was developed to enable the analysis of complete IMS experiments. These experiments initially consist of two different types of data:

- *Measurement files* which contain the spectra and a header with basic information.
- A *meta information file* which contains additional information to the sampled probes like classes, labels, and parameters.

The analysis of an experiment requires to throughly examine and compare the analytes of all measurements. To enable this, two additional types of data were introduced:

- *Peaklists* contain information about the peaks of a measurement and can be generated automatically and edited through a user interface. Each peak is described through the three parameters intensity, $1/k_0$ position and RT position.
- *Area annotations* are setup to define centre and tolerance boundaries of a specific part of a measurement to select the peaks from

all measurements which satisfy those values. They can be imported from a list of known substances and also defined through a user interface.

These four data types are combined and managed in the IPHEx user interface by the *project browser* and visualised in a table format.

Two secondary types of data, the *pre-processed spectra* and the *heatmap images*, are necessary to enable the annotation of areas and peaks and gain a visual impression of the measurement. They are designed for explorative visual usage and are available through the IPHEx user interface by the *heatmap explorer*.

5.2. User interface and primary components

The IPHEx user interface consists of different *internal window components* which can flexibly be arranged on the IPHEx desktop, depending on type and aim of a data analysis (see Figure 5.2). Information can be exchanged and integrated into the windows by dragging them from the source window and dropping them to the target window.

The software initially starts with an opened file system view which is read from the operating system and integrated in an internal window. Selecting files from this *internal file chooser* and dropping them on the IPHEx desktop opens different windows, depending of the type of the files.

Beside the *internal file chooser*, three primary windows are used during a standard data analysis while several secondary windows are available for special purposes.

- The *project browser* contains a combined table which is generated from the four previously described data types to manage and explore all measurements of an experiment.

- The *heatmap explorer* visualises single measurements together with peak lists and area annotations. It provides functionality to explore the measurement in detail and to create and edit peak lists and areas.

- The *chart viewer* is used every time a chart display is requested by other compounds and is automatically adjusting its visualisation, depending of the type of data to display.

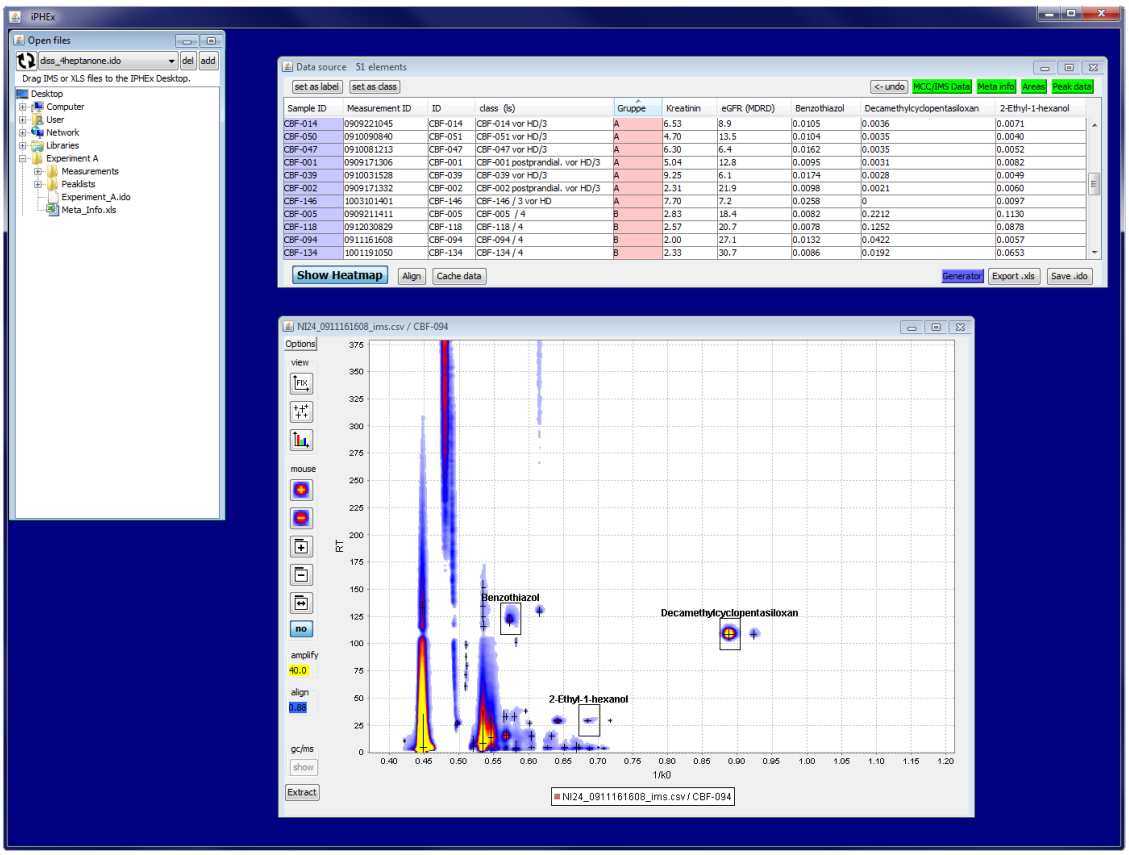


Figure 5.2.: The IPHEx user interface with three visible components. In the top left corner a file system view is provided for opening projects or integrating files into the analysis via drag and drop. The top window shows a view of the current project with various integrated information from different sources. The large window shows the *heatmap explorer* which visualises one single measurement as a heatmap and offers various analyis and annotion features.

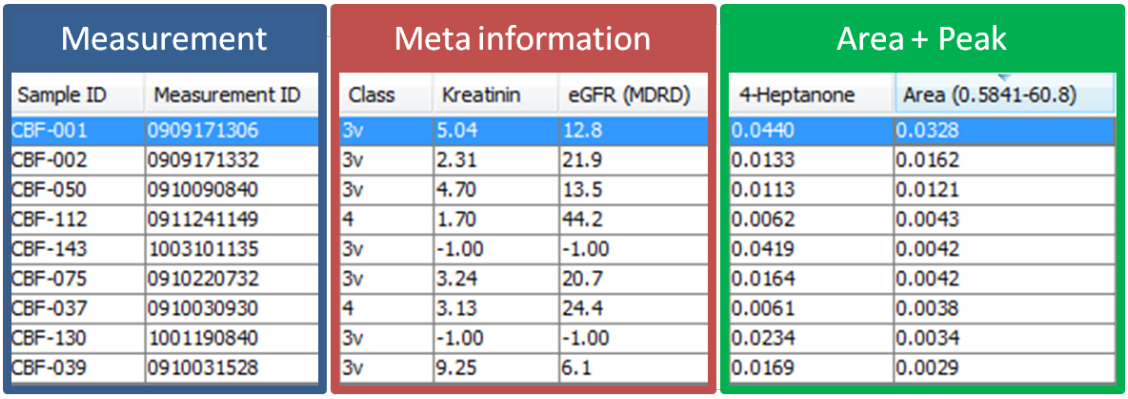
Several projects can be opened in parallel and information can be transfered between them. Additional tools provide specialised visualisations and analysis features to enable a comparison of peak intensities and measurement parameters. Furthermore an intuitive drag and drop interface enables a fast exchange of information between the components.

5.2.1. Drag and drop interface

Target	Dropped data	Performed action
IPHEx desktop	IMS measurement(s) (*ims.csv)	Project browser
	IMS Set (*IMSSet*.xls)	Project browser
	IPHEx file (*.ido)	Project browser
	Excel file (*.xls)	Table viewer
	Strings (tab and/or return delimited)	Table viewer
	Numbers (return delimited)	Chart viewer
Project browser	IMS measurement(s) (*ims.csv)	Add measurement(s)
	Excel file (*.xls)	Add meta information
	Peaklist file(s) (*.pl)	Add Peaklist(s)
	String + 4 numbers tab separated	Add area information
Table viewer	Strings (tab delimited)	Add line
	IMS measurement(s) (*ims.csv)	Data source browser
Chart viewer	Strings (return delimited unique)	Add labels
	Strings (return delimited non-unique)	Add classes

Table 5.1.: Interrelation of the basic IPHEx compounds and the drag & drop listeners. A selection of a table is represented by strings, where columns are delimited by tabulators and lines are delimited by newline characters to enable a flexible transport of information through the drag & drop interface.

IPHEx consequently makes use of a drag and drop listener concept to open, combine, and exchange all types of data between the different IPHEx compounds. This reduces the general time needed for an analysis by avoiding the otherwise necessary amount of file chooser dialogs. It interprets the dropped data without regarding the source of the associated previously performed drag action. This enables a flexible exchange and combination of data from different sources. The interface transports strings which are delimited by tabulators and newlines. Files which can be selected from the internal file chooser by dragging one or more files or folders are converted to a newline delimited list of their absolute paths. In case a folder occurs within the selection, all files within this folder are listed and transported instead. Selections from tables are also converted to tab and newline delimited strings.



Measurement		Meta information			Area + Peak	
Sample ID	Measurement ID	Class	Kreatinin	eGFR (MDRD)	4-Heptanone	Area (0.5841-60.8)
CBF-001	0909171306	3v	5.04	12.8	0.0440	0.0328
CBF-002	0909171332	3v	2.31	21.9	0.0133	0.0162
CBF-050	0910090840	3v	4.70	13.5	0.0113	0.0121
CBF-112	0911241149	4	1.70	44.2	0.0062	0.0043
CBF-143	1003101135	3v	-1.00	-1.00	0.0419	0.0042
CBF-075	0910220732	3v	3.24	20.7	0.0164	0.0042
CBF-037	0910030930	4	3.13	24.4	0.0061	0.0038
CBF-130	1001190840	3v	-1.00	-1.00	0.0234	0.0034
CBF-039	0910031528	3v	9.25	6.1	0.0169	0.0029

Figure 5.3.: Screenshot of the IPHEx data source browser table. Each row represents one measurements while the columns contain related information retrieved from different sources.

The drop listeners of all IPHEx compounds interpret the incoming data by the number of containing tab and newline delimiters as well as the type of each part of the string. For example, the pattern of a string with four tab delimiters where the first string consists of letters while the following four strings consist of numbers, only matches the definition of areas and thus is used to build an area data type and added to the targeted compound. An overview of the possible drag and drop actions as well as their usage is given in Table 5.1.

5.2.2. *Project browser*

The IPHEx *project browser* combines all data types of an experiment and displays it in a simplified table view. Every row represents one measurement. The first columns contain Sample ID and Measurement ID and possibly more information directly gathered from the measurement files (columns one and two in Figure 5.3). The following columns contain additional information provided by a meta information file (columns three to five in Figure 5.3). All following columns contain information about specific analytes, either imported from a list of known substances or selected via the *heatmap explorer* (columns six and seven in figure 5.3). The values inside these columns are peak intensities from peaks whose positions are inside the tolerance values of the defined area. Selecting one peak intensity while a *heatmap explorer* is opened results in a direct jump to the respective measurement and a zoom to the specific position in the heatmap, allowing a fast visual control of the peak. Every row can be sorted by a click on the header of a column which enables a determination of maximal or non existing peak intensities.

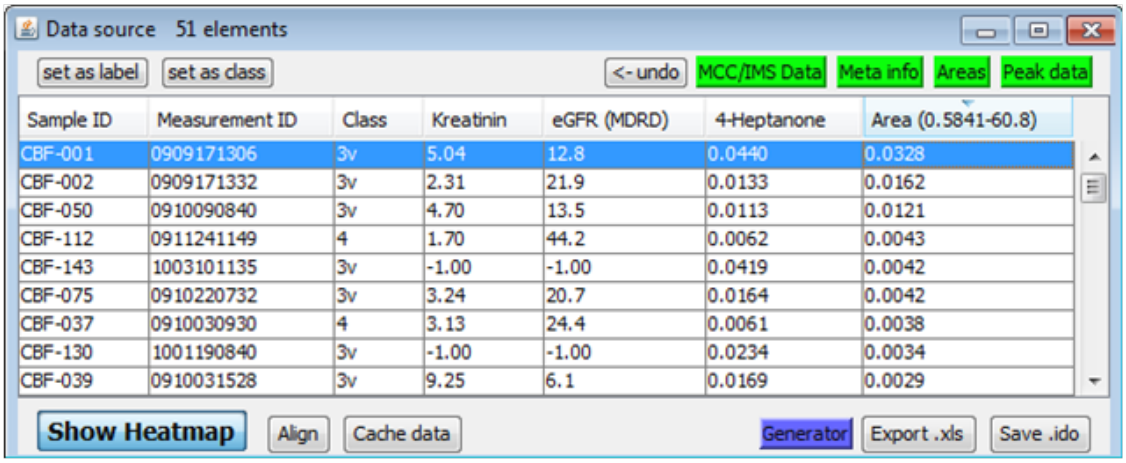


Figure 5.4.: Screenshot of the complete IPHEx data source browser. Several options and buttons above the table allow the organisation and editing of the containing data types. Buttons below the table provide functionality for exploration, analysing and exporting the data.

Various options for editing those data types are available through an interface in the upper right corner of the explorer (see Figure 5.4). Columns that contain information about classes and columns with identifiers can be marked to include this information in further visualisations and analysis using the buttons *set as label* and *set as class* in the upper left corner. The single measurements can be displayed as a heatmap together with all available information by opening the *heatmap explorer*. Different analyses and visualisations can be generated and are described in the following Sections.

The *project browser* is started and data is added to it using a drag and drop based interface which is described in section 5.2.1.

5.2.3. *Heatmap explorer*

The IPHEx *heatmap explorer* enables the detailed examination of single measurements and provides functionality to add and remove area annotations as well as peaks to the *project browser*. Every heatmap can be zoomed, panned and an amplification factor can be defined to alternate the colour assignment, resulting in an accentuation of low signals.

Whenever a heatmap image needs to be displayed the first time, the measurement file is opened, the raw spectra are read, and the pre-processing pipeline is started. After that, the pre-processed spectra are saved to a folder inside the IPHEx directory to speed up every subsequent usage of the spectrum. Whenever a measurement needs

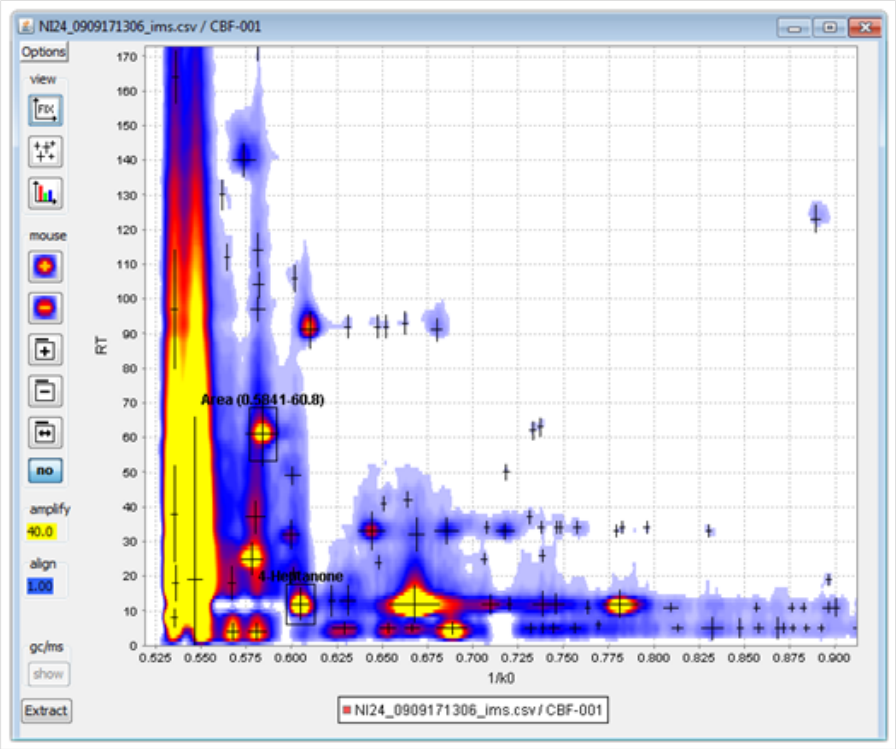


Figure 5.5.: The IPHEx heatmap explorer shows heatmaps of measurements selected through the data source browser. It provides tools for editing peaklists and generation of areas as well as general functionality to zoom, pan and explore the measurement.

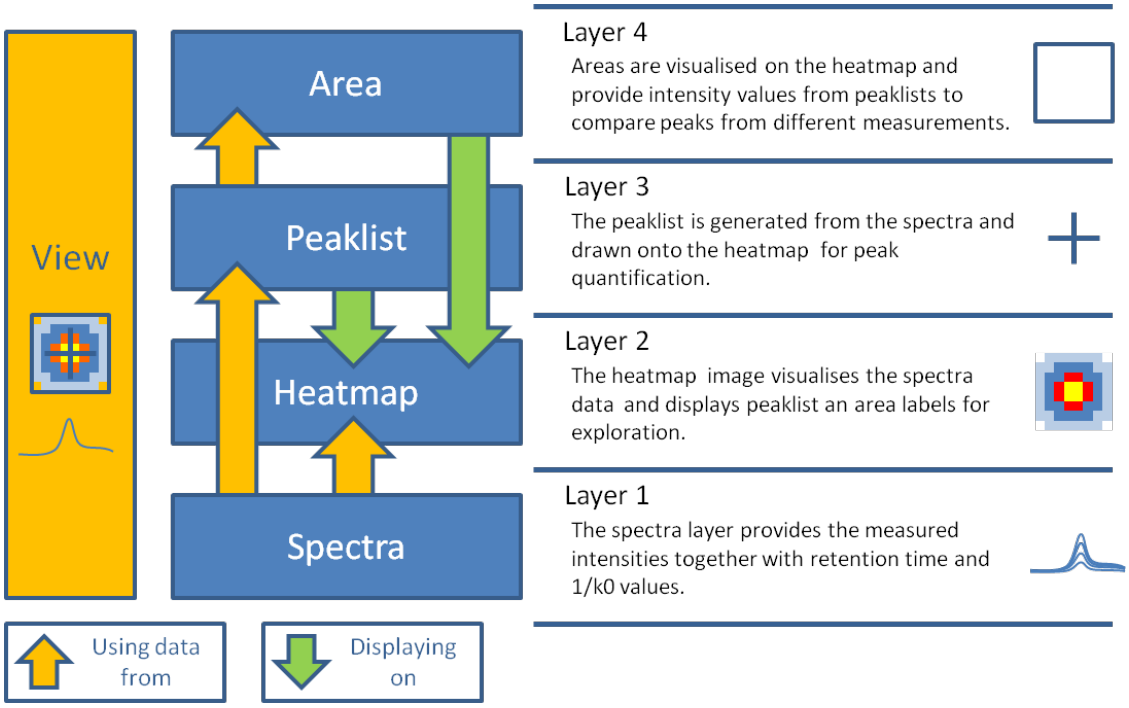


Figure 5.6.: Visualisation layer of the IPHEx *heatmap explorer*.

to be visualised as a heatmap, this folder is checked if it contains a pre-processed version of this spectrum which can be used instead of the original file. Heatmap images are not saved directly as image files because the parameter of the visualisation can change according to aims of the experiment and types of measurements.

The generated peaklists are displayed directly onto the heatmaps and peaks can be added to and removed from a peaklist. All area annotations registered in the *project browser* are automatically displayed in the *heatmap explorer* as rectangles together with the name of the area above them. They can be moved, resized, and new areas can be defined. Defining and altering peaks and areas directly affects the respective values in the *project browser*. A peaklist is always assigned to one single measurement while areas are always assigned to the whole experiment and thus visible on all heatmaps.

Because of the high number of measurements an experiment usually contains, mechanisms were developed to decrease the time needed for an analysis. Therefore only some measurements need to be regarded in detail with the *heatmap explorer* to setup the area annotations and alignment procedure. The majority of the measurements can be investigated by using the *project browser* to analyse them based on the values of their quantified peaks. This enables the sorting of the measurements based on the peak intensity values of specific peaks

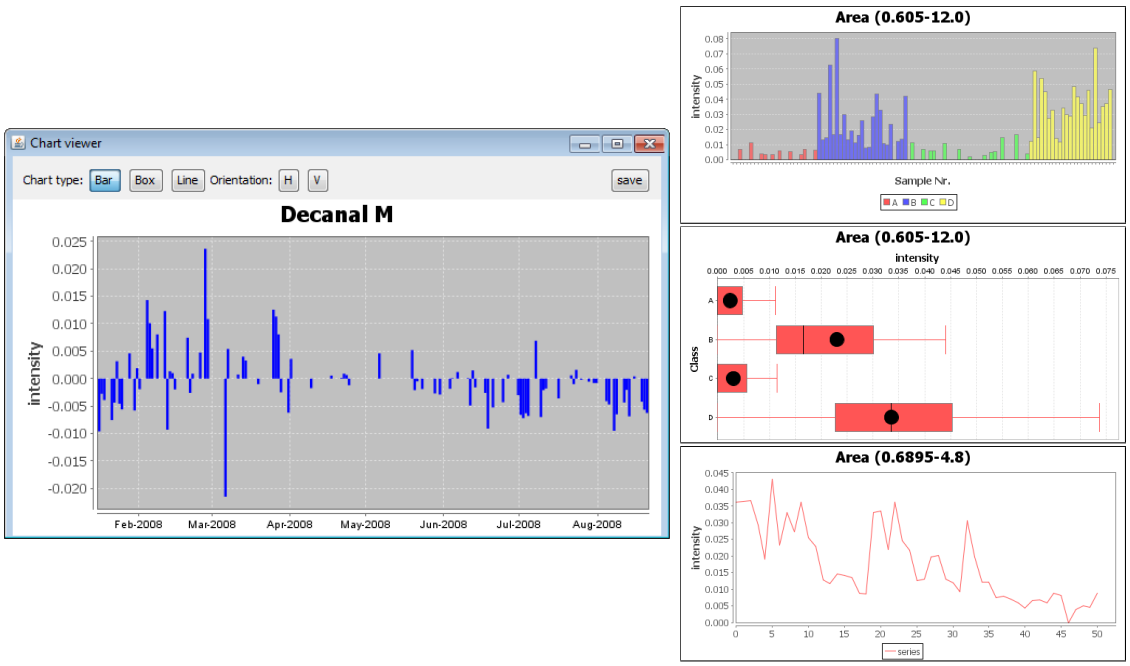


Figure 5.7.: The chart viewer of the IPHEx software, opened with a time depended bar chart and three further visualisations (box plot, class based bar chart, and line plot) which can be selected using the buttons in the top left corner.

and thus a separation of measurements with expected average intensities from those of special interest.

This supports the goal of IPHEx to enable an analysis without the need to consider every single measurement in detail. Together with the described buffering techniques which allow fast skimming through heatmap images it drastically reduces the time needed to perform an IMS data analysis.

5.2.4. *Chart viewer*

The *chart viewer* is a flexible part of the IPHEx software to provide different kinds of chart visualisations and is used for various tasks. Upon startup it displays a standard bar chart with vertically orientated bars and can be altered using the option buttons for chart type and orientation in the top left.

If it is started via the *project browser* by selecting one or more columns, and selecting *show chart*, it displays the values of this column according to the order of the table. If columns are marked as class and/or label columns this information is used by the chart viewer to assign the same colour to values from the same class and label them

according to the selected label column. Selecting the Measurement ID column for labelling sets the x-axis to show the dates of the measurements, if the Sample ID column is selected for labelling, all selected columns are displayed in one chart.

Instead of setting area annotations in the *heatmap explorer* and plotting the intensities afterwards, it can also be used to visualise peak intensities in an area directly by selecting the *Live chart* button in the *heatmap explorer* and moving the mouse over a heatmap. The *chart viewer* then opens and shows the intensities of the peaks which are closest and within the shown tolerance boundaries, instantly updating while the mouse moves. This visualisation can be used in parallel to the area annotation feature to quickly identify and select interesting peaks.

It is also used by the correlation viewer to show detailed information about the peak intensities.

5.3. Specialised analysis tools

Beside the three primary components of the user interface which are available for general analysis and project management, there exist several additional tools and extensions for specialised and detailed analyses.

5.3.1. Correlation viewer

The correlation viewer enables the detection of dependencies and similar behaviour between analytes or meta information and analytes. The Pearson correlation coefficients are visualised as circles inside the diagram while the diameter refers to the value of a coefficient. Positive correlations are represented in blue and negative values in red colour. Clicking on a circle opens two *chart viewer*s to compare the intensity values in detail.

The *correlation viewer* is started by the *project browser* via selecting all columns that should be correlated against each other and selecting the *correlation viewer* option. All types of columns that contain numerical content can be chosen. It provides functionality to zoom and pan inside the diagram which is especially important when comparing a large number of columns. To improve the readability, only circles with a correlation factor of at least 0.10 are labelled with the correlation value.

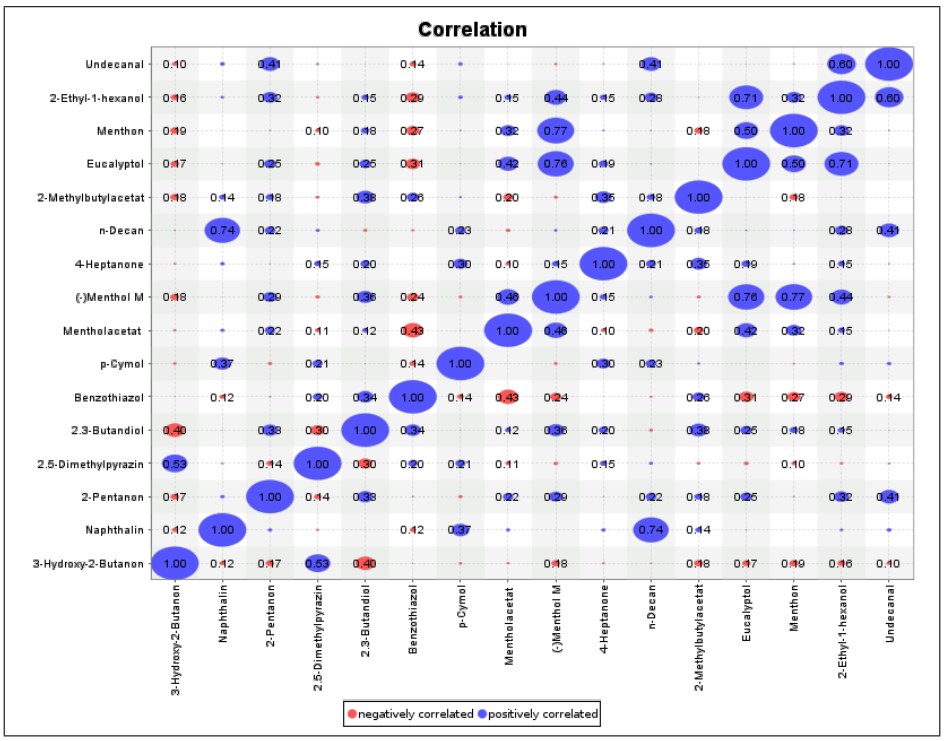


Figure 5.8.: Correlation plot of the intensity of 16 known substances determined in 158 measurements.

This visualisation method will become more and more important in the future because it allows an easy analysis and interpretation of a unlimited number of measurements.

5.3.2. Principal component analysis

A principal component analysis (PCA) can by performed by selecting all substances from the *project browser* which should be included and selecting the *PCA* option from the *generator* menu in the bottom right corner (Figure 5.9). The tool allows to choose the principal components which are displayed on the x- and y-axis and uses the same symbols for measurements which were assigned to identical classes with the *project browser*.

5.3.3. GC/MS comparison

An important task when analysing MCC/IMS measurements is the identification of the measured substances. One method to achieve this is to record a sample with a GC/MS in addition and compare it to the MCC/IMS measurement. Dr. Melanie Jünger, who was part of the

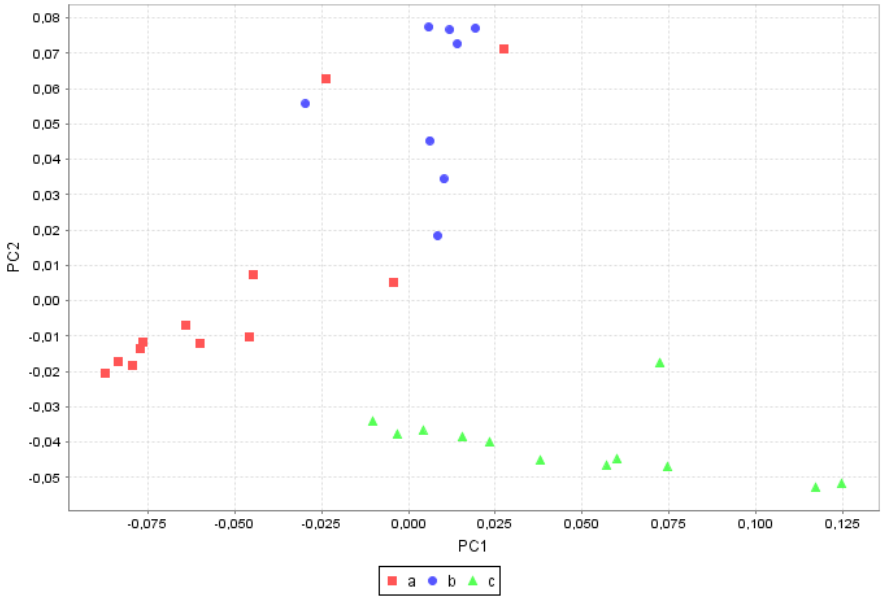


Figure 5.9.: Visualisation of a principal component analysis using a set of substances from a project with previously labelled classes.

metabolomics group at the ISAS, investigated the characteristics of sixteen substances with both devices [41]. Experimentally determined MCC and GC retention times of all substances were aligned and their relation was determined as a result.

Based on this work, GC/MS and MCC/IMS measurements can be visualised in parallel by the *heatmap explorer*. To include GC/MS measurements to the *project browser* they need to be stored or exported to the *NetCDF* data format and either the name of the file or one of the description fields has to contain the filename of the associated MCC/IMS measurement. When these files are added to a project by simply dragging them from the *internal file chooser* to the *project browser*, a new column is created which shows the filenames of the GC/MS measurements. When measurements which are visualised in the *heatmap explorer* have an assigned GC/MS measurement available, a button in the bottom left inside the *heatmap explorer* is active which allows the parallel visualisation (Figure 5.10).

5.3.4. Spot table

The spot table visualisation allows comparison of areas in different measurements on a visual level. It can be used to display many measurements in parallel while concentrating only on specific areas. When interesting peaks have been found with the intensity value based IPHEx

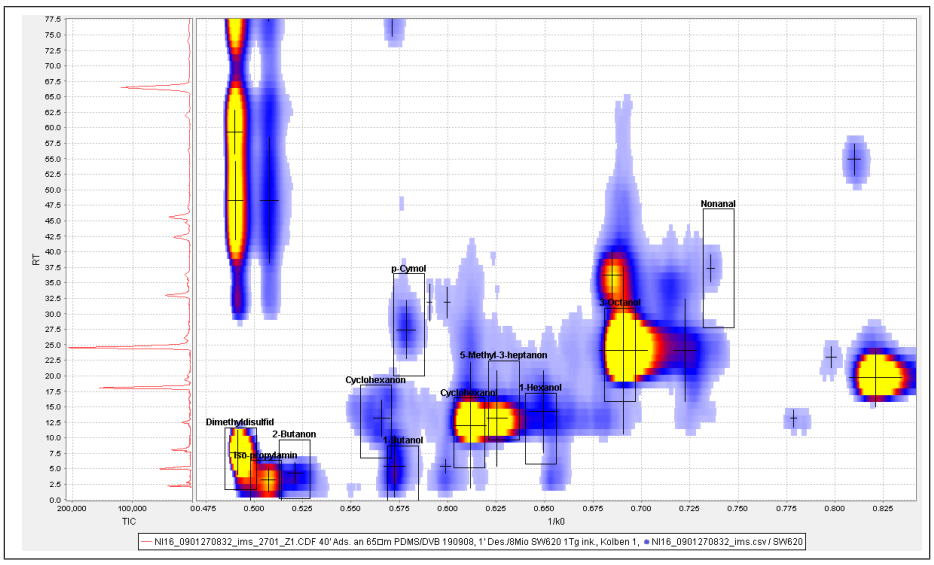


Figure 5.10.: A parallel plot of an MCC/IMS and a GC/MS measurement. The total ion count plotted against the gas chromatographic retention time is shown at the left. The axis is transformed to enable a direct comparison to the MCC/IMS Heatmap at the right

methods, the spot table can be used to prove that the intensity values are caused by proper signals. Another application for this kind of visualisation is the supervision of the alignment procedures and the determination of the general quality of a measurement regarding specific reference areas. It is started by the *project browser* by selecting measurements and selection the *spot table* option.

5.3.5. Reference subtraction

In many projects there exist some kind of reference measurements. In case of bacteria or cell culture experiments which are using headspace samples, these are measurements that contain only substances produced by the medium, in case of breath analysis these are measurements taken from the ambient surrounding air. They can be subtracted by including and selecting in them in the *project browser* and using the *subtract reference* option from the *generator* menu in the bottom right corner. Afterwards all displayed peak intensities show the result of the subtraction from the original sample and the reference measurement and can be used for all further analyses. An additional column is displayed in the *project browser* which shows the time difference between both measurements. Due to the consequent usage of the peaklist concept in IPHEx an automatic reference subtraction is possible for the

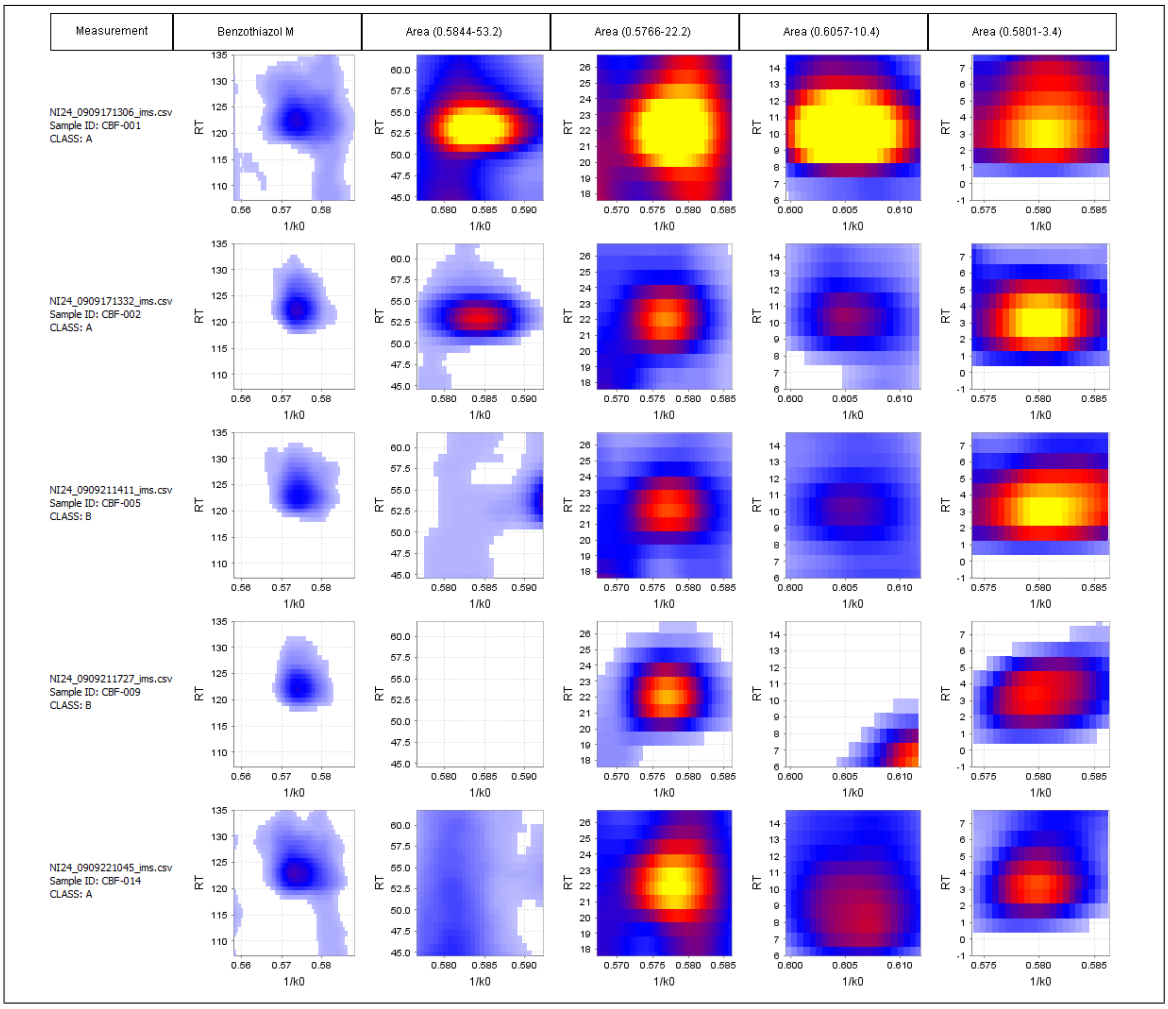


Figure 5.11.: Spot table visualisation of five different areas in five measurements.

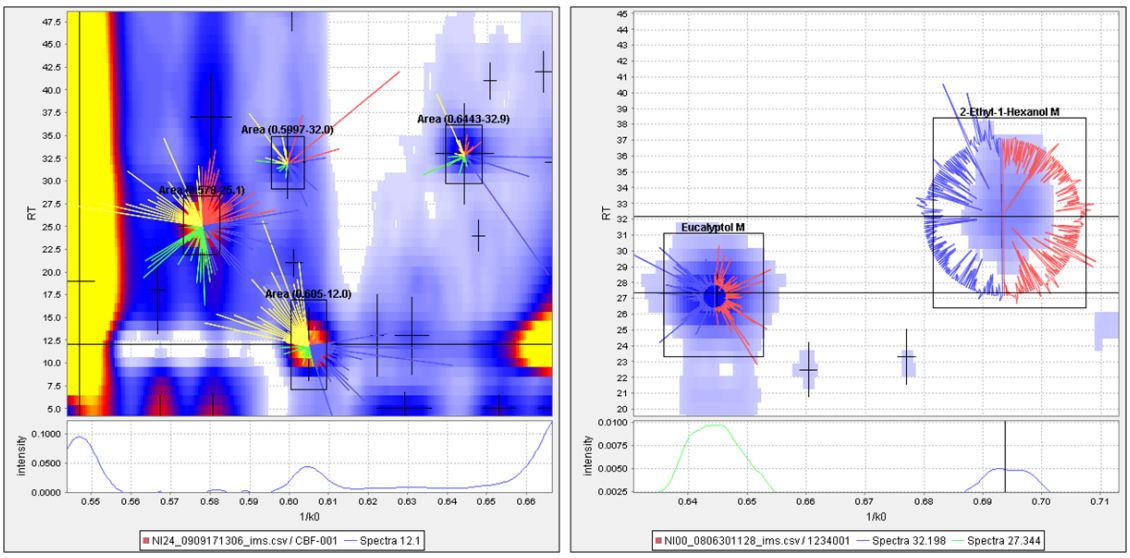


Figure 5.12.: Two different examples for polar glyphs. The left picture shows polar glyphs on 4 selected areas, each displaying the intensities of 113 measurements. In this case, only positive values occur. The right picture shows an experiment consisting of 448 measurements where 224 measurements were samples of ambient air and subtracted from the breath samples. In this case also negative values occur. It is easy to see that in this experiment Eucalyptol has mostly positive values and thus is part of the breath sample while 2-Ethyl-1-Hexanol is mostly caused by the ambient air and therefore has negative values. The lower parts of both images show selected single spectra.

first time in MCC-IMS analyses. The usage of reference measurements is a key technique to create reasonable analysis results but was often skipped or briefly dealt with in the past because of the complexity and high expenditure of time. With the integration of this feature into IPHEx this problem no longer exists.

5.3.6. Polar glyph visualisation

A method for parallel visualisation of peak intensities in all measurements are the polar coordinate plots. Every angle represents a measurement and the radius encodes the peak intensities. These plots can be scaled down to a small size, while still remaining interpretable which enables parallel visualisation directly onto the heatmaps as shown in Figure 5.12. They are positioned on the heatmap, using the position of the area annotation as centre of the polar plot to combine the position and intensity information. The visualisation of these glyphs can be en-

abled by selecting the respective entry from the options menu in the top left corner of the *heatmap explorer*.

5.4. Workflow and general usage of IPHEx

The following paragraphs describe a standard workflow for an IMS data analysis with the IPHEx software. The first step, starting the *project browser*, is always necessary. All further steps can be ordered or skipped depending on the type and target of the analysis but are listed here in a typically applied order.

5.4.1. Selecting files and starting the *project browser*

The first step in an analysis of an IMS experiment with IPHEx is the selection of measurements. This can be done in two different ways, depending on the existence of a meta information file.

- If a meta information file for the experiment exists it can be selected in the internal file chooser and dragged to the IPHEx desktop. It is displayed in table form, where each line contains a filename or sample-id of a measurement at the first position and different meta informations in the following positions. Entries of this table can be edited by simply editing the single cells or deleting lines by dragging them to the bin icon in the lower right corner. If only a small subset of this information is needed, a selection of cells can be dragged to the IPHEx desktop to create a new meta information table. Once the customisation of the meta information table is finished, measurement files or a folder containing those files can be dragged directly from the *internal file chooser* to the table. The *project browser* opens with all files which were dragged to the meta information table and are listed within this table. Additionally, if filenames exist that were listed in the meta information table but could not be found in the file system, a window opens containing a list of those filenames.
- If no meta information file exists, the *project browser* is started by selecting measurement files from the *internal file chooser* and dropping them directly to the IPHEx desktop. Also files can be added to an already existing *project browser* in the same way.

The *project browser* is the central element of the IPHEx software and combines all experiment related information. It can be started by selecting measurements as described above or by dragging a previously

saved experiment file from the internal file chooser to the IPHEx desktop. The *project browser* has always measurement files registered, all further informations are optional.

5.4.2. Defining meta information

If a meta information file exists, it can be integrated by selecting it in the *internal file chooser* and dropping it on the *project browser*. If no meta information file is available or additional information needs to be described it can be added by choosing the *add/edit info for selection* in the *meta info* menu in the upper right corner.

5.4.3. Starting the *heatmap explorer*

Single measurements can be displayed inside the *heatmap explorer* by selecting one and pressing the *Show heatmap* button in the lower left corner. The *heatmap explorer* visualises single measurements as a heatmap and provides functionality to explore and analyse the data (see Section 5.2.3). It also shows all area annotation based information from the *project browser* and the peaklist for a measurement if available. Different tools are part of this explorer to enable peaklist editing, area annotation, and various top level analyses and will be described in the following workflow as they become necessary.

5.4.4. Generating peaklists

In most cases of an IMS data analysis, quantitative information of peaks is needed for further investigation. To provide this information a peaklist can be generated from every measurement by an automatic procedure by selecting all necessary measurements in the *project browser* and choosing the *generate peaklist for selection* option from the respective menu in the upper right. Options to export and delete those are also provided within the menu and they can be visualised and refined manually using the *heatmap explorer*.

Peak intensities are visible and can be investigated in detail in the *project browser* when analyte information is defined.

5.4.5. Setting up analyte information and retention time alignment

If the analysis is focused on a specific analyte that should be detected in all measurements of an experiment, it can be directly imported

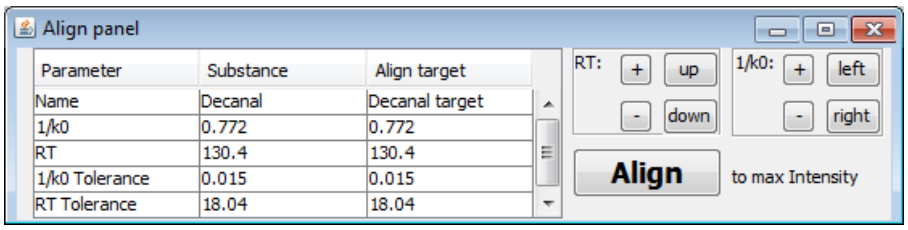


Figure 5.13.: Screenshot of the *alignment parameter window*. The second column of the table shows the substance which was chosen for the alignment, the third column shows the parameters of the target area which is used to detect peaks for the alignment. The parameters can be altered by the buttons on the right side, and are directly visualised in the *heatmap explorer* to aid the parameter setup process.

by opening a table file which contains the respective parameters and dragging the name of the analyte together with the four needed position and tolerance values to the *project browser*. The dropped information is used to form a new area annotation column which is generated by searching all peaklists for peaks within the defined area and extracting its intensity. The position and tolerance boundaries of this are also visible inside the *heatmap explorer*.

The *heatmap explorer* allows to identify peaks of a measurement based on a peaklist and a list of known analytes. Choosing the option *identify known analytes* uses a file from the IPHEx directory (SubstanceDB.xls as default), matches all peaks from the viewed measurement against this file and adds occurring analytes as area information to the *project browser*.

Areas can also be added, deleted and moved manually within the *heatmap explorer* using the button panel at the left side to investigate so far unknown peaks and refine the parameters of existing area annotations. Furthermore they can be renamed and their parameters can be fitted manually by selecting the *edit area* option from the *area* menu in the upper right of the *project browser* .

A very important step is using the retention time alignment function to compensate the variations between measurements caused by the pre-separation technique. When a list of known analytes is available and should be used to identify peaks of an experiment it is also necessary to make this list applicable by aligning the measurements of an experiment to the existing analyte information. Both can be achieved by searching an analyte in the experiment which occurs in as many measurements as possible, and additional in the list of known analytes with the above described methods.

Afterwards the respective analyte area column is marked in the *project browser* and the *align* button in the lower left is chosen which opens an additionally *alignment parameter window* (Figure 5.13). The *heatmap explorer* now shows the selected area annotation from the defined analyte together with a second annotation called *target annotation*. The position and tolerance boundaries of this *target annotation* can be altered by selecting the respective buttons in the *alignment parameter window*. The retention time axis of all measurements which have a peak inside this *target annotation* window will be altered to move this peak to the retention time position of the originally selected analyte from the *project browser*.

5.4.6. Top level analysis

After the previously described steps were performed, an investigation of the project based on the determined peak intensities can be performed. A typical start is to select one column from the *project browser* and mark it as label column, which will include the entries in this column into all analyses as identifiers. If information about different groups or classes is available in the *project browser*, it is also marked as class column to enable a class based comparison. Buttons to mark columns as class or label column are available in the top left of the *project browser* and columns which are selected by this features are highlighted in red and blue. Now single substances or areas of the measurement can be visualised with the *chart viewer* by selecting the respective column from the *project browser* and selecting the *Show Chart* entry from the *Generator* menu in the bottom right. Details about the *chart viewer*, are shown in Section 5.2.4. More complex analysis and visualisation for example correlation analysis, pca, GC/MS comparisons, and subtracting available reference measurements are easyly accessible based on the here composed project information and described in detail in the previous Section 5.3.

5.5. Summary

IPHEx is an analysis environment is, which can perform all necessary tasks for a comprehensive investigation of large amounts of MCC-IMS data with relatively low expenditure of time. Due to the complex and unstructured type of MCC-IMS data, a fully automatic analysis without user input is not feasible. A fully manual analysis is also impossible due to the amount and complexity of the data. The IPHEx software system solves this task by performing all possible operations automat-

ically (pre-processing,heatmap generation, peak detection, reference subtraction) and supporting all user required tasks for analysis, management, and visualisation as much as possible to reduce the necessary expenditure of time.

CHAPTER 6

Application examples

The previous chapters showed the methods and concepts developed for the analysis of IMS data as well as the integration into a Java based software framework. The developed system allows a detailed and fast analysis of large data sets and an integration of various existing information. In the following chapter examples of metabolomic experiments and their analysis using the IPHEx software are described.

6.1. MCC/IMS signals in human breath related to sarcoidosis-results of a feasibility study using an automated peak finding procedure

The first use of the IPHEx software was the analysis of exhaled air measurements from patients with sarcoidosis and was recorded at the lung hospital Hemer. Volatile metabolites occurring in human exhaled air are postulated to correlate directly to different kinds of diseases. An MCC-IMS was used to identify potential volatile metabolites occurring in human breath within less than 500 seconds and without any pre-concentration. The IMS investigations are based on different drift times of ion swarms from metabolites formed directly in air at ambient pressure using about 10 ml of human exhaled breath. The aim was to find sectors of interest inside the measurements which can support

the diagnosing process. Therefore different data processing steps and procedures including detection and comparison of peaks in different measurements were used to generate a list of peaks which are possibly related to the disease. First results of this work performed with early parts of the IPHEx software, were published in the *Journal of Breath Research* [42].

6.1.1. Data

For this experiment, twenty measurements were available, were eleven originate from patients that had a confirmed sarcoidosis (group A) and nine measurements were used as control group (group B). Both groups showed no significant difference in age ($A: 46.5 \pm 15.5 years; B: 51.8 \pm 12.6 years; p = 0.44$), Body Mass Index (BMI) ($A: 28.4 \pm 7.8; B: 27.6 \pm 6.2 kgm-2; p = 0.56$), Forced Expiratory Volume (FEV) 1 ($A: 2.9 \pm 1.5 L; B: 3.0 \pm 1.0 L; p = 0.84$), Virtual Capacity (VC) ($A: 3.9 \pm 1.7 L; B: 4.4 \pm 1.0 L; p = 0.49$), and diffusion capacity (TLCO) ($A: 62.1 \pm 18.0\%; B: 78.2 \pm 14.4\%; p = 0.09$). Four patients with sarcoidosis and two patients of the control group were never smokers.

6.1.2. Analysis and visualisations

After the application of pre-processing methods to the measurements which were described in Section 4.1 including filtering, baseline correction and alignment, peaklists were created and heatmaps were visualised (Figure 6.1).

In this study, the aim was to develop an automatic procedure which detects one or more peaks that are class-specific, which means that they occur much more frequently in the disease group than in the control group or vice versa. To identify these specific peaks, a value or score needs to be calculated. This score should indicate how effectively one peak can separate the two groups. Since the number of peaks can be vast, the score must be obtained automatically. The peaks with the highest scores are assumed to be possibly class-specific. To achieve this, the occurrences of all peaks in class A and in class B are counted. Peaks were considered the same if they do not exceed a tolerance value of 10% of the retention time plus a fixed value of 5 s on the y-axis and 0.015 $1/k_0$ on the x-axis. Based on these occurrences, the sensitivity, specificity and accuracy can be determined for every peak in the list. The sensitivity is equal to the number of datasets where a peak occurred in class A, divided by the number of all datasets in class A. The specificity is equal to one minus the number of datasets where a peak occurred in class B, divided by the number

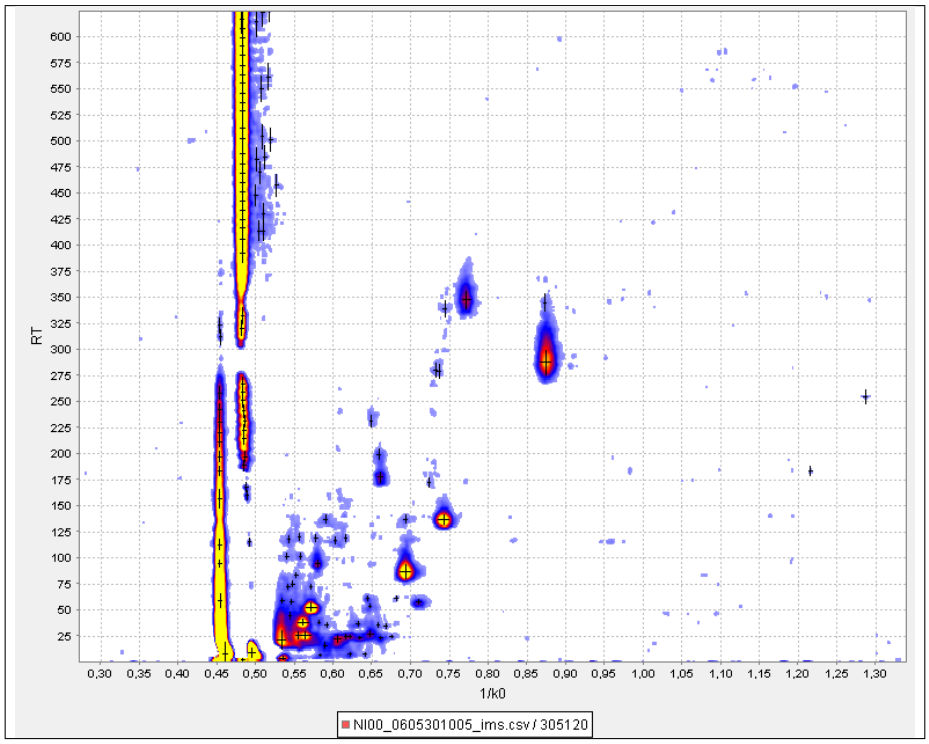


Figure 6.1.: Heatmap image of one measurement, obtained from a sarcoidosis patient. Filtering, alignment and baseline correction methods were applied, peaks were detected and marked with a cross by an automatic procedure.

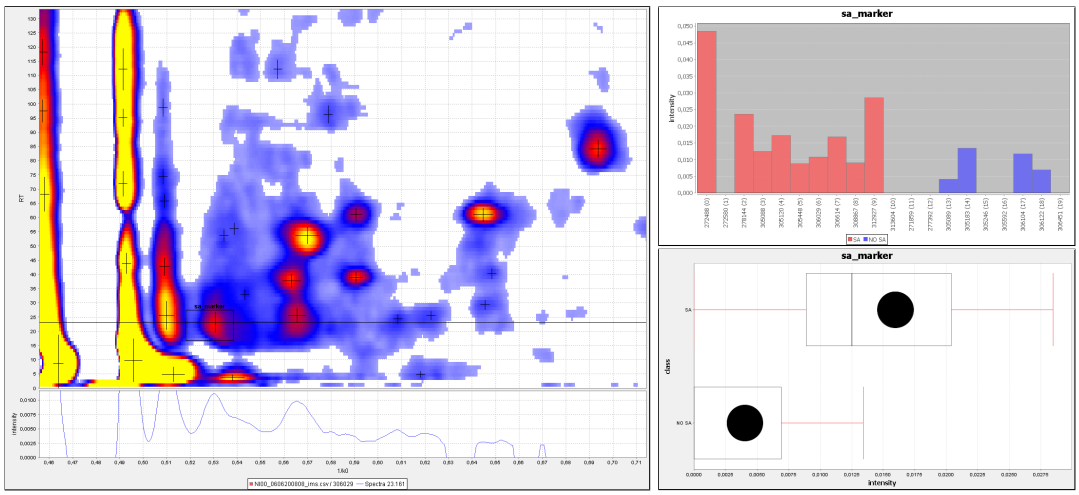


Figure 6.2.: One possible marker region for sarcoidosis which was determined by a scoring procedure. The heatmap on the left shows the position of the marker, bar and box plots on the right show the intensity distributions in the sarcoidois (red) and control (blue) groups.

of all datasets in class B. The accuracy is computed by adding the number of datasets where a peak occurred in class A to the number of datasets where the peak did not occur in class B, divided by the size of the whole pre-classified sample. The accuracy was used as a score to find and sort possible class-specific peaks. This procedure was later on extended to the *optimal intensity threshold procedure* shown in Section 4.3.2 to include intensity levels as well.

When the analysis of this project started in 2008, the first pre-processing and peak detection methods were ready to use. The original analysis was done based on scoring the values from the peaklists, visualisations, visual control of the results and plots had to be done by hand with the help of Excel and BBImsAnalyse. One region where a potential marker for the sarcoidosis group occurs was identified as a result of the study. Figure 6.2 shows the position of the region on one exemplary heatmap together with two plots of the intensities in all members of the experiment, repeated with the current version of IPHEx. Since pre-processing, alignment and peak detection methods were improved over the time, the here visualised result differs slightly from that of the original publication, while the overall result remains similar. A complete comparison of all regions done with the spot table visualisation (see Section 5.3.4) is presented in Figure 6.3. The results of retention time alignment using the Decanal are shown in the first column. All measurements contain Decanal and the vertical position of Decanal and thus the whole retention time is aligned properly in all

measurements. The second column shows the region which was used for discrimination.

6.1.3. Results and discussion

The aim of this study was to detect possible marker regions for sarcoidosis completely automatically based on previously generated peaklists. To verify the results and provide a visualisation method, the spot table visualisation was created and later on included into the IPHEx software.

The application of the scoring procedure based on the usage of classification accuracy to all datasets of patients with confirmed sarcoidosis and a control group shows a potential marker region characterised by the following parameters: inverse mobility $1/k_0$ 0.53 ± 0.01 V s cm^2, retention time 22 ± 5 s. The visual control of the marker region shows that eventually different analytes within the region are responsible for the peak intensity value which is not optimal. Regarding the small number of available data for this disease and possible error rates in peak detection and signal processing steps, these results should be considered preliminary and need to be confirmed once more data becomes available. Results of this work were published in the *International Journal for Ion Mobility Spectrometry* [42] and in the *Journal of Breath Research* [43] in 2009.

During this study the need of a close connection between analysis and visualisation became obvious and was one reason why the IPHEx software was developed.

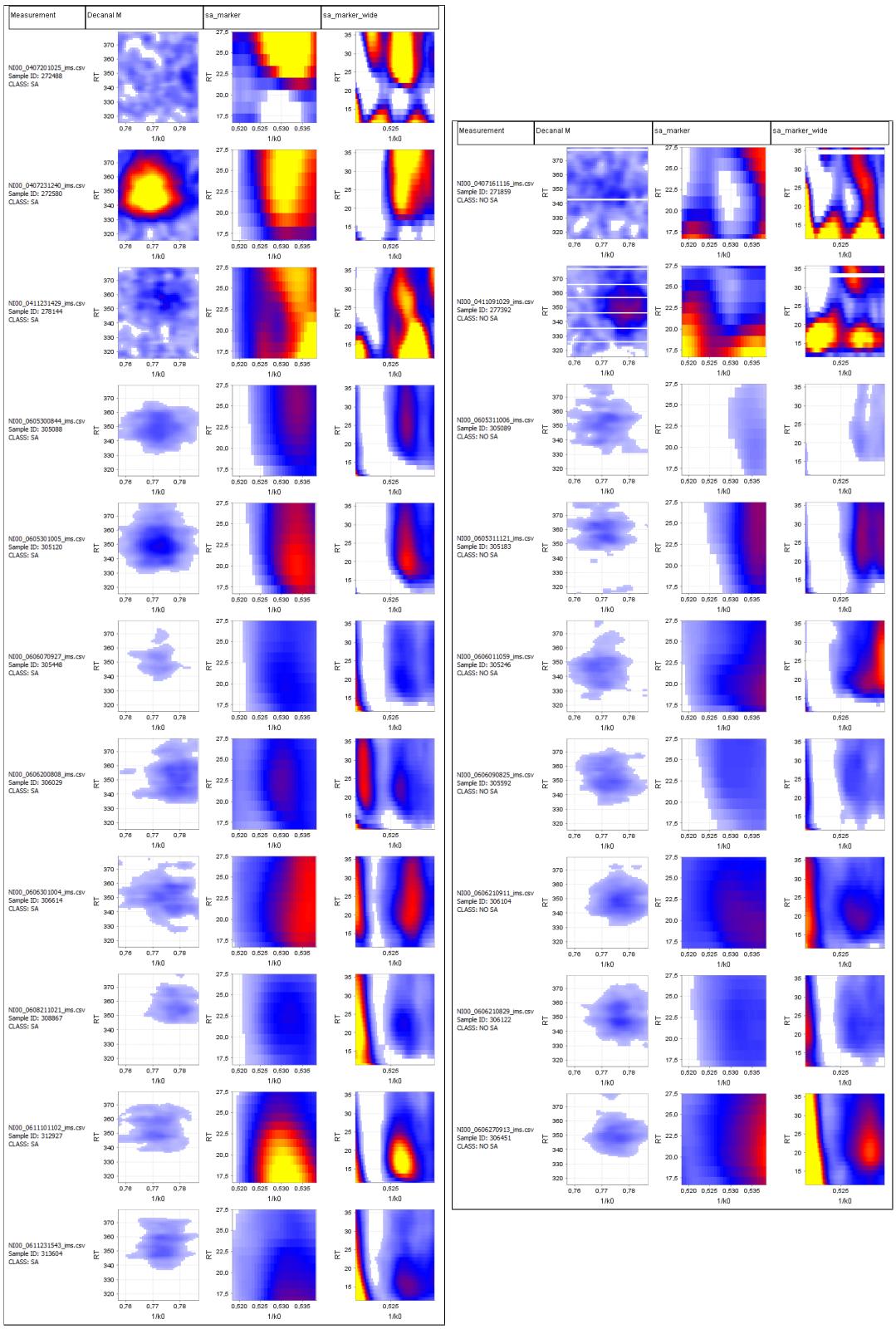


Figure 6.3.: A visualisation of the marker region for sarcoidosis in all measurements. The left part contains measurements from sarcoidosis, the right part shows measurements which were used as control group. The first column shows a substance which was used to validate the alignments, column two shows the marker region, and column three a zoomed out view of the marker region.

6.2. Ion mobility spectrometry of human pathologic bacteria

This study was performed to determine volatile organic compounds in the headspace above eight clinically important bacterial species (Klebsiella pneu., Pseudomonas aeruginosa, Serratia marcescens, Staph. aureus, Strept. Pneumoniae, E. coli, Enterobacter) and human pathogenic yeast (Candida albicans, Aspergillus fumigatus) cultured on Columbia blood agar media using a MCC-IMS.

6.2.1. Background and objectives

Suspected bacterial infection is empirically treated with broad spectrum antibiotics because definitive diagnosis and antibiotic sensitivities require time consuming cultures. These delays significantly increase patient mortality, while empiric antibiotics result in greater expense and increasing antimicrobial resistance. A multi-capillary column equipped ion mobility spectrometer (MCC-IMS) was used to measure complex mixtures of gases regardless of the water vapour content in real time without sample preparation or pre-concentration. The general aim was to detect peaks in the MCC-IMS data which are typical for a specific bacterial infection and thus can be used to aid the diagnostic process. Appropriate, early antibiotic treatment results in lower mortality rates than when given once culture results are known. This study was done in cooperation with the Department of Anesthesiology, Emergency and Intensive Care Medicine, University Göttingen, Germany.

6.2.2. Analysis and visualisation

A total of 128 measurements labeled with 11 different classes were used to perform the analysis. They were pre-processed, a peak detection was performed, and they were combined in the *project browser*. After labeling them with their respective classes, the *Live Chart Modus* could be used to detect all class related regions in an easy and very time efficient visual exploration procedure. The close connection between visualisation, exploration and comparison of the intensity distribution together with the fast response from the live chart viewer allowed a precise determination of relevant regions (see Figure 6.4).

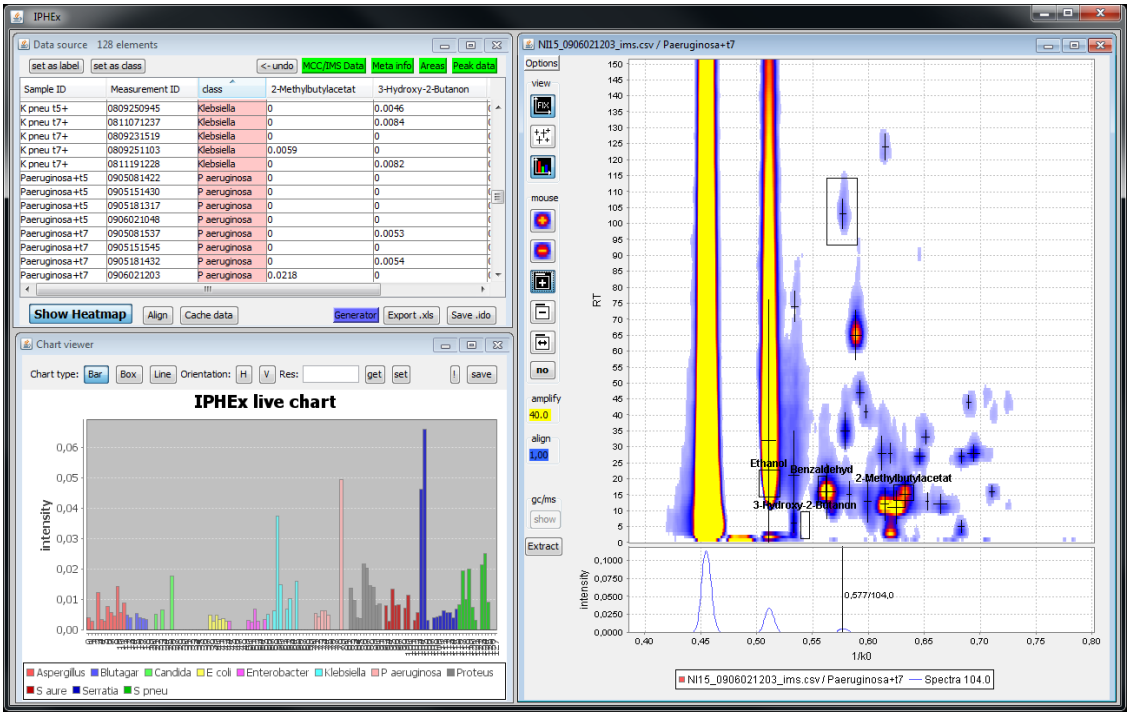


Figure 6.4.: A screenshot of the IPHEx software with a *project browser* on the top left, a *chart viewer* on the bootom left and a *heatmap explorer* on the right side. The *chart viewer* is set to *Live Chart Modus*. Moving the mouse cursor over the heatmap directly updates the values of the *chart viewer* and shows the intensity of a peaks at this position in all measurements.

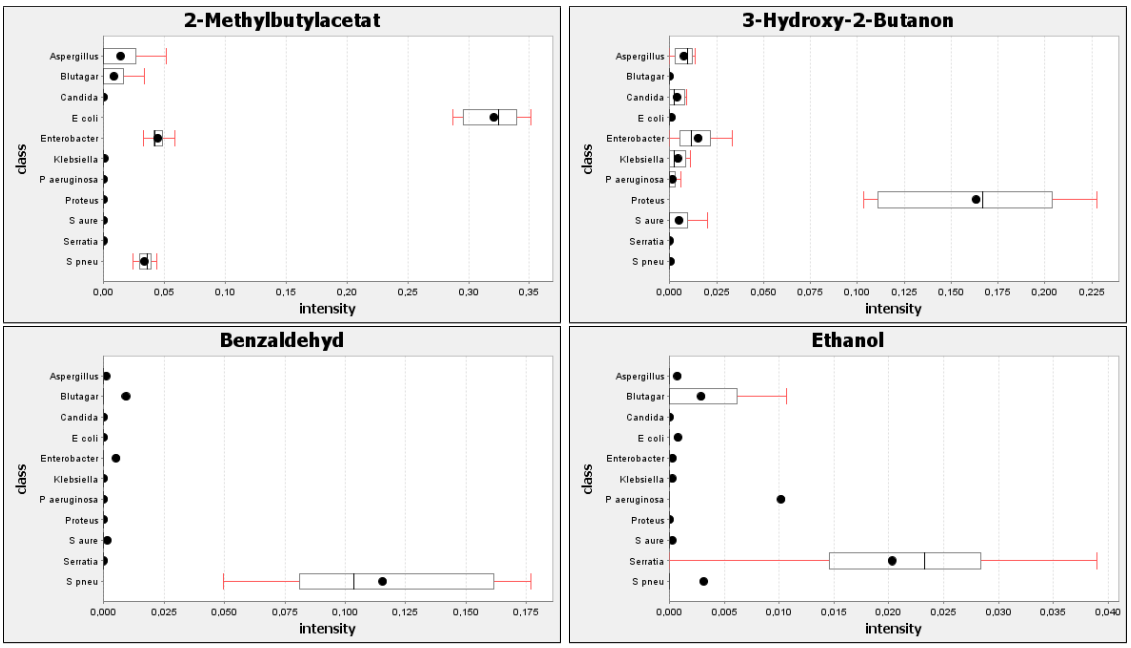


Figure 6.5.: Box-plots of four different identified substances which are related to a specific sample class.

6.2.3. Results and discussion

Several discriminating regions could be identified as a result of this study. With the help of CG/MS measurements performed in parallel, the seven substances ethanol, 3-hydroxy-2-butanon, 2-methylbutylacetat, cyclohexanon, benzaldehyd, benzeneethanol and tricyclodecan-1-amin could be assigned to these regions.

The distribution of four of the substances in the eleven different classes are visualised exemplarily in Figure 6.5 as boxplots. More substances can possibly be identified once the number of identified and described substances for IMS data grows. During the analysis of this experiment, boxplots were integrated to the IPHEx software to enable the visualisation of a growing number of measurement and different classes per project. Furthermore the direct comparison to GC/MS measurements was integrated into the software later on as a result of the experiments gathered while performing this study.

6.3. Large scale time series investigation

Ion mobility Spectrometry is used to detect volatile analytes within human breath directly. Many volatile organic compounds (VOC) show significant day-to-day variation in the signal height related to the concentration of the analyte, although the breath collection had been performed under the same conditions with respect to similar sampling procedure, similar measurement time, and measurement conditions. For the first time, a high number of MCC-IMS measurements, collected over a time period of one year were used to investigate those variations. The results of this work were published in the *International Journal for Ion Mobility Spectrometry* in 2010 [44].

6.3.1. Background and objectives

In this study the development of a non-invasive and easy method for early diagnosis or therapy monitoring should be supported by identifying substances in the breath of patients and comparing them to the ambient air. Recently, Thekedar et al. [45] reported that to verify the potential of breath analysis with respect to medical questions the variability in measurement of exhaled VOC must be known. Therefore, in the study the exhaled breath of a single healthy person in the same room environment over a time period of nearly one year was investigated and compared to previously taken room air samples using direct sampling with the same ion mobility spectrometer.

6.3.2. Analyis and visualisation

After starting the IPHEx *project browser* by selecting the measurement files of this experiment, the peak detection was used to create peak lists (for details see Section 5.4). In this case, aligning the retention times of the data was very important since the measurements were recorded over a long period of time. To find a substance for the alignment, the detection of an analyte which occurs in as many measurements as possible. This was done by starting the *heatmap explorer* and using the live chart visualisation to track peak intensities within adjustable boundaries (see 5.4.6). Using this method, Decanal was chosen and the retention time alignment was performed for all measurements and peak lists. Since only the measurement date and the type of the sample (sample/reference) were needed for this experiment and can be imported directly from the measurement files, no meta information file was needed. Information about the substances

occurring in this experiment was imported from a list of known analytes and integrated as areas in the data source browser.

To compare the variability of different analytes between exhaled air and ambient air, the column *sample type* was selected as class column using the *set class* functionality from the data source browser. Afterwards the intensity distributions of either the known imported substances or areas defined through the heatmap explorer could be visualised. Due to the high number of measurements the interpretation of the bar chart visualisations became difficult because of the high number and thus small size of displayed bars. Therefore the chart viewer was set to generate boxplots. Especially the live chart functionality using this kind of visualisations enables a fast identification of analytes which occur in breath samples in a different concentration than in the room air samples.

For the investigation of the variations over a time period under consideration of the ambient air, a second approach was chosen. The IPHEx software offers a method for direct detection and subtraction of reference measurements from the breath samples. Therefore all ambient samples have to be selected in the *project browser* and subtracted from the breath samples using the *subtract reference* option. Afterwards all displayed peak intensities are updated, show the result of the subtraction and can be used for the generation of charts. If the date and time stamp column of the measurements is chosen as a label column via the *set label* option, the axis of bar and line charts switch to the time line format.

Combining the reference subtraction procedure with the substance based boxplots enables the generation of one ultimate visualisation which combines all information in one chart.

Examples of the different resulting charts are given in the following section.

6.3.3. Results and discussion

During this analysis 30 different areas were selected and regarded in detail. A total number of 24 of these region could be assigned to known substances while 6 regions in which peaks occurred could not yet be identified.

The differences between exhaled breath and room air for the peaks formed by those substances are shown in Figure 6.6. Overall, Camphene, 3-Caren, Eucalyptol, Benzothiazole, for example show higher values in the exhaled air compared to the room air. In contrast, for 3-Hydroxy-2-Butanone, Benzaldehyd, Phenole, Octanal, 2-Ethyl-1-Hexanol, Pelargonaldehyde(Nonanal), Decamethylcyclopentasiloxane, for ex-

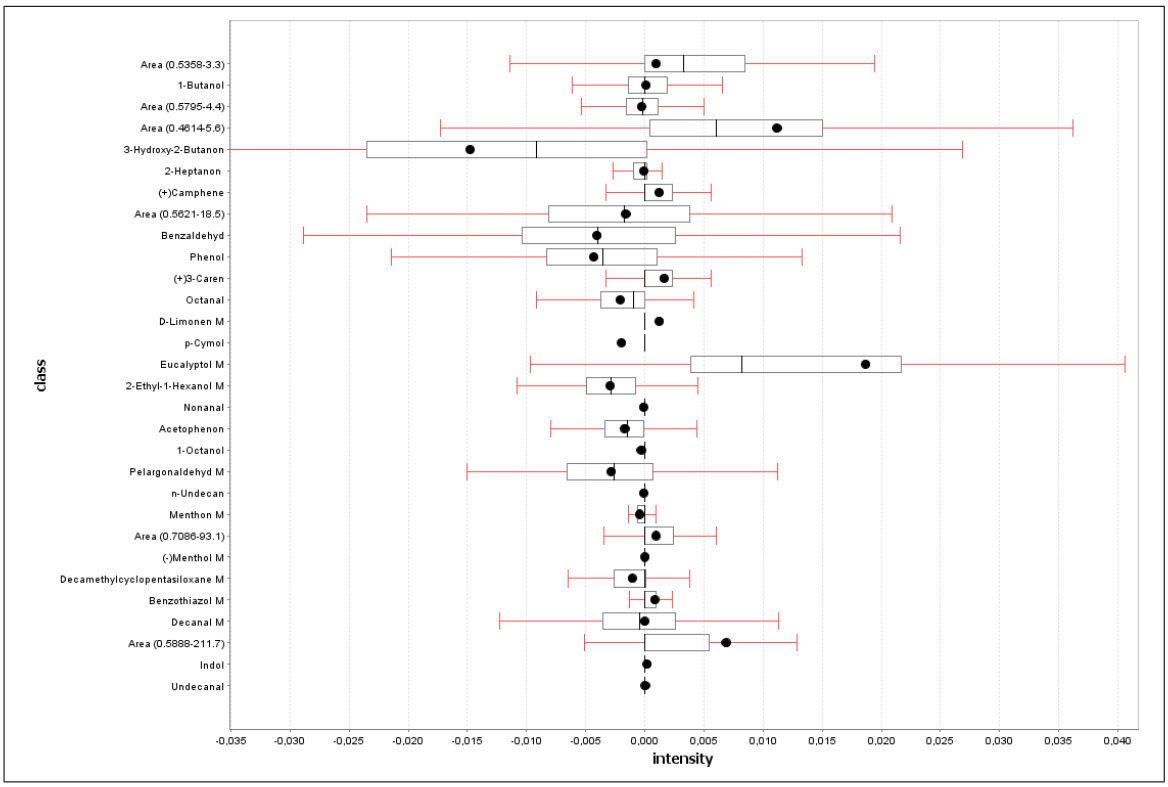


Figure 6.6.: Differences in intensity between exhaled breath and room air, visualised as box plots using 30 different areas. Positive values show that the substance is emitted by the person, in case of negative values the lung acts as a cleaning system.

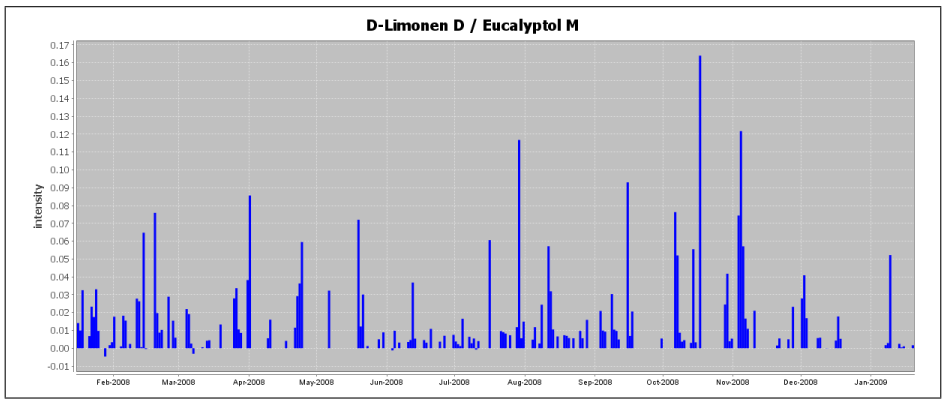


Figure 6.7.: Intensity of the peak formed by the Eucalyptol monomer and the Limonene dimer tracked over one year in the exhaled breath of a single healthy person. A room air sample was taken before each breath sample and the respective peak intensities were subtracted.

ample the concentration in the exhaled air is lower than the surrounding room air in this experimental series.

Figure 6.7 shows the signal related to Eucalyptol and the dimer of Limonene at 0.573/121.1 (both signals are unresolved in the present case) are higher in most of the cases in the exhaled air than in the room air.

In contrast 2-Ethyl-1-Hexanol was observed to be at a higher concentration in most of the days during the time period in the room air compared to the exhaled air (see Figure 6.8). In this case, the lung seems to be active as cleaning system.

Recently, Bödeker et al. [61] considered the changes between the room air concentration of specific analytes using MCC/IMS during approximately one week of contiguous investigation. The effect of the climatic ventilation system is shown and the sometimes rapid increase or decrease of single analytes was found. Such rapid changes would affect the difference between exhaled and inhaled air of the patients as well as the healthy controls, measured at different times of the day as healthy controls. The strategy for clinical trials should be considered with respect to the analytes and the changes both in inhaled and in exhaled air.

To improve the possibilities of the analysis of exhaled breath using the characterisation of volatile metabolites with respect to medical diagnosis not only the number of analytes, but also the ambient air as well as the variability of each substance should be taken into account.

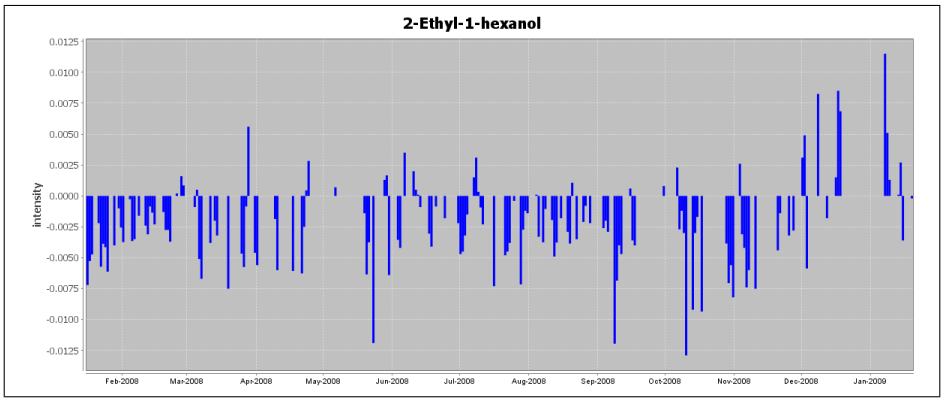


Figure 6.8.: Intensity of the peak formed by 2-Ethyl-1-Hexanol tracked over one year in the exhaled breath of a single healthy person. A room air sample was taken before each breath sample and the respective peak intensities were subtracted.

CHAPTER 7

Discussion and conclusion

IPHEx is the first software system supporting the analysis, management, and visualisation of large amounts of MCC-IMS measurements in parallel. It is currently used for the investigation of metabolomic experiments with a focus on the analysis of exhaled air, headspace samples of cell and bacteria cultures, as well as general screening of ambient air. While the main methods of IPHEx are designed to process three dimensional data obtained from different MCC-IMS devices, it also handles GC-MS based data for comparison and substance identification as well as several other information obtained from flat and Excel files. It became the standard analysis platform at the *Leibniz-Institut für Analytische Wissenschaften ISAS e.V.* for MCC-IMS data and showed its potential during the examination of experiments performed in cooperation with the *Lungenklinik Hemer - Zentrum für Pneumologie und Thoraxchirurgie*, the *University Göttingen - Department of Anesthesiology, Emergency and Intensive Care Medicine*, the *Charité - Universitätsmedizin Berlin* and several others. It is also used for experimental purposes at the *Korea Institute of Science and Technology, Saarbrücken* and the *B&S Analytik GmbH, Dortmund*.

The application to many different experiments and tasks demonstrates that the requirements have successfully been addressed and the software and therefore the underlying methods and concepts are suited to analyse large amounts of IMS data in an efficient way.

The definition and design of the four data concepts which are consequently used by the different parts of the software system enables a structured and reproducible analysis of projects. The use of peak-

lists as an intermediate layer between the original spectra data and the area annotations make repetitive compute-intensive quantification steps obsolete during the project investigation. Together with the storage of pre-processed image data to rapidly visualise the measurement files as heatmap images, the design of new interactive analysis features became possible. Especially the live chart viewer is frequently used to explore and analyse MCC-IMS data in an unprecedented short amount of time. The meta information data concept enables a flexible integration of additional information which is available or necessary for further analysis. This information typically consisting of categories and parameters can be used to partition the measurements into different classes, compare parameters to quantified peak, or label analysis results such as charts. Meta information entries that contain numeric values are treated equally to quantified peak values in the analysis framework. This allows the integration and even the exclusive usage of those values for correlation, PCA and all other available analyses. The possibility to assign and subtract reference measurements resulting from previously taken ambient air or medium headspace samples, enables direct comparison between the levels of the quantified peaks. Using this feature, all subsequent analyses are based on the difference between sample and reference. Negative peak intensities are possible in this case to indicate a dominant occurrence of a peak in the reference measurement.

The *project browser* allows a structured storage, organisation and filtering of all data belonging to one experiment and all associated analysis results. A project can be saved at any stage during the analysis process and it can be extended with further measurements, annotations and meta informations. Different projects can be opened in parallel to compare them or transfer data from one to another.

To enable a detailed investigation of single measurements a lot of effort was put in the development of a dynamically exploreable and adjustable visual representation of the spectra data. The developed methods to achieve this were combined to the *heatmap explorer*. On the visualisation side it provides zoom and pan functionality for the retention time and $1/k_0$ axis to investigate different areas of the heatmap in detail. The intensity values are the third dimension of the data which is encoded by colour and can be investigated by either altering the colour space or visualising the single spectra at the bottom of the heatmap plot. Beside the visualisation task the *heatmap explorer* is also necessary to set up, validate, and edit area and peak annotations.

The software was developed to handle, analyse, and visualise a large amount of measurements under the awareness that the number of measurements per project will most likely grow further in the near future. The relatively low cost of the device combined with low time

needed to record a sample and general flexibility promote the usage as well as the need to gather as many data as possible to consider biological viability and enable the use for medical diagnostics. During this work the *Large scale time series investigation* project (in Chapter 6.3) with about 500 measurements was analysed without any noticeable loss of performance. For testing issues projects with data from up to 2000 different measurements were created and visualised without any problems. The pre-processing and peak detection methods are done within five to twenty seconds per measurement, highly depending on size and quality of the data as well as the available computational power, but these are required only once per measurement. While the fast handling techniques are convenient for projects with several hundreds of measurements they are absolutely required when dealing with projects that contain several thousands of measurements. Obviously visualisations of intensities in such projects are displayable only as box plots, correlation plots, PCAs, or using date axis, since single intensity visualisations like bar charts exceed the available space on a standard monitor.

During this work, many different methods to filter, process, analyse and visualise MCC-IMS data were developed. Only a carefully chosen subset of them is available in the currently released IPHEx software. This was done to reduce the number of parameters, options, and buttons which can or need to be altered by the user and thus guarantees that every analysis of a project is reproducible and leads to the same results. The only parameter which has a high impact on the analysis results is the chosen substance for the retention time alignment. At the moment, the best practice to retain the reproducibility regarding this is to include the substance which was chosen for alignment into the *project browser* and keep it there as a reference.

With IPHEx, for the first time a complete analysis environment exists, which offers a solution for all analysis, management, and visualisation tasks which are necessary to perform a comprehensive investigation of large amounts of MCC-IMS data.

CHAPTER 8

Outlook

The creation of the IPHEx platform enables a structured, efficient and large-scale analysis of MCC-IMS metabolomic experiments and closes the gap between raw MCC-IMS data and analysis results. Different extensions to the system are conceivable for the future which found on the methods and software presented here.

IPHEx is designed as a stand alone software and thus has to be installed locally on a computer. Due to the continuously increasing data throughput rates of the internet, distributed approaches become feasible. One possible development could be a central storage of MCC-IMS measurements on a server infrastructure and the usage of IPHEx as a client to access them. Furthermore it can be combined with a user management system and a database to store analysis results and offer a platform to discuss and analyse projects with several cooperation partners in parallel. Furthermore information about identified substances or dependencies between specific classes of measurements and peaks could be provided and accessed in a global way. Even a completely web based analysis of MCC-IMS may become possible. First steps towards this approach were done in cooperation with the developers of the BioImax analysis platform [46].

Currently the measurements are recorded and analysed with different software systems and typically on different computers. An extension to the existing system could be to integrate the recording process into the software as well. The most interesting benefit could be an immediate response from the system if previously defined substances or combination of substances occur, further reducing the time needed

for a diagnosis. Also the quality of the data could be controlled during the recording process to eventually repeat a measurement in the case of malfunctions or contaminations. Following this thought, the analysis software could even adjust parameter during the recording process to optimise the analysis results or demand a second sampling process with altered parameters if a special substance or combination of substances occurred.

Danksagung

An dieser Stelle möchte ich mich herzlich bei allen Menschen bedanken, die mich im Laufe meiner Promotion unterstützt haben.

Ein besonderer Dank gilt:

- Prof. Dr. Jens Stoye für die freundliche und ausgezeichnete Betreuung und die Möglichkeit meine Dissertation anfertigen zu können.
- Dem *Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.* für die Finanzierung meiner Promotion und der gesamten Arbeitsgruppe Metabolomics für eine wunderschöne gemeinsame Arbeitszeit.
- PD. Dr. Jörg Ingo Baumbach für die Bereitstellung des Projektes und seinen intensiven Einsatz ohne den die Dissertation nicht möglich gewesen wäre.
- Der Arbeitsgruppe Genome Informatics für ein angenehmes Arbeitsklima, die tolle Zusammenarbeit, die netten Gespräche, die schönen Arbeitsgruppenausflüge und die vielen Kuchen.
- Luzia Seifert für die grenzenlose Hilfestellung bei allen chemischen, physikalischen und psychologischen Problemen.
- Christian, Jochen, Sebastian, Kolja und Sascha für Diskussionen, Ratschläge und Korrekturen.
- Und natürlich meinen Eltern, sowie meiner Frau Ariane und meinem kleinen Sohn Jonas.

APPENDIX A

APPENDIX

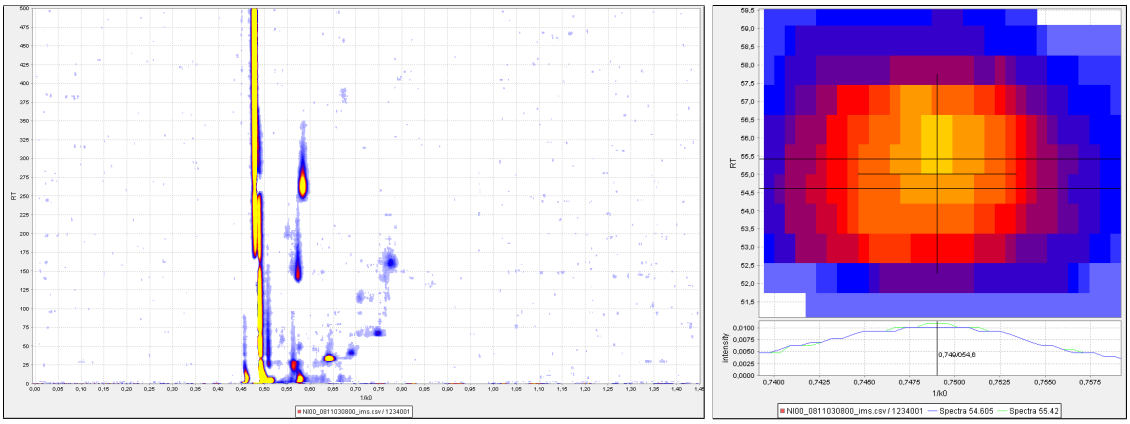


Figure A.1.: Two heatmap images of a measurement with compensated RIP and an median filter applied only. The left part shows a complete measurement while the right part shows a detailed zoomed view of one single peak.

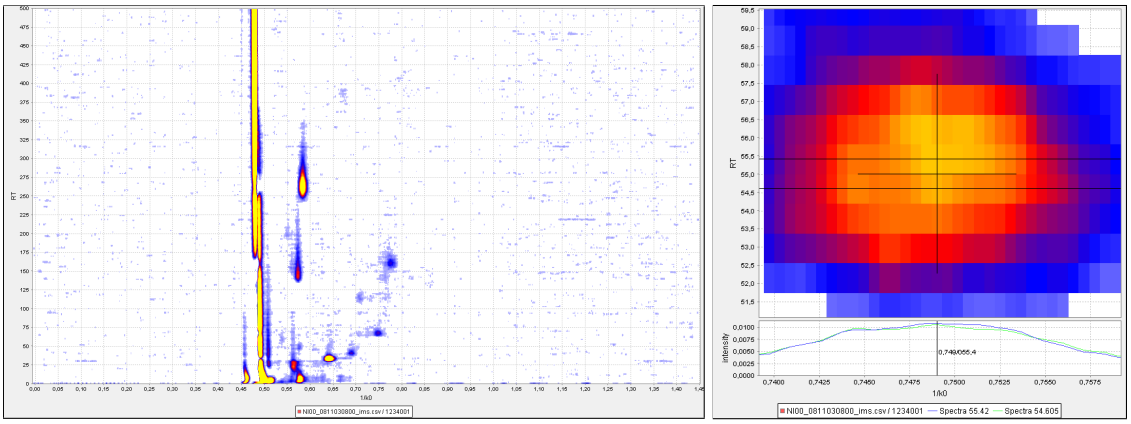


Figure A.2.: Two heatmap images of a measurement with compensated RIP and an gauss filter applied only. The left part shows a complete measurement while the right part shows a detailed zoomed view of one single peak.

List of Figures

2.1. IMS with closed ion shutter and a probe in the drift region. The sample enters the device trough the gas inlet on the left side and moves towards the faraday plate at the right side. 14
2.2. Complete experimental setup of an IMS coupled to a multi-capillary chromatographic column and a sample loop to allow the recording of headspace and breath sample under controlled conditions. 15
2.3. One IMS spectrum, averaged over 10 single scans. The x-axis shows $1/k_0$, while the y-axis is ued to display the signal intensity. 17
2.4. Screenshot of the BBImsAnalyse software user interface. A heatmap in the centre visualises one measurement, the bottom and right parts show representation of single spectra which were selected. 19
2.5. Peak comparison of a marked area in 20 different samples of human breath ordered by signal intensity from the upper left (highest value) to lower right (no relevant signal intensity) [36]. 20

3.1. Example of a heatmap created from a MCC-IMS measurement with labeled peaks and annotations. 25
3.2. Two possible visualisations of a series of thirteen substance values taken from different measurements. 26
3.3. A bar chart visualisation of a series of thirteen substance values, using two different colours to distinguish assigned classes 27

3.4. A box plot to visualise the intensitiy distribution of one substance in four different classes. 27

4.1. Heatmap with detail view of a selected single spectrum. The heatmap consists of 500 single spectra which are pseudo coloured using a heatmap paint scale. 32

4.2. Influence of the RIP compensation filter on an IMS heatmap. Left: Raw measurement; Right: Compensated RIP 34

4.3. Two heatmap images of a measurement with compensated RIP *without* applying any filter methods. The left part shows a complete measurement while the right part shows a detailed zoomed view of one single peak. 35

4.4. Two heatmap images of a measurement with compensated RIP after applying the filter pipeline. The left part shows a complete measurement while the right part shows a detailed zoomed view of one single peak. 37

4.5. The left part shows the peak intensities of 4-Heptanone in two different classes. A relation between classes and peak intensity seems probable. The right part shows the distribution of the peak positions. While most of the peaks are very close to the proposed position of 4-Heptanone in the centre, some differ within defined tolerance boundaries 40

4.6. Peak intensities of 4-Heptanone displayed as bars and labeld with previously known classes A and B. The intensity is shown on the left axis. The separation accuracy graph shows the percentaged number of correctly assigned classes when assigning all measurements containing peaks with an intensity equal to or higher than this value to class A and all others to class B. The right axis shows this percentage. 41

4.7. Automatic centering effect of the scoring procedure. The black crosses mark the positions of the peaks. The purple cross is the center used to form the subset shown in the left diagram, the green cross is the center for the subset shown in the right diagram. The maximal achieved score is written in each diagram, showing that the subset formed by the peak in case 2 is a better separation criterion than the subset formed by the peak in case 1. . . 42

4.8. Colour space used to generate the heatmap visualisation. 43

4.9. Visualisation of the intensity distribution of one single peak. The bar chart and the boxplot on the left show the distribution of the selected analyte on the right in 4 different classes. The intensity of the selected analyte is higher in classes B and D than in classes A and C. The lower part of the right picture shows one selected spectrum, marked with a black line in the heatmap. 44

5.1. Schema of the different data types and processing steps. 48

5.2. Screenshot of the basic IPHEx user interface 50

5.3. Screenshot of the IPHEx data source browser table. Each row represents one measurements while the columns contain related information retrieved from different sources. . 52

5.4. Screenshot of the complete IPHEx data source browser. Several options and buttons above the table allow the organisation and editing of the containing data types. Buttons below the table provide functionality for exploration, analysing and exporting the data. 53

5.5. The IPHEx heatmap explorer shows heatmaps of measurements selected through the data source browser. It provides tools for editing peaklists and generation of areas as well as general functionality to zoom, pan and explore the measurement. 54

5.6. Visualisation layer of the IPHEx *heatmap explorer*. 55

5.7. The chart viewer of the IPHEx software, opened with a time depended bar chart and three further visualisations (box plot, class based bar chart, and line plot) which can be selected using the buttons in the top left corner. 56

5.8. Correlation plot of the intensity of 16 known substances determined in 158 measurements. 58

5.9. Visualisation of a principal component analysis using a set of substances from a project with previously labelled classes. 59

5.10 A parallel plot of an MCC/IMS and a GC/MS measurement. The total ion count plotted against the gas chromatographic retention time is shown at the left. The axis is transformed to enable a direct comparison to the MCC/IMS Heatmap at the right 60

5.11 Spot table visualisation of five different areas in five measurements. 61

5.12. Two different examples for polar glyphs. The left picture shows polar glyphs on 4 selected areas, each displaying the intensities of 113 measurements. In this case, only positive values occur. The right picture shows an experiment consisting of 448 measurements where 224 measurements were samples of ambient air and subtracted from the breath samples. In this case also negative values occur. It is easy to see that in this experiment Eucalyptol has mostly positive values and thus is part of the breath sample while 2-Ethyl-1-Hexanol is mostly caused by the ambient air and therefore has negative values. The lower parts of both images show selected single spectra. 62

5.13. Screenshot of the *alignment parameter window*. The second column of the table shows the substance which was chosen for the alignment, the third column shows the parameters of the target area which is used to detect peaks for the alignment. The parameters can be altered by the buttons on the right side, and are directly visualised in the *heatmap explorer* to aid the parameter setup process. . . 65

6.1. Heatmap image of one measurement, obtained from a sarcoidosis patient. Filtering, alignment and baseline correction methods were applied, peaks were detected and marked with a cross by an automatic procedure. 71

6.2. One possible marker region for sarcoidosis which was determined by a scoring procedure. The heatmap on the left shows the position of the marker, bar and box plots on the right show the intensity distributions in the sarcoidois (red) and control (blue) groups. 72

6.3. A visualisation of the marker region for sarcoidosis in all measurements. The left part contains measurements from sarcoidosis, the right part shows measurements which were used as control group. The first column shows a substance which was used to validate the alignments, column two shows the marker region, and column three a zoomed out view of the marker region. 74

6.4. A screenshot of the IPHEx software with a *project browser* on the top left, a *chart viewer* on the bootom left and a *heatmap explorer* on the right side. The *chart viewer* is set to *Live Chart Modus*. Moving the mouse cursor over the heatmap directly updates the values of the *chart viewer* and shows the intensity of a peaks at this position in all measurements. 76

6.5. Box-plots of four different identified substances which are related to a specific sample class. 77
6.6. Differences in intensity between exhaled breath and room air, visualised as box plots using 30 different areas. Positive values show that the substance is emitted by the person, in case of negative values the lung acts as a cleaning system. 80
6.7. Intensity of the peak formed by the Eucalyptol monomer and the Limonene dimer tracked over one year in the exhaled breath of a single healthy person. A room air sample was taken before each breath sample and the respective peak intensities were subtracted. 81
6.8. Intensity of the peak formed by 2-Ethyl-1-Hexanol tracked over one year in the exhaled breath of a single healthy person. A room air sample was taken before each breath sample and the respective peak intensities were subtracted. 82

A.1. Two heatmap images of a measurement with compensated RIP and an median filter applied only. The left part shows a complete measurement while the right part shows a detailed zoomed view of one single peak. 91
A.2. Two heatmap images of a measurement with compensated RIP and an gauss filter applied only. The left part shows a complete measurement while the right part shows a detailed zoomed view of one single peak. 92

List of Tables

3.1. One simple table structure which includes basic information to analyse a project. 26

5.1. Interrelation of the basic IPHEx compounds and the drag & drop listeners. A selection of a table is represented by strings, where columns are delimited by tabulators and lines are delimited by newline characters to enable a flexible transport of information through the drag & drop interface. 51

Bibliography

[1] G. A. Eiceman and Z. Karpas, *Ion mobility spectrometry*. CRC Press, 2005.

[2] M.J.Cohen and F. Karasek, “Plasma chromatography - a new dimension for gas chromatography and mass spectrometry,” *J. Chrom. Sci.*, vol. 8, no. 6, p. 330, 1970.

[3] M. A. Baim and H. H. Hill, “Tunable selective detection for capillary gas chromatography by ion mobility monitoring,” *Analytical Chemistry*, vol. 54, pp. 38–43, Jan. 1982.

[4] M. Phillips, J. Herrera, S. Krishnan, M. Zain, J. Greenberg, and R. N. Cataneo, “Variation in volatile organic compounds in the breath of normal humans,” *Journal of Chromatography. B, Biomedical Sciences and Applications*, vol. 729, p. 75–88, June 1999. PMID: 10410929.

[5] J. Baumbach, A. Davies, P. Lampen, and H. Schmidt, “Jcamp-dx. a standard format for the exchange of ion mobility spectrometry data,” *Pure Appl. Chem*, vol. 73, no. 11, pp. 1765–1782, 2001.

[6] W. Vautz, B. Bödeker, S. Bader, and J. Baumbach, “Recommendation of a standard format for data sets from gc/ims with sensor-controlled sampling,” *International Journal for Ion Mobility Spectrometry*, vol. 11, no. 1, pp. 71–76, 2008.

[7] W. Vautz, B. Bödeker, J. Baumbach, S. Bader, M. Westhoff, and T. Perl, “An implementable approach to obtain reproducible reduced ion mobility,” *International Journal for Ion Mobility Spectrometry*, vol. 12, no. 2, pp. 47–57, 2009.

[8] L. W. Sumner, P. Mendes, and R. A. Dixon, “Plant metabolomics: large-scale phytochemistry in the functional genomics era,” *Phytochemistry*, vol. 62, p. 817–836, Mar. 2003.

[9] J. Kopka, “Current challenges and developments in GC-MS based metabolite profiling technology,” *Journal of Biotechnology*, vol. 124, p. 312–322, June 2006.

[10] C. D. Broeckling, D. V. Huhman, M. A. Farag, J. T. Smith, G. D. May, P. Mendes, R. A. Dixon, and L. W. Sumner, “Metabolic profiling of medicago truncatula cell cultures reveals the effects of biotic and abiotic elicitors on metabolism,” *Journal of Experimental Botany*, vol. 56, p. 323–336, Jan. 2005. PMID: 15596476.

[11] A. Amann, P. Spanel, and D. Smith, “Breath analysis: the approach towards clinical applications.,” *Mini Rev Med Chem*, vol. 7, p. 115–129, Feb. 2007.

[12] M. Phillips, R. N. Cataneo, A. R. C. Cummin, A. J. Gagliardi, K. Gleeson, J. Greenberg, R. A. Maxfield, and W. N. Rom, “Detection of lung cancer with volatile markers in the breath.,” *Chest*, vol. 123, p. 2115–2123, June 2003.

[13] J. D. Pleil and A. B. Lindstrom, “Measurement of volatile organic compounds in exhaled breath as collected in evacuated electropolished canisters.,” *J Chromatogr B Biomed Appl*, vol. 665, p. 271–279, Mar. 1995.

[14] M. Mieth, J. K. Schubert, T. Groger, B. Sabel, S. Kischkel, P. Fuchs, D. Hein, R. Zimmermann, and W. Miekisch, “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*, vol. 82, no. 6, p. 2541–2551, 2010.

[15] J. Beauchamp, F. Kirsch, and A. Buettner, “Real-time breath gas analysis for pharmacokinetics: monitoring exhaled breath by online proton-transfer-reaction mass spectrometry after ingestion of eucalyptol-containing capsules,” *Journal of Breath Research*, vol. 4, no. 2, p. 026006, 2010.

[16] C. Warneke, J. Kuczynski, A. Hansel, A. Jordan, W. Vogel, and W. Lindinger, “Proton transfer reaction mass spectrometry (PTR-MS): propanol in human breath,” *International Journal of Mass Spectrometry and Ion Processes*, vol. 154, p. 61–70, May 1996.

[17] D. Smith, P. Španěl, B. Enderby, W. Lenney, C. Turner, and S. J. Davies, "Isoprene levels in the exhaled breath of 200 healthy pupils within the age range 7–18 years studied using SIFT-MS," *Journal of Breath Research*, vol. 4, no. 1, p. 017101, 2010.

[18] M. J. Seeley, W. Hu, J. M. Scotter, M. K. Storer, and G. M. Shaw, "In vitro SIFT-MS validation of a breath fractionating device using a model VOC and ventilation system," *Journal of Breath Research*, vol. 3, no. 1, p. 016001, 2009.

[19] P. Silkoff, "History, technical and regulatory aspects of exhaled nitric oxide," *Journal of Breath Research*, vol. 2, no. 3, p. 037001, 2008.

[20] A. Gelperin and A. T. C. Johnson, "Nanotube-based sensor arrays for clinical breath analysis," *Journal of Breath Research*, vol. 2, no. 3, p. 037015, 2008.

[21] B. P. J. d. L. Costello, R. J. Ewen, and N. M. Ratcliffe, "A sensor system for monitoring the simple gases hydrogen, carbon monoxide, hydrogen sulfide, ammonia and ethanol in exhaled breath," *Journal of Breath Research*, vol. 2, no. 3, p. 037011, 2008.

[22] P. J. Mazzone, J. Hammel, R. Dweik, J. Na, C. Czich, D. Laskowski, and T. Mekhail, "Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array," *Thorax*, vol. 62, p. 565 –568, July 2007.

[23] K. Toda, J. Li, and P. K. Dasgupta, "Measurement of ammonia in human breath with a Liquid-Film conductivity sensor," *Analytical Chemistry*, vol. 78, p. 7284–7291, Oct. 2006.

[24] I. Horváth, Z. Lázár, N. Gyulai, M. Kollai, and G. Losonczy, "Exhaled biomarkers in lung cancer," *European Respiratory Journal*, vol. 34, p. 261 –275, July 2009.

[25] S. Dragonieri, J. T. Annema, R. Schot, M. P. v. d. Schee, A. Spanevello, P. Carratú, O. Resta, K. F. Rabe, and P. J. Sterk, "An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD," *Lung Cancer*, vol. 64, p. 166–170, May 2009.

[26] J. I. Baumbach, D. Berger, J. W. Leonhardt, and D. Klockow, "Ion mobility sensor in environmental analytical chemistry—concept and first results," *International Journal of Environmental Analytical Chemistry*, vol. 52, no. 1, p. 189–193, 1993.

[27] V. Ruzsanyi, S. Sielemann, and J. I. Baumbach, “Determination of vocs in human breath using ims,” *Int. J. Ion Mobility Spectrom.*, vol. 5, p. 45–48, 2002.

[28] F. Li, Z. Xie, H. Schmidt, S. Sielemann, and J. I. Baumbach, “Ion mobility spectrometer for online monitoring of trace compounds,” *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 57, p. 1563–1574, Oct. 2002.

[29] J. I. Baumbach, “Process analysis using ion mobility spectrometry,” *Analytical and Bioanalytical Chemistry*, vol. 384, no. 5, p. 1059–1070, 2005.

[30] W. Vautz, J. I. Baumbach, and E. Uhde, “Detection of emissions from surfaces using ion mobility spectrometry,” *Analytical and Bioanalytical Chemistry*, vol. 384, no. 4, p. 980–986, 2006.

[31] W. Vautz and J. I. Baumbach, “Exemplar application of multicapillary column ion mobility spectrometry for biological and medical purpose,” *International Journal for Ion Mobility Spectrometry*, vol. 11, no. 1-4, p. 35–41, 2008.

[32] S. Bader, W. Urfer, and J. I. Baumbach, “Reduction of ion mobility spectrometry data by clustering characteristic peak structures,” *Journal of Chemometrics*, vol. 20, pp. 128–135, Mar. 2006.

[33] S. Bader, W. Urfer, and J. I. Baumbach, “Preprocessing of ion mobility spectra by lognormal detailing and wavelet transform,” *International Journal for Ion Mobility Spectrometry*, vol. 11, pp. 43–49, June 2008.

[34] B. Bödeker, *Entwicklung eines Verfahrens zur Klassifikation von Ionenmobilitätsspektrometerdaten*. Algorithm Engineering report, Univ., Algorithm Engineering, 2007.

[35] B. Bödeker, W. Vautz, and J. I. Baumbach, “Visualisation of MCC/IMS-data,” *International Journal for Ion Mobility Spectrometry*, vol. 11, pp. 77–81, Nov. 2008.

[36] B. Bödeker, W. Vautz, and J. I. Baumbach, “Peak comparison in MCC/IMS-data—searching for potential biomarkers in human breath data,” *International Journal for Ion Mobility Spectrometry*, vol. 11, pp. 89–93, Nov. 2008.

[37] M. Westhoff, P. Litterst, S. Maddula, B. Bödeker, S. Rahmann, A. Davies, and J. Baumbach, “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*, vol. 13, pp. 131–139, 2010. 10.1007/s12127-010-0049-2.

[38] V. Bessa, K. Darwiche, H. Teschler, U. Sommerwerck, T. Rabis, J. Baumbach, and L. Freitag, "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*, vol. 14, pp. 7–13, 2011. 10.1007/s12127-011-0060-2.

[39] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of the 1996 IEEE Symposium on Visual Languages*, (Washington, DC, USA), p. 336–, IEEE Computer Society, 1996.

[40] S. Wegner, A. Sahlström, K. P. Pleissner, H. Oswald, and E. Fleck, "Eine hierarchische wasserscheidentransformation für die spotdetektion in 2D Gelelektrophorese-Bildern," *Mechanik elastischer Körper und Strukturen*, p. 134, 2002.

[41] M. Jünger, B. Bödeker, and J. I. Baumbach, "Peak assignment in multi-capillary column–ion mobility spectrometry using comparative studies with gas chromatography–mass spectrometry for VOC analysis," *Analytical and Bioanalytical Chemistry*, vol. 396, no. 1, p. 471–482, 2009.

[42] A. Bunkowski, B. Bödeker, S. Bader, M. Westhoff, P. Litterst, and J. I. Baumbach, "MCC/IMS signals in human breath related to sarcoidosis—results of a feasibility study using an automated peak finding procedure," *Journal of Breath Research*, vol. 3, p. 046001, Dec. 2009.

[43] A. Bunkowski, B. Bödeker, S. Bader, M. Westhoff, P. Litterst, and J. Baumbach, "Signals in human breath related to sarcoidosis.results of a feasibility study using mcc/ims," *International Journal for Ion Mobility Spectrometry*, vol. 12, no. 2, pp. 73–79, 2009.

[44] A. Bunkowski, S. Maddula, A. N. Davies, M. Westhoff, P. Litterst, B. Bödeker, and J. I. Baumbach, "One-year time series of investigations of analytes within human breath using ion mobility spectrometry," *International Journal for Ion Mobility Spectrometry*, 2010.

[45] B. Thekedar, W. Szymczak, V. Höllriegl, C. Hoeschen, and U. Oeh, "Investigations on the variability of breath gas sampling using

PTR-MS," *Journal of Breath Research*, vol. 3, no. 2, p. 027007, 2009.

[46] C. Loyek, A. Bunkowski, W. Vautz, and T. W. Nattkemper, "Web2. 0 paves new ways for collaborative and exploratory analysis of chemical compounds in spectrometry data.," *Journal of integrative bioinformatics*, vol. 8, no. 2, p. 158, 2011.