

# Signal processing for GC-MS measurements for biomarker identification

**Citation for published version (APA):**

D' Angelo, M., & Technische Universiteit Eindhoven (TUE). Stan Ackermans Instituut. Design and Technology of Instrumentation (DTI) (2011). *Signal processing for GC-MS measurements for biomarker identification*. [EngD Thesis]. Technische Universiteit Eindhoven.

**Document status and date:**

Published: 01/01/2011

**Document Version:**

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

**Please check the document version of this publication:**

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

**Take down policy**

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

providing details and we will investigate your claim.

SAI DTI - PHILIPS RESEARCH

---

# Signal processing for GC-MS measurements for biomarker identification

---

*Author:*  
Ir. Marina D'Angelo

*Supervisors:*  
Tamara Nijssen  
Anton Vink  
Arthur de Jong



**PHILIPS**

# Executive Summary

Breath analysis is a technique that is gaining importance in both industry and academia. Potentially, it is a non-invasive technique that will allow screening, diagnosing and monitoring of patients.

Many studies have been performed in an attempt to make a distinction between healthy and sick patients by only studying their breath. It was proven successful for detecting lung and breast cancer, for identifying transplant rejection and for diagnosing liver disease among others.

Ideally, it is Philips goal to develop a device that can take breath, process it and classify the patient as healthy or sick. Initially, this would be done for respiratory diseases, including asthma and sepsis with respiratory complications.

However, processing breath is not simple. There is a spectrum of possible devices for analysis. Electronic noses are a great bedside alternative, while gas chromatography is ideal for research studies where the nature of the biomarkers should be found.

Philips is involved in several studies within the next couple of years, for asthma and sepsis among others, and will process the samples with gas chromatography-mass spectrometry (GC-MS).

My objective was to provide a reliable software workflow for the analysis of the very complex GC-MS data, which could identify the molecules present in them and provide a reliable list of possible biomarkers as an output. This list would in the future be used to train classifiers for the mentioned diseases.

The result of this project was a complete processing workflow, beginning with the use of a third party peak extraction software, followed by the customized design of a filtering and alignment solution.

This combination provides a highly sensitive compound detection algorithm, a reliable peak quality filter and an accurate solution for comparison of multiple samples. Results are provided in a flexible manner for comprising a variety of classifier design possibilities.

This solution can greatly contribute to the analysis GC-MS data for biomarker identification.

# Contents

<b>1</b>	<b>Introduction</b>	<b>9</b>
1.1	Breath Analysis . . . . .	9
1.1.1	Breath Markers . . . . .	9
1.1.2	Analysis Techniques . . . . .	10
1.2	Gas Chromatography-Mass Spectrometry (GC-MS) . . . . .	10
1.2.1	Main Structure . . . . .	11
1.2.2	Data . . . . .	11
<b>2</b>	<b>Results</b>	<b>13</b>
2.1	Signal Processing . . . . .	13
2.2	Commercial Software Alternatives . . . . .	14
2.2.1	AMDIS vs. MassHunter . . . . .	14
2.2.2	Mass Profiler Professional, by Agilent . . . . .	15
2.3	Matlab Tool . . . . .	15
2.3.1	Quality Score Filtering . . . . .	15
2.3.2	Alignment . . . . .	17
2.3.3	Output . . . . .	17
2.4	Processing Examples . . . . .	17
2.4.1	Pilot Experiments . . . . .	17
2.4.2	Analysis of asthma data . . . . .	20
<b>3</b>	<b>Conclusions</b>	<b>22</b>
<b>A</b>	<b>Introduction to Breath Analysis</b>	<b>23</b>
A.1	Introduction . . . . .	23
A.2	Breath Markers . . . . .	23
A.3	Sample Collection . . . . .	23
A.4	Analysis Methods . . . . .	24
A.4.1	Electronic Nose . . . . .	25
A.4.2	Gas chromatography-Mass Spectrometry . . . . .	25
A.5	Conclusions . . . . .	26
<b>B</b>	<b>Summary of GC-MS techniques</b>	<b>29</b>
B.1	Introduction . . . . .	29
B.2	Main Structure . . . . .	29
B.2.1	Inlet . . . . .	30
B.2.2	Column . . . . .	30
B.2.3	Oven . . . . .	31
B.2.4	Detector (Mass Spectrometer) . . . . .	31
B.3	Data . . . . .	32

B.4	Used Setup . . . . .	32
<b>C</b>	<b>GC-MS data processing</b>	<b>35</b>
C.1	Introduction . . . . .	35
C.2	Processing Workflow . . . . .	35
C.2.1	Raw Data Collection . . . . .	36
C.2.2	Peak Extraction . . . . .	36
C.2.3	Alignment . . . . .	36
C.2.4	Filtering of Exogenous Compounds . . . . .	37
C.2.5	Statistical Analysis . . . . .	38
C.3	Conclusions . . . . .	38
<b>D</b>	<b>Analysis of AMDIS, MassHunter and Mass Profiler</b>	<b>39</b>
D.1	Introduction . . . . .	39
D.2	Procedure . . . . .	39
D.3	Results . . . . .	40
D.3.1	AMDIS . . . . .	40
D.3.2	MassHunter . . . . .	42
D.3.3	Mass Profiler Professional . . . . .	45
D.3.4	WEKA . . . . .	49
D.4	Conclusions . . . . .	50
<b>E</b>	<b>Analysis of GC-MS repeatability on simple experiment</b>	<b>52</b>
E.1	Introduction . . . . .	52
E.2	Results . . . . .	52
E.3	Conclusions . . . . .	57
<b>F</b>	<b>Analysis of pilot study data</b>	<b>58</b>
F.1	Introduction . . . . .	58
F.2	Methods . . . . .	58
F.3	Results . . . . .	59
F.3.1	Dry Nitrogen . . . . .	59
F.3.2	Wet Nitrogen . . . . .	61
F.3.3	Dry VOCs . . . . .	63
F.3.4	Wet VOCs . . . . .	66
F.3.5	Breath . . . . .	69
F.3.6	Stored Nitrogen . . . . .	71
F.3.7	Comparison of dry and wet nitrogen experiments . . . . .	73
F.3.8	Comparison of dry and wet VOCs experiments . . . . .	73
F.4	Air quality at the ICU of Amsterdam AMC hospital . . . . .	75
F.5	Conclusions . . . . .	78
<b>G</b>	<b>Analysis of GC-MS repeatability on pilot study data</b>	<b>79</b>
G.1	Introduction . . . . .	79
G.2	Method . . . . .	79
G.3	Results . . . . .	79
G.3.1	Dry Nitrogen . . . . .	80
G.3.2	Wet Nitrogen . . . . .	81
G.3.3	Dry VOCs . . . . .	81
G.3.4	Wet VOCs . . . . .	82
G.3.5	Breath . . . . .	83

G.3.6	Stored Nitrogen . . . . .	86
G.4	Conclusions . . . . .	87
<b>H</b>	<b>Analysis of stability with time</b>	<b>88</b>
H.1	Introduction . . . . .	88
H.2	Analysis . . . . .	88
H.3	Results . . . . .	88
H.3.1	One-Way Repeated Measures ANOVA . . . . .	89
H.4	Conclusions . . . . .	89
H.5	Figure appendix . . . . .	91
<b>I</b>	<b>Software tool for GC-MS data processing</b>	<b>98</b>
I.1	Introduction . . . . .	98
I.2	Design . . . . .	98
I.2.1	AMDIS Files . . . . .	98
I.2.2	Quality Filtering . . . . .	99
I.2.3	Alignment . . . . .	101
I.2.4	Interface . . . . .	105
I.2.5	Results . . . . .	106
I.3	Conclusions . . . . .	107
<b>J</b>	<b>User Guide</b>	<b>108</b>
J.1	Introduction . . . . .	108
J.2	AMDIS Processing . . . . .	108
J.2.1	Description . . . . .	108
J.2.2	Loading Files . . . . .	108
J.2.3	Configuration . . . . .	108
J.2.4	Batch Processing . . . . .	112
J.2.5	Export . . . . .	112
J.3	Matlab Tool . . . . .	113
J.3.1	Loading Files . . . . .	113
J.3.2	Quality Score Filtering . . . . .	114
J.3.3	Alignment . . . . .	115
J.3.4	Other Output Files . . . . .	115

# List of Figures

1.1	Philips focus over a variety of possible applications of breath analysis . . . . .	9
1.2	Basic elements of a gas chromatograph . . . . .	11
1.3	Typical gas chromatogram . . . . .	12
1.4	Mass spectrum of a single peak (Ethanol) . . . . .	12
2.1	Overview of the processing workflow for GC-MS data . . . . .	13
2.2	Detailed diagram of the data processing workflow . . . . .	15
2.3	Quality filtering tool . . . . .	16
2.4	Results for the 9 compound experiment using the Matlab tool . . . . .	18
2.5	Chromatogram of conditioned Tenax tubes with dry nitrogen . . . . .	18
2.6	Results of processing conditioned tubes with dry nitrogen with the Matlab tool .	19
2.7	Chromatogram of a mixture of known VOCs on conditioned tubes . . . . .	19
2.8	Results of processing a mixture of known VOCs with the Matlab tool . . . . .	19
2.9	Chromatogram of breath sample on conditioned tubes . . . . .	20
2.10	Results for the breath experiment that was part of the pilot study using the Matlab tool . . . . .	20
A.1	On-line breath collection for e-nose analysis . . . . .	24
A.2	Off-line breath collection for GC-MS analysis . . . . .	24
A.3	Breath collection setup . . . . .	24
A.4	Setup for breath adsorption in Tenax tubes . . . . .	25
A.5	Photography of Cyranose 320 E-Nose and a typical sensor pattern . . . . .	25
A.6	Photography of Agilent 6890N GC and a typical breathogram . . . . .	26
B.1	Basic elements of a gas chromatograph . . . . .	29
B.2	Diagram of the interactions within a capillary column . . . . .	30
B.3	Inner view of a capillary column . . . . .	31
B.4	Diagram of a mass spectrometer . . . . .	31
B.5	3D diagram of a mass spectrometer . . . . .	32
B.6	Sample GC-MS results, including a chromatogram and one mass spectrum per data point . . . . .	33
B.7	Temperature cycle of the GC-MS oven . . . . .	33
B.8	GC-MS setup used for processing breath samples at MiPlaza . . . . .	34
C.1	Chromatogram and color plot of mass spectral information of a breath sample .	35
C.2	Overview of the processing workflow for GC-MS data . . . . .	36
C.3	Simple chromatogram and results of peak extraction stage . . . . .	36
C.4	Portion of a breath chromatogram showing time shifting of two peaks . . . . .	37
D.1	Color legend for marking true and false positives . . . . .	41

D.2	Peak abundances as measured by MassHunter. Red and blue show the different mixtures . . . . .	44
D.3	Quality Score Filtering definition . . . . .	45
D.4	Results of the alignment stage in Mass Profiler Professional . . . . .	46
D.5	Peak abundance across different groups . . . . .	47
D.6	Results of the alignment stage in Mass Profiler Professional . . . . .	48
D.7	Peak abundance across different groups . . . . .	49
E.1	9 compound mixture used for repeatability testing . . . . .	52
E.2	Chromatographic overlay of 6 runs of testing mixture . . . . .	54
E.3	Overlay of 6 runs for peak with worst area deviation, Hexadecane . . . . .	55
E.4	Overlay of 6 runs for peak with worst retention time deviation, Toluene . . . . .	55
F.1	Overlay of 3 runs of dry nitrogen . . . . .	59
F.2	Zoom into the toluene peak for the 3 measurements . . . . .	59
F.3	Zoom into 2 siloxanes for the 3 measurements . . . . .	60
F.4	Results obtained by processing the dry nitrogen samples with the Matlab tool . . . . .	60
F.5	Overlay of 3 runs of wet nitrogen . . . . .	61
F.6	Zoom into 2 siloxanes for the 3 measurements . . . . .	62
F.7	Results obtained by processing the wet nitrogen samples with the Matlab tool . . . . .	62
F.8	Overlay of 3 runs of a dry mixture of known VOCs . . . . .	63
F.9	Zoom into the toluene peak for the 3 measurements . . . . .	63
F.10	Zoom into the two VOCs added to the mixture . . . . .	64
F.11	Zoom into 2 siloxanes for the 3 measurements . . . . .	64
F.12	Results obtained by processing the known dry VOCs samples with the Matlab tool . . . . .	65
F.13	Overlay of 3 runs of a wet mixture of known VOCs . . . . .	66
F.14	Zoom into the toluene peak for the 3 measurements . . . . .	66
F.15	Zoom into the two VOCs added to the mixture . . . . .	67
F.16	Zoom into 2 siloxanes for the 3 measurements . . . . .	67
F.17	Results obtained by processing the known wet VOCs samples with the Matlab tool . . . . .	68
F.18	Overlay of 3 runs of a breath sample . . . . .	69
F.19	Zoom into phenol peak . . . . .	69
F.20	Zoom into 2 siloxanes for the 3 measurements . . . . .	70
F.21	Results obtained by processing breath samples with the Matlab tool . . . . .	70
F.22	Overlay of 3 runs of dry nitrogen on tubes stored for 14 days . . . . .	71
F.23	Zoom into the toluene peak for the 3 measurements . . . . .	71
F.24	Zoom into 2 siloxanes for the 3 measurements . . . . .	72
F.25	Results obtained by processing the stored conditioned tubes with the Matlab tool . . . . .	72
F.26	Zoom into the toluene peak for dry and wet cases . . . . .	73
F.27	Zoom into 2 siloxanes for dry and wet measurements . . . . .	73
F.28	Zoom into the toluene peak for dry and wet cases . . . . .	74
F.29	Zoom into 2 siloxanes for dry and wet measurements . . . . .	74
F.30	Zoom into the two VOCs added to the mixture . . . . .	74
F.31	Sample of the compressed air administered to ICU patients at the hospital . . . . .	75
F.32	Air sample of Hamilton ventilators . . . . .	75
F.33	Air sample of Maquet ventilators . . . . .	76
F.34	Overlay of Hamilton and Maquet ventilators air . . . . .	76
F.35	Overlay of ventilator air and a breath sample . . . . .	77
G.1	Overlay of 3 runs of dry nitrogen on Tenax tubes . . . . .	80



G.2	Overlay of 3 runs of wet nitrogen on Tenax tubes . . . . .	81
G.3	Overlay of 3 runs of dry nitrogen and 2 VOCs on Tenax tubes . . . . .	81
G.4	Overlay of 3 runs of wet nitrogen and two VOCs on Tenax tubes . . . . .	82
G.5	Overlay of 3 runs of a breath sample . . . . .	83
G.6	Overlay of 3 runs of dry nitrogen on tubes stored for 2 weeks . . . . .	86
H.1	Evolution of carbon dioxide abundances over three weeks of storage . . . . .	91
H.2	Evolution of acetaldehyde abundances over three weeks of storage . . . . .	91
H.3	Evolution of 2-methyl-1-propene abundances over three weeks of storage . . . . .	91
H.4	Evolution of ethanol abundances over three weeks of storage . . . . .	92
H.5	Evolution of acetone abundances over three weeks of storage . . . . .	92
H.6	Evolution of isoprene abundances over three weeks of storage . . . . .	92
H.7	Evolution of dimethylsulfide abundances over three weeks of storage . . . . .	93
H.8	Evolution of carbon disulfide abundances over three weeks of storage . . . . .	93
H.9	Evolution of 1-propanol abundances over three weeks of storage . . . . .	93
H.10	Evolution of trimethylsilanol abundances over three weeks of storage . . . . .	94
H.11	Evolution of 2-butenal abundances over three weeks of storage . . . . .	94
H.12	Evolution of 2-methyl-1,3-dioxalane abundances over three weeks of storage . . . . .	94
H.13	Evolution of benzene abundances over three weeks of storage . . . . .	95
H.14	Evolution of heptane abundances over three weeks of storage . . . . .	95
H.15	Evolution of toluene abundances over three weeks of storage . . . . .	95
H.16	Evolution of hexamethylcyclotrisiloxane abundances over three weeks of storage . . . . .	96
H.17	Evolution of N,N-dimethylacetamide abundances over three weeks of storage . . . . .	96
H.18	Evolution of benzaldehyde abundances over three weeks of storage . . . . .	96
H.19	Evolution of octamethylcyclotetrasiloxane abundances over three weeks of storage . . . . .	97
H.20	Evolution of limonene abundances over three weeks of storage . . . . .	97
H.21	Evolution of decamethylcyclopentasiloxane abundances over three weeks of storage . . . . .	97
I.1	Block diagram . . . . .	98
I.2	Information contained in an *.elu file for a single peak . . . . .	99
I.3	Quality filtering tool . . . . .	100
I.4	Illustration of the concept of optimal separability . . . . .	101
I.5	Illustration of the concept of sample “Alignment” . . . . .	102
I.6	First step of alignment process . . . . .	103
I.7	Second step of alignment process . . . . .	103
I.8	Final step of alignment process . . . . .	104
I.9	Quality filtering tool . . . . .	105
I.10	Graphical user interface for the aligner software . . . . .	105
I.11	Graphic output of the alignment process . . . . .	106
I.12	Excel output of the alignment process . . . . .	106
I.13	Reduced example of output in Weka’s arff file format for statistical analysis of asthma data . . . . .	107
J.1	Analyze pop-up menu . . . . .	108
J.2	Analysis settings menu . . . . .	109
J.3	Analysis settings menu . . . . .	109
J.4	Analysis settings menu . . . . .	110
J.5	AMDIS main window . . . . .	111
J.6	Batch processing menu . . . . .	112
J.7	File selection menu . . . . .	113
J.8	File management menu . . . . .	113

J.9 Quality filtering tool . . . . . 114

# 1 Introduction

## 1.1 Breath Analysis

The concept of breath testing has existed for years. Physicians know that certain odours may be strong indicators of disease. For example, a fruity breath could suggest ketoacidosis, or an ammonia-like smell could indicate kidney failure.

In some areas, breath analysis already has enormous commercial applicability, such as in the case of breath alcohol monitors. However, its commercial applications can go far beyond these simple devices. They have the potential to develop into tools for screening, diagnosing and monitoring disease.

Following the advances of technology and research, breath testing has evolved towards the study of volatile organic compounds (VOCs) present in breath. Recent studies, in conjunction with the increasing understanding of disease processes and biomolecules, propose exhaled breath analysis as a safe, non-invasive method that can provide additional information to the traditional blood and urine studies.

It is Philips aim to ideally develop a device that can take a patient's breath and classify him as healthy or sick, initially for two diseases: sepsis and asthma. Still, in order to eventually develop this device, much research needs to be done, so as to discover which are the biomarkers that can act as predictors of these diseases.

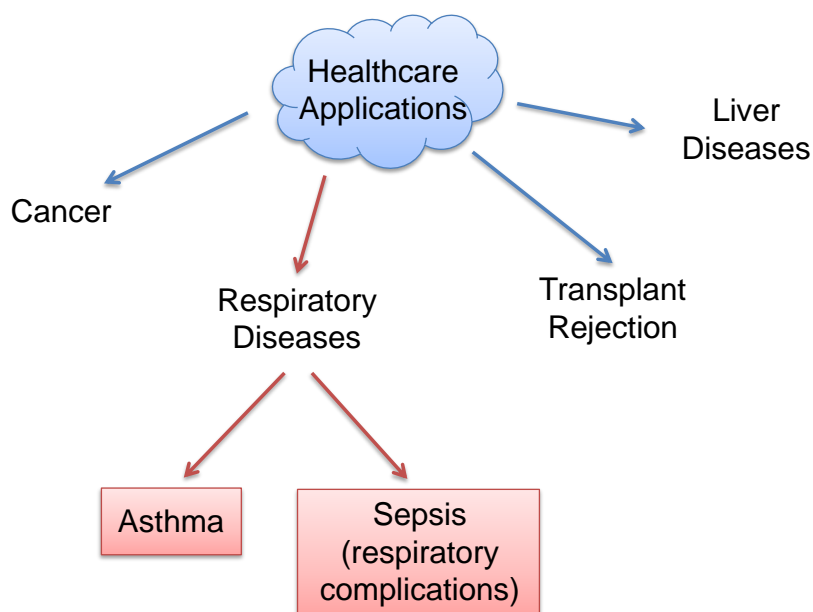


Figure 1.1: Philips focus over a variety of possible applications of breath analysis

### 1.1.1 Breath Markers

Volatile organic compounds (VOCs) are small molecules which evaporate from liquids or solids into air, reaching an equilibrium. This process provides a simple method to study the content of liquid or solids without entering into contact with them, by analysing the composition of the headspace or surrounding air.

Usually, we release hundreds of VOCs in every breath, which are the result of the metabolic

processes occurring in the body. However, their concentrations are in the picomolar range, so special techniques are required for their collection and analysis.

Over the last 20 years, breath analysis has greatly evolved. It is now understood that VOCs are usually the result of the fractioning of larger biomolecules, so they should be studied as a pattern rather than individually.

Several thousand VOCs have been observed in breath samples so far, though barely 1% of them can be found in all males. The remaining 99% is influenced by environmental or lifestyle factors. The goal of Philips, and of other research institutes throughout the world, is to eventually link some of these VOCs to certain diseases.

### 1.1.2 Analysis Techniques

The analysis methods for breath range from laboratory techniques such as gas chromatography to bedside alternatives, such as the use of electronic nose. This spectrum satisfies different needs: laboratory techniques are more suitable for research and biomarker discovery, while bedside devices are ideal useful for diagnosis and monitoring.

One of the most popular techniques is the use of artificial olfactory systems (also called electronic noses), which can translate an odour into a pattern produced by broadly selective chemical sensors. Electronic noses (E-noses) are more suitable for diagnostic assessment and monitoring in clinical environments, given their small size and portability. Through the use of pattern recognition with sufficient training data, an e-nose can learn to distinguish between healthy and sick sensor patterns. Ideally, any hospital could have an e-nose for diagnostic purposes, since its price is quite low compared to other breath testing methods.

Gas chromatography-mass spectrometry (GC-MS) allows for the separation and identification of different compounds. Given its size and high cost, they are most suitable for clinical research. GC-MS is generally considered the golden standard for breath analysis. In the following section, we describe this technique, which was the one used in this project.

## 1.2 Gas Chromatography-Mass Spectrometry (GC-MS)

Gas chromatography-mass spectrometry (GC-MS) is an instrumental technique that combines the features of gas chromatography and mass spectrometry to accurately separate and identify different substances within a test sample.

Chromatography is a methodology developed around 50 years ago, which provided an unparalleled separation power of a sample mixture and a great ease of use. It consists of two distinct phases: a stationary phase, which can be either solid or liquid, and a moving gaseous phase. The rate of interaction between analyte and stationary phase will define the degree of separation (or elution) of the compounds.

The eluted molecules are then introduced into the mass spectrometer where they are ionized, accelerated, deflected, and detected separately. This results in a spectrum of masses that are a “fingerprint” of the compounds present in the original test sample.

The GC-MS technique combines the best of the two instruments, providing the proper separation of compounds required by the mass spectrometer in order to avoid overlapping results, and mass spectrometry’s great identification power.

### 1.2.1 Main Structure

Figure 1.2 shows the main structure of a GC-MS.

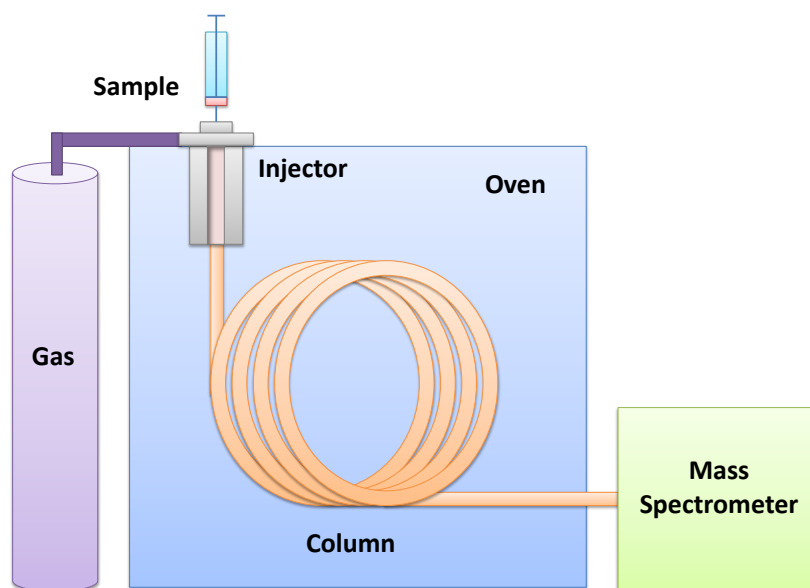


Figure 1.2: Basic elements of a gas chromatograph

The sample is injected into the column, carried by nitrogen gas. It interacts with the inner lining of the column. Since different molecules interact differently with the stationary phase, they travel at different speeds. Therefore, they exit the column (or elute) at various times. In this manner, at the end of the column, molecules are separated according to their type. Every type of molecule appears as a different peak in the chromatogram, which is a plot that shows abundance vs. time.

In a later stage, the already separated molecules are ionized and the fragments are detected by a mass spectrometer. This provides a specific pattern for every point in time, and it is what allows for the identification of the components of every peak in the chromatogram.

### 1.2.2 Data

A GC-MS system produces a 3D dataset. Figure 1.3 shows a typical total ion count (TIC) chromatogram, which represents the integration of all the mass spectral information for every point in time. The two axes that form the chromatogram are the retention time, which are the times at which compounds elute, and the abundance of such components at the detector. The latter is normally measured in arbitrary units, but can be calibrated with internal standards added to the measured sample.

Since a chromatogram has an average duration of 35 minutes for breath samples and about 3 mass spectra are processed every second, about 7000 mass spectra are produced per sample. This means that a large amount of data must be processed to extract useful information from the measurement.

The GC-MS instrument produces output files which contain both time and spectral information, along with instrument and configuration details. One of the advantages of GC-MS compared to other separation techniques is the wealth of existing mass spectral information. With the aid of a special library search software, it is possible to compare a compound's spectrum against spectral libraries, and find the identity of that compound.

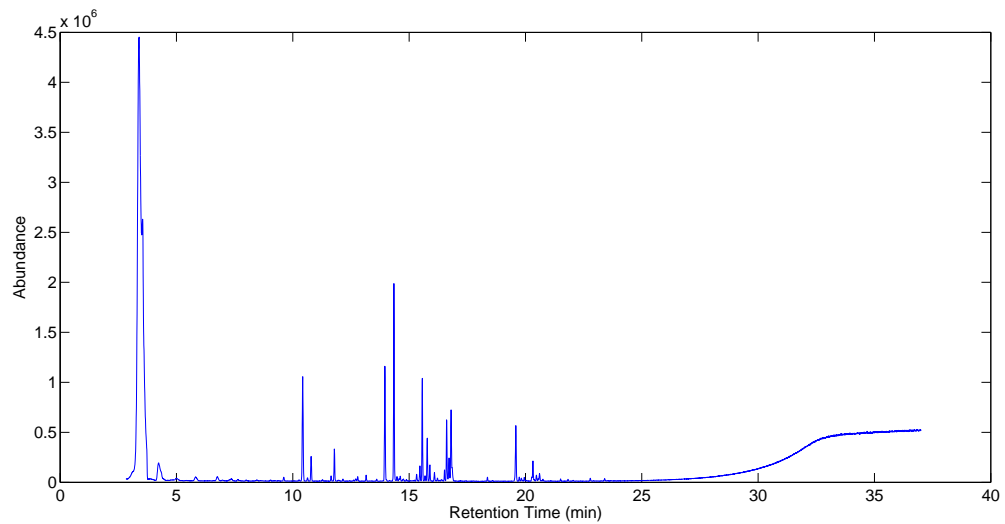


Figure 1.3: Typical gas chromatogram

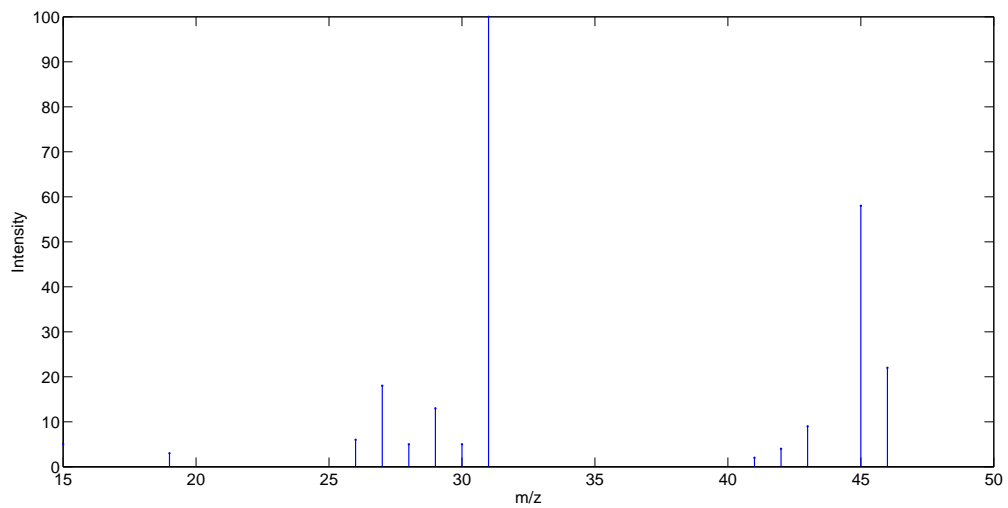


Figure 1.4: Mass spectrum of a single peak (Ethanol)

**The aim this project is to develop a software workflow which takes gas chromatography-mass spectrometry data of breath samples as an input and returns a list of the components present in that sample.**

## 2 Results

### 2.1 Signal Processing

A GC-MS system produces a 3D dataset. The three axes that compose the dataset are time, mass-to-charge ratio ( $m/z$ ) and abundances. For every point in time, there is a corresponding mass spectrum, which contains the information of the ion fragments present at that point in time.

The processing of the GC-MS data can be divided into 5 steps, shown in figure 2.1.

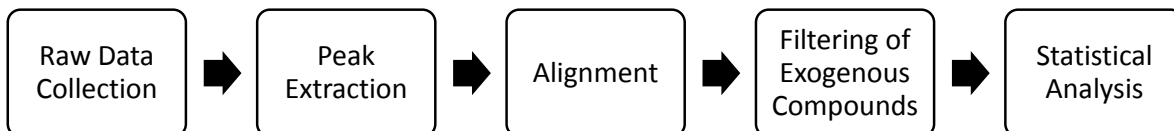


Figure 2.1: Overview of the processing workflow for GC-MS data

The instrument generates output files, which contain the raw information of the measurement.

The following step is the extraction of features out of the dataset, which means finding all the different peaks present in the chromatogram. In order to perform this feature extraction, the total ion count is found, by integrating all the mass spectra. The total ion count is then processed in order to detect all the peaks present in the curve. It is not a simple process, because some compounds may elute very close in time, so the algorithm must be smart enough not to miss any components. This is performed by the deconvolution algorithm, which in our case is AMDIS. Its choice is explained in the following section. The result of this stage is a list of peaks, where each represents a single compound, and their particular characteristics, such as retention time, area and mass spectrum.

However, no two chromatographic measurements are ever the same. This means that the same compound may elute at a slightly different retention times every time. This presents a challenge when working with multiple samples, because since the final objective is to compare them, we need to be certain that the component eluting at  $x$  minutes in sample 1 is the same as component eluting at  $x$  minutes in sample 2. This is what is called “alignment”. Since retention time is not enough to unambiguously identify a certain compound, more information, i.e. the mass spectra, needs to be taken into account for the comparison. The truth about the identity of a peak always lies in its mass spectral information.

Therefore, in order to make a study with different patient samples, it is necessary to “align” the compounds to make sure that we are comparing the same compound in each sample, no matter their retention time.

There are two possible ways to perform the alignment: prior to the peak extraction by adjusting the non linear shifts of the time axis, or after the peak extraction, by working with peak lists. In our case, we worked in the second manner, because given its discrete nature it can be much faster than working with complete chromatographic curves.

Still, since every sample contains a few hundred peaks, it is necessary to develop a fast alignment algorithm. In our case, we use a parameter to optimize the speed of the alignment software: the retention time window. This is explained in more detail in Appendix I. The retention time window basically limits the search of the peak in a list by setting the maximum expected shift for that peak. This means that the software would only look for it (in a second sample) in a window around its retention time in the first sample. Thus, the mass spectrum comparison only

needs to be performed against a few compounds. By working in this way, we are independent of the nature of the shifts, which are generally non-linear.

Once everything is processed, we may need to eliminate compounds that have a non-endogenous origin. We have identified a number of contaminants in the samples that are setup-related, and even more may be found in the future. Such is the case of phenol or N,N-dimethyl acetamide, which originate in the sampling bag. It is essential to remove these elements so as to ensure the validity of the conclusions that may be drawn out of the data. They have no value for identifying disease and may even affect the performance of a classifier.

The last stage of GC-MS data processing is the statistical analysis. In this step, a classifier is built from the available data which allows to classify new patients as healthy or sick. There are several software alternatives for this stage, but it is out of the scope of this project.

The most important part of GC-MS data processing is to ensure the quality of the biomarkers found. This can only be achieved by optimizing every step of the process, improving extraction and alignment algorithms, and properly defining filtering steps.

## 2.2 Commercial Software Alternatives

In order to test the capabilities of the software package that will later be used for the asthma and sepsis projects at Philips, a small experiment was planned.

Both projects, asthma and sepsis, propose to find and identify biomarkers required to correctly classify patients as healthy or ill. For this purpose, a software package developed by Agilent was evaluated. However, given the complexity of breath mixtures it is difficult to assess the quality of the processing by analysing patient data.

For that reason, a small scale test was carried out with known data. In this way, several concerns could be analysed, such as the effectiveness of the peak extraction, the feature finding procedures, the quality of the alignment of peaks and the statistical analysis conclusions.

Two mixtures were made in the lab, each containing the same 9 compounds. Of these 9 compounds, 5 were in the same concentration and 4 were present in a different concentration. In this way, we had two different controlled groups. The goal was to process them in the same manner we would process our breath samples, and study if they could be correctly classified into these two different categories.

### 2.2.1 AMDIS vs. MassHunter

The peak extraction procedure was performed with AMDIS software, which is a free program from the National Institute of Standards and Technology (USA) and with MassHunter, from Agilent. Each software applied their deconvolution algorithm to our test dataset. AMDIS found an average of 13 compounds per sample (were only 9 were real peaks) while the average for MassHunter was 19. This meant that AMDIS had a sensitivity of 100% and a positive predictive value of 68.35%, while MassHunter also had a 100% sensitivity, but a lower positive predictive value of 47.37%. One other factor that supported the software choice was that the commercial alignment software, Mass Profiler Professional, had the possibility to filter false positives only for the AMDIS case. Furthermore, MassHunter extracted peak areas were unstable. For many substances, it was known that their concentration did not change from sample to sample. However, MassHunter showed an inexplicable variation in the extracted areas for these peaks, rendering it completely unreliable. Since AMDIS had shown to be superior in terms of positive predictive value, some of its false positives could be removed by MPP in the following stage and its extracted peak abundances were stable for repeated measurements, it was an obvious choice over MassHunter. Thus, we discarded Agilent's deconvolution software and chose the free alternative for all future processing.



## 2.2.2 Mass Profiler Professional, by Agilent

Initially, Mass Profiler Professional was a solid possibility for sample alignment. It could either work in combination with AMDIS or with MassHunter. Nonetheless, combining it with AMDIS had a serious benefit: it allowed the user to apply quality filtering on the data. It was a rudimentary filtering since only a constant could be set, but it was already a great improvement from the raw data.

The results with the 9 component mixtures were excellent. All false positives were eliminated, leaving just 9 peaks per sample, and these peaks were correctly aligned. The software behaved as expected. Still, it was clear that even though it was enough for processing simple mixtures such as our test samples, it would be much harder to filter out false positives out of complex breath data. The formula used for filtering had an inherent dependence on peak abundance. This was not suitable for our case, since we could not guarantee that good biomarkers were necessarily highly abundant. Thus, it was decided that our application required a solution similar to Mass Profiler, but adapted to our needs, in particular for the quality filtering stage. The software developed is described in the following section.

## 2.3 Matlab Tool

After analysing the available commercial tools for GC-MS data processing, we determined that none of the studied alternatives fulfilled our exact requirements, in particular for the quality filtering and alignment stages. It was decided that it would be more useful to develop a specific software solution for our needs. This tool was created with Matlab, for its great flexibility for analysing and plotting data, and its ease for making modifications to the source code.

A more detailed description of the tool can be found in Appendix I.

Figure 2.2 shows a more detailed view of the data workflow and the limits of the Matlab tool.

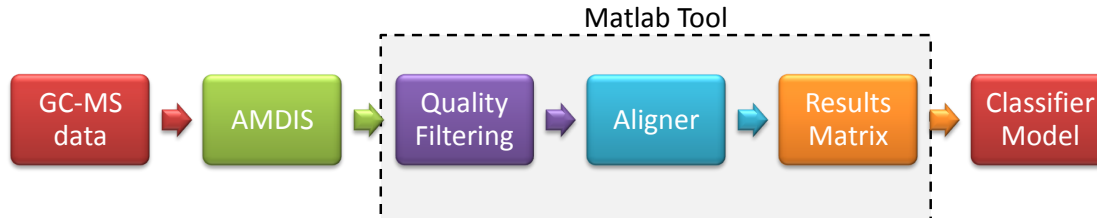


Figure 2.2: Detailed diagram of the data processing workflow

### 2.3.1 Quality Score Filtering

The main drawback of the alignment software we tested was its inability to cope with false positives. Deconvolution algorithms usually produce false positives as a consequence of their high sensitivity to peaks. It is not unusual for a deconvolution software to find between 30% and 50% false results. This happened for both pieces of software analyzed.

When it was decided to develop a custom solution for processing breath data, this became one of the first requirements. Since breath datasets are so large and complex, it was necessary to ensure the quality of the analysis results by removing false positives or low quality compounds. The solution was to implement a type of filtering that removes components that are suspected to be false positives or that simply do not meet quality criteria, by having for instance a poor signal-to-noise ratio. In this way, unreliable peaks could be eliminated.

AMDIS provides a set of characteristics along with every extracted peak. Some of these characteristics, such as the signal-to-noise ratio have a very clear link to quality. Still, this link

depends on the experimental configuration. Our intention was to create a tool that could allow the user to train a filter for poor quality data.

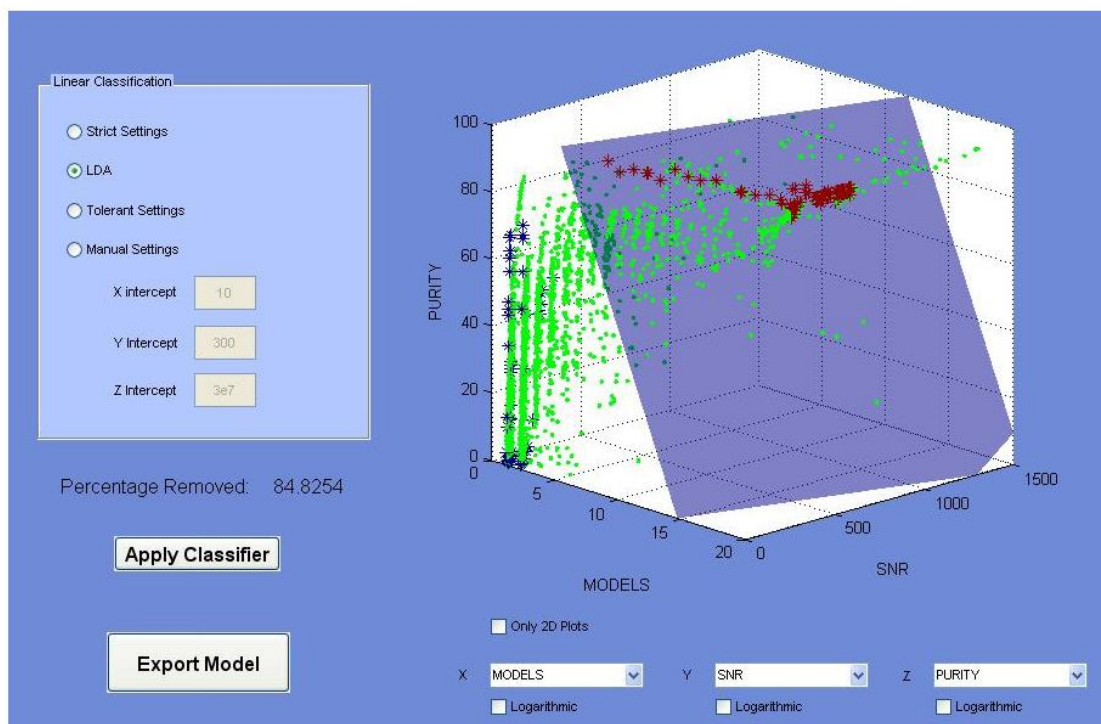


Figure 2.3: Quality filtering tool

Even though this type of filtering was considered in Mass Profiler Professional, it was poorly documented and depended on the user blindly setting filtering thresholds. Our tool overcame this by not only allowing for multiple filtering parameter selection, but also by providing visual and quantitative outputs to the operator, so that he could be in complete control of how much data is filtered out and what remains. In this manner, only good quality peaks can be preserved. Less unreliable peaks in the data translates into a better performance of the alignment algorithm in the next processing stage.

Quality score filtering can always be disabled if the user intends to work with pure, raw data, at the risk of considering background noise as potential biomarkers.

In order to create the filter, a library of known true and false positives was necessary. We obtained it through a series of controlled experiments where we knew exactly what was inside the mixtures, and what necessarily had to be the consequence of algorithmic artifacts.

The tool can plot two or three peak parameters (from the six parameters available: models, SNR, abundance, purity, width and amount) and find a linear decision surface to divide true and false positives. In the future, other surfaces could be implemented, but in our case it did not seem necessary or justifiable at this point.

Figure 2.3 shows an image of the user interface. The user can load a library of known true and false positives (blue and red stars in the plot) and overlay it with his current data (green dots). In this case, the unclassified data come from an asthma experiment with 19 patients.

The resulting filter can be exported back into the main software tool. A more detailed explanation of the quality score filter can be found in Appendix I.

### 2.3.2 Alignment

As mentioned in previous sections, in order to compare the contents of different samples, it is necessary to perform an “alignment” of the peak lists. Since non-linear time shifts are expected, but they are known to be quite small, it is possible to determine a maximum retention time window in which a given chemical compound can be found. We take advantage of this fact in order to speed up the alignment process.

The basic functioning of the algorithm is as follows. All the compounds in the first sample are added to a partial matrix, where every row is a different compound. This list is compared against the second sample. If a compound in the partial list is found in the second sample, then its abundance is added to the matrix. All the compounds present in sample 2, but that were not originally in the partial matrix are added. This improved partial matrix is now compared against sample 3, and the process is repeated for all samples. In the end the list will contain all the compounds found in all samples.

In order to find out whether two compounds are the same it is necessary to compare their mass spectra. Since we do not intend to perform hundreds of comparisons per compound in the partial matrix, we simply search for it in a window around its retention time. The similarity of the spectra is calculated by finding their correlation. This is further explained in Appendix I.

### 2.3.3 Output

So as to maximize the possibilities for data analysis, we provide results in four ways. Firstly, they are saved as a Matlab matrix and a bar plot, where the bars are clustered by group. The user, at the beginning, can set the group to which every sample belongs, for example “Asthma” or “Healthy”.

Data is also stored as an excel file, where the same information that is plotted is stored as an array. Furthermore, that array is also stored in Weka ARFF format for statistical processing.

## 2.4 Processing Examples

In order to test the capabilities of the software package that will later be used for the asthma and sepsis projects at Philips, a series of experiments were planned.

### 2.4.1 Pilot Experiments

Both projects, asthma and sepsis, intend to find and identify biomarkers in breath samples. However, given the complexity of breath mixtures it is difficult to assess the quality of the processing by analysing patient data. Furthermore, it is necessary to understand the effects that the setup may have on the measurements.

For these reasons, several controlled tests were carried out with known data. In this way, several concerns regarding the setup and the software could be analysed.

#### 9 component experiment

This experiment, which was originally performed to compare different software packages, was repeated with our Matlab tool. Two different mixtures, composed by the same 9 compounds in different amounts were prepared. The aim was to see if they were properly extracted and aligned. The results obtained with the combined workflow of AMDIS + our Matlab tools were exactly what we expected: no peaks were lost and there was no confusion in the alignment. Figure 2.4 shows a bar plot of the results.

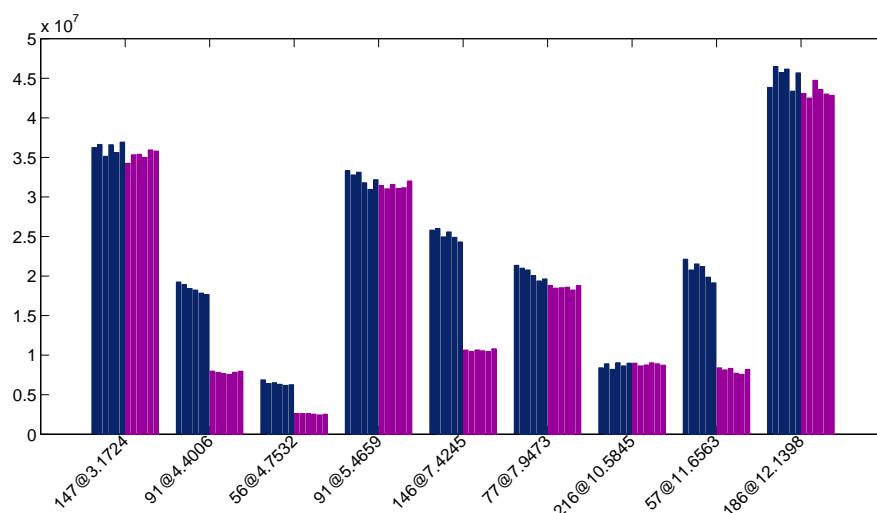


Figure 2.4: Results for the 9 compound experiment using the Matlab tool

### Pilot Study

The pilot study mainly focused on studying the effects of the sampling setup on the measurements. In some cases, it also provided a chance to test the Matlab tool. This is explained in detail in Appendix F. In this section, however, we show some of the results obtained. When possible, we also provide the results after processing with the Matlab tool.

Figure 2.5 shows the results obtained when analysing a blank conditioned tube that only had nitrogen flowed through.

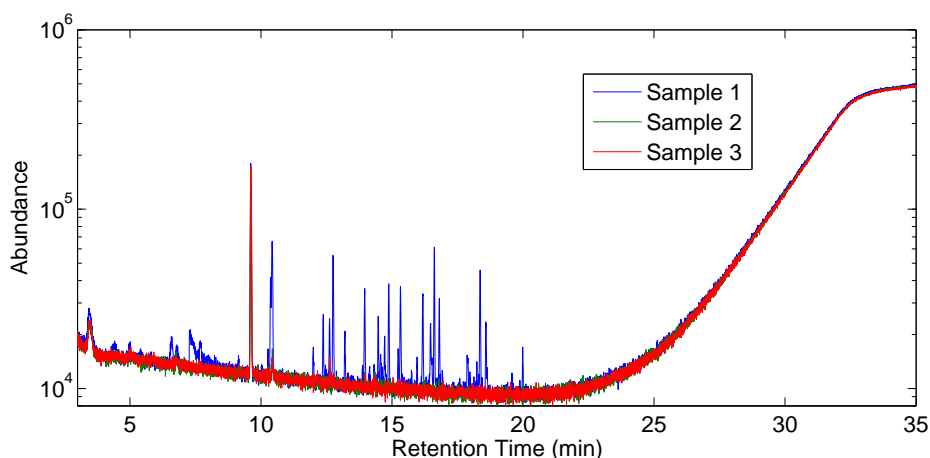


Figure 2.5: Chromatogram of conditioned Tenax tubes with dry nitrogen

Figure 2.6 shows the results obtained by processing the previous chromatograms with AMDIS and the Matlab tool. It is clear how the noisy background is ignored and only one peak is successfully found. This peak is Toluene, and it was intentionally added to the sample as an internal standard.

Figure 2.7 is the chromatogram of a known VOC mixture that was stored in conditioned tubes. This mixture contained 2 VOCs and toluene as an internal standard.

Again, the results obtained with the software workflow were good. Figure 2.8 shows that a total of 3 peaks were identified, and they correspond to the 2 VOCs and toluene.

In order to also analyse a more complex mixture, a breath sample was collected. It was

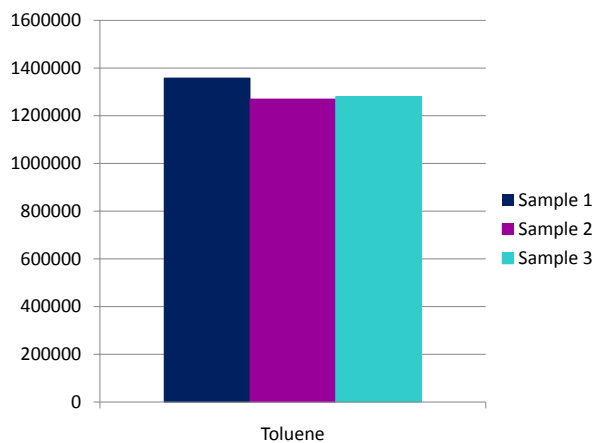


Figure 2.6: Results of processing conditioned tubes with dry nitrogen with the Matlab tool

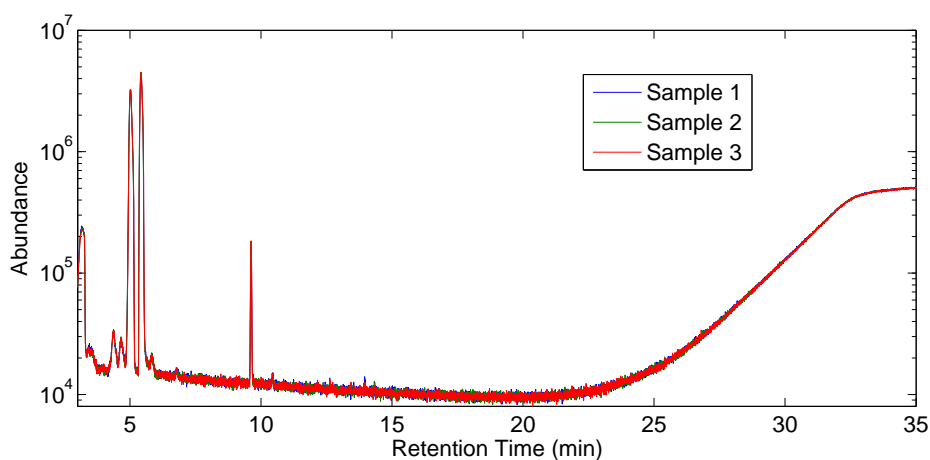


Figure 2.7: Chromatogram of a mixture of known VOCs on conditioned tubes

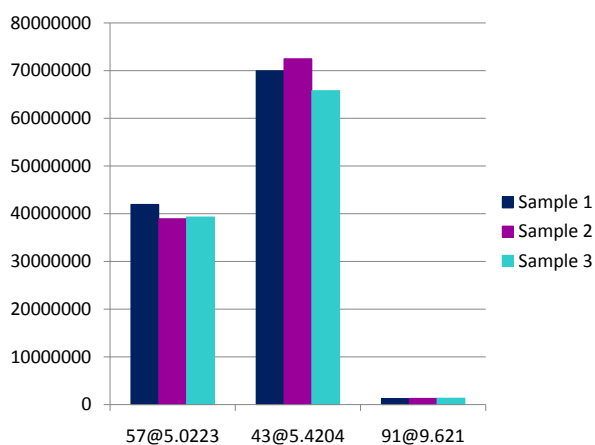


Figure 2.8: Results of processing a mixture of known VOCs with the Matlab tool

then stored in 3 conditioned Tenax tubes and analysed. Figure 2.8 shows the chromatogram obtained.

Figure 2.10 contains the peaks detected by the software. The quality score filter that was

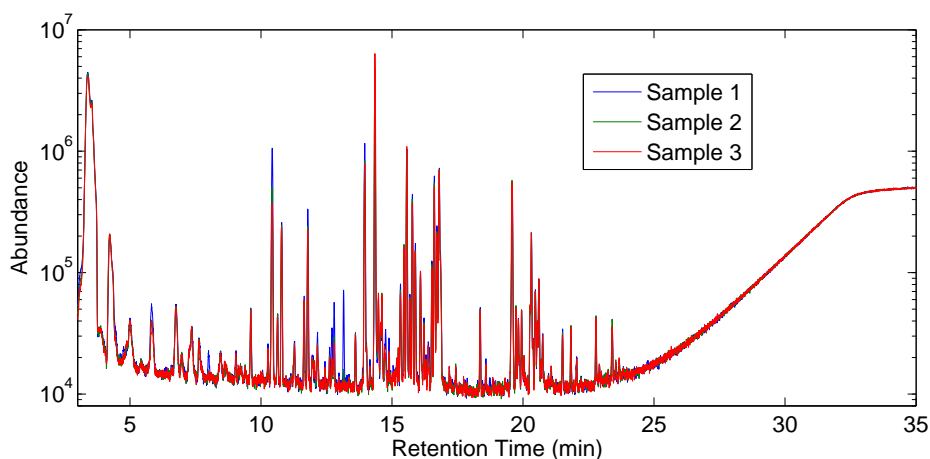


Figure 2.9: Chromatogram of breath sample on conditioned tubes

applied was calibrated with data from the previous 9 compound experiment and the decision surface was chosen with linear discriminant analysis. The performance was very good, with the only error appearing in the last component. That peak was not detected for sample 1, because it was removed by a too strict filtering stage. However, in the future this can be adjusted to obtain perfect results.

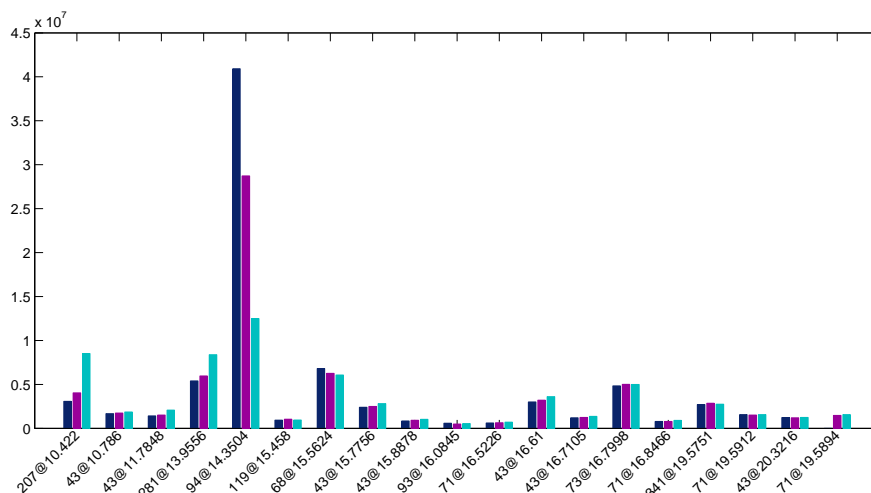


Figure 2.10: Results for the breath experiment that was part of the pilot study using the Matlab tool

## 2.4.2 Analysis of asthma data

We wanted to take the software testing on step further. We had analysed breath obtained in a controlled manner in the lab, but it was necessary to evaluate the performance on real clinical data.

There was one small dataset of patient breath samples available from the asthma study. The dataset consisted of 19 patient samples (10 controls and 9 with asthma), where each was sampled 3 times. Each of this samples was processed with 1 week difference with each other.

All these files were processed with AMDIS and the Matlab tool. From the results matrix, a group random of components were extracted, spanning the entire range of retention times. The

evolution of these components over time was calculated, and these results can be found in detail in Appendix H.

It was found that some components were inconsistent with time, while other were stable. Most of the compounds that behaved erratically were precisely those whose origin we could not explain, i.e. silicones. However, the compounds of known, endogenous origin did not vary considerably with time.

These results were published as part of a paper that was written in cooperation between Amsterdam AMC and Philips.

## 3 Conclusions

The project was successful in terms of achieving a final design product that satisfies the customer's original requirements. The software fulfilled the need to easily process the GC-MS data and provide an accurate and reliable list of possible biomarkers present in a breath sample.

After studying and testing different commercial alternatives, it was discovered that none satisfied the specified needs. Breath samples are very particular compared to other samples, since they are very complex and the most important components for patient discrimination are not necessarily the largest ones.

In comparison to the commercial programs in the market, the reliability was improved through the development of a filtering system for poor quality compounds. This system allows the user to remove potential false positives or markers that do not meet quality criteria easily and with complete control over the process. The user is always aware of how much information is being lost because of the filtering, but also knows that what remains is reliable enough to draw statistical conclusions.

The alignment stage also demonstrated to work properly, yielding adequate results for both test and clinical data samples.

The final result is a flexible software toolbox that in the future can be used for analysing other complex breath datasets.



# A Introduction to Breath Analysis

## A.1 Introduction

The concept of breath testing exists since early times. In the past, physicians would associate a particular odor to certain diseases. For instance, a fruity breath could suggest ketoacidosis, or an ammonia-like smell could indicate kidney failure.

In present times, breath analysis has proved to have an enormous commercial applicability, as exemplified by the common alcohol testing devices available in the market.

Following the advances of technology, breath testing has evolved towards the study of volatile organic compounds (VOCs) present in breath. Recent studies, in conjunction with the increasing understanding of disease processes and biomolecules, propose exhaled breath analysis as a safe, non-invasive method that can provide additional information to the traditional blood and urine studies.

Ideally, breath analysis can be used for screening, diagnosis and monitoring of disease.

## A.2 Breath Markers

Volatile organic compounds (VOCs) are small molecules which evaporate from liquids or solids into air until they reach an equilibrium. This process provides a non-invasive method to study the content of liquid or solids, by analysing the composition of the surrounding air.

Normally, the human body releases hundreds of different VOCs that are the result of various metabolic processes. However, they are present in picomolar concentrations, thus special techniques are required for their collection and analysis.

Over the last 20 years, breath analysis has greatly evolved. It is now understood that VOCs are usually the result of the fractioning of larger biomolecules, so they should be studied as a pattern rather than individually. This type of study has been benefited by the use of artificial olfactory systems, which are discussed in a later section.

Today, thanks to the technological advances in this field, several thousand VOCs have been observed in breath samples. Only about 1% of them can be found in all males, while the remaining 99% is influenced by environmental or lifestyle factors. Hopefully, these VOCs could be linked to different clinical conditions.

The results of different research studies are summarized in tables A.1 and A.2. These studies used gas chromatography-mass spectrometry to identify the substances found in breath.

## A.3 Sample Collection

The electronic nose can work by directly sampling exhaled breath. GC-MS analysis, on the other hand, are off-line, since the instrument is not at the hospital. Therefore, in this case breath is captured initially in Tedlar bags and the contents are later captured in sorbent tubes, which are then transported to the laboratory.

Figure A.1 shows the online measurement of a child's breath. Figure A.2 instead, shows the breath collection setup for GC-MS analysis. The same setup can also be used for offline e-nose studies. A detailed view of the collection device can be observed in figure A.3. The patient breathes into a two-way mouthpiece. The inspired air is free from environmental VOCs thanks to an inspiratory VOC filter. Exhaled air goes through a silica filter that absorbs moisture and is then stored in a Tedlar bag.

Normally, the patient is required to breathe VOC free air for about 5 minutes and then one single breath, at expiratory vital capacity, is collected.



Figure A.1: On-line breath collection for e-nose analysis



Figure A.2: Off-line breath collection for GC-MS analysis

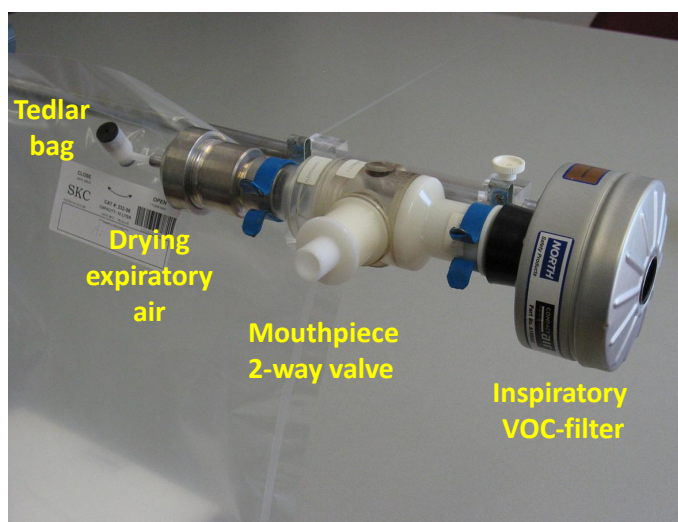


Figure A.3: Breath collection setup

Since samples need to be transported, in particular for the GC-MS case, breath should be transferred to a more suitable container, rather than moving Tedlar bags. For this reason, the gas contained in the bags is extracted and flowed through adsorption tubes that capture the VOCs present. Figure A.4 shows a diagram of this setup.

A pump extract air out of the Tedlar bag and makes it go through the Tedlar tube. A mass flow controller measures the exact volume going through the setup, in order to ensure VOCs are captured by the tube.

## A.4 Analysis Methods

The analysis methods for breath range from laboratory techniques such as gas chromatography to bedside alternatives, such as the use of electronic nose. This spectrum satisfies different needs: laboratory techniques are more suitable for research and biomarker discovery, while

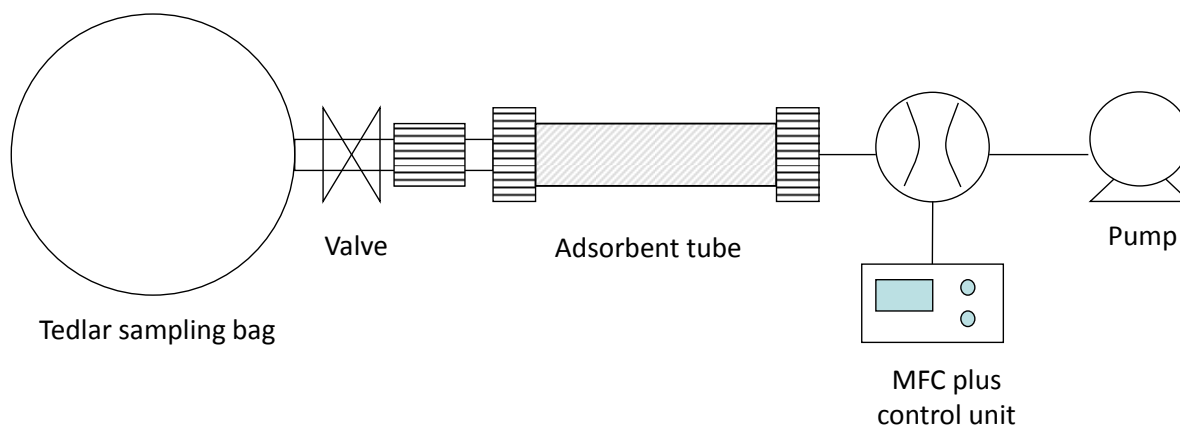


Figure A.4: Setup for breath adsorption in Tenax tubes

bedside devices ideal useful for diagnosis and monitoring. In the following section, we describe the two main techniques for exhaled breath analysis.

#### A.4.1 Electronic Nose

One of the most popular techniques is the use of artificial olfactory systems (also called electronic noses), which can translate an odour into a pattern produced by broadly selective chemical sensors. The sensor pattern is a fingerprint of the smell, though identification is only possible by comparison against known sensor patterns. Figure A.5 shows an electronic nose and a typical sensor pattern.

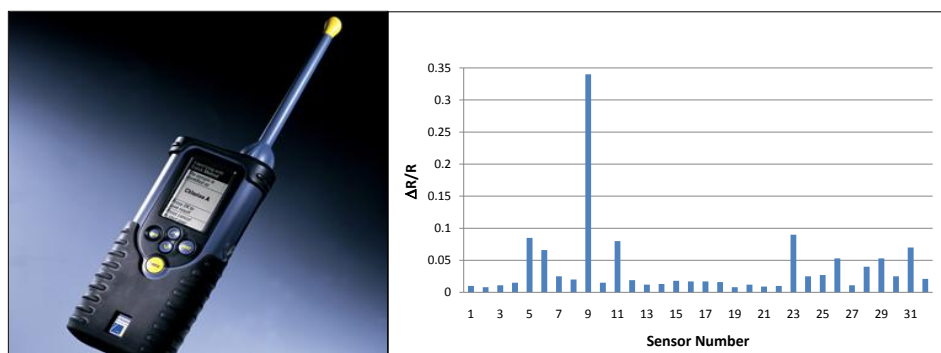


Figure A.5: Photography of Cyranose 320 E-Nose and a typical sensor pattern

Electronic noses (E-noses) are more suitable for diagnostic assessment and monitoring in clinical environments, given their small size and portability. Through the use of pattern recognition with sufficient training data, an e-nose can learn to distinguish between healthy and sick sensor patterns. Ideally, any hospital could have an e-nose for diagnostic purposes, since its price is quite low compared to other breath testing methods.

#### A.4.2 Gas chromatography-Mass Spectrometry

Gas chromatography-mass spectrometry (GC-MS) allows for the separation and identification of different compounds. Given its size and high cost, they are most suitable for clinical research. GC-MS is generally considered the golden standard for breath analysis.

It provides a chemical profile that can be used for the development of a classifier and for the identification of every compound present in the sample.

The main difference when comparing this technique with electronic noses, is that in GC-MS the compounds present in breath can be properly identified and named, while in e-noses only patterns of smell are found rather than specific compounds. Therefore, if the aim is to study the connection between biomarker and disease, it is necessary to have full knowledge of the chemical identity of the marker.

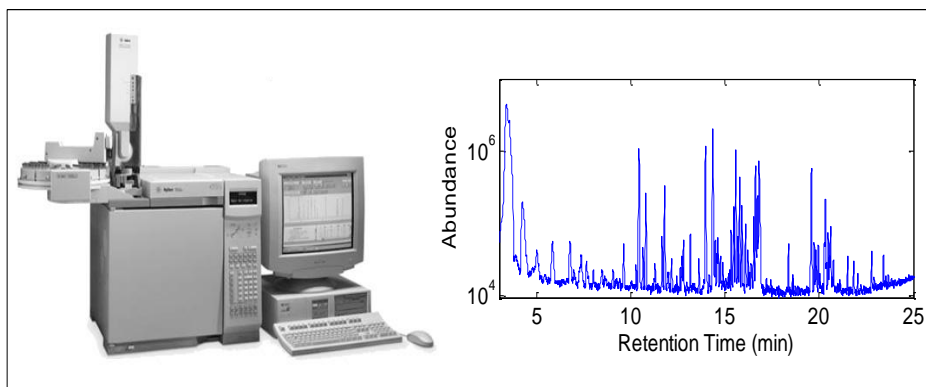


Figure A.6: Photography of Agilent 6890N GC and a typical breathogram

## A.5 Conclusions

There is an enormous potential in sampling markers from exhaled air. Research has already shown positive results for many different diseases. However, it is important to remember that diagnosis will not occur with a single biomarker but with a set of markers that constitute the so-called breathprint.

Breathprints may provide useful information for screening, diagnosis and continuous monitoring of disease, in a non-invasive manner.

Table A.1: Performance After Post Filtering

	COPD	Cystic Fibrosis	Oxidative Stress	Tuberculosis	H. Pylori	Liver Disease	Transplant Rejection	Breast Cancer	Lung Cancer
Isoprene	x								
C16 hydrocarbon	x								
4,7-Dimethyl-undecane	x								
2,6-Dimethyl-heptane	x								
4-Methyl-octane	x								
Hexadecane	x								
3,7-Dimethyl 1,3,6-octatriene	x								
2,4,6-Trimethyl-decane	x								
Hexanal	x								
Benzonitrile	x								
Octadecane	x								
Undecane	x								
Terpineol	x								
Pentane		x	x						
DMS		x							
2-propanol		x						x	
Ethane			x						
Nitric Oxide			x						
Oxetane, 3-(1-methylethyl)-				x					
Dodecane, 4-methyl-				x					
Bis-(3,5,5-trimethylhexyl) phthalate				x					
Benzene, 1,3,5-trimethyl-				x					
Decane, 3,7-dimethyl-				x					
Tridecane				x					
1-Nonene, 4,6,8-trimethyl				x					
Heptane, 5-ethyl-2-methyl				x					
1-Hexene, 4-methyl-				x					
Carbon dioxide					x				
Acetone						x			
2-butanone						x			
2-pentanone						x			



# B Summary of Gas Chromatography-Mass Spectrometry (GC-MS) techniques

## B.1 Introduction

Gas chromatography-mass spectrometry (GC-MS) is an instrumental technique that combines the features of gas chromatography and mass spectrometry to accurately separate and identify different substances within a test sample.

Chromatography is a methodology developed around 50 years ago, which provided an unparalleled separation power of a sample mixture and a great ease of use. It consists of two distinct phases: a stationary phase, which can be either solid or liquid, and a moving gaseous phase. The rate of interaction between analyte and stationary phase will define the degree of separation (or elution) of the compounds.

The eluted molecules are then introduced into the mass spectrometer where they are ionized, accelerated, deflected, and detected separately. This results in a spectrum of masses that are a “fingerprint” of the compounds present in the original test sample.

The GC-MS technique combines the best of the two instruments, providing the proper separation of compounds required by the mass spectrometer in order to avoid overlapping results, and mass spectrometry’s great identification power.

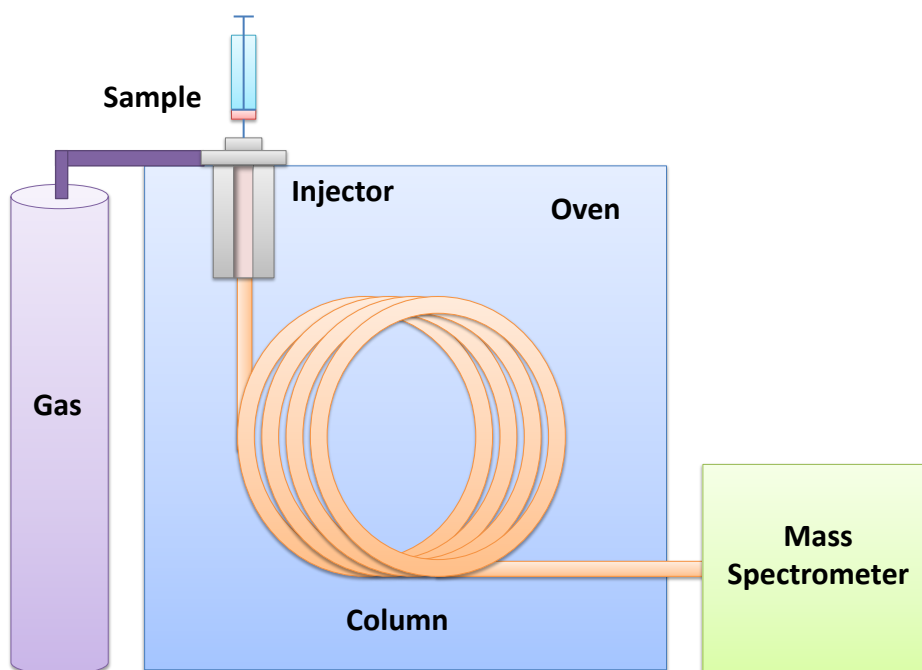


Figure B.1: Basic elements of a gas chromatograph

## B.2 Main Structure

Figure B.1 shows a diagram of a gas chromatograph. The main elements comprised in a gas chromatograph are: inlet, column, oven and detector, which in the case of GC-MS is the mass spectrometer itself.

### B.2.1 Inlet

The inlet is a key element in the gas chromatograph, since it is the portion with which the analyst interacts the most. For some columns sample injection a syringe can easily fit into the column. However, for most columns (i.e. capillary columns, which are explained in the next section), samples are injected into a chamber, vaporized and then transferred into the column in the vapor phase.

### B.2.2 Column

Columns are the “heart” of gas chromatography, since they are responsible for the separation of compounds. There are mainly two types of columns used in GC: the packed column, used for particular applications such as the analysis of fixed gases, and the capillary column, which is present in 90% of modern chromatographs.

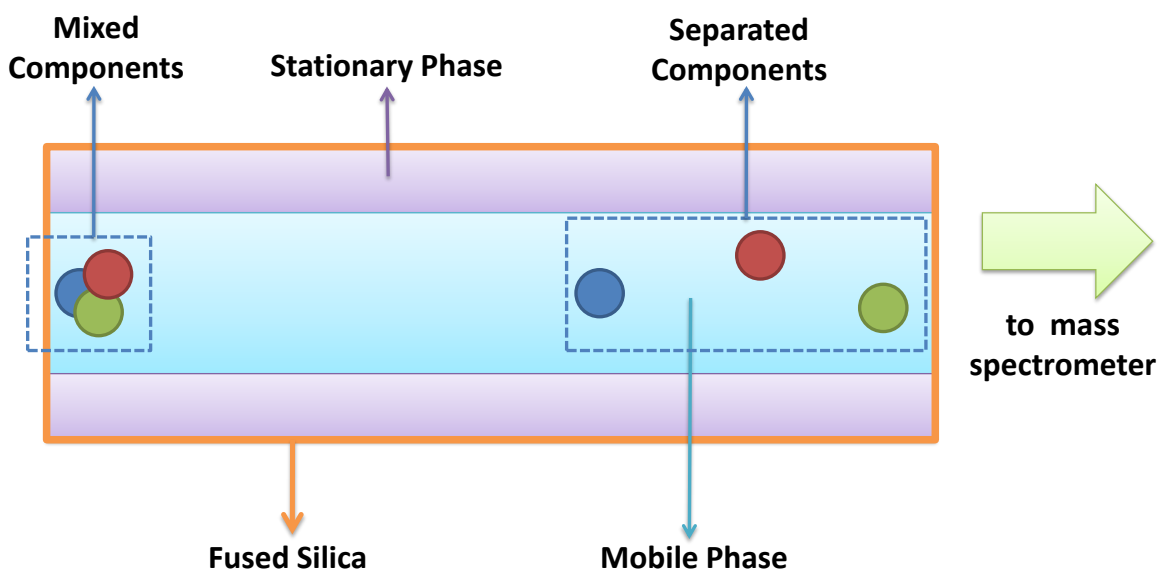


Figure B.2: Diagram of the interactions within a capillary column

As shown in Fig. 2, the different components in a mixture interact at different rates with the stationary phase. This results in different transit times across the length of the column, which produce an effective separation (or elution) of the compounds.

#### Packed Columns

Packed columns have an internal diameter between 2mm and 4mm and a length between 1m and 4m. These columns are internally packed with an adsorbent. Since the ability to separate is strongly dependant on the interaction between analyte and stationary phase interactions, many different packing materials are available. The tubing can either be made of glass or of stainless steel.

#### Capillary Columns

Capillary columns are the most commonly employed columns. Their length ranges from 10m to 100m and their diameter varies from  $100\mu\text{m}$  to  $500\mu\text{m}$ . They contain no packing materials. The stationary phase is coated on the internal wall of the column as a film  $0.1\mu\text{m}$  to  $5\mu\text{m}$ .



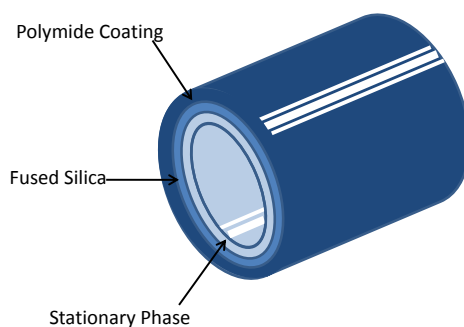


Figure B.3: Inner view of a capillary column

### B.2.3 Oven

The main user controlled variable in the entire setup is temperature. The column is contained in temperature controlled oven that operates between 5°C and 400°C, with an accuracy of around 0.1°C. It can also control the gradients with which temperature varies. The oven and column have low thermal masses in order to allow for rapid heating and cooling of the system.

### B.2.4 Detector (Mass Spectrometer)

Mass spectrometry identifies substances by electrically charging sample molecules, accelerating them through a magnetic field, breaking them up into charged fragments and finally detecting these charged pieces. Fig. B.4 shows the basic structure of a quadrupole mass spectrometer, which is one of the most common varieties of the device.

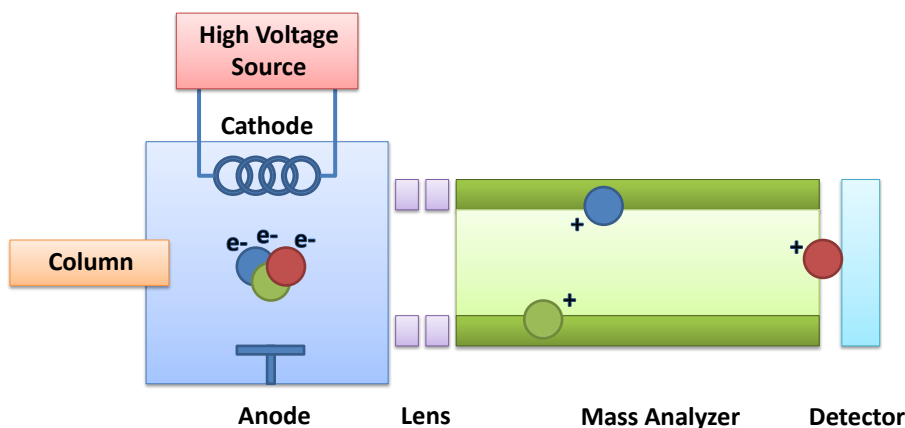


Figure B.4: Diagram of a mass spectrometer

The main elements of a mass spectrometer are: the inlet, which in the GC-MS case it is directly connected to the column; the electron impact ionizer, which ionizes sample molecules; the lens; the quadrupole analyzer; and the detector.

The mass spectrometer acts on the separated molecules that exit the column. These molecules are ionized by impacting them with an electron beam. The positive ions are accelerated by an electric field and then sorted by their mass to charge ratio ( $m/z$ ). The entire

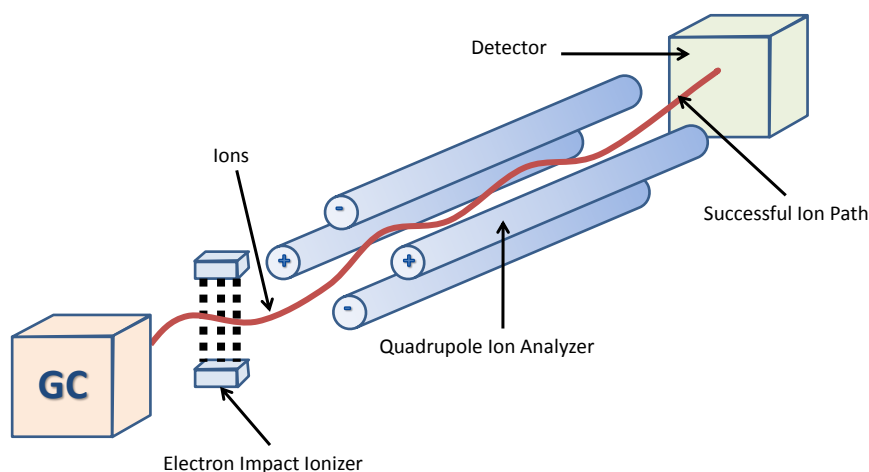


Figure B.5: 3D diagram of a mass spectrometer

process takes place in vacuum. Finally, the ions are detected and counted, and the results are digitally processed.

The output of the mass spectrometer is a plot of mass/charge ratios vs. abundances. This spectrum is characteristic of the particular substance under analysis. Therefore, by comparing the spectrum against an electronic database of thousands of plots, a technician may conclusively determine the identity of the compound.

### B.3 Data

Unlike traditional GC detectors, a GC-MS system produces a 3D dataset. Figure B.6 shows the output of the GC-MS. Plot a) shows a typical chromatogram, where the amplitude at each time represents the integration over the entire mass spectrum at that particular time. The area of each peak is proportional to the concentration of that substance. For each time point, there is a corresponding mass spectrum; an example of these can be seen in plot c). The color plot b) shows the entire dataset obtained from one measurement. Since chromatogram has an average duration of 35 minutes for breath samples and about 3 mass spectra are processed every second, about 7000 mass spectra are produced per sample. This means that a large amount of data must be processed to extract useful information from the measurement.

The GC-MS instrument produces output files of raw data, which in our case are Agilent \*.d files. These files contain both time and spectral information, along with instrument and configuration details. One of the advantages of GC-MS compared to other separation techniques is the wealth of existing mass spectral information. A special software from the manufacturer of the equipment can process the raw data and compare it against spectral libraries, providing identification of the compounds in the sample.

### B.4 Used Setup

Normally, breath samples are contained in sorbent tubes, which trap the volatile organic compounds. These tubes require to be thermally desorbed in order to release the VOCs into the chromatograph. In our setup, the tubes are taken by an autosampler, which enables a thermal desorption system (manufactured by Gerstel) to perform automatic processing of the samples.

The samples are heated and captured in a cold trap (also by Gerstel) in order to minimize band broadening. A capillary gas chromatograph (Agilent 6890N) is used, with a column 30m

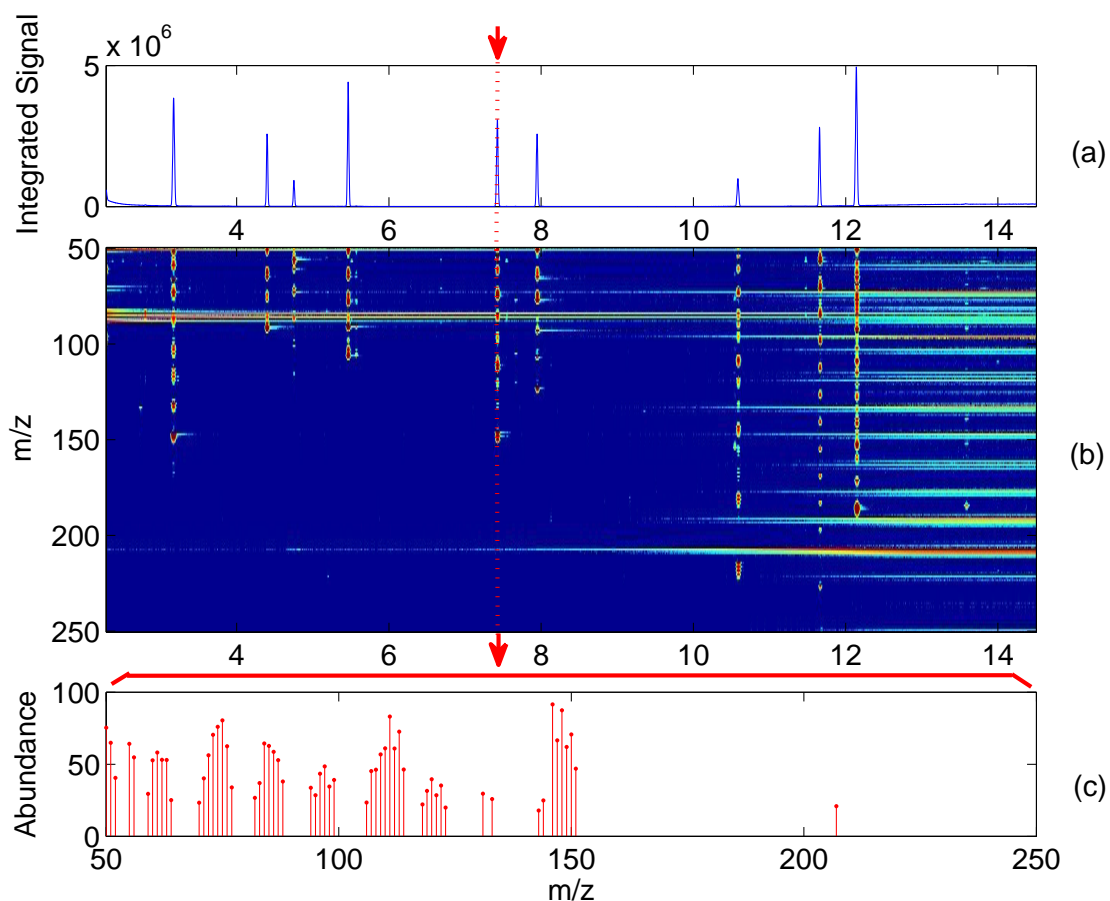


Figure B.6: Sample GC-MS results, including a chromatogram and one mass spectrum per data point

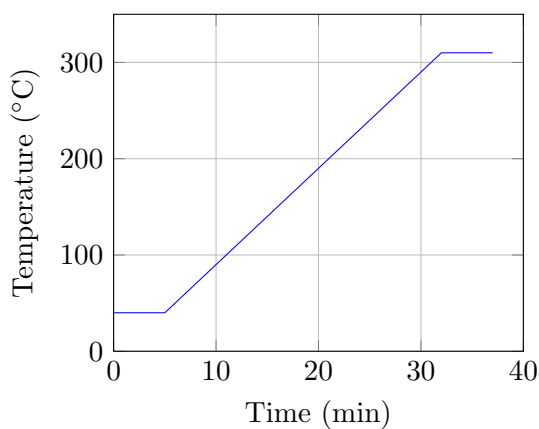


Figure B.7: Temperature cycle of the GC-MS oven

long and 0.25mm diameter, 100% dimethylpolysiloxane. The temperature cycle can be observed in figure B.7. The mass spectrometer (Agilent 5975 MSD) is used in electron ionization mode at 70eV, with a scan range of  $m/z$  29-450 Da.

Figure B.8 shows an image of the entire setup in the laboratory. The autosampler can be observed on top, together with the thermal desorption system (TDS). The large rectangular door is the oven, and the instrument on the right is the mass spectrometer.

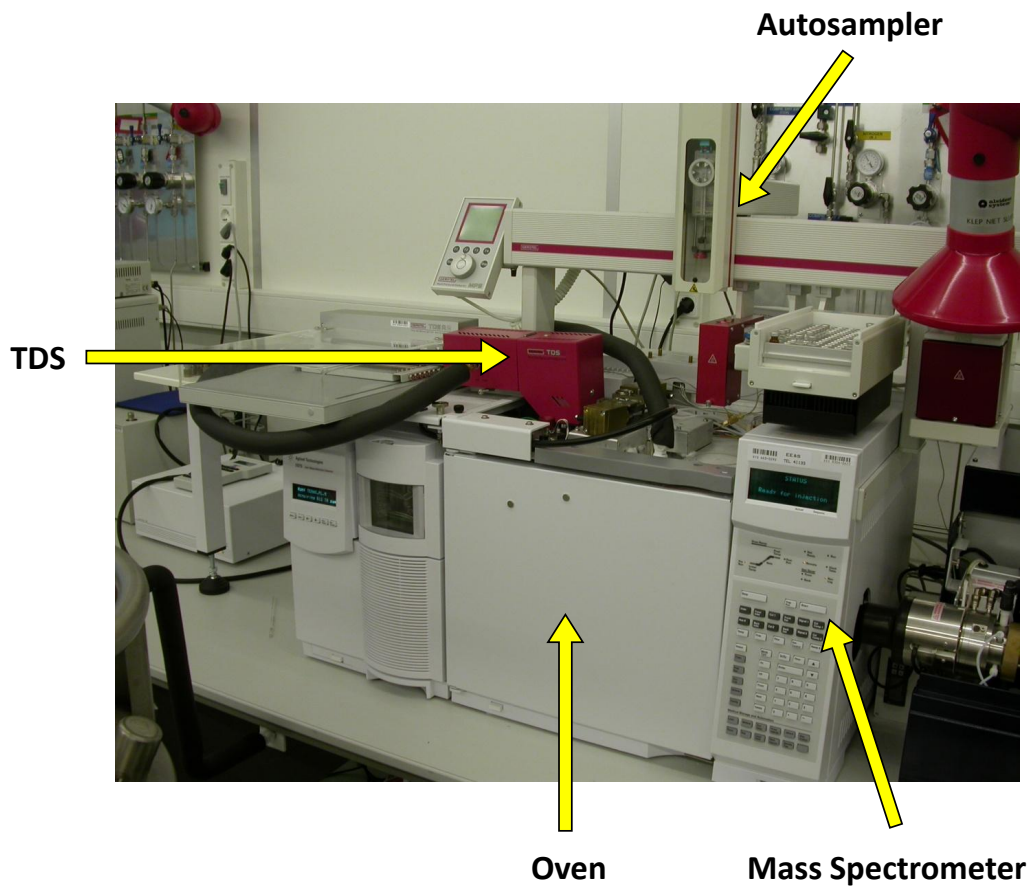


Figure B.8: GC-MS setup used for processing breath samples at MiPlaza

# C Gas chromatography-mass spectrometry data processing

## C.1 Introduction

Unlike traditional gas chromatography detectors, a GC-MS system produces a 3D dataset. For every point in time, there is a corresponding mass spectrum, so the three dimensions that compose the output are time, mass-to-charge ratio ( $m/z$ ) and abundance. A typical dataset can be observed in figure C.1. Since chromatogram has an average duration of 35 minutes for breath samples and about 3 mass spectra are processed every second, about 7000 mass spectra are produced per sample. This means that there is a large amount of data that must be processed in order to extract useful information from the measurement.

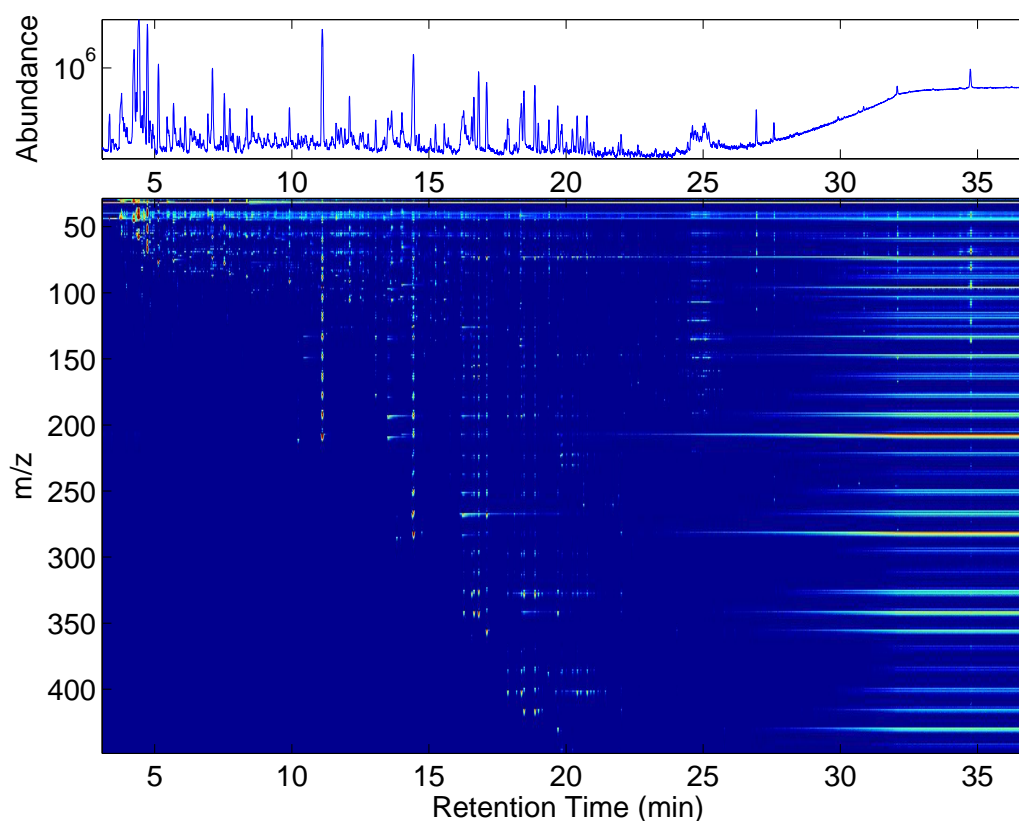


Figure C.1: Chromatogram and color plot of mass spectral information of a breath sample

In the following section, we describe the steps required to transform the 3D dataset into a list of compounds present in the sample, which may eventually be disease biomarkers.

## C.2 Processing Workflow

The processing of the GC-MS data can be divided into 5 steps, shown in figure C.2.

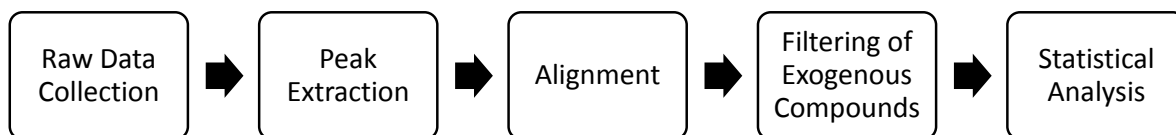


Figure C.2: Overview of the processing workflow for GC-MS data

### C.2.1 Raw Data Collection

Even though the instrument automatically generates the output files, which in this case are Agilent \*.d files, proprietary filetypes are a challenge. The fact that Agilent does not distribute the structure of their files and that converters in the market are quite expensive, seriously limits the possibilities for processing. In fact, only two pieces of software were found that could handle our raw data (AMDIS and Agilent's MassHunter), and they are discussed in a separate report.

### C.2.2 Peak Extraction

In a second step, features are extracted out of the raw data. This step implies the integration of all the mass spectra available for the creation of a TIC (total ion count) plot, which represents signal strength vs. time. The main objective at this stage is to locate all the peaks present in the chromatogram. However, some compounds may be coeluting and can be hard to distinguish from the chromatogram only. Therefore, to solve this situation, a deconvolution algorithm is applied. The algorithm attempts to discriminate compounds that are eluting at very close times and may even be indistinguishable to the naked eye.

Figure C.3 shows a typical peak extraction situation. The arrows above the chromatogram show the positions where a compound was found.

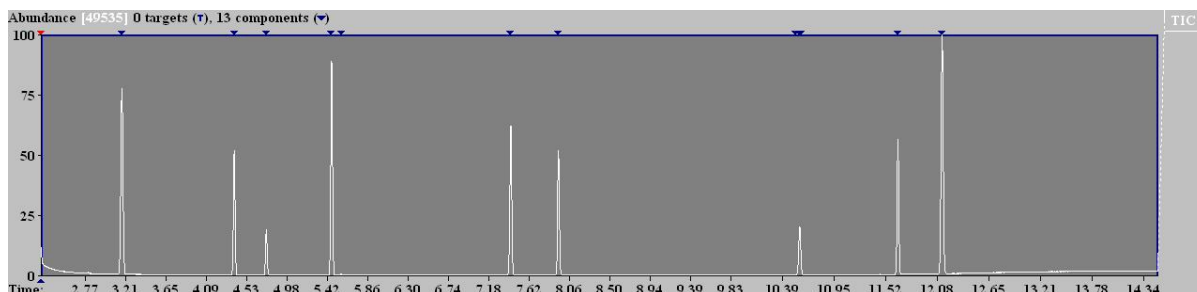


Figure C.3: Simple chromatogram and results of peak extraction stage

The result of this stage is a list of peaks, each representing a single compound, with their identifying data (retention time, height, area, etc.) and their corresponding mass spectra.

### C.2.3 Alignment

The same compound may elute at different retention times in different runs. This time difference may be very small if the measurements were done subsequently, while large delays in processing may lead to larger shifts. The main problem, however, is that these time shifts are not linear, so a compound at early retention times may have shifted a fraction of a second, while a compound at higher retention times may have moved a few seconds. This needs to be considered for data processing, since the retention time of a compound is not enough to confirm the identity of that compound. The truth about the identity of a peak always lies in its mass spectral information.

Figure C.4 shows an example of the time shift suffered by peaks in a breath sample. Each run was carried out one week after the previous, which caused the time shift to be larger than in samples that are processed in a narrow time window. The third sample (red) is around 0.05 min shifted towards lower retention times.

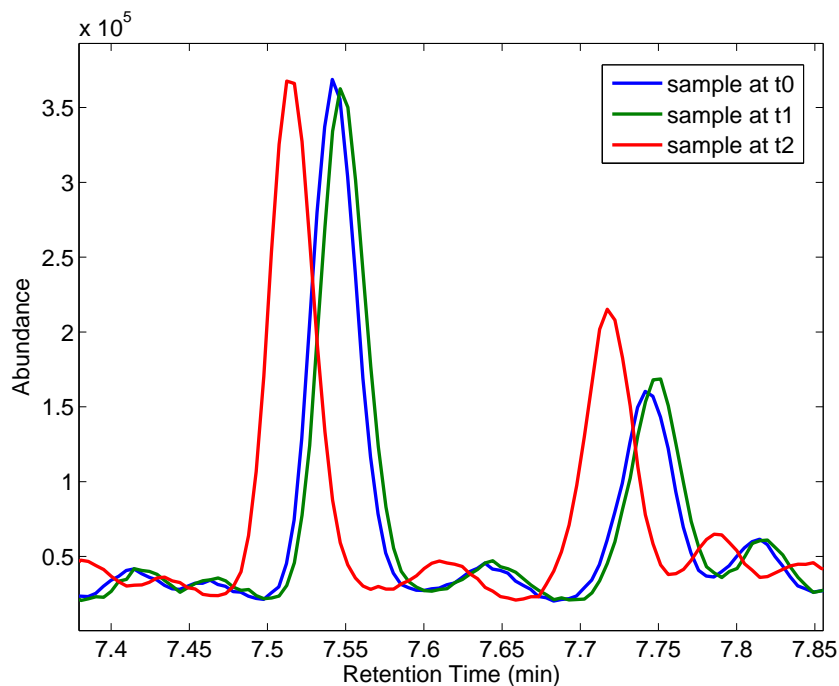


Figure C.4: Portion of a breath chromatogram showing time shifting of two peaks

Therefore, in order to make a study with different patient samples, it is necessary to “align” the compounds to make sure that we are comparing the same compound in each sample, no matter their retention time.

This can be done in two different ways. One possibility is to align the chromatographic curves, prior to peak extraction, and once they are aligned run the peak finding algorithm. However, these algorithms tend to be slow when dealing with multiple samples. Another possibility is to first run the peak extraction algorithm, and then work with the different peak lists. In this way, the dataset is smaller (a few hundred peaks per breath sample), and with certain considerations such as setting the maximum expected time shift, it is possible to align the lists fast and efficiently.

In our case, either the Agilent MPP software or our own Matlab code can perform the alignment by taking two user-defined parameters: the retention time window and the match factor. The retention time window is the maximum time shift expected in the data, which as we mentioned before is useful to speed up the alignment process. The match factor is a threshold used to determine the similarity of different peaks, by comparing their mass spectra. By working in this manner, we are independent of the nature of the shifts, which are generally non-linear.

#### C.2.4 Filtering of Exogenous Compounds

Once everything is processed, we may need to eliminate compounds that have a non-endogenous origin. Through various studies performed on test data, several setup-related compounds have been identified. For example, some compounds were found to be originated by the sampling

bag, such as phenol or N,N-dimethyl acetamide. These and many others should be removed since they have no value for identifying disease and may even confuse the classifier at a later stage.

In small studies, a statistical tool may believe some of these non endogenous compounds have some predictive value for a disease. However, if we have prior knowledge of their origin and remove them before running a statistical analysis, we can avoid drawing mistaken conclusions.

### C.2.5 Statistical Analysis

Finally, a statistical analysis is performed. There are many software alternatives for this stage. Agilent's software can rank the compounds by defining the sample groups, extracting the minimal number of peaks required for the classification and training a classifier.

On the other hand, there are better, more flexible alternatives available. It is possible to use a statistical toolbox such as Weka or Matlab itself, especially in early stages of development, since it provides much better possibilities for defining and training classifiers, as well as for analysing the available data.

## C.3 Conclusions

The aim of GC-MS data processing is not only to extract all information out of raw files, but to do so in a reliable manner, so as to improve the quality of research results from the start.

There is a large amount of data in any breath sample. However, it is important to ensure the quality of the biomarkers found. For instance, making sure that a marker is not actually setup-related, is essential if we want valid statistical conclusions for disease diagnosis.

This can only be achieved by optimizing every step of the process, improving extraction and alignment algorithms, and properly defining filtering steps.



# D Analysis of the behavior of AMDIS, MassHunter and Mass Profiler Professional software on test data

## D.1 Introduction

The analysis of gas chromatography-mass spectrometry (GC-MS) data requires preprocessing in order to effectively extract the information contained in the samples. This preprocessing includes an initial peak extraction step, run for every sample under analysis, which determines which peaks are indeed present, at what retention times and in what amount. Since slight variations in the position of the same peak in different samples can occur, then all identified peaks need to be aligned across all samples to make comparisons possible. Once all peaks are identified and are matched between different samples, then statistical analysis may be performed.

In order to test the capabilities of the software package that will later be used for the asthma and sepsis projects at Philips, a small experiment was planned.

Both projects, asthma and sepsis, propose to find and identify biomarkers required to correctly classify patients as healthy or ill. For this purpose, a software package developed by Agilent is evaluated. However, given the complexity of breath mixtures it is difficult to assess the quality of the processing by analyzing patient data.

For that reason, a small scale test is carried out with known data. In this way, several concerns can be analyzed, such as the effectiveness of the peak extraction, the feature finding procedures, the quality of the alignment of peaks and the statistical analysis conclusions.

## D.2 Procedure

The experiment consists of two different gas mixtures, each containing the following 9 compounds:

- HMDS (3.174 min)
- Toluene (4.402 min)
- Butylacetate (4.754 min)
- Ethylbenzene (5.467 min)
- Dichlorobenzene (7.426 min)
- Nitrobenzene (7.947 min)
- TCB (10.587 min)
- Hexadecane (11.657 min)
- Diphenylsulfide (12.142 min)

However, four compounds (toluene, butylacetate, dichlorobenzene and hexadecane) are present in twice the concentration in mixture 1 than in mixture 2.

There are 6 replicates for each mixture, giving a total of 12 sample files. All replicates are analyzed in a GC-MS instrument, obtaining their chromatograms and mass spectra.

The 12 files are processed with AMDIS software in order to perform feature extraction. Once this step is completed, the peaks detected are aligned across all different files and a statistical analysis is performed.

In order to assess the performance of the software under evaluation we introduce a few statistical parameters. Since in this experiment the original components of the mixture are known, the number of true positives, false positives and false negatives are known. However, given the nature of our data, the number of true negatives cannot be calculated. Thus only a few statistical measures can be applied. We define them as follows:

$$\text{Sensitivity} = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Negatives}} \quad (\text{D.1})$$

$$\text{Positive Predictive Value (PPV)} = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Positives}} \quad (\text{D.2})$$

$$\text{False Discovery Rate (FDR)} = \frac{\text{Number of False Positives}}{\text{Number of True Positives} + \text{Number of False Positives}} \quad (\text{D.3})$$

## D.3 Results

### D.3.1 AMDIS

The peak extraction procedure is performed with AMDIS software, which is a free program from the National Institute of Standards and Technology (USA). This software is intended for peak extraction and identification of GC-MS data from a variety of sources, including the Agilent GC-MS used at Philips MiPlaza.

The peak extraction process basically attempts to find all components (peaks) present in a certain mixture. For clearly defined peaks it is relatively simple, but the software also need to be able to make a distinction between coeluting peaks, combining its knowledge of time and spectral information. This is done by a deconvolution algorithm which can identify different peaks, with different mass spectra that overlap in time.

The setup of the software is essential to avoid the extraction of false peaks (false positives) and the proper finding of the present compounds (to make sure no compounds are missed). Of particular importance are the deconvolution parameter settings.

The component width was set to 20, since it is approximately the number of samples available per peak. The adjacent peak subtraction option is insignificant since it only exist for identification processes which we do not perform within this software. Instead of running a library search for all the compounds found in a breath sample, we only do so for the output biomarkers (the ones of interest), and in that case we identify them by using the raw data files using the library search feature in the Chemstation software, which is provided with the Agilent instrument. However, the three main parameters (resolution, sensitivity and shape requirements) are set to the low option. This is done to avoid finding too many compounds in a simple mixture (many false positives). Still, the sensitivity or the resolution could be improved for more complicated samples such as in the case of breath analysis, and this will need to be studied later on.

The resulting output of AMDIS is a list of 13 compounds on average per source file, with differing peak qualities. The retention times of these peaks are can be found in table D.1.

We used a color legend to classify the peaks found into three categories: true positive (from the original 9 compounds), true positive (resulting from contamination, which was not

intentionally added to the mixture but is definitely present) and false positive (peak detected where there is nothing present). The color legend is shown in figure D.1.

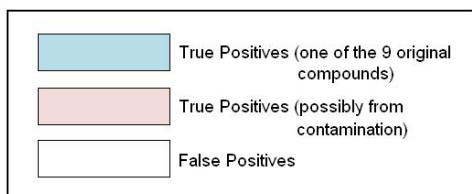


Figure D.1: Color legend for marking true and false positives

SAMPLE 1	SAMPLE 2	SAMPLE 3	SAMPLE 4	SAMPLE 5	SAMPLE 6
2.7994	2.2905	2.2906	2.2905	2.2912	2.2904
3.1724	3.1717	2.8010	2.8002	2.8034	3.1729
4.4006	3.1793	3.1733	3.1727	3.1721	4.4016
4.7532	4.4001	4.4015	4.4001	4.4006	4.7539
5.4659	4.7527	4.7543	4.7525	4.7534	5.4658
5.5683	5.4649	5.4663	5.4657	5.4651	5.5102
7.4245	5.5698	5.5719	5.5678	5.5728	5.5703
7.9473	7.4247	7.4246	7.4237	7.4237	7.4242
10.5401	7.9463	7.9479	7.9465	7.9465	7.9369
10.5845	10.5364	10.5707	10.5404	10.5824	7.9465
10.5999	10.5835	10.5838	10.5840	11.6550	10.5754
11.6563	10.5999	10.5907	10.5903	12.1380	10.5832
12.1398	11.6566	11.6559	11.6559		10.5928
	12.1393	12.1386	12.1386		11.6565
		12.1462			12.1402
					13.5798
SAMPLE 7	SAMPLE 8	SAMPLE 9	SAMPLE 10	SAMPLE 11	SAMPLE 12
3.1743	3.1763	2.8032	2.2908	3.1759	2.2915
4.4023	4.4030	3.1734	2.7450	4.4026	2.7434
4.7538	4.7548	4.4019	3.1735	4.7557	3.1752
5.4671	5.4667	4.7539	3.1808	5.4666	4.4026
5.5710	5.5724	5.4661	4.4017	5.5746	4.7549
7.4244	7.4238	5.5726	4.7538	7.4234	5.4669
7.9467	7.9463	7.4232	5.4665	7.9460	5.5748
10.5824	10.5831	7.9473	5.5706	10.5336	7.4237
10.5935	10.5931	10.5825	7.4243	10.5752	7.9455
11.6560	11.6549	11.6556	7.9466	10.5835	10.5360
12.1387	12.1391	12.1398	10.5841	11.6568	10.5830
13.5828		13.5778	10.5974	12.1398	11.6540
			11.6574		12.1380
			12.1406		

Table D.1: Compounds identified by AMDIS

Even though there were clearly more peaks identified than the original 9 compounds in the mixture (see table D.1), in the following stage in Mass Profiler Professional the quality of these peaks will be assessed and only reliable peaks will remain.

Table D.2 shows the results for peak identification, using a strict definition for true positive: only those compounds that were intentionally added to the mixture. In this calculation, only 108 true positives are expected (12 samples with 9 peaks each). However, it will obviously have the highest number of false positives, and thus a lower positive predictive value.

However, if we consider a broader definition of true positives and take into account those peaks that were not *intentionally* added to the mixture but are indeed present in the chromatogram, the total number of false positives will decrease. However, this would also lead to

		CONDITION	
		POSITIVE	NEGATIVE
OUTCOME	POSITIVE	108	50
	NEGATIVE	0	-

Sensitivity 100%  
 PPV 68.35%  
 FDR 31.65%

Table D.2: Statistical results of peak identification in AMDIS, with only compounds in the mixture considered as true positives

the appearance of false negatives, since these small contamination peaks are not always successfully identified. The results were recalculated using the broad definition of true positive and they are presented in table D.3.

		CONDITION	
		POSITIVE	NEGATIVE
OUTCOME	POSITIVE	133	25
	NEGATIVE	23	-

Sensitivity 85.2%  
 PPV 84.18%  
 FDR 15.82%

Table D.3: Statistical results of peak identification in AMDIS, considering contamination as true positives

### D.3.2 MassHunter

We run the same feature extraction process using Agilent’s MassHunter software. This software is not as intuitive as AMDIS, but gives more freedom to the user for parameter selection. Still, these parameter selection possibilities are poorly documented and the user must rely on the manufacturer’s suggestions.

The average compounds found per sample is significantly higher than with AMDIS, with an average of 19 compounds each. Thus, with a great number of false positives, the positive predictive value is lower. However, just like in the previous case, true positives were counted using a strict and a broad definition for them.

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
2.299	2.297	2.802	2.801	2.298	3.174
2.799	2.373	3.151	3.173	2.802	4.402
3.149	2.801	3.172	4.400	3.156	4.753
3.173	3.159	3.174	4.403	3.173	5.466
3.177	3.173	4.401	4.752	3.185	5.510
4.401	3.193	4.754	5.466	4.401	5.570
4.753	4.400	5.466	5.572	4.753	7.425
5.466	4.753	5.573	7.424	5.465	7.946
5.517	5.465	7.425	7.947	5.504	10.584
5.568	5.570	7.947	10.541	5.573	10.641
7.425	7.425	8.004	10.586	7.424	11.656
7.947	7.946	10.534	11.655	7.946	12.142
10.536	10.537	10.586	11.660	10.535	13.581
10.586	10.585	11.656	11.676	10.584	13.995
10.640	11.656	12.142	11.684	10.597	14.400
11.656	11.662	13.579	12.135	11.651	
12.140	11.692	13.782	12.140	11.655	
13.580	12.141	13.994	12.161	12.139	
13.988	13.578	14.445	12.188	12.193	
14.420	13.991		13.580	13.581	
	14.309		13.994	13.981	
			14.456	14.405	

Sample 7	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12
2.301	2.803	2.238	2.803	2.299	2.302
2.804	3.177	2.803	3.174	2.803	2.803
3.149	4.403	3.174	3.198	3.177	3.176
3.175	4.754	4.402	4.402	4.403	4.403
4.402	5.467	4.753	4.753	4.755	4.754
4.754	5.573	5.466	5.467	4.757	5.467
5.467	7.425	5.574	5.517	5.467	5.571
5.571	7.946	7.424	5.571	5.575	7.424
7.426	9.160	7.947	7.424	7.424	7.945
7.947	10.540	10.539	7.946	7.946	10.536
9.164	10.584	10.584	9.157	10.534	10.584
10.537	10.629	11.655	10.542	10.585	10.633
10.578	11.654	12.141	10.585	10.639	11.654
10.584	12.140	13.579	11.656	11.656	11.681
11.656	12.158	13.989	12.142	11.673	12.139
11.679	13.576	14.395	13.583	12.136	13.580
12.140	14.011		13.988	12.141	14.289
13.576	14.102		14.160	13.580	
13.989				13.981	
14.302				14.346	

Table D.4: Compounds identified by MassHunter

Table D.5 shows the results considering only the 9 compounds in the mixture as true positives. Given the high number of false positives under this definition, the positive predictive value is below 50%.

On the other hand, if contamination peaks are also considered true positives, since we cannot truly say that their identification is a software error, we obtain the results shown in table D.6.

Furthermore, these results cannot be improved any further considering that Mass Profiler has no quality score filtering for MassHunter data.

Still, the worst problem we have encountered with MassHunter is its variability when ex-

		CONDITION	
		POSITIVE	NEGATIVE
OUTCOME	POSITIVE	108	120
	NEGATIVE	0	-

Sensitivity 100%  
 PPV 47.37%  
 FDR 52.63%

Table D.5: Statistical results of peak identification in MassHunter, with only compounds in the mixture considered as true positives

		CONDITION	
		POSITIVE	NEGATIVE
OUTCOME	POSITIVE	154	74
	NEGATIVE	2	-

Sensitivity 98.72%  
 PPV 67.54%  
 FDR 32.46%

Table D.6: Statistical results of peak identification in MassHunter, considering contamination as true positives

tracking peak abundance.

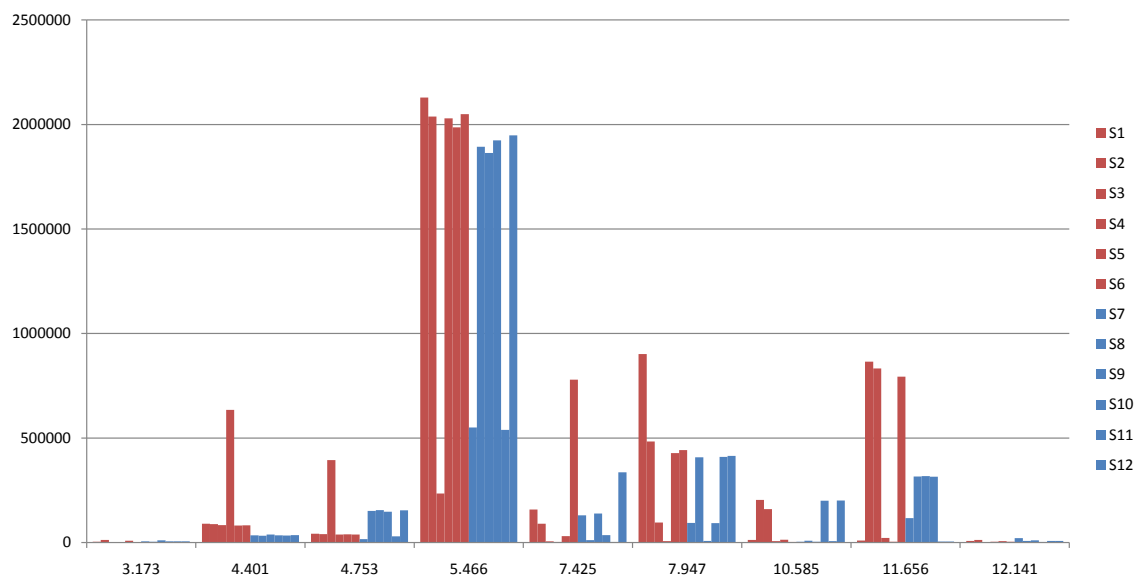


Figure D.2: Peak abundances as measured by MassHunter. Red and blue show the different mixtures

Even though it calculates peak area properly, agreeing with Chemstation results, the calculation for abundance and its results change depending on configuration. In our experiment, even when the configuration parameters were the same for analyzing all files, peaks belonging to the same mixture had different abundances measured.

The plot in figure D.2 shows how abundances vary for different compounds. We expected to find similar abundances within each group (red or blue), and the only four cases where differences should have been present between groups were for the components at 4.4, 4.75, 7.42 and 11.66 minutes. However, all cases show significant variations within mixtures in all types of compounds (with or without concentration change).

### D.3.3 Mass Profiler Professional

Mass Profiler Professional is a special software by Agilent that receives a set of samples in the form of peak lists as an input, runs an alignment algorithm and then allows the user to perform all sorts of filtering of the data. It has multiple data visualization tools and it also includes a statistical processing toolbox for finding features (peaks) of interest. Thus, it can go from raw peak lists to statistical conclusions of biomarkers of interest.

#### After AMDIS peak extraction

The choice of Mass Profiler parameters was also fundamental for the correct selection and alignment of the peaks found in the previous stage. The settings the user is allowed to filter compounds by their abundance (absolute or relative to the largest peak), by choosing only the first n largest compounds, by only considering peaks within a retention time window, by setting a minimum number of ions a peak should contain or by applying a quality score filter.

We chose to perform no abundance filtering since we did not want to remove identified peaks only on the basis of their abundance. The key parameter in this configuration is the choice of the minimum quality score which will help eliminate peaks that do not have a minimum level of confidence.

The definition of minimum quality score is explained in figure D.3.

**Minimum Quality Score Filter**

- You can set the Minimum Quality Score (default value = 40) to any value between 0 and 100. The Compound Quality Score Filter is only available for AMDIS experiments.
- If the Quality of any Compound is below the Minimum Quality Score then it would be filtered out.
- $Quality = a * MO + b * SNR + c * \text{Log}_{10}(\text{abundance})$ ; where MO is the number of model ions, SNR is the signal to noise ratio, and the coefficients (default value = 1, for each) a, b and c are configurable from Tools - Options - Configuration Dialog - Miscellaneous - AMDIS Compound Quality Score Parameters.

Figure D.3: Quality Score Filtering definition

In our case, the quality score was set to the default value of 40, as suggested by the manufacturer, given that no further documentation was available regarding the normalization of

such value (the selectable value ranges from 0 to 100).

For the alignment algorithm, the retention time window, i.e. the maximum time shift a peak is expected to undergo, was set to 0.1 min (more than enough for the current application which showed little time shifting) and the match factor was set to 0.6 in the alignment properties. The match factor is a threshold for the similarity score used by the alignment algorithm to compare two peaks. The similarity score is a weighted sum of factors such as retention time, abundance and mass spectra.

The output of the alignment stage showed 9 compounds were detected per sample file (exactly the number of substances in the mixture) and they were all completely aligned. Figure D.4 shows a scatter plot of retention time vs. principal mass (two elements that can almost certainly identify a compound) and their frequency. All points are plotted in yellow because they all have a frequency of 12, which means that every peak (characterized by its retention time and principal mass) was found 12 times in the 12 samples. This is exactly what we expected for the known mixtures. All the false positives disappeared by this stage, and only good quality peaks remained.

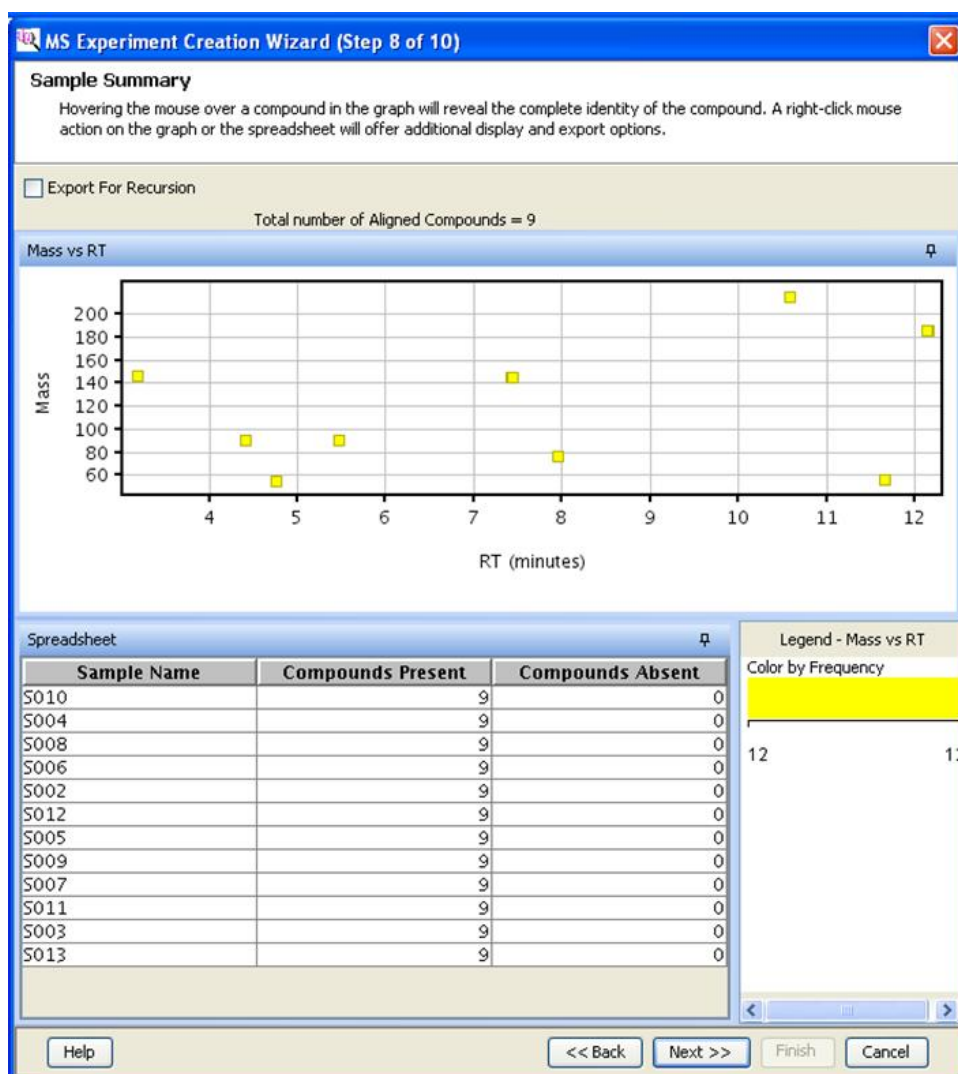


Figure D.4: Results of the alignment stage in Mass Profiler Professional

Another visualization is possible within Mass Profiler, and this is shown in figure D.5.



Every color (or letter) represents a given peak. Its abundance is plotted across all samples, though they are separated into two groups, one on the left and another on the right. It can be observed how for certain peaks the abundance shows no variation from one group to the other (for instance peaks a, b, c, f and h) while for others there is a clear change in abundance between groups (see peaks d, e, g and i).

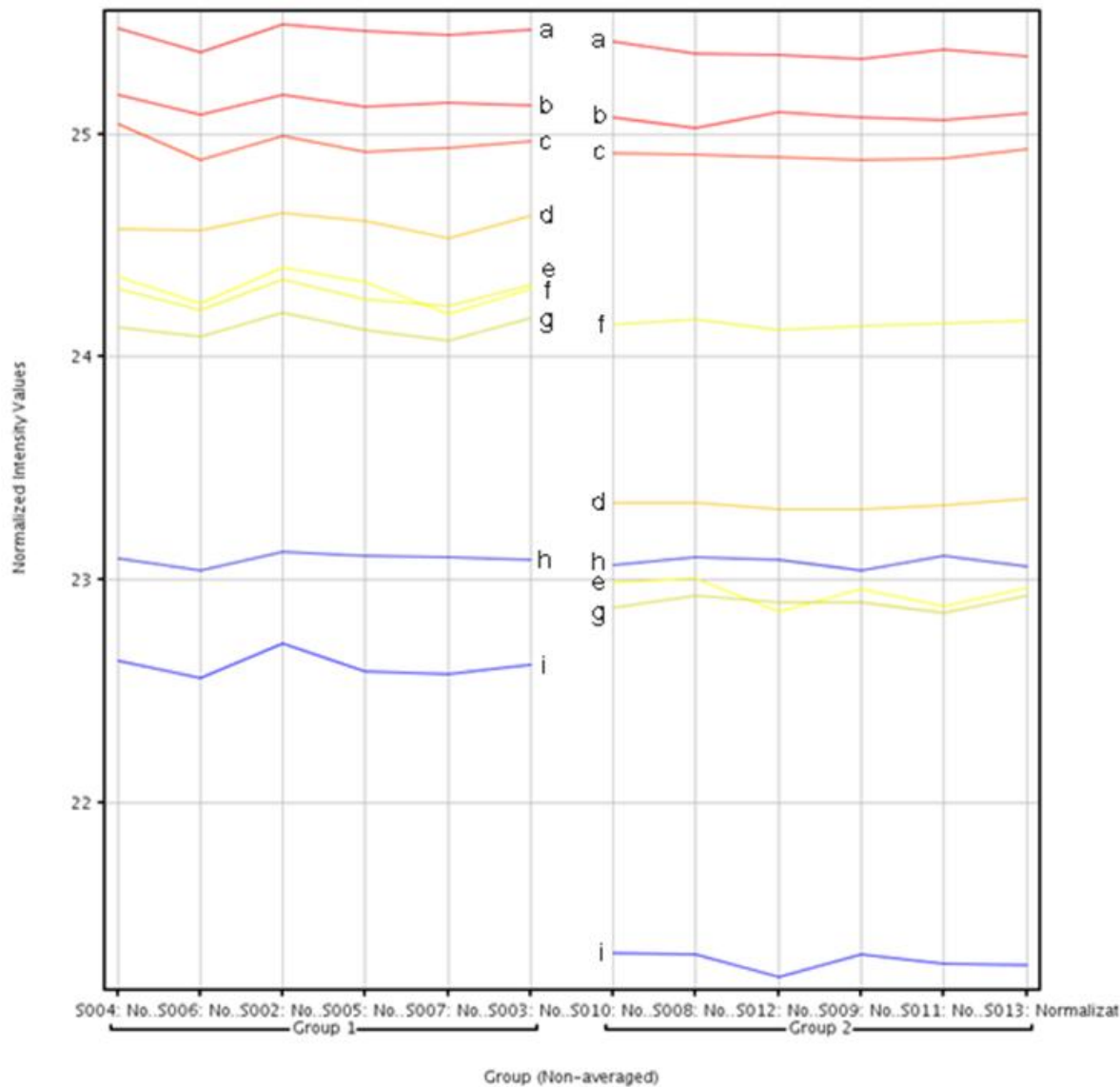


Figure D.5: Peak abundance across different groups

Results can also be seen in table form, containing the found compounds and their abundances in the different samples.

### After MassHunter peak extraction

We run the output files from MassHunter in Mass Profiler and performed the same process as in 3.3.1. However, the option quality score filtering is not available for this type of preprocessing, which means that any false positives found by the deconvolution software cannot be eliminated, especially if their abundances are high.

If we visualize the alignment process (fig. D.6), the results are different than in the previous case. Now, not all points appear in yellow (frequency = 12) but there are also points in red (frequency = 2). This means that on top of the true positives, many false positives still remain. Moreover, the peak at 10.585 min, which is a true positive, is only detected as present in 4 samples and the same happens for the peak at 12.14 min is missing in one sample.

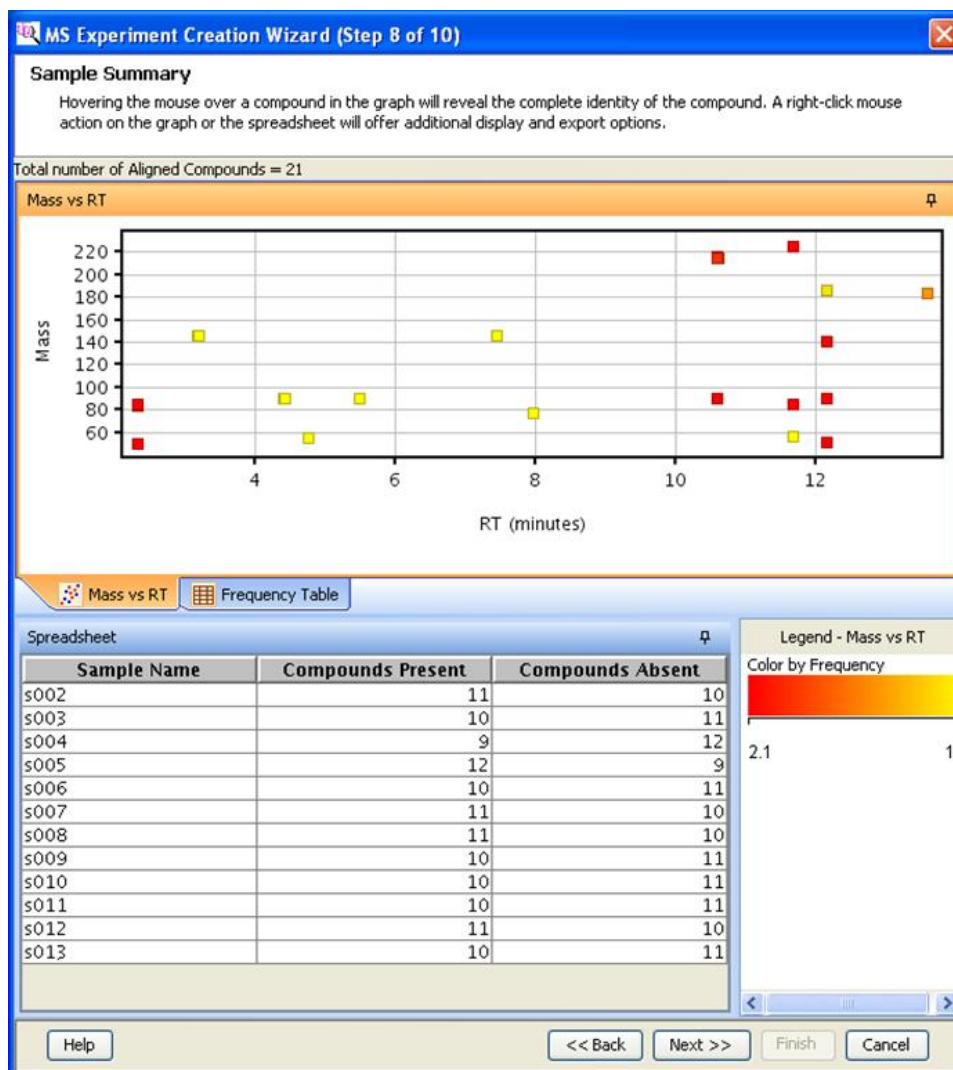


Figure D.6: Results of the alignment stage in Mass Profiler Professional

Furthermore, it can be observed in figure D.7 how the abundances are not as stable as the ones observed in the AMDIS experiment in figure D.5. Considering that samples within a group have the exact same concentrations of components in the mixture, the extraction can be considered poor given the high variation of abundance of any component within a group.

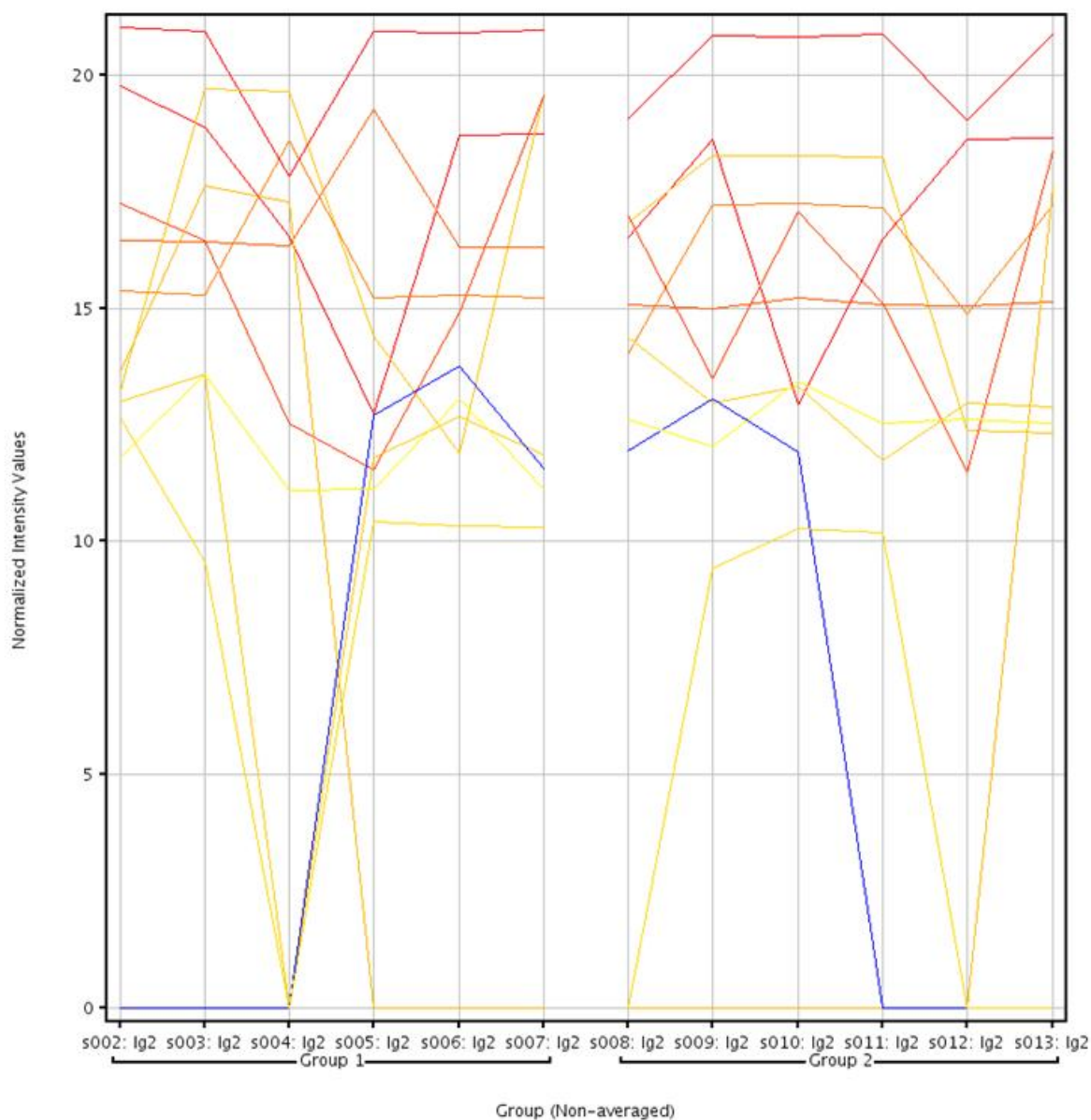


Figure D.7: Peak abundance across different groups

Peak abundance variations are so large, it is impossible to observe from the plot in figure 4 which are the peaks with different concentrations in the two mixtures. Not only the number of identified peaks is larger (only 12 compounds are plotted, but there are 21 in total) but the abundances are unreliable.

### D.3.4 WEKA

In order to check the success of the processing steps, the results were passed through a statistical toolbox. If everything was processed correctly, then the 4 compounds that changed their levels between the two groups should be identified as the ones with the largest discrimination power.

To perform the statistical analysis we used the files that were processed with AMDIS and MPP since they were the most accurate. Results preprocessed with MassHunter could not be considered reliable enough to continue with further processing.

Two ranking functions were used: InfoGainAttributeEval, which ranks each feature according to its individual prediction power, and SVMAttributeEval, which trains a support vector machine and ranks features according to the weights they were assigned by the classifier. These two ranking algorithms were chosen based on literature research, and may not be the optimal choice. Further studies should be performed to choose the best algorithm for the current application.

The results can be seen in tables D.7 and D.8.

average merit	average rank	attribute
0.995 +- 0.002	1.5 +- 1.5	6 dichlorobenzene
0.995 +- 0.002	2.2 +- 0.4	3 toluene
0.995 +- 0.002	2.8 +- 0.4	4 butylacetate
0.995 +- 0.002	4 +- 0	1 hexadecane
0.995 +- 0.002	5.2 +- 0.6	7 nitrobenzene
0.662 +- 0.111	6.1 +- 0.54	9 diphenylsulfide
0.662 +- 0.111	6.2 +- 1.78	2 HDMS
0 +- 0	8.3 +- 0.46	8 TCB
0.19 +- 0.29	8.7 +- 0.46	5 ethylbenzene

Table D.7: Attribute ranking by information gain algorithm

average merit	average rank	attribute
8.9 +- 0.3	1.1 +- 0.3	6 dichlorobenzene
7.7 +- 0.458	2.3 +- 0.46	3 toluene
7.4 +- 0.663	2.6 +- 0.66	4 butylacetate
6 +- 0	4 +- 0	1 hexadecane
4.9 +- 0.3	5.1 +- 0.3	7 nitrobenzene
3.6 +- 1.114	6.4 +- 1.11	2 HDMS
2.8 +- 0.748	7.2 +- 0.75	5 ethylbenzene
2.3 +- 0.64	7.7 +- 0.64	9 diphenylsulfide
1.4 +- 0.8	8.6 +- 0.8	8 TCB

Table D.8: Attribute ranking by SVM algorithm

The top four features are the same for both types of analysis, and they match the peaks that were present in different concentrations on the mixtures.

## D.4 Conclusions

AMDIS showed to be efficient in the extraction of peaks, with a sensitivity of 100% and a positive predictive value of 68% before the second filtering stage applied by MPP.

MassHunter on the other hand had a sensitivity of 100% but a positive predictive value of 47%. Besides, without the extra filtering stage in MPP (which is not available for this software) it cannot be improved. However, the worst situation with MassHunter is that the calculated abundances varied enormously within groups, which is unacceptable.

The combination of AMDIS and Mass Profiler Professional proved to be accurate in detecting and aligning peaks. Even though the AMDIS stage, working at the lowest resolution, found a number of false positives in the mixtures, the quality score filtering performed by Mass Profiler Professional completely removed them. All components were properly aligned, and the basic classification algorithm applied yielded the expected result: the four most powerful classifiers were the ones we knew were different in the two mixtures.

For the combination of MassHunter and MassProfiler were poor, with unstable and unreliable abundance values that lost the essence of the information contained in the samples. The abundance difference between the two groups for the four compounds with changed concentration was unclear.

Perhaps for the MassHunter the parameter settings are not ideal, but the parameters to be set are not transparent and user friendly. The results could not be improved, even after following the suggestions from Agilent's Tech Support.

Therefore for the analysis of GC-MS the combination of AMDIS and MPP seems to be the most reliable choice, considering the stable abundances and the disappearance of false positives thanks to the quality score filtering stage in MPP.

In the future, it might be of interest to carry out a similar small scale test, with a more complex known matrix, though not as complex as a typical breath sample. This way, the resolution capabilities of AMDIS could be tested, as well as the ability of MPP to align components that are closer in time.

# E Analysis of the repeatability of the GC-MS measurements using the 9 compound experiment data

## E.1 Introduction

For our study, it is essential to quantify the repeatability of our GC-MS measurements. In this analysis, we will only examine the repeatability of the measurement itself, without taking into account factors such as the setup, sample collection procedures, etc.

In order to perform this analysis, a standard mixture was prepared. The mixture contained 9 dissolved compounds, and its chromatogram can be observed in Figure E.1.

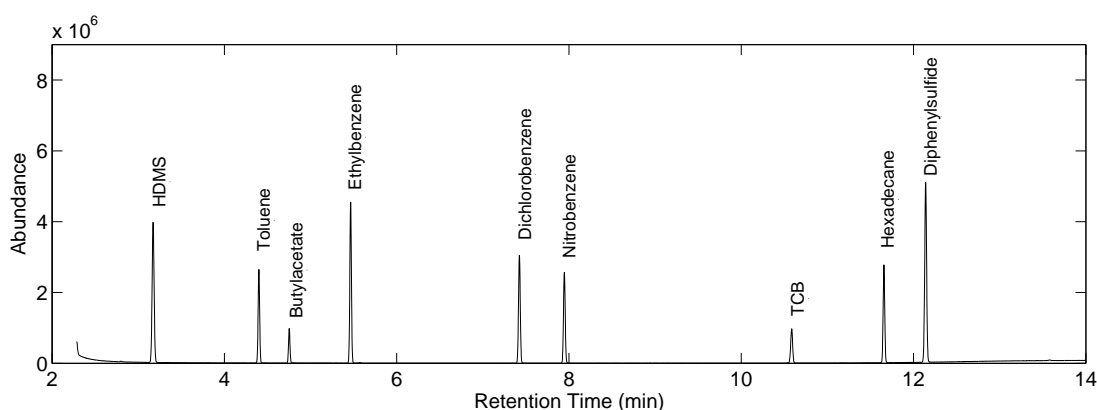


Figure E.1: 9 compound mixture used for repeatability testing

The sample was measured 6 times with the GC-MS. The analyst then measured the area and the retention time of each of the 9 peaks. The results are shown in the next section.

## E.2 Results

Tables E.1 and E.2 show the different peak area measurements for the 9 peaks. The average, standard deviation and relative standard deviation were calculated for each one.

The maximum relative standard deviation can be observed for hexadecane, at 5.30%. This is an acceptable value for the breath measurements, though it considers only the instrument and not the repeatability of the method.

Tables E.3 and E.4 show the retention time measurements and their variation. In this aspect, the instrument is expected to be more precise and with less deviation.

	Peak Area				
	HDMS	Toluene	Butylacetate	Ethylbenzene	Dichlorobenzene
1	36235864	19252502	6878658	33336618	25812672
2	36637964	18916004	6422838	32782286	26016309
3	35145391	18424430	6513922	33119767	24954436
4	36603740	18229633	6305572	31803450	25575427
5	35604348	17842600	6174296	30954019	24892497
6	36940287	17666168	6264802	32180927	24305584
Mean:	36194599	18388556	6426681	32362845	25259488
Std. Dev.:	688615	611729	251587	897541	650581
%RSD:	1.90	3.33	3.91	2.77	2.58

Table E.1: Abundances of HDMS, toluene, butylacetate, ethylbenzene and dichlorobenzene for all 6 measurements, calculated by a technician.

	Peak Area			
	Nitrobenzene	TCB	Hexadecane	Diphenylsulfide
1	21350717	8400226	22127731	43846959
2	20992463	8910714	20770632	46496060
3	20775759	8209585	21528092	45723424
4	20057437	9044669	21193561	46169432
5	19389743	8641744	19857550	43380622
6	19635614	8990952	19146705	45687642
Mean:	20366956	8699648	20770712	45217357
Std. Dev.:	788895	341061	1101380	1286109
%RSD:	3.87	3.92	5.30	2.84

Table E.2: Abundances of nitrobenzene, TCB, hexadecane and diphenylsulfide for all 6 measurements, calculated by a technician.

	Peak Area				
	HMDS	Toluene	Butylacetate	Ethylbenzene	Dichlorobenzene
1	3.175	4.4055	4.7605	5.4725	7.4285
2	3.175	4.4045	4.7575	5.4745	7.4315
3	3.174	4.407	4.7605	5.4735	7.4335
4	3.177	4.4025	4.762	5.4725	7.4285
5	3.173	4.4095	4.7585	5.4735	7.4285
6	3.175	4.4045	4.7615	5.4715	7.4295
Mean:	3.175	4.406	4.760	5.473	7.430
Std. Dev.:	0.0013	0.0024	0.0017	0.0010	0.0021
%RSD:	0.04	0.05	0.04	0.02	0.03

Table E.3: Retention times of HDMS, toluene, butylacetate, ethylbenzene and dichlorobenzene for all 6 measurements, calculated by a technician.

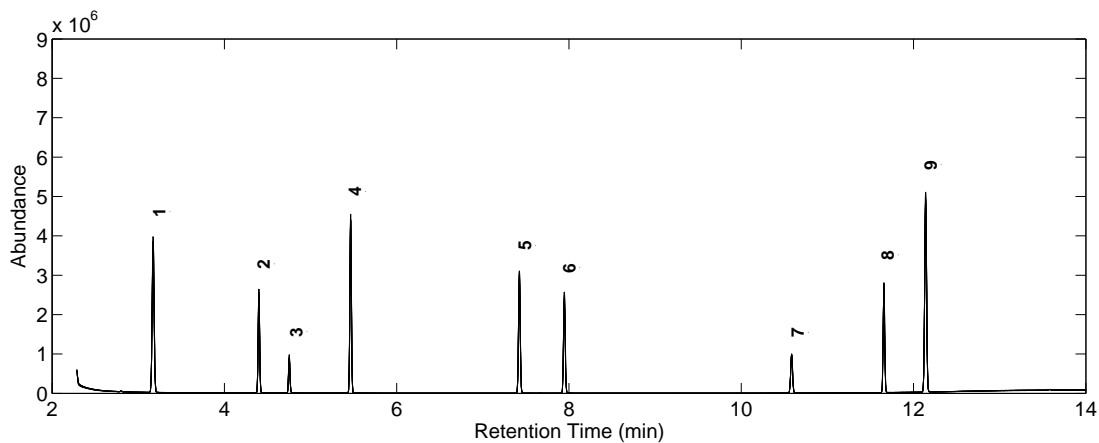
For retention times, the largest relative standard deviation was 0.05%. This is completely acceptable, given that any alignment software can easily handle such low deviations without any conflicts. These differences are absorbed by the alignment process and do not affect the analysis any further.

Figure E.2 shows an overlay of the chromatograms measured in the 6 runs. Visually, hardly any differences can be observed. Figures E.3 and E.4 zoom into the two peaks with the highest

	Peak Area			
	Nitrobenzene	TCB	Hexadecane	Diphenylsulfide
1	7.9545	10.592	11.659	12.141
2	7.9535	10.590	11.664	12.144
3	7.9535	10.592	11.659	12.140
4	7.9485	10.592	11.663	12.144
5	7.9480	10.592	11.661	12.138
6	7.9495	10.595	11.664	12.139
Mean:	7.951	10.592	11.662	12.141
Std. Dev.:	0.0029	0.0018	0.0023	0.0025
%RSD:	0.04	0.02	0.02	0.02

Table E.4: Retention times of nitrobenzene, TCB, hexadecane and diphenylsulfide for all 6 measurements, calculated by a technician.

relative standard deviation of area and retention time.



- |                    |                    |
|--------------------|--------------------|
| 1. HDMS            | 6. Nitrobenzene    |
| 2. Toluene         | 7. TCB             |
| 3. Butylacetate    | 8. Hexadecane      |
| 4. Ethylbenzene    | 9. Diphenylsulfide |
| 5. Dichlorobenzene |                    |

Figure E.2: Chromatographic overlay of 6 runs of testing mixture

The same data files were automatically analyzed by AMDIS. This way, it was also possible to consider the influence of the automatic software in the measurements, as opposed to the measurements performed by a technician (shown in tables E.1, E.2, E.3 and E.4), given that for processing breath all area calculations will be performed automatically.

AMDIS measurements are shown in tables E.5, E.6, E.7 and E.8.



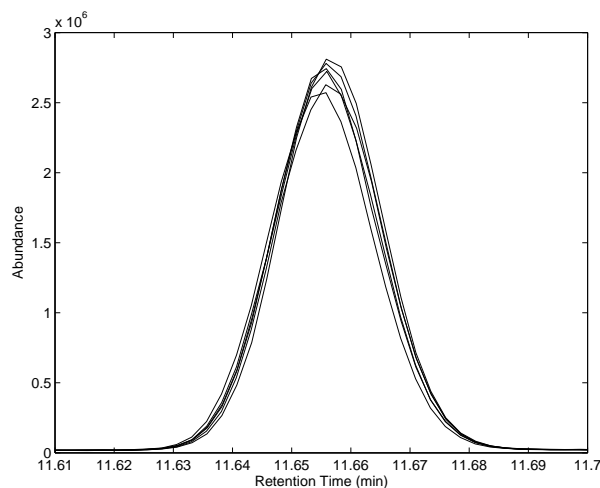


Figure E.3: Overlay of 6 runs for peak with worst area deviation, Hexadecane

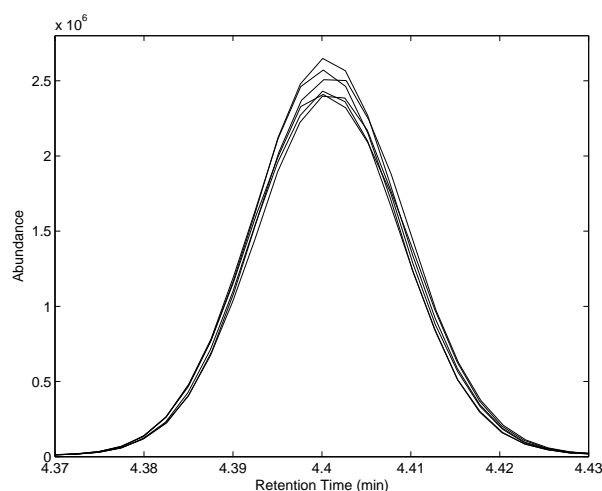


Figure E.4: Overlay of 6 runs for peak with worst retention time deviation, Toluene

	Peak Area				
	HMDS	Toluene	Butylacetate	Ethylbenzene	Dichlorobenzene
1	36235835	19252504	6878655	33336618	25812672
2	33675985	18914854	6422849	32782287	25772552
3	35145362	18424428	6513913	33119766	24954436
4	35181004	18227463	6215741	31803448	24968684
5	32764630	17842602	6174309	30954017	24892497
6	34266070	17666170	6271040	32180928	24308075
Mean:	34544814	18388004	6412751	32362844	25118153
Std. Dev.:	1235036	611644	261974	897542	577388
%RSD:	3.58	3.33	4.09	2.77	2.30

Table E.5: Abundances of HDMS, toluene, butylacetate, ethylbenzene and dichlorobenzene for all 6 measurements, calculated by an automated algorithm.

Again, the maximum relative standard deviation for the integrated signal is 5.30%. The automatic calculation neither introduced nor reduced the average abundance deviation. However, significant improvements can be observed in the retention time deviations, where the maximum

	Peak Area			
	Nitrobenzene	TCB	Hexadecane	Diphenylsulfide
1	21350717	8400226	22127731	43846959
2	20567710	8607208	20770632	43024452
3	20775759	8209589	21528092	45723424
4	20052628	8386089	21193561	43727654
5	19389743	8661539	19857551	43384375
6	19635614	8802163	19146705	42361214
Mean:	20295362	8511136	20770712	43678013
Std. Dev.:	739332	216987	1101380	1136841
%RSD:	3.64	2.55	5.30	2.60

Table E.6: Abundances of nitrobenzene, TCB, hexadecane and diphenylsulfide for all 6 measurements, calculated by an automated algorithm.

	Peak Area				
	HMDS	Toluene	Butylacetate	Ethylbenzene	Dichlorobenzene
1	3.1724	4.4006	4.7532	5.4659	7.4245
2	3.1717	4.4001	4.7527	5.4649	7.4247
3	3.1733	4.4015	4.7543	5.4663	7.4246
4	3.1727	4.4001	4.7525	5.4657	7.4237
5	3.1721	4.4006	4.7534	5.4651	7.4237
6	3.1729	4.4016	4.7539	5.4658	7.4242
Mean:	3.173	4.401	4.753	5.466	7.424
Std. Dev.:	0.0006	0.0007	0.0007	0.0005	0.0004
%RSD:	0.018	0.015	0.014	0.010	0.006

Table E.7: Retention times of HDMS, toluene, butylacetate, ethylbenzene and dichlorobenzene for all 6 measurements, calculated by an automated algorithm.

	Peak Area			
	Nitrobenzene	TCB	Hexadecane	Diphenylsulfide
1	7.9473	10.5845	11.6563	12.1398
2	7.9463	10.5835	11.6566	12.1393
3	7.9474	10.5854	11.6559	12.1386
4	7.9465	10.5840	11.6559	12.1386
5	7.9465	10.5824	11.6550	12.1380
6	7.9465	10.5832	11.6565	12.1402
Mean:	7.947	10.584	11.656	12.139
Std. Dev.:	0.0005	0.0010	0.0006	0.0008
%RSD:	0.006	0.010	0.005	0.007

Table E.8: Retention times of nitrobenzene, TCB, hexadecane and diphenylsulfide for all 6 measurements, calculated by an automated algorithm.

relative standard deviation is 0.018%.

### **E.3 Conclusions**

The maximum standard deviation of the measured area was 5.30%, calculated by both a technician and a computer. In the case of retention time, it was 0.05% for the human operator and 0.018% for the computer measurements.

The addition of the peak extraction software improves the repeatability for retention times, and has a similar performance as manual measurements for the integrated signal.

Though these results are optimistic because they show a high repeatability of GC-MS measurements, including the automatic computer calculations, further testing is required to calculate an overall repeatability of the entire breath collection and measurement procedures.

# F Analysis of pilot study data

## F.1 Introduction

In order to further understand some observations made on asthma data, a series of experiments were planned.

Firstly, it was intended to explain the origin of the numerous siloxanes showing up in breath measurement. The main suspicion was that they were somehow setup-related and not endogenous. Since no tubing in the setup contained silicones, a possibility was that water could be interacting with the GC column and releasing the siloxanes. For this purpose, we required to study the effects of water on different samples.

Moreover, it was also necessary to find out whether water could also affect the detected intensities of different components.

Furthermore, it was planned to study the effects of storage time on conditioned tubes. Since tubes are conditioned at Philips MiPlaza, and are then transported to the hospitals with a certain delay, it was essential to know that the quality of the conditioning was not lost with the passing of time.

## F.2 Methods

A total of 7 experiments were carried out. These are described in the following list.

- Conditioned tubes with dry nitrogen  
3 tubes were conditioned. Dry nitrogen was flowed through them, and they were later inserted into the GC-MS.
- Conditioned tubes with wet nitrogen  
3 tubes were conditioned. Humidified nitrogen (at 100% relative humidity) was flowed through them, and they were later inserted into the GC-MS.
- Conditioned tubes with a mixture of known VOCs and dry nitrogen  
3 tubes were conditioned. Dry nitrogen passing through 2 permeation tubes (sources of known VOCs) was flowed through them, and they analysed with the GC-MS.
- Conditioned tubes with a mixture of known VOCs and wet nitrogen  
3 tubes were conditioned. Wet nitrogen (100% relative humidity) passing through 2 permeation tubes was flowed through them, and the analysis was performed.
- Conditioned tubes with breath  
3 tubes were conditioned. Breath was collected in a Tedlar bag. 3 500ml samples were extracted out of the bag, and each was flowed through the tubes. They were then processed in the GC-MS.
- Conditioned tubes stored for 14 days with dry nitrogen  
3 tubes were conditioned and stored. Dry nitrogen (was flowed through them, and they were later inserted into the GC-MS.
- Analysis of the ventilator air at the ICU  
Samples were taken of the compressed air that is given to patients at the ICU, and air was also sampled from the different ventilators available.

## F.3 Results

### F.3.1 Dry Nitrogen

Figure F.1 shows an overlay of the three measurements. Even though they are all supposed to be blanks, sample 1 shows a higher background signal. It shows many more peaks than the other two, which match perfectly with each other. Still, if we compare this value with the column bleed, the extra peaks are around the noise level.

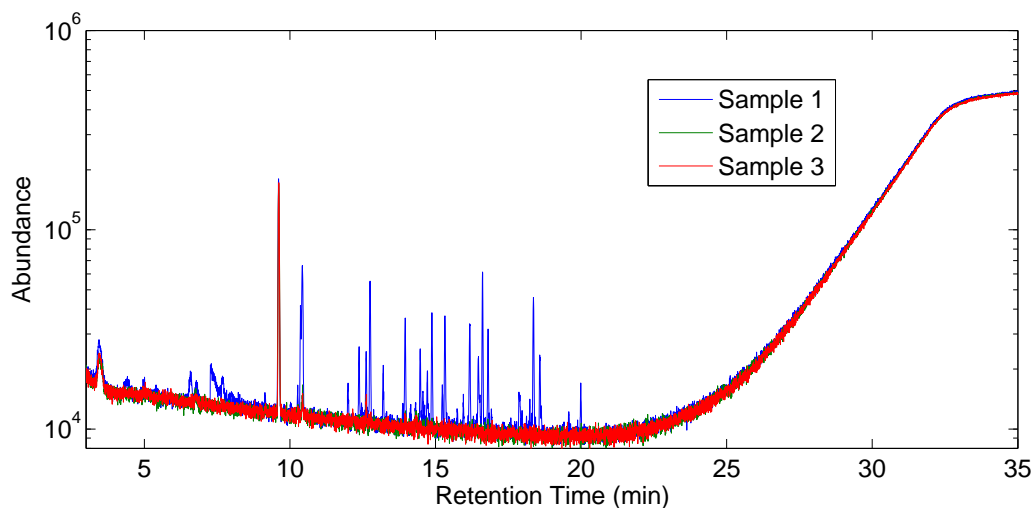


Figure F.1: Overlay of 3 runs of dry nitrogen

The level of the toluene standard added to the mixture is stable for all three.

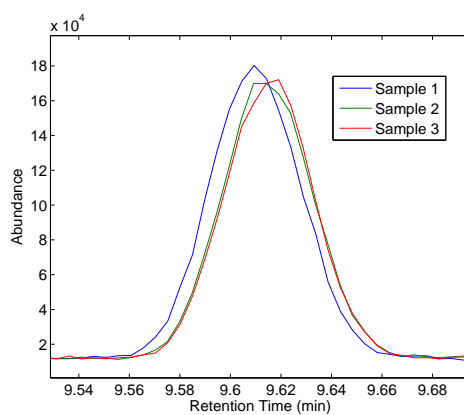


Figure F.2: Zoom into the toluene peak for the 3 measurements

Furthermore, as figures F.3(a) and F.3(b) show, sample 1 appears to have higher levels of the two siloxanes than the other two samples.

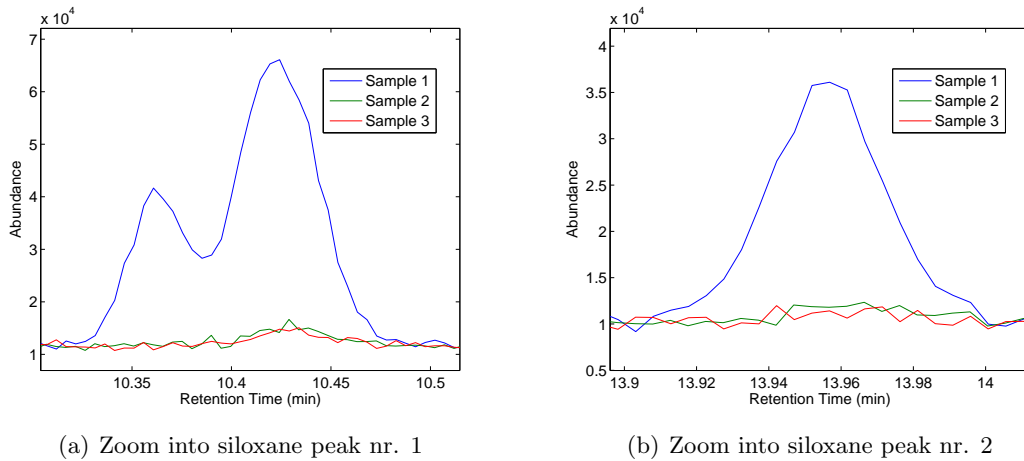


Figure F.3: Zoom into 2 siloxanes for the 3 measurements

However, when we running the Matlab software with quality filtering, none of the noise peaks remain. Figure F.4 shows that the only one compound found by the tool in all three samples is toluene.

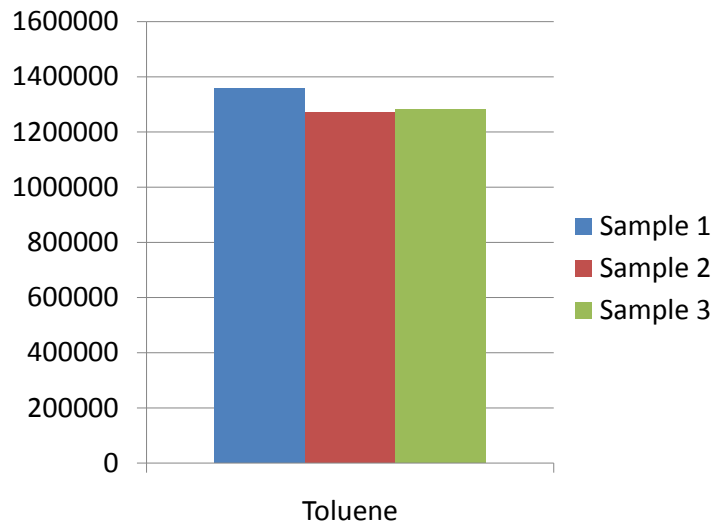


Figure F.4: Results obtained by processing the dry nitrogen samples with the Matlab tool

### F.3.2 Wet Nitrogen

In this experiment something went wrong. This can be observed in figure F.5, where the 3 curves do not match each other for the lower retention times. Samples 1 and 2 have a peak around 4.5 min which is in all likelihood an artifact, given that it does not even match between the two of them. Figure F.6(b) shows a closer view of this artifact.

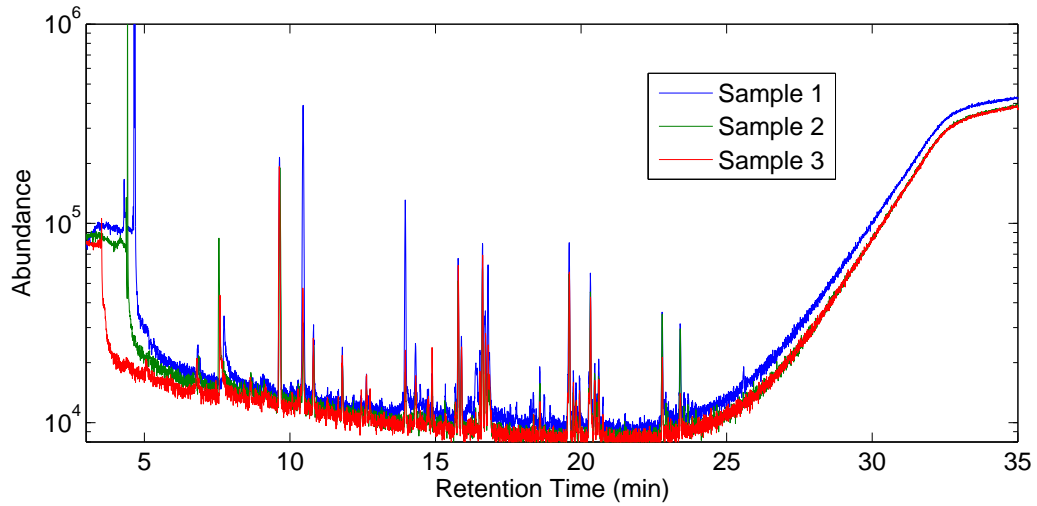
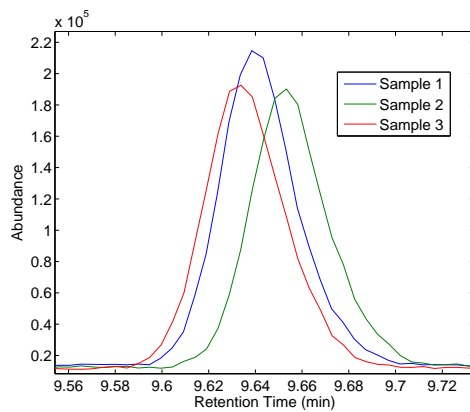
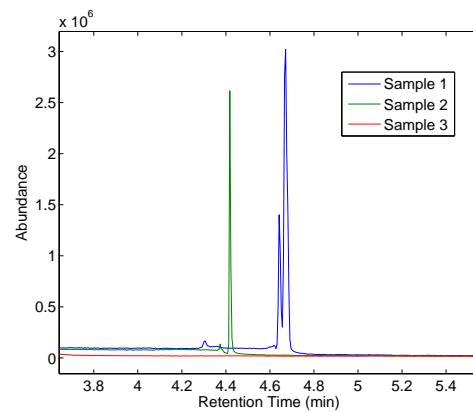


Figure F.5: Overlay of 3 runs of wet nitrogen

The toluene standard is matched well, though a slight drift in time can be observed.



(a) Zoom into the toluene peak



(b) Zoom into artifact in low retention times

Moreover, Sample 1 shows much higher levels of the two siloxanes than the other samples.

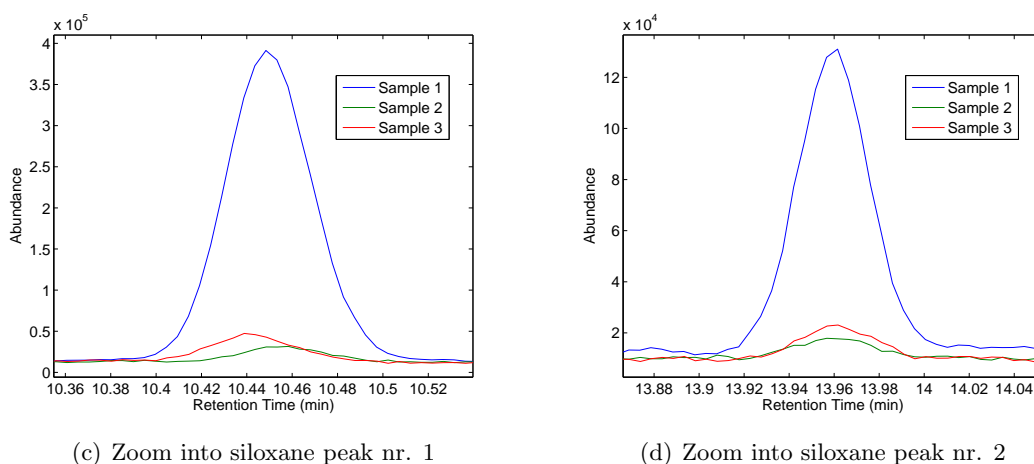


Figure F.6: Zoom into 2 siloxanes for the 3 measurements

The software was unable to overcome the matching problems. However, the chemist in the project mentioned that when a case like this is found, it is usually discarded, so there is no expectation for the software to work with poor data. Nonetheless, the toluene standard which was the only reliable peak in the entire measurement was properly detected and aligned by our tool.

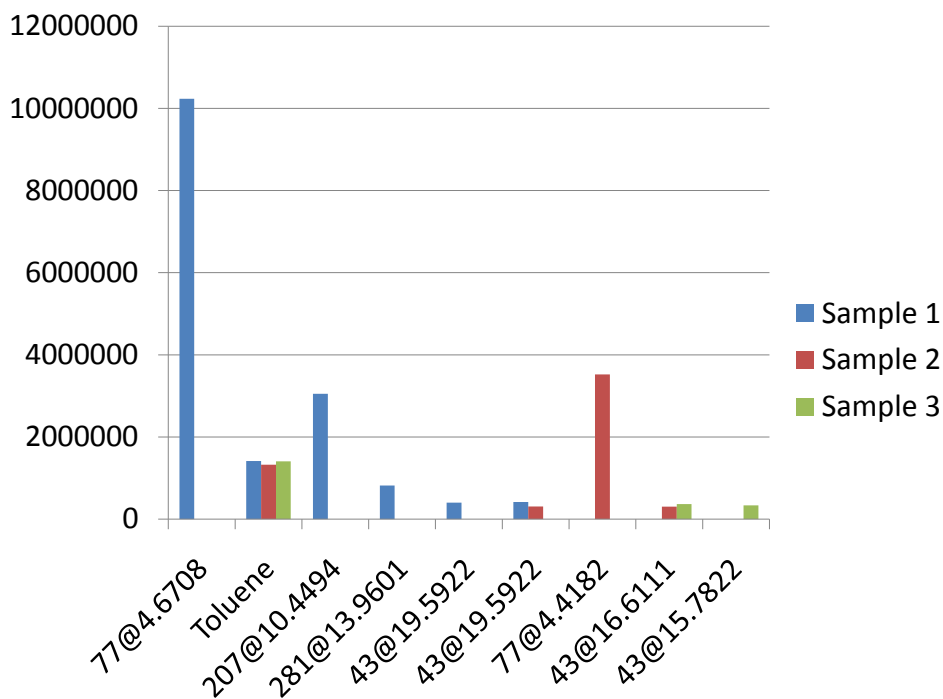


Figure F.7: Results obtained by processing the wet nitrogen samples with the Matlab tool



### F.3.3 Dry VOCs

All the samples showed good repeatability. If we observe closely the toluene peak and the two VOCs, they coincide almost perfectly both in abundance and in retention time.

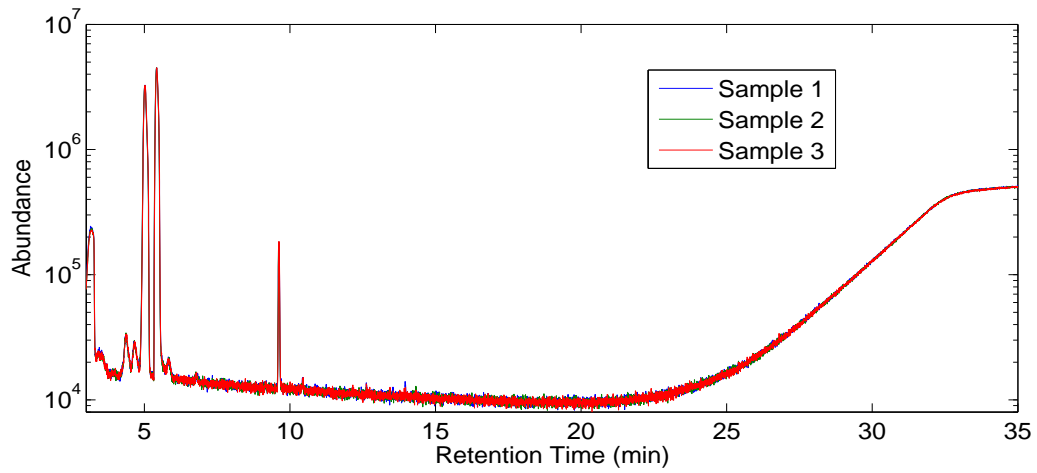


Figure F.8: Overlay of 3 runs of a dry mixture of known VOCs

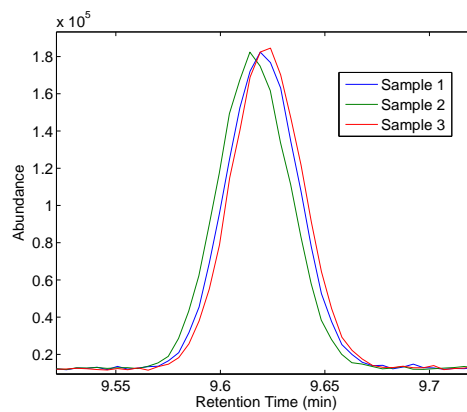
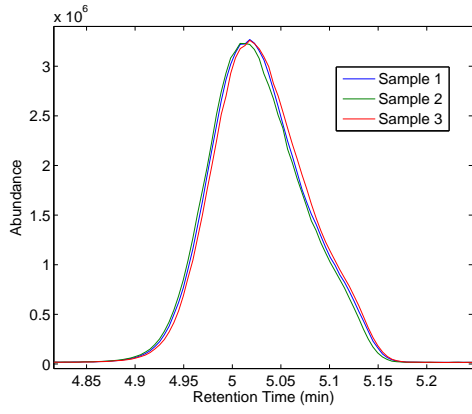
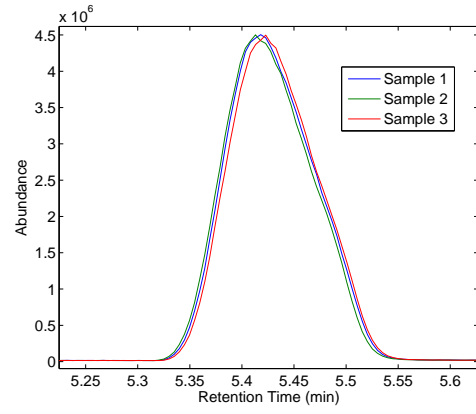


Figure F.9: Zoom into the toluene peak for the 3 measurements



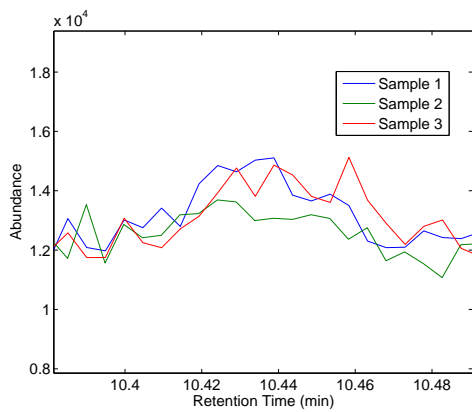
(a) Zoom into Hexane



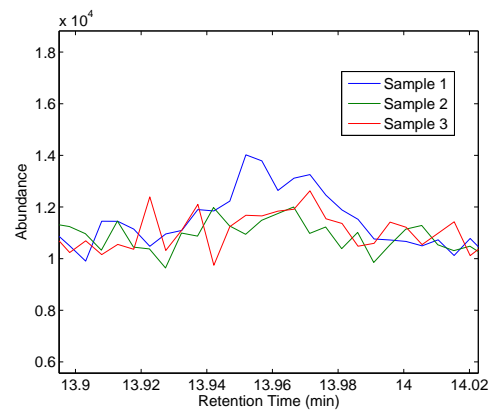
(b) Zoom into Ethylacetate

Figure F.10: Zoom into the two VOCs added to the mixture

No siloxanes could be observed above the baseline, and figures F.11(a) and F.11(b) zoom into the retention time area where they would normally be found.



(a) Zoom into siloxane peak nr. 1



(b) Zoom into siloxane peak nr. 2

Figure F.11: Zoom into 2 siloxanes for the 3 measurements

When processed with AMDIS and the alignment software only the three expected peaks (toluene and the two VOCs) are detected and aligned.

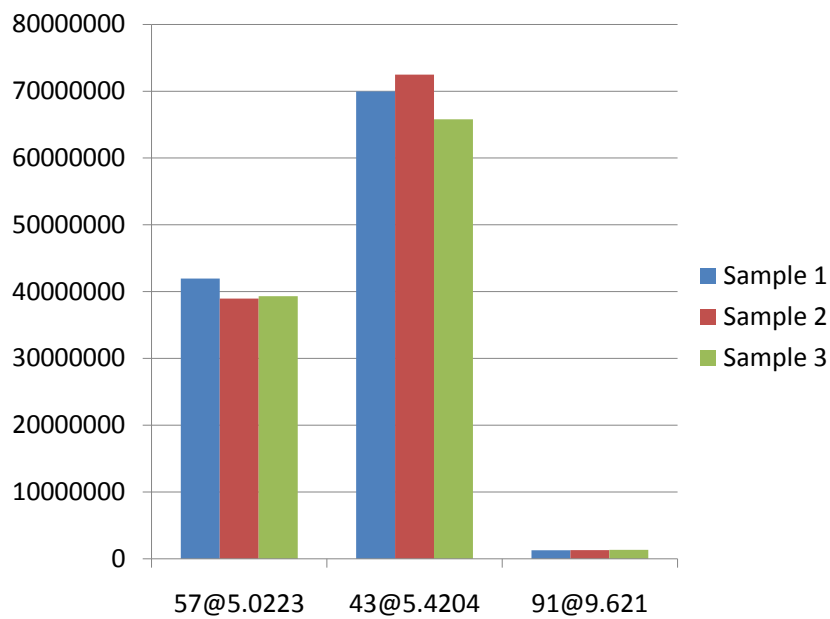


Figure F.12: Results obtained by processing the known dry VOCs samples with the Matlab tool

### F.3.4 Wet VOCs

The three samples also show good matching in the wet case. In figure F.13 the two VOCs are very clear peaks, and the toluene standard is also visible.

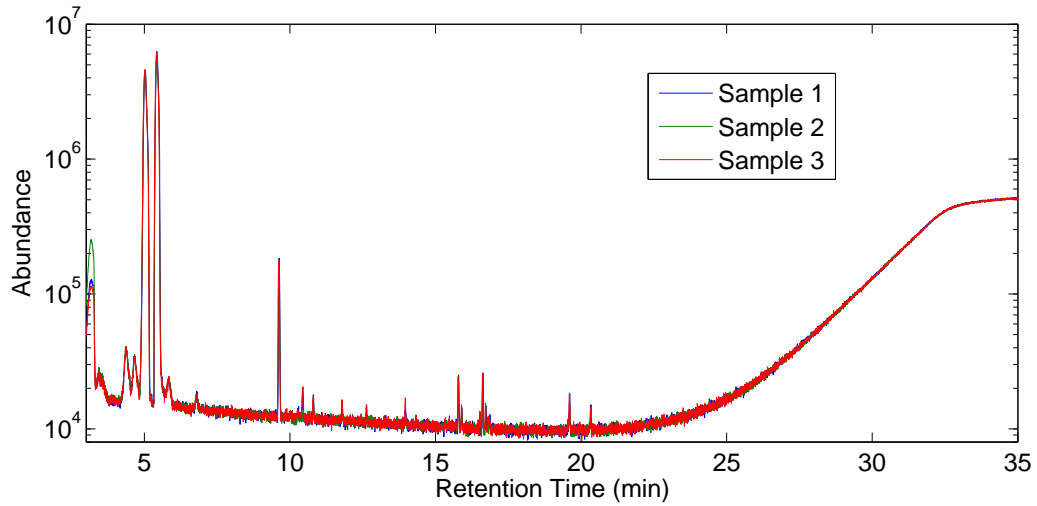


Figure F.13: Overlay of 3 runs of a wet mixture of known VOCs

If we look closely at the toluene peak, water does not seem to have affected its detection. However, a better comparison against the dry case is performed later on.

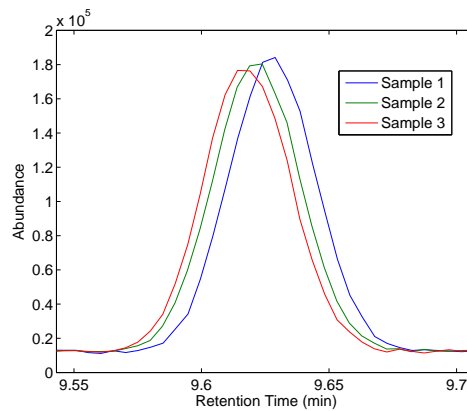
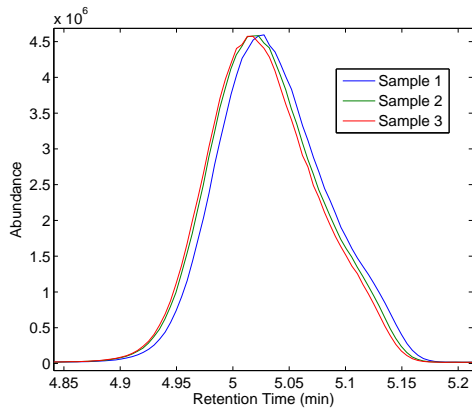
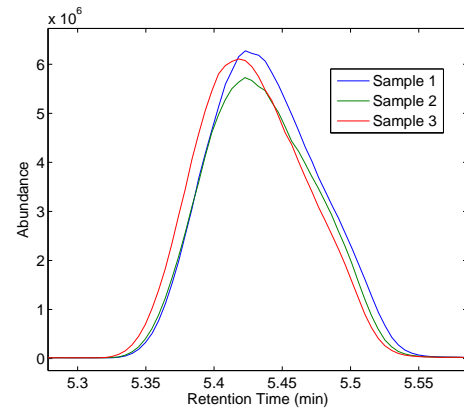


Figure F.14: Zoom into the toluene peak for the 3 measurements



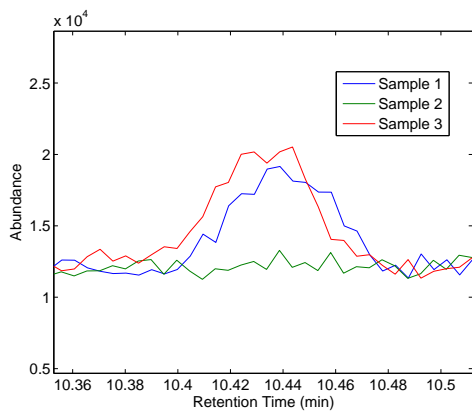
(a) Zoom into Hexane



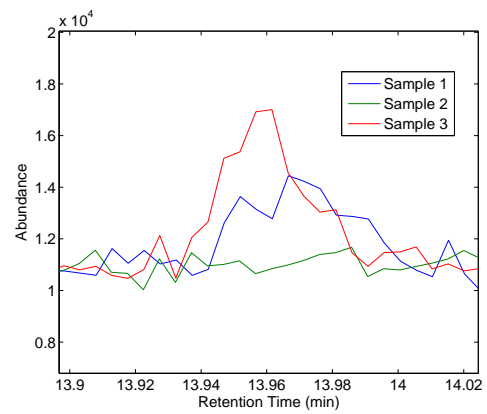
(b) Zoom into Ethylacetate

Figure F.15: Zoom into the two VOCs added to the mixture

Siloxanes can be observed barely above the noise level. There does not seem to be a significant increase in siloxanes due to the presence of water.



(a) Zoom into siloxane peak nr. 1



(b) Zoom into siloxane peak nr. 2

Figure F.16: Zoom into 2 siloxanes for the 3 measurements

The software only identified these three main peaks (after quality filtering) and they showed good repeatability.

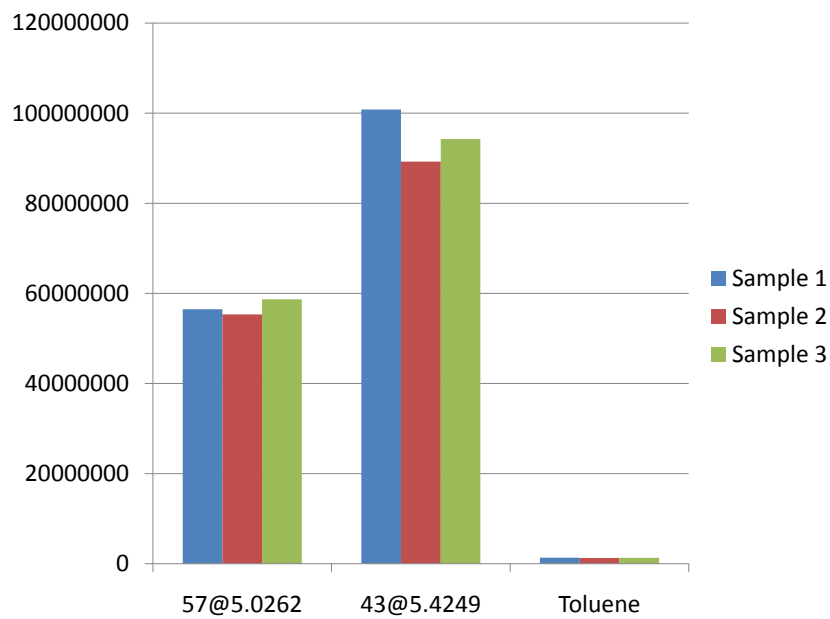


Figure F.17: Results obtained by processing the known wet VOCs samples with the Matlab tool

### F.3.5 Breath

The three breath samples showed good repeatability. Except for a few cases, which will be explained, there matching is adequate.

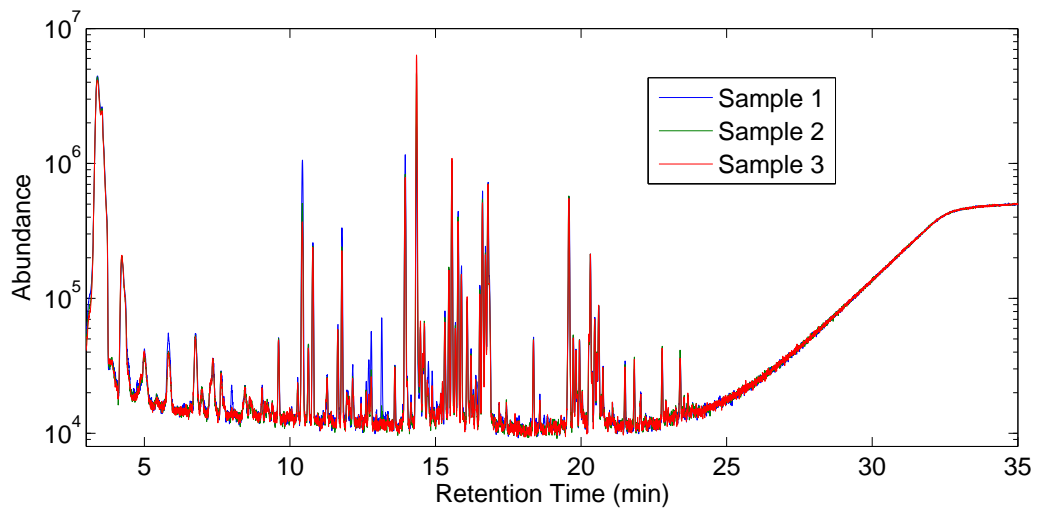


Figure F.18: Overlay of 3 runs of a breath sample

There are three peaks that seem to behave in a strange manner. Phenol, on one hand, tends to increase with time (sample 1 was the first to be extracted out of the Tedlar bag, while sample 3 was the last). However, it is known that phenol is a contaminant originated in Tedlar bags, so its increasing trend could be linked to the time a sample spent inside the bag.

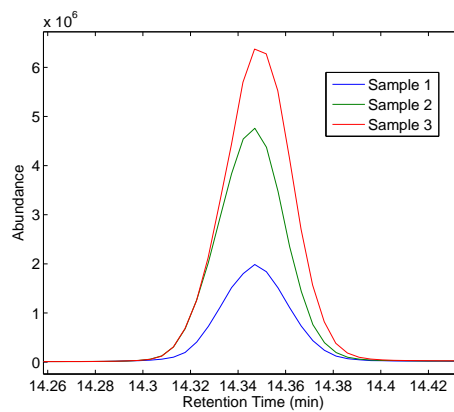


Figure F.19: Zoom into phenol peak

The siloxanes, on the other hand, vary their abundance from sample to sample, apparently decreasing with time. It should be noted that they are significantly higher than the levels measured in the blank runs.

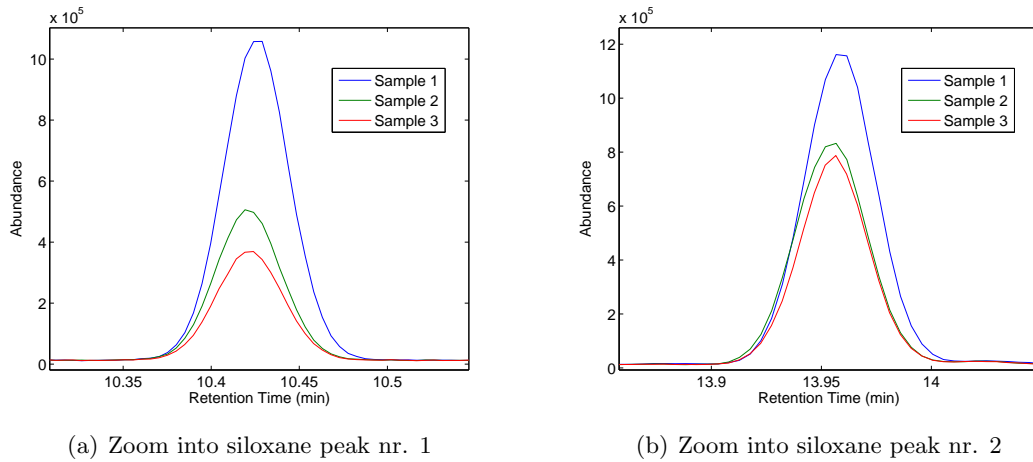


Figure F.20: Zoom into 2 siloxanes for the 3 measurements

The software identified 22 compounds, where only one was not found in one of the samples. This is the result of a too strict quality filtering. A simple change of this decision threshold can correct this defect.

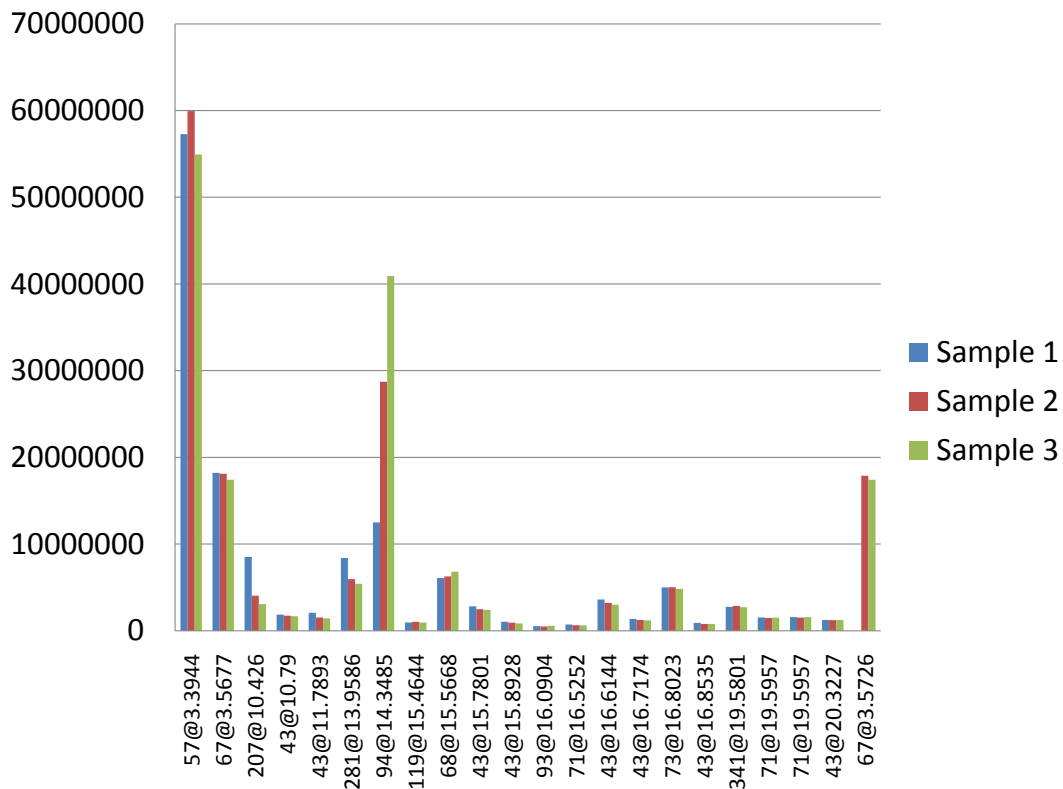


Figure F.21: Results obtained by processing breath samples with the Matlab tool



### F.3.6 Stored Nitrogen

Again, the three measurements had good matching. They appear to have higher background levels than in the case with no storage. However, these peaks are still well below the column bleed level.

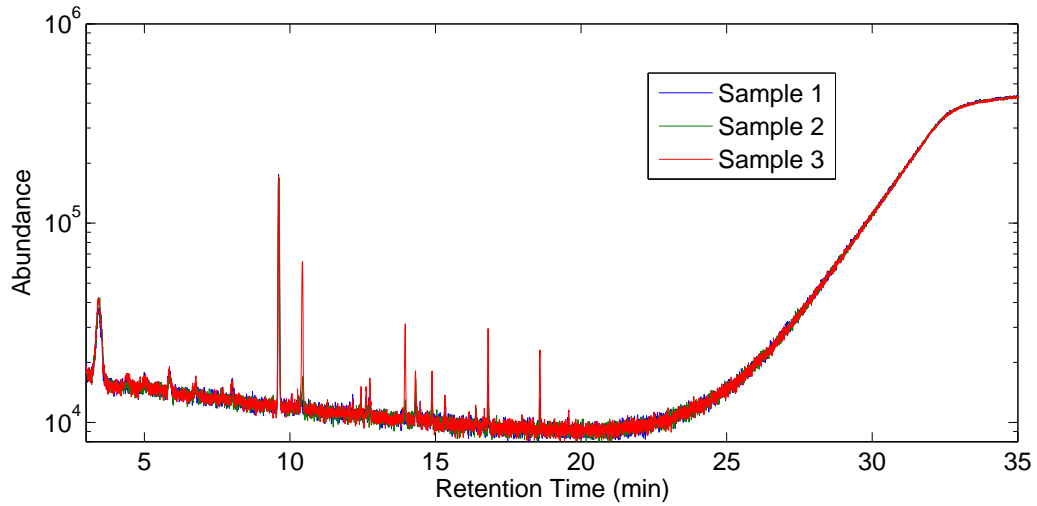


Figure F.22: Overlay of 3 runs of dry nitrogen on tubes stored for 14 days

If we observe the toluene peak in detail, there is only a small shift in time between runs.

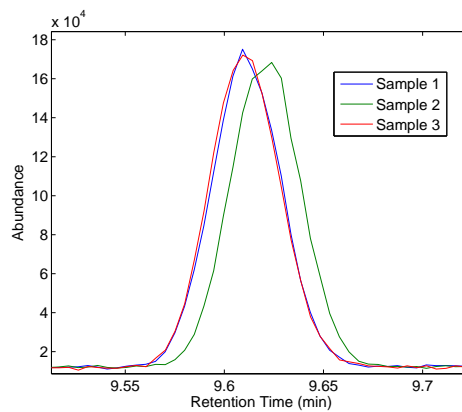


Figure F.23: Zoom into the toluene peak for the 3 measurements

Only one of the samples, number 3, shows an increase in siloxane levels. Since it does not occur in the other two samples, we cannot confirm that storage is the cause of their presence.

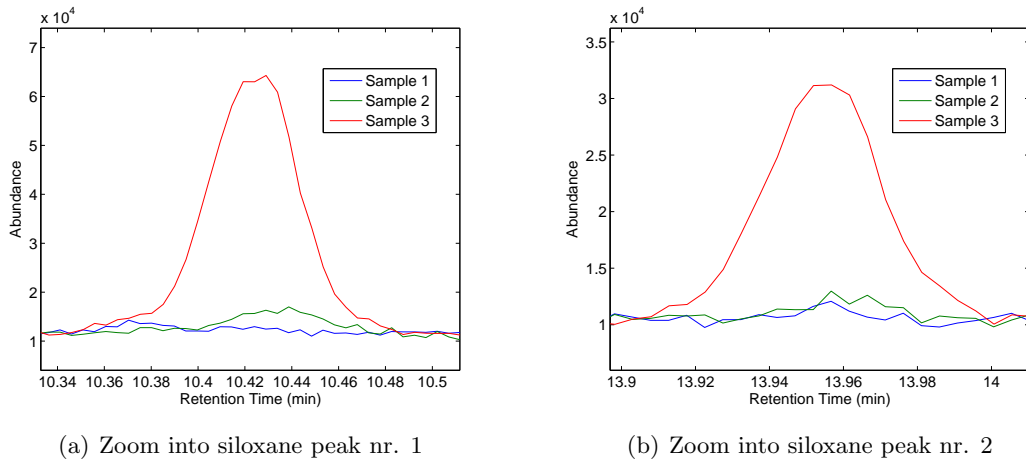


Figure F.24: Zoom into 2 siloxanes for the 3 measurements

When processing the samples with the Matlab tool, all the background peaks were eliminated and only toluene remained.

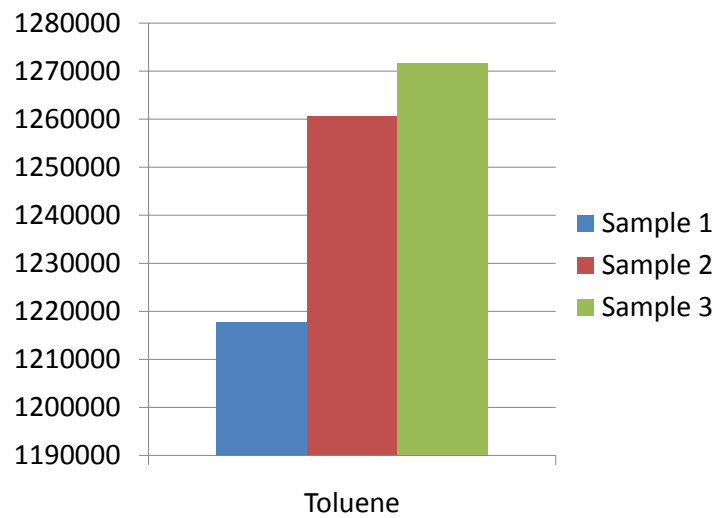


Figure F.25: Results obtained by processing the stored conditioned tubes with the Matlab tool

### F.3.7 Comparison of dry and wet nitrogen experiments

When comparing wet and dry blank tubes, the retention time for toluene is slightly shifted (it has a lower retention time for dry samples).

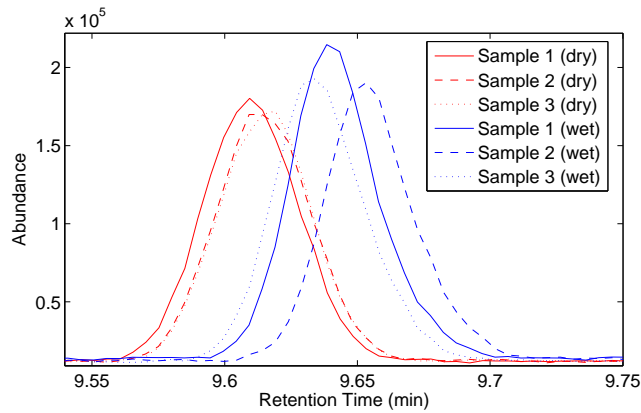


Figure F.26: Zoom into the toluene peak for dry and wet cases

The siloxane levels are higher on average for the wet group, though there is one sample in the dry set that has higher levels than two of the wet samples. Therefore, there is no clear time dependence of the abundance of siloxanes.

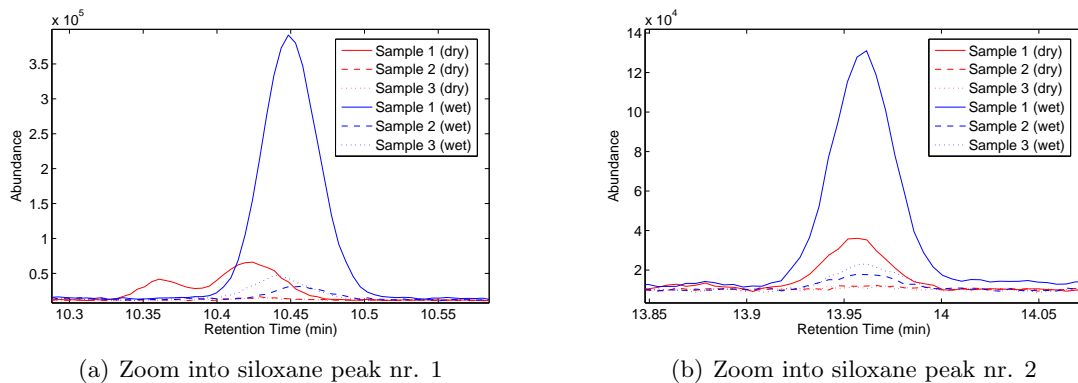


Figure F.27: Zoom into 2 siloxanes for dry and wet measurements

### F.3.8 Comparison of dry and wet VOCs experiments

When comparing dry and wet voc mixtures, the toluene standard does not show any significant variation. The siloxanes are around the same level, possibly higher in the wet case but not very significantly.

However, what is remarkable about this experiment is that the VOCs seem to have been affected by the presence of water. The abundances for the dry samples are considerably lower than for the wet samples. An explanation was found for this phenomenon: the presence of water affects the release of VOCs from the permeation tube. So water is not really interfering with the detection and measurement, but with the generation of the mixtures.

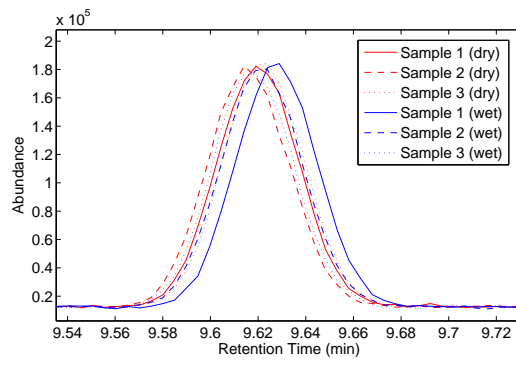
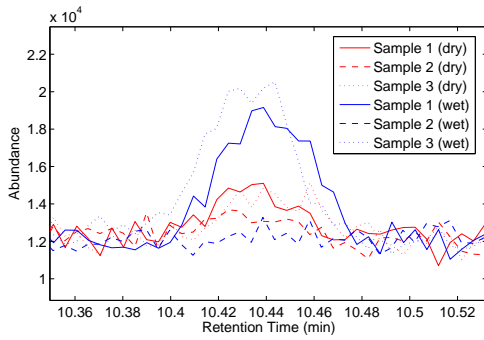
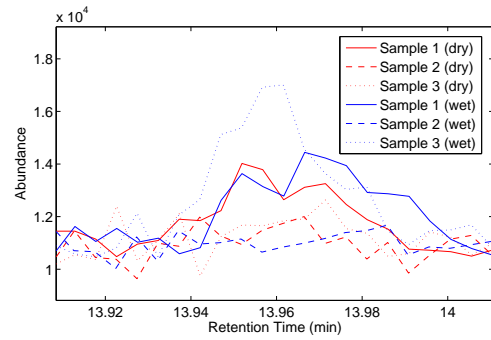


Figure F.28: Zoom into the toluene peak for dry and wet cases

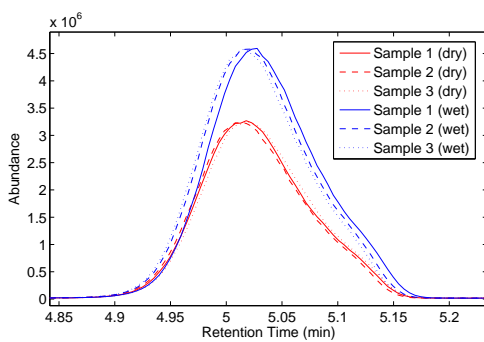


(a) Zoom into siloxane peak nr. 1

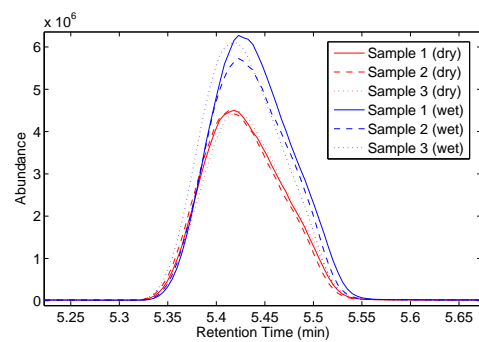


(b) Zoom into siloxane peak nr. 2

Figure F.29: Zoom into 2 siloxanes for dry and wet measurements



(a) Zoom into Hexane



(b) Zoom into Ethylacetate

Figure F.30: Zoom into the two VOCs added to the mixture

## F.4 Air quality at the ICU of Amsterdam AMC hospital

As a part of our pilot study, it was important to assess the quality of the air that will be provided to patients in the ICU, for the sepsis study. Before sampling breath from intubated ICU patients it is necessary to identify compounds that are not originating from patients breath but are coming from the ventilator setup or from the air that the patient is provided (compressed air). To this end we sampled 3 tubes containing compressed air, 3 tubes of air coming from a Hamilton Galileo ventilator and 3 tubes of air coming from a Maquet, Servo-I ventilator.

### Compressed air

All 3 tubes gave reproducible measurements. It is positive that the compressed air used by the hospital is very clean, and does not significantly contribute to the background.

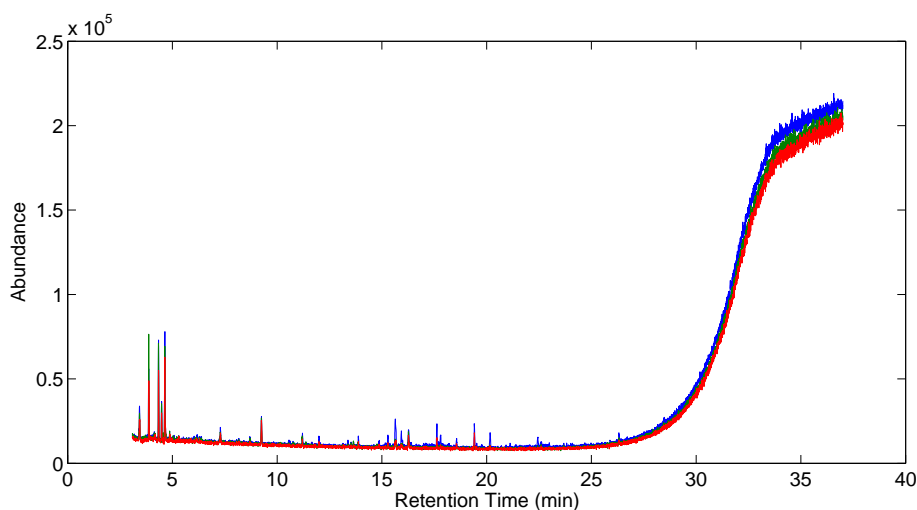


Figure F.31: Sample of the compressed air administered to ICU patients at the hospital

### Ventilators

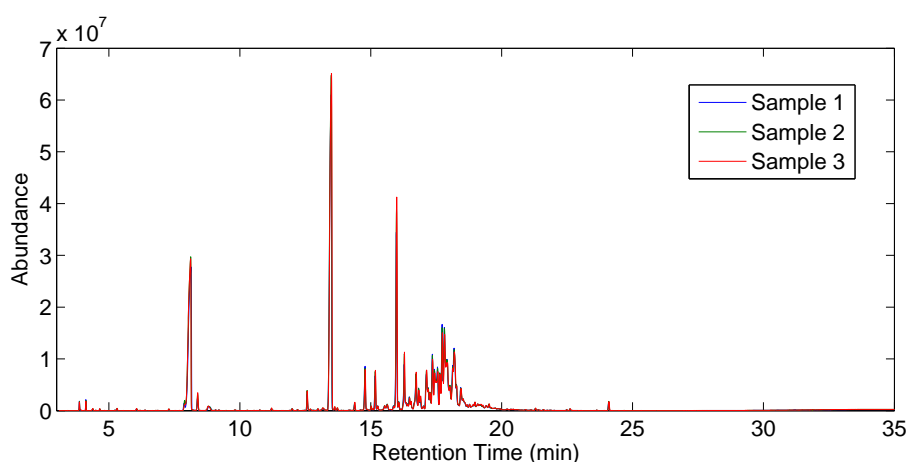


Figure F.32: Air sample of Hamilton ventilators

Figure F.32 and Figure F.33 show the chromatograms from air that went through two

different types of mechanical ventilators. Both ventilators give a clear contamination signature that is consistent between all samples.

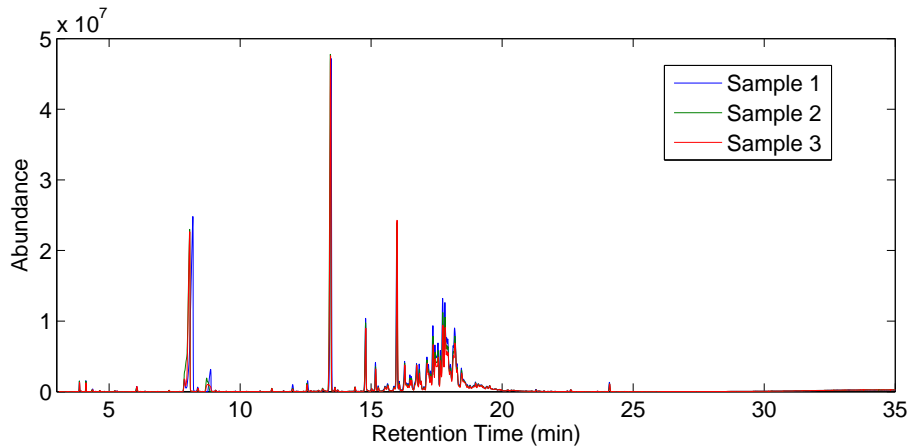


Figure F.33: Air sample of Maquet ventilators

In the retention times below 15 minutes there are some distinct peaks present. However, for retention times between 15 and 20 minutes, there is a large continuous portion of contaminants, where the peaks are almost indistinguishable. Figure F.34 shows that fortunately, the GC signature of both ventilators is the same. This supports the theory that the contamination is mainly coming from the tubing.

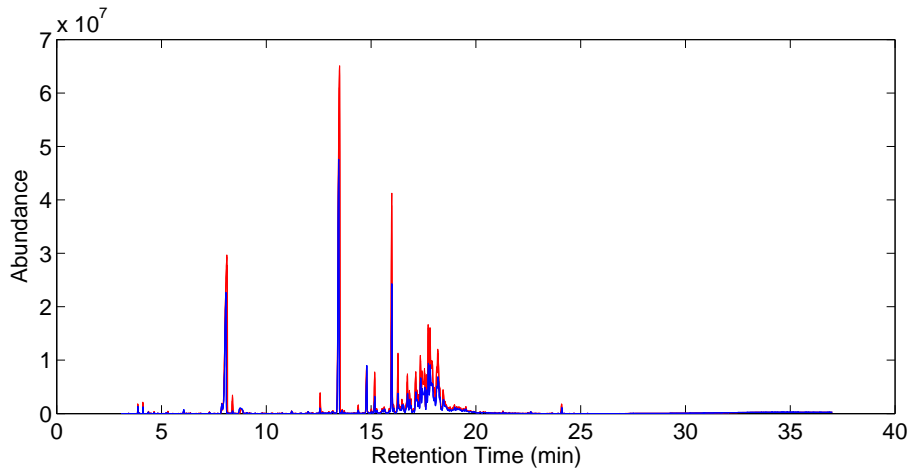


Figure F.34: Overlay of Hamilton and Maquet ventilators air

### Comparison of ventilator signal with breath sample

Figure F.35 shows a spectrum coming from the ventilator and a breath sample in the same plot. The column bleed differences should be ignored, since they are the result of some amplitude adjustments made to the breath signal to compensate for a lower sampled volume with respect to the ventilator air.

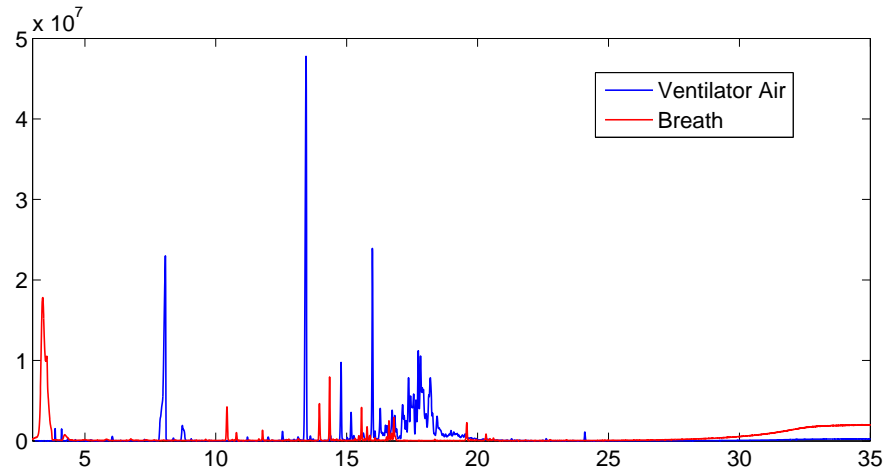


Figure F.35: Overlay of ventilator air and a breath sample

Even though both signals are within the same order of magnitude, it would be possible to work with samples from ventilated patients if all components between 15 and 20 minutes retention time were ignored, and the discrete peaks below 15 minutes were removed as we do for any non-endogenous components.

## F.5 Conclusions

There was no apparent connection between the presence of water and the appearance of siloxanes in the samples. Water did seem to produce some shifts in retention times, but nothing that the retention time window of the software tool could not overcome. Background signal levels did not change with the addition of water.

The wet VOC samples had variation in their abundance levels with respect to the dry case, but this was found to be caused by the setup itself and not to water interacting with the VOCs.

The test of stored condition tubes showed that if they are used within 2 weeks of their conditioning, the background levels are still low, well below the column bleed. There would be no problem in transporting tubes from Philips to hospitals within this time window, as long as they are properly closed.

The levels of the siloxanes in breath are much higher than the levels in blank tubes. This means that we cannot completely rule out the possibility that they are also endogenous in origin.

Regarding the sampling of ventilator air, despite the fact that the samples were far more contaminated than expected, it is possible to find a solution. As we mentioned before, all components between 15 and 20 minutes retention time could be ignored, and the discrete peaks below 15 minutes could be individually removed. If we look at normal contents of a breath sample, only about 30% of the information would be lost by ignoring this window. Thus, we conclude that sepsis studies with the given setup are feasible.



# G Analysis of the repeatability of the GC-MS measurements using the pilot experiments data

## G.1 Introduction

Since all of the experiments in the pilot study were run in triplicates, then it is possible to repeat the study on repeatability we performed for the 9 compound experiment. In this way, we can find out whether the repeatability of the system (GC-MS + Software) is within the range of our previous findings. In the past, we found that the maximum differences in abundance were 5.30% and for retention time 0.018%.

The pilot study experiments include runs of nitrogen and nitrogen with VOCs (both at 0% and 100% relative humidity), breath and dry nitrogen on stored tubes. This allows us to evaluate repeatability under different conditions, such as moisture level and complexity of the mixture.

## G.2 Method

For every experiment, we processed the data using AMDIS for peak extraction and then aligned the samples using our Matlab-based software. Quality score filtering was applied, in order to remove the large number of false positives present. The results from the alignment stage were used for comparing abundances, while the retention times were directly obtained from the AMDIS files, since this information is no longer available post-alignment.

## G.3 Results

In this section we calculated the repeatability for every case, both for abundance and retention time. The process is repeated for the 6 different experiments.

### G.3.1 Dry Nitrogen

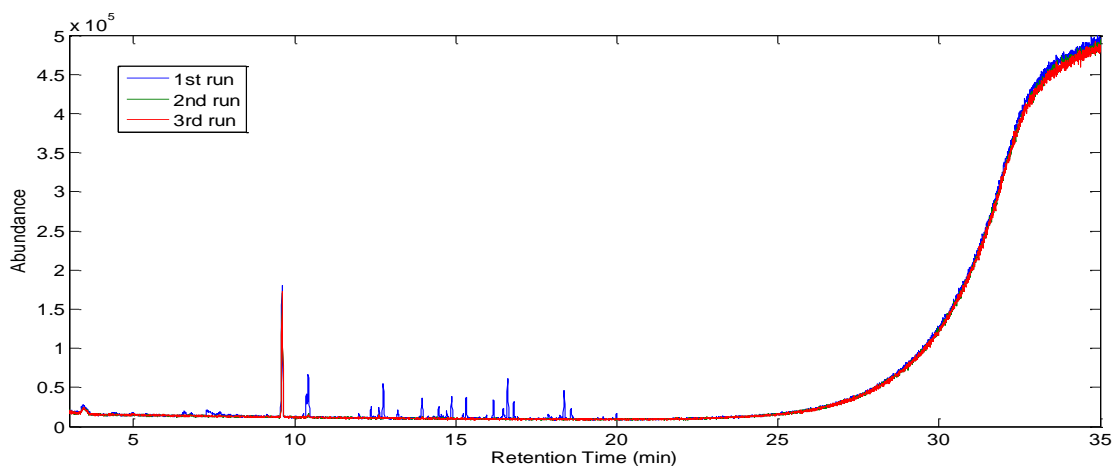


Figure G.1: Overlay of 3 runs of dry nitrogen on Tenax tubes

Figure G.1 shows some differences between the three runs. The first run showed more peaks in the background. However, most of these peaks were considered of poor quality and eliminated by the alignment software.

After processing, only toluene remained. Tables G.1 and G.2 show the repeatability values for abundance and retention time for this peak.

Peak Area		Retention Time	
Toluene		Toluene	
1	36235864	1	9.6176
2	36637964	2	9.6141
3	35145391	3	9.6108
Mean:	36194599	Mean:	9.6142
Std. Dev.:	688615	Std. Dev.:	0.0034
%RSD:	1.90	%RSD:	0.0354

Table G.1: Peak area for toluene, as calculated by the software

Table G.2: Retention time for toluene, as calculated by the software

Table G.1 shows that the peak area was highly repeatable under these conditions, with only 1.90% relative standard deviation (RSD). The retention time was stable as well, with 0.03% RSD.

Even though the three measurements were not perfect matches, as shown in the overlay of the chromatograms, the quality score filtering overcomes this problem by removing the low quality background peaks.

### G.3.2 Wet Nitrogen

In this case, nitrogen at 100% relative humidity was flowed through 3 Tenax tubes, to study whether the humidity affected the background level.

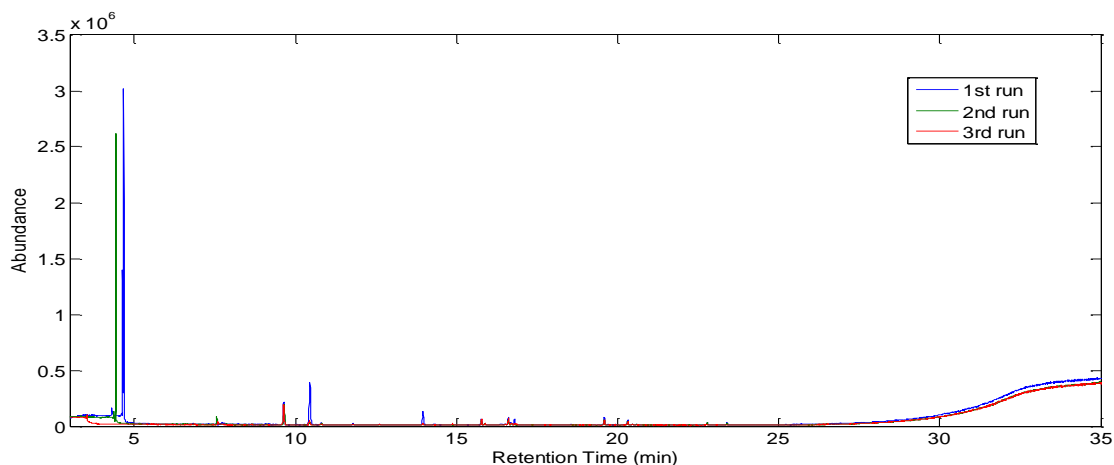


Figure G.2: Overlay of 3 runs of wet nitrogen on Tenax tubes

Several anomalies can be observed in figure G.2. None of the three runs seem to match at early retention times (<math>5</math> min) and in two of them there is an artifact of high abundance in that same range. This large peak should not be present, given that this was just a wet blank run. Its origin cannot be explained.

Since the disturbances in in this experiment are so large that it cannot be considered an useful sample for repeatability calculations. Thus, no further analysis is performed for this case.

### G.3.3 Dry VOCs

A mixture of 2 known VOCs was prepared. These two components together with nitrogen gas were flowed through Tenax tubes.

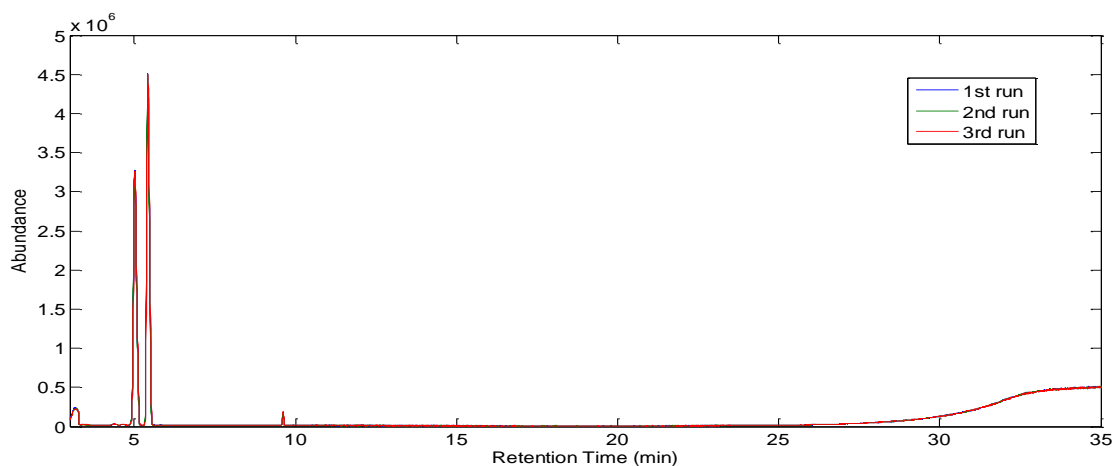


Figure G.3: Overlay of 3 runs of dry nitrogen and 2 VOCs on Tenax tubes

Peak area was repeatable, always below 5% relative standard deviation, and so were retention times, with a RSD always below 0.1%.

<b>Peak Area</b>			
	Toluene	VOC 1	VOC 2
1	1264386	41949291	69975420
2	1300783	38961485	72484109
3	1343967	39308520	65789774
Mean:	1303045	40073099	69416434
Std. Dev.:	39838	1634069	3381993
%RSD:	3.06	4.08	4.87

Table G.3: Abundances of toluene and VOCs, calculated by the aligner software

<b>Retention Time</b>			
	Toluene	VOC 1	VOC 2
1	9.6230	5.0209	5.4229
2	9.6166	5.0149	5.4140
3	9.6210	5.0223	5.4204
Mean:	9.6202	5.0194	5.4191
Std. Dev.:	0.0033	0.0039	0.0046
%RSD:	0.0340	0.0783	0.0847

Table G.4: Retention times of toluene and VOCs, calculated by AMDIS

### G.3.4 Wet VOCs

The same VOCs as in the previous case were flowed together with nitrogen at 100% relative humidity through Tenax tubes.

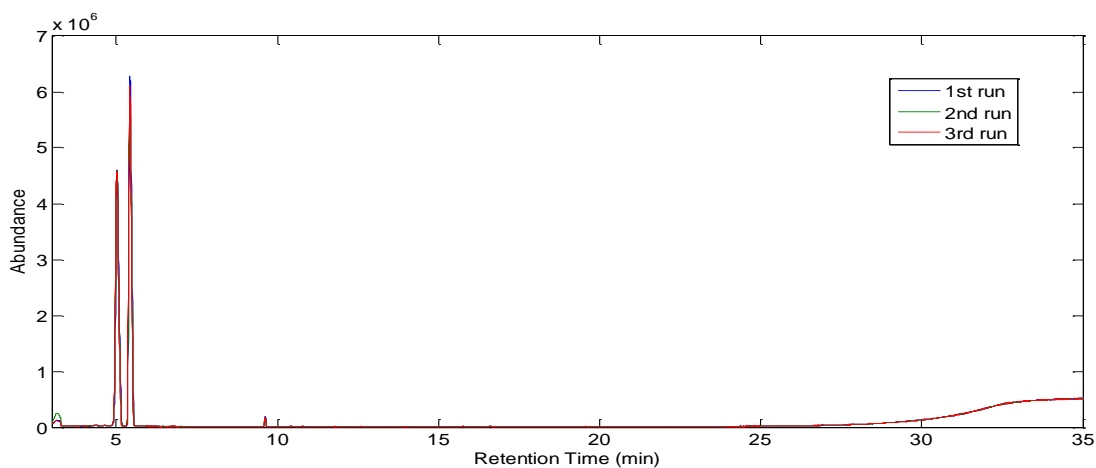


Figure G.4: Overlay of 3 runs of wet nitrogen and two VOCs on Tenax tubes

Peak area repeatability was at the same levels as in the previous case, despite the addition of moisture. Relative standard deviations of retention times were below 0.1% as well.

	Peak Area		
	Toluene	VOC 1	VOC 2
1	1340780	56491180	100800645
2	1286785	55331645	89248945
3	1310662	58702711	94244394
Mean:	1312742	56841845	94764661
Std. Dev.:	27057	1712672	5793397
%RSD:	2.06	3.01	6.11

Table G.5: Abundances of toluene and VOCs, calculated by the aligner software

	Retention Time		
	Toluene	VOC 1	VOC 2
1	9.6185	5.0178	5.4204
2	9.6208	5.0211	5.4241
3	9.6294	5.0262	5.4249
Mean:	9.6229	5.0217	5.4231
Std. Dev.:	0.0057	0.0042	0.0024
%RSD:	0.0597	0.0843	0.0443

Table G.6: Abundances of toluene and VOCs, calculated by the aligner software

### G.3.5 Breath

A breath sample was collected in a Tedlar bag. From this bag, three 500ml samples were taken and flowed through three Tenax tubes.

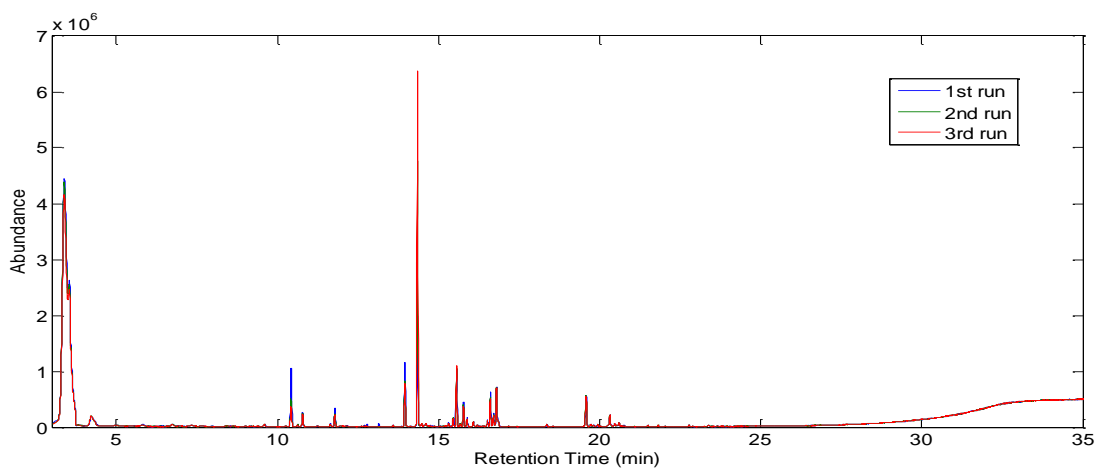


Figure G.5: Overlay of 3 runs of a breath sample

	Peak Area						
	C1	C2	C3	C4	C5	C6	C7
1	8514398	1863864	2076050	8380039	12497166	949647	6074921
2	4035237	1742158	1518265	5961625	28720671	1036786	6264961
3	3073246	1676752	1423252	5403719	40914725	935987	6813079
Mean:	5207627	1760925	1672522	6581794	27377521	974140	6384320
Std. Dev.:	2903861	94957	352679	1582112	14256313	54681	383281
%RSD:	55.76	5.39	21.08	24.03	52.07	5.61	6.00

Table G.7: Abundances of breath components, calculated by the aligner software

	Peak Area						
	C8	C9	C10	C11	C12	C13	C14
1	2820886	1027657	538161	710270	3612844	1363615	4996041
2	2487468	918589	492025	640716	3209747	1242016	5007101
3	2393855	839510	581354	606307	3004128	1202806	4827204
Mean:	2567403	928585	537180	652431	3275573	1269479	4943449
Std. Dev.:	224457	94470	44672	52962	309650	83848	100822
%RSD:	8.74	10.17	8.32	8.12	9.45	6.60	2.04

Table G.8: Abundances of breath components, calculated by the aligner software (cont.)

	Peak Area			
	C15	C16	C17	C18
1	906765	2747214	1532320	1240000
2	788878	2861103	1473240	1207504
3	785825	2707828	1515174	1236279
Mean:	827156	2772048	1506911	1227928
Std. Dev.:	68960	79598	30394	17785
%RSD:	8.34	2.87	2.02	1.45

Table G.9: Abundances of breath components, calculated by the aligner software (cont.)

The results for peak area repeatability are more complex than for previous cases. There are three compounds that should not be taken into account because they are consequence of the sample collection methodology. Compounds C1 and C4 are siloxanes, which are thought to be unstable in the presence of water. C5 is a contaminant from the Tedlar bag. For the rest of the compounds, the relative standard deviation is around or below 10%, with the exception of C2, whose unstable behaviour cannot be explained at the moment.

<b>Retention Time</b>							
	C1	C2	C3	C4	C5	C6	C7
1	10.4260	10.7900	11.7893	13.9586	14.3485	15.4644	15.5668
2	10.4202	10.7856	11.7815	13.9543	14.3475	15.4561	15.5620
3	10.4220	10.7860	11.7848	13.9556	14.3504	15.4580	15.5624
Mean:	10.4227	10.7872	11.7852	13.9562	14.3488	15.4595	15.5637
Std. Dev.:	0.0030	0.0024	0.0039	0.0022	0.0015	0.0043	0.0027
%RSD:	0.0285	0.0226	0.0332	0.0158	0.0103	0.0281	0.0171

Table G.10: Retention times of breath components, calculated by AMDIS

<b>Retention Time</b>							
	C8	C9	C10	C11	C12	C13	C14
1	15.7801	15.8928	16.0904	16.5252	16.6144	16.7174	16.8023
2	15.7738	15.8884	16.0865	16.5198	16.6106	16.7130	16.7979
3	15.7756	15.8878	16.0845	16.5226	16.6100	16.7105	16.7998
Mean:	15.7765	15.8897	16.0871	16.5225	16.6117	16.7136	16.8000
Std. Dev.:	0.0032	0.0027	0.0030	0.0027	0.0024	0.0035	0.0022
%RSD:	0.0206	0.0172	0.0187	0.0163	0.0144	0.0209	0.0131

Table G.11: Retention times of breath components, calculated by AMDIS (cont.)

<b>Retention Time</b>				
	C15	C16	C17	C18
1	16.8535	19.5801	19.5957	20.3227
2	16.8472	19.5738	19.5894	20.3179
3	16.8466	19.5751	19.5912	20.3216
Mean:	16.8491	19.5763	19.5921	20.3207
Std. Dev.:	0.0038	0.0033	0.0032	0.0025
%RSD:	0.0227	0.0170	0.0166	0.0124

Table G.12: Retention times of breath components, calculated by AMDIS (cont.)

Retention times, on the other hand, were highly repeatable, below 0.05% RSD in all cases.

### G.3.6 Stored Nitrogen

In this case, conditioned tubes were stored for 2 weeks. They were then used normally along with dry nitrogen.

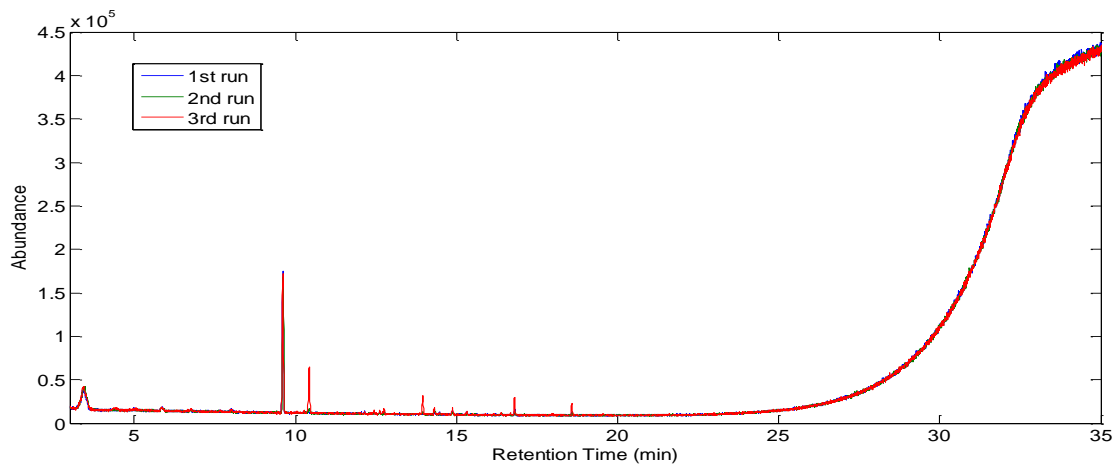


Figure G.6: Overlay of 3 runs of dry nitrogen on tubes stored for 2 weeks

Peak Area		Retention Time	
	Toluene		Toluene
1	1217716	1	10.7591
2	1260728	2	10.7672
3	1271649	3	10.7600
Mean:	1250031	Mean:	10.7621
Std. Dev.:	28513	Std. Dev.:	0.0044
%RSD:	2.28	%RSD:	0.0413

Table G.13: Peak area for toluene, as calculated by the software

Table G.14: Retention time for toluene, as calculated by the software

Peak area for toluene was very repeatable, at 2.28% relative standard deviation. The same is true for its retention time, with 0.04% RSD.



## G.4 Conclusions

For all the simple experiments, i.e. with the exception of the breath measurements, the obtained values for repeatability of peak area matched those of the previous study (*Analysis of the repeatability of the GC-MS measurements using the 9 compound experiment data*). The worst repeatability of peak area was for one of the VOCs at 6.11%, in the same order of magnitude of the previous results.

For breath, however, repeatability results were not as good. Approximately 20% of the components had a relative standard deviation of the area of less than 5%. More than 50% of the compounds had a RSD of between 5% and 11%. One component had an inexplicably large RSD of 21% and three others are not considered due to their methodological origin. However, if we only take into consideration the good, breath related components, 93% of them have a RSD of less than 11% which is still a reasonable value, in the order of magnitude of our previous findings.

Retention times, on the other hand, were repeatable in all cases. This is expectable, since the experiments were performed in a short time window, so the GC column did not suffer any changes in that period. This results also include the effects of AMDIS, and the software does not appear to add any extra error when calculating the times. For all cases, the relative standard deviation was below 0.1%.

In conclusion, repeatability for GC-MS measurements in combination with the software was good. The system behaved well with and without moisture, and despite suffering some deterioration, it was still reasonable for complex mixtures such as breath.

# H Analysis of component abundance stability with time for 19 asthma and control samples

## H.1 Introduction

Breath samples are stored in Tenax tubes until they are processed by the gas chromatograph-mass spectrometer (GC-MS). It is known that time has an effect on samples that can be observed mainly in two ways: a time drift of the peaks (the time axis suffers non-linear shifts) and change in abundances. Our intention is to study, from the limited dataset we currently have, the effects of storage on samples.

In a previous analysis we concluded that the best combination of software for preprocessing GC-MS files was the use of AMDIS and either Mass Profiler Professional or our own alignment tool. For this analysis we only need to be confident in the abundance calculation performed by AMDIS and in the subsequent alignment with the Matlab tool. Since in the previous study we determined that AMDIS high positive predictive value and the combination with a quality score filtering algorithm yielded good alignment of test samples, AMDIS and the Matlab tool were chosen to analyse 19 available asthma and control files.

## H.2 Analysis

A total of 19 patients were studied: 10 control (healthy) patients, and 9 asthma patients. For each patient 3 bags of breath were collected and then transferred to Tenax tubes. The first tube was analysed after one week (t0), the second after two weeks (t1) and the third after three weeks (t2). This yielded a total of 57 samples for GC-MS analysis. The 57 GC-MS output files are initially processed with AMDIS in order to extract the peaks and were later filtered and aligned with the Matlab tool.

However, since there is an average of 120 compounds found per breath sample, it would take too long to analyse the behaviour with time of each one. Therefore, we picked a number of components to study, based on their frequency (the components studied were commonly present in patients' breath), their abundance and their retention time.

Table H.1 shows the list of compounds chosen.

The chosen peaks span the entire spectrum, some are more distinct than others, they have different abundances and others are present in complicated areas of the spectrum, particularly the 4.5 min region where many compounds are co eluting. Furthermore, the peaks included both known breath related compounds, such as acetone and isoprene, and compounds of unknown origin.

## H.3 Results

In the obtained plots, which due to their number are shown in the Figure appendix section, the evolution of peak abundances over time can be observed. Each plot represents one component, each group of three bars is a single patient, and each color is a specific time of analysis, i.e. one, two or three weeks.

Mainly two types of behaviour can be observed. Some compounds appear to behave in a relatively stable manner (for instance Isoprene or Limonene) while others show an erratic

Name	Retention Time
Carbon Dioxide	3.36
Acetaldehyde	3.72
2-methyl-1-propene	3.79
Ethanol	4.27
Acetone	4.40
Isoprene	4.73
Dimethylsulfide	4.82
Carbon Disulfide	5.13
1-propanol	5.48
Trimethylsilanol	5.71
2-butenal	6.94
2-methyl-1,3-dioxalane	7.10
Benzene	7.54
Heptane	8.55
Toluene	9.92
Hexamethylcyclotrisiloxane	11.11
N,N-dimethylacetamide	11.59
Benzaldehyde	13.65
Octamethylcyclotetrasiloxane	14.44
Limonene	15.24
Decamethylcyclopentasiloxane	17.11

Table H.1: Compounds for which time stability was analysed

behaviour (such as Hexamethylcyclotrisiloxane or Octamethylcyclotetrasiloxane).

In the case of the compounds with erratic behaviour we can observe a common pattern that appears to be linked to each patient. For example, if we have a look at Octamethylcyclotetrasiloxane and Decamethylcyclopentasiloxane, some patients seem to have a relatively stable behaviour (such as HEY35, TUY and REE). On the other hand, other patients show a variation pattern that appears to repeat in every case, for example ALB has a high concentration at t0, a very low abundance at t1 and again a medium abundance at t2. Or STE, who seems to only have a noticeable abundance at t1, for all the inconsistent compounds.

### H.3.1 One-Way Repeated Measures ANOVA

In order to determine if the means varied significantly over time, we ran a one-way repeated measures ANOVA analysis. The null hypothesis was that the group means (for t0, t1 and t2) were the same. In several cases, this hypothesis was rejected. The results are summed up in table H.2.

## H.4 Conclusions

Most of the compounds that behave erratically are precisely those whose origin we cannot explain, i.e. silicones. A first step towards understanding the reasons for this unstable behaviour would be explaining the origin of these compounds.

The ANOVA analysis, however, showed that several compounds have an unstable behaviour. This is something that will need to be studied further, in order to ensure that samples coming from different hospitals in Europe, with different storage times, are valid. Still, most compounds of known endogenous origin were stable.

Retention Time	ID	F	p-value
3.36	Carbon dioxide	0.9497	0.3963
3.72	Acetaldehyde	4.7549	0.0147
3.79	2-methyl-1-propene	0.0949	0.9097
4.27	Ethanol	0.1403	0.8696
4.40	Acetone	1.0411	0.3634
4.73	Isoprene	5.6958	0.0071
4.82	Dimethylsulfide	1.6463	0.2070
5.13	Carbon disulfide	3.1473	0.0550
5.48	1-propanol	7.6006	0.0018
5.71	Trimethylsilanol	5.6786	0.0072
6.94	2-butenal	0.1952	0.8235
7.10	2-methyl-1,3-dioxalane	1.5238	0.2316
7.54	Benzene	0.9655	0.3904
8.55	Heptane	0.5695	0.5708
9.92	Toluene	0.7154	0.4958
11.11	Hexamethylcyclotrisiloxane	6.2425	0.0047
11.59	N,N-dimethylacetamide	0.1992	0.8203
13.65	Benzaldehyde	0.4546	0.6383
14.40	Siloxane	2.6327	0.0857
14.44	Octamethylcyclotetrasiloxane	0.3632	0.6980
15.24	Limonene	5.9623	0.0058
16.81	Siloxane	5.0064	0.0121
17.11	Decamethylcyclopentasiloxane	0.2936	0.7473

Table H.2: Calculated F and p values for the dataset using one-way repeated measures ANOVA

In future testing, we will also evaluate if this variability in abundances has to do with compounds interacting with water. If this was the case, it would explain the link with specific patients rather than compounds, since we do remove moisture in the setup, but have no control or measurement on that amount.

Repeatability of GC-MS measurements should be studied in the near future. This way, it would be possible to make a distinction between possible instrument related variations and differences that may be related to storage time. We cannot truly define a compound as stable with time without first knowing fully the behaviour of the instrument without any time differences. Once the repeatability of the instrument is known, we can deduce how much of the variation is due to storage.

### H.5 Figure appendix

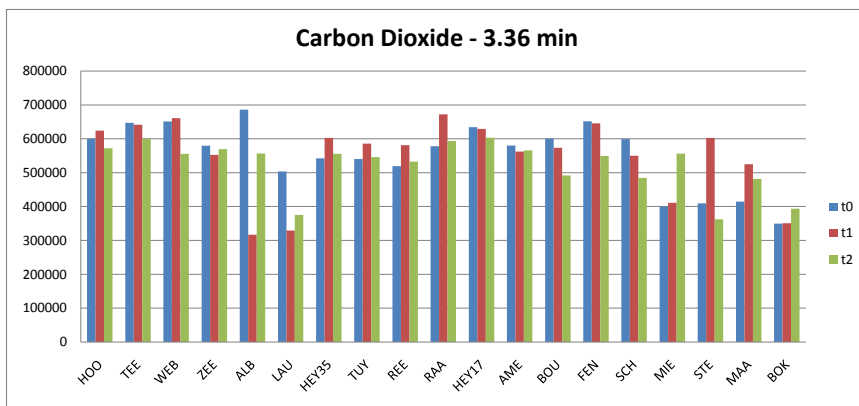


Figure H.1: Evolution of carbon dioxide abundances over three weeks of storage

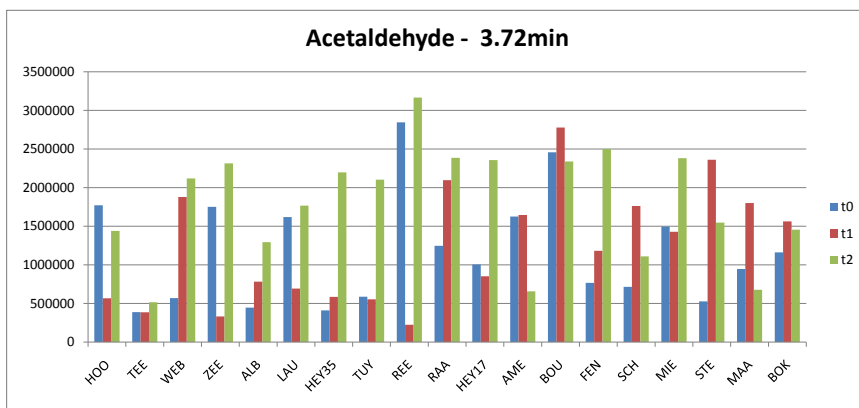


Figure H.2: Evolution of acetaldehyde abundances over three weeks of storage

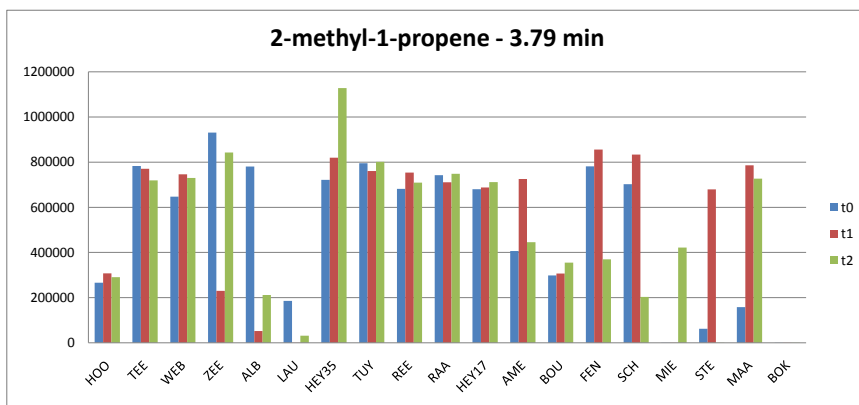


Figure H.3: Evolution of 2-methyl-1-propene abundances over three weeks of storage

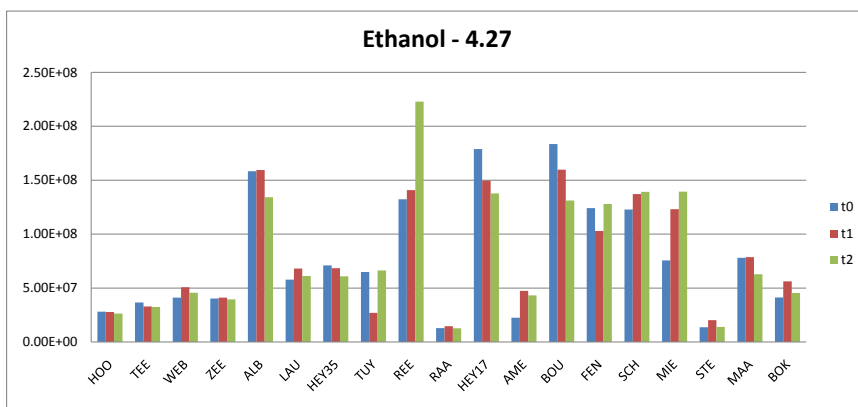


Figure H.4: Evolution of ethanol abundances over three weeks of storage

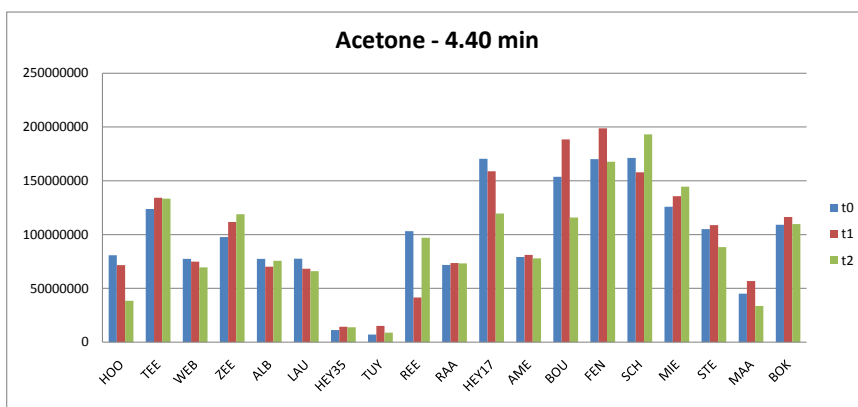


Figure H.5: Evolution of acetone abundances over three weeks of storage

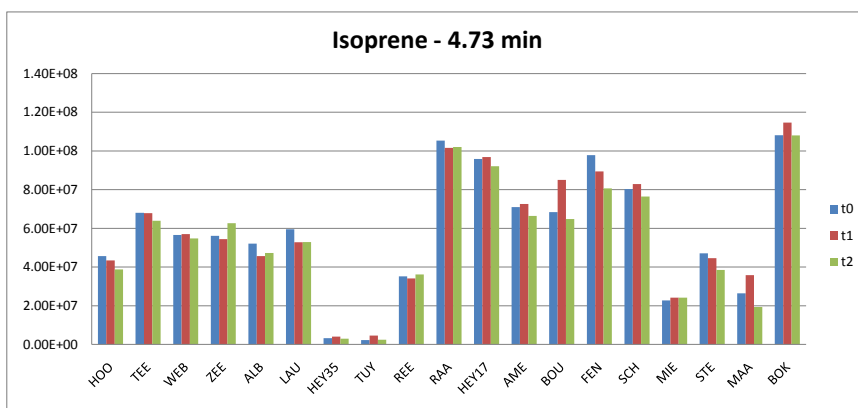


Figure H.6: Evolution of isoprene abundances over three weeks of storage

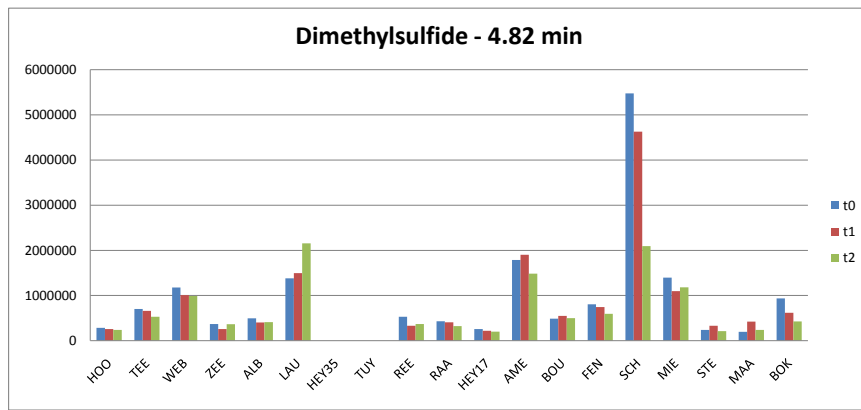


Figure H.7: Evolution of dimethylsulfide abundances over three weeks of storage

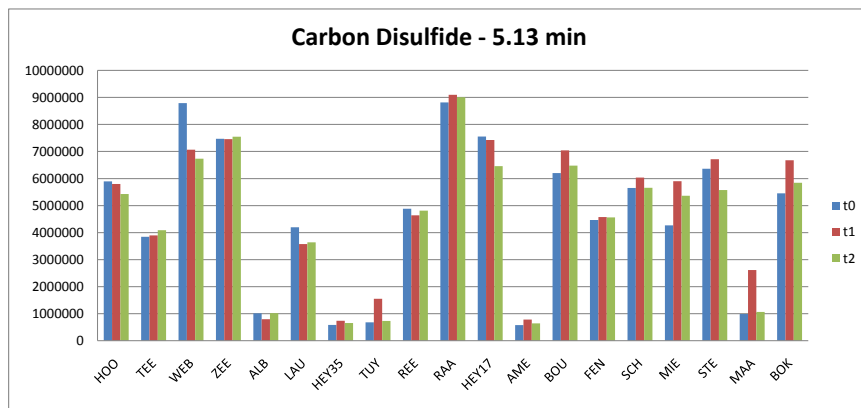


Figure H.8: Evolution of carbon disulfide abundances over three weeks of storage

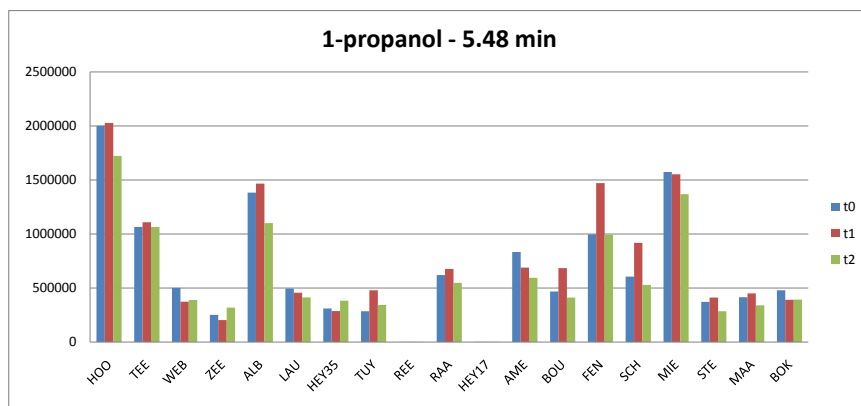


Figure H.9: Evolution of 1-propanol abundances over three weeks of storage

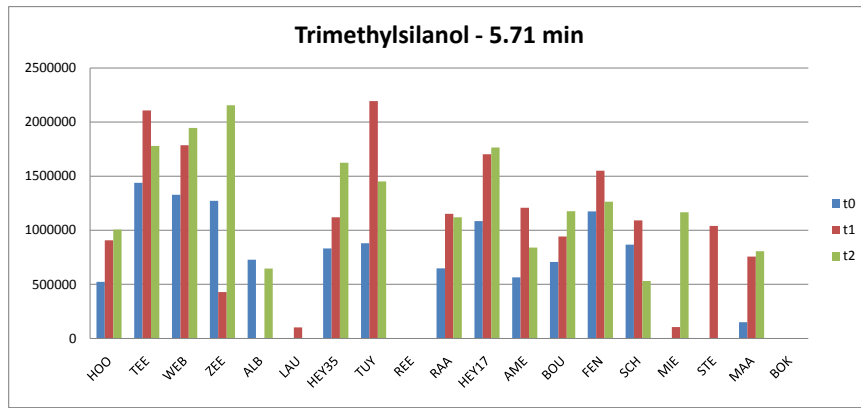


Figure H.10: Evolution of trimethylsilanol abundances over three weeks of storage

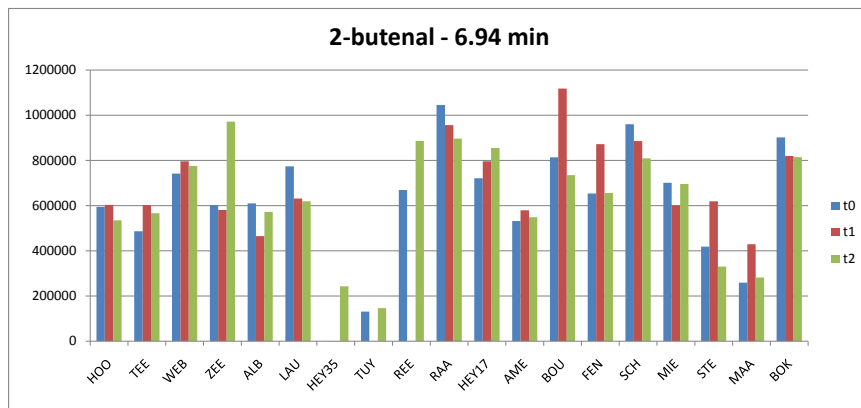


Figure H.11: Evolution of 2-butenal abundances over three weeks of storage

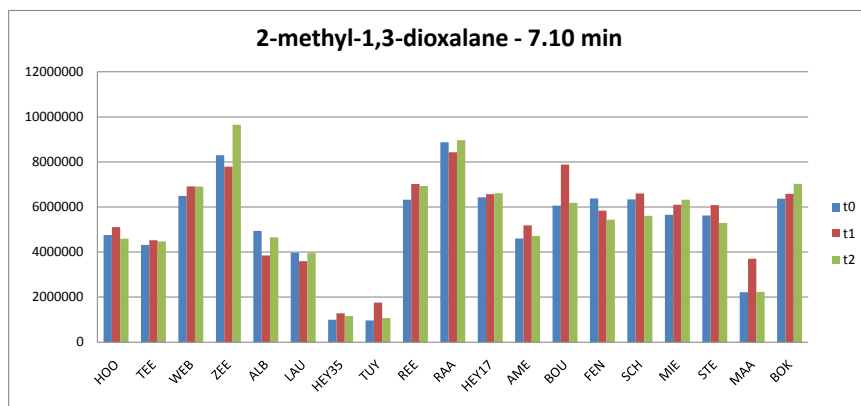


Figure H.12: Evolution of 2-methyl-1,3-dioxalane abundances over three weeks of storage



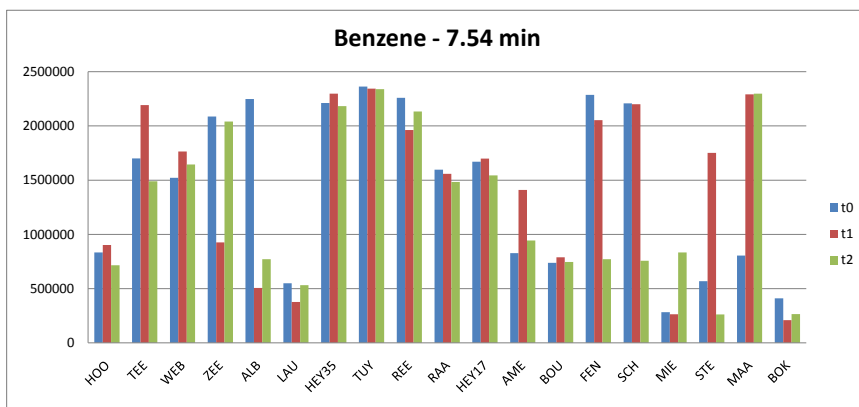


Figure H.13: Evolution of benzene abundances over three weeks of storage

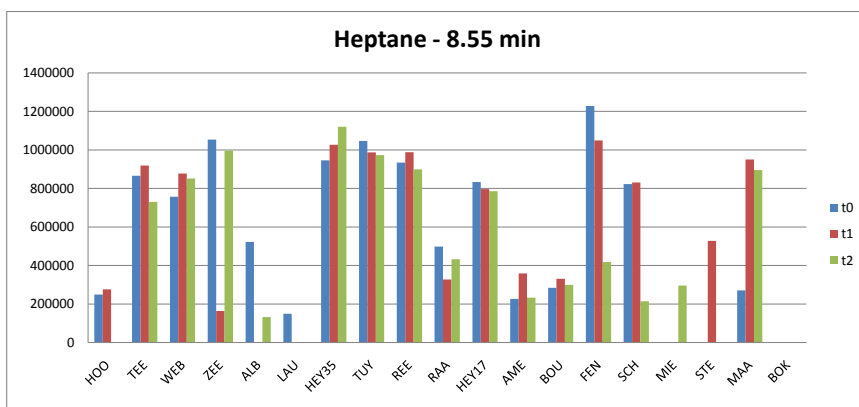


Figure H.14: Evolution of heptane abundances over three weeks of storage

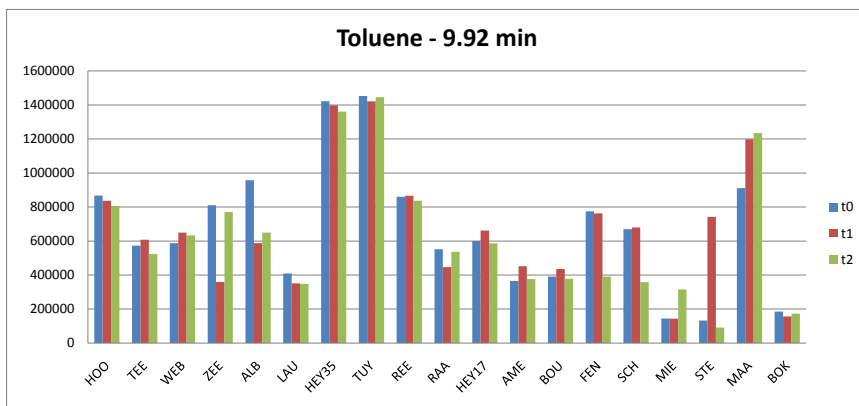


Figure H.15: Evolution of toluene abundances over three weeks of storage

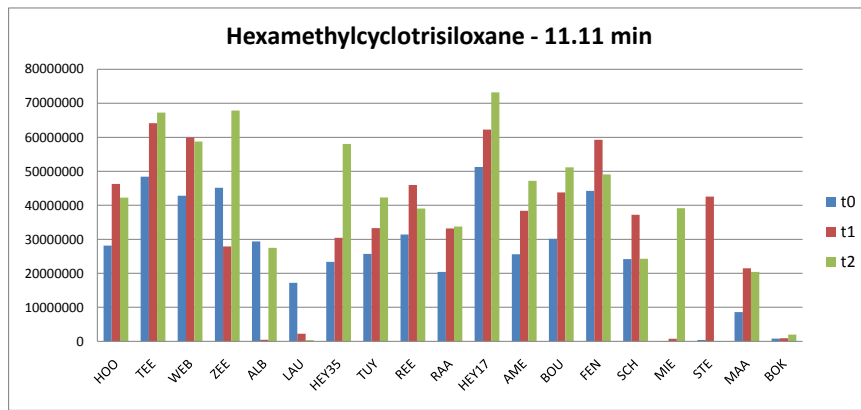


Figure H.16: Evolution of hexamethylcyclotrisiloxane abundances over three weeks of storage

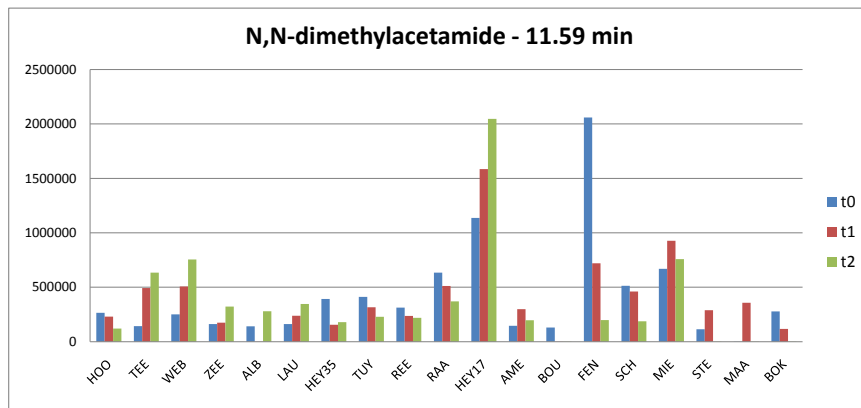


Figure H.17: Evolution of N,N-dimethylacetamide abundances over three weeks of storage

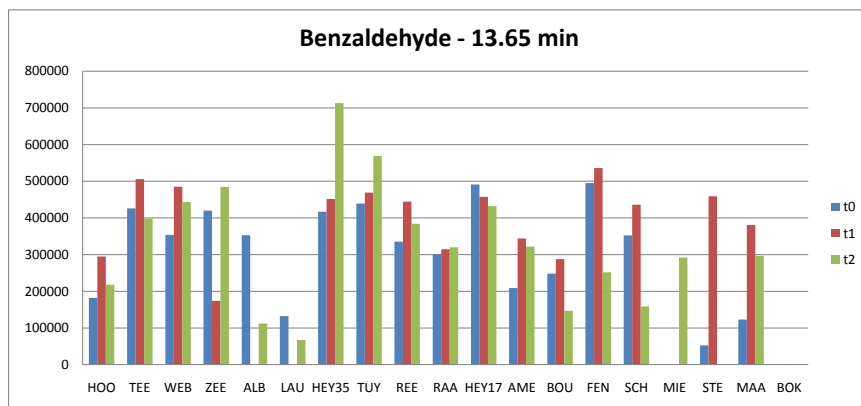


Figure H.18: Evolution of benzaldehyde abundances over three weeks of storage

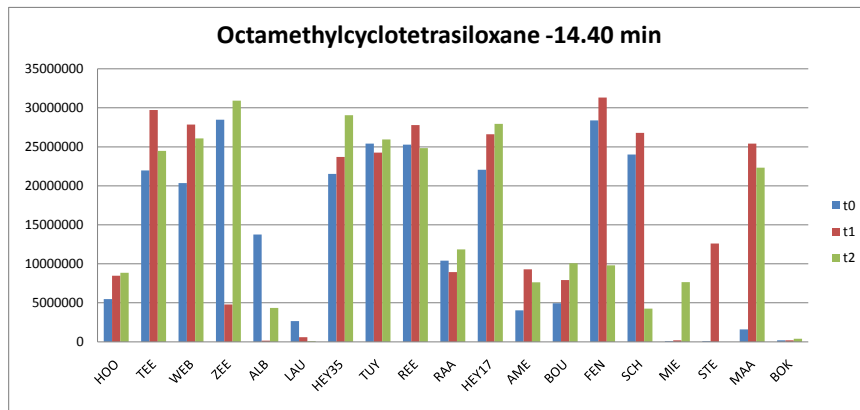


Figure H.19: Evolution of octamethylcyclotetrasiloxane abundances over three weeks of storage

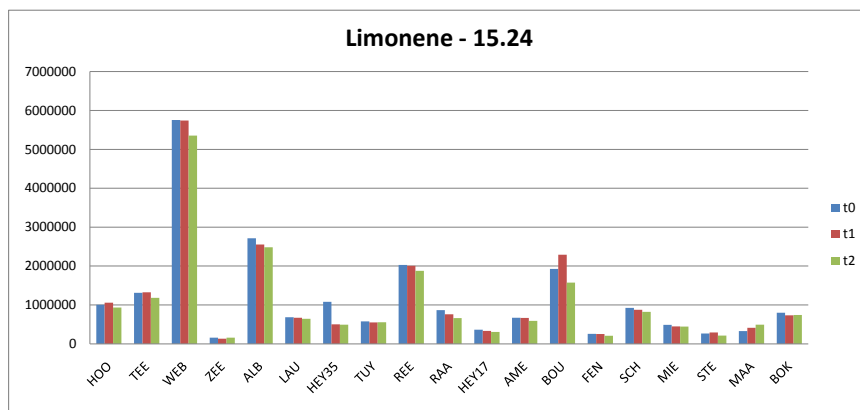


Figure H.20: Evolution of limonene abundances over three weeks of storage

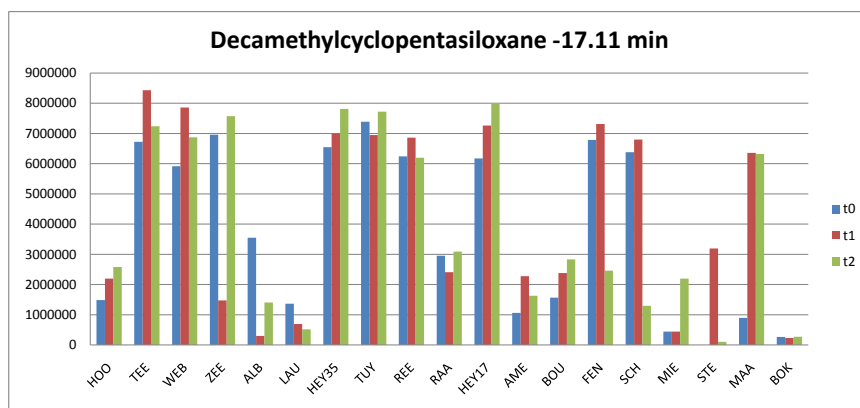


Figure H.21: Evolution of decamethylcyclopentasiloxane abundances over three weeks of storage

# I Software tool for GC-MS data processing

## I.1 Introduction

After analysing the available commercial tools for GC-MS data processing, we determined that none of the studied alternatives fulfilled our exact requirements. Therefore, a simple though very specific solution could be designed for the problem at hand.

Figure I.1 shows the different processing stages a breath sample needs to go through before any statistical analysis is possible. The Matlab tool designed covers 3 of these stages: quality filtering, alignment and result reporting. It obtains its data straight out of AMDIS, which was discussed in a previous report, and the results could be directly sent to a statistical analysis software.

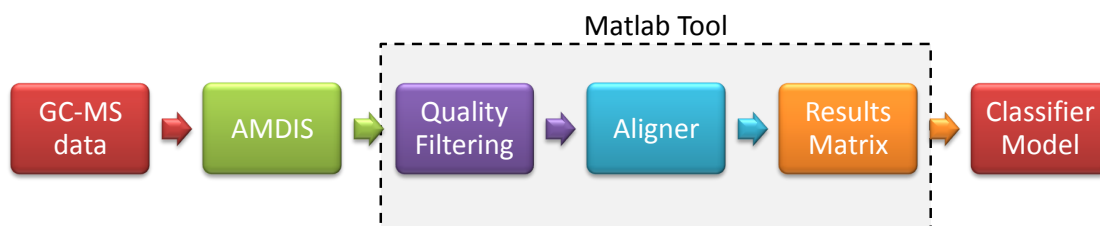


Figure I.1: Block diagram

## I.2 Design

### I.2.1 AMDIS Files

The first challenge of developing this tool was the compatibility with AMDIS files. The standard files from AMDIS have the extension \*.elu. There is little documentation on their structure, but it could be deduced out of the analysis of several files we had available.

Figure I.2 shows an example of the information contained in an \*.elu file for a single peak. Normally, \*.elu files contain information for a large number of peaks, so they can have even larger sizes.

The header contains several parameters of the peak, many of which will be very important in the next sections. Some of the parameters contained include:

- Retention Time (RT)
- Purity (PC): The percentage of the total ion signal at the components maximum intensity scan that belongs to the deconvoluted component. AMDIS determines this by first extracting all of the ions associated with a component and then summing them to yield the total ion signal of the component.
- Signal to Noise Ratio (SN): The total signal-to-noise value as measured by utilizing all ions in a component.
- Width (WD): The full width at half maximum height of the chromatographic component peak; where the width is given in scans.
- Amount (RA): The area of the deconvoluted component relative to the total ion count for the entire chromatogram, expressed as a percentage.

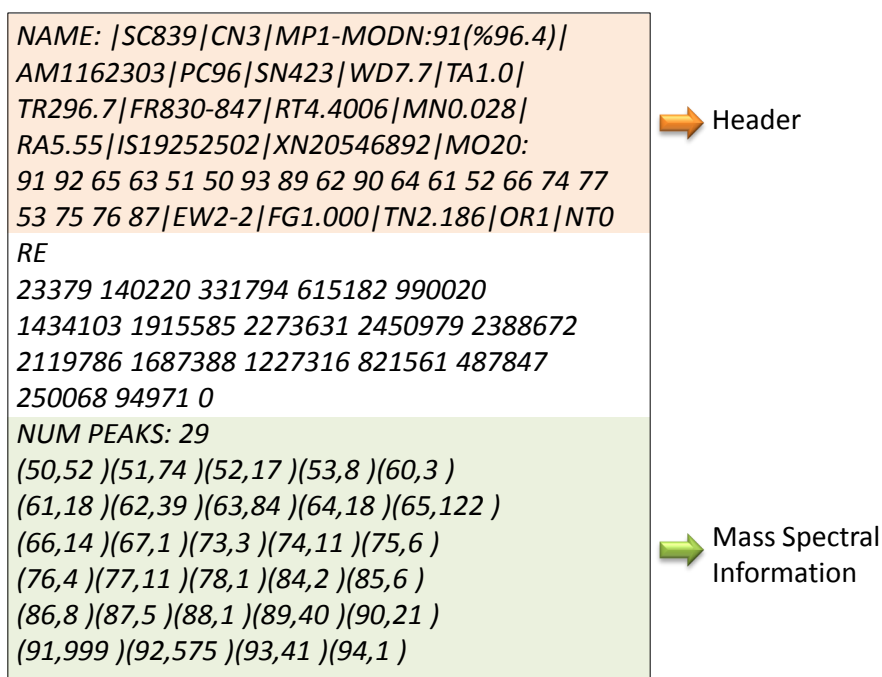


Figure I.2: Information contained in an \*.elu file for a single peak

- Integrated Signal (IS): Sum of all the ions associated with a component.
- Models (MO): The number of ions whose shape matches that of the total ion count (chromatographic peak)

The mass spectral information section contains all fragments found for that specific peak. Each one of the points listed contains the mass to charge ratio of the fragment and its abundance.

### I.2.2 Quality Filtering

Deconvolution algorithms have the drawback that they may produce false positives as an output. In general, when working with a reasonable sensitivity, these types of software tend to over identify peaks in a given chromatogram. This is the case of AMDIS, and so was the case of MassHunter which we evaluated in a previous analysis. For instance, when studying a simple 9 component mixture in AMDIS, an average of 13 compounds were found every time. This defect means that the GC-MS processing tool should be able to handle these false positives appropriately.

However, the samples that will be processed with the Matlab tool are unknown, in the sense that the user has no previous knowledge of the contents. Therefore, the false positives cannot be unequivocally identified. The solution for this situation is to implement a type of filtering that removes components that are suspected to be false positives. In this way, unreliable peaks could be eliminated.

Every peak identified by AMDIS possesses certain inherent characteristics, which were mentioned in the previous section. Some of these could help distinguish between true and false positives because they can be linked to peak quality. Still, this link would be dependent on the current experimental configuration. Thus, we needed to develop a tool that would aid the user to train an optimal filter given a certain database of known true and false positives, and would remove all unreliable peaks from the current data. Less unreliable peaks in the data translates into a better performance of the alignment algorithm in the next processing stage.

The main requirements of the quality filtering stage were that it should be tunable, from completely disabled to very strict filtering, it should have several decision surface alternatives, it should provide a visual output to facilitate the understanding of what is being applied to the data and it should provide a quantitative output of the effects of the filtering on the data, i.e. how much information is being removed.

In our current situation, we had a library of known true and false positives from a set of controlled experiments that were performed. With this knowledge, we were able to develop a tool that can find a filtering surface and show how it affects the experimental data.

Figure I.3 shows an image of the user interface. The user can load a library of known true and false positives (blue and red stars in the plot) and overlay it with his current data (green dots). In this case, the unclassified data come from an asthma experiment with 19 patients.

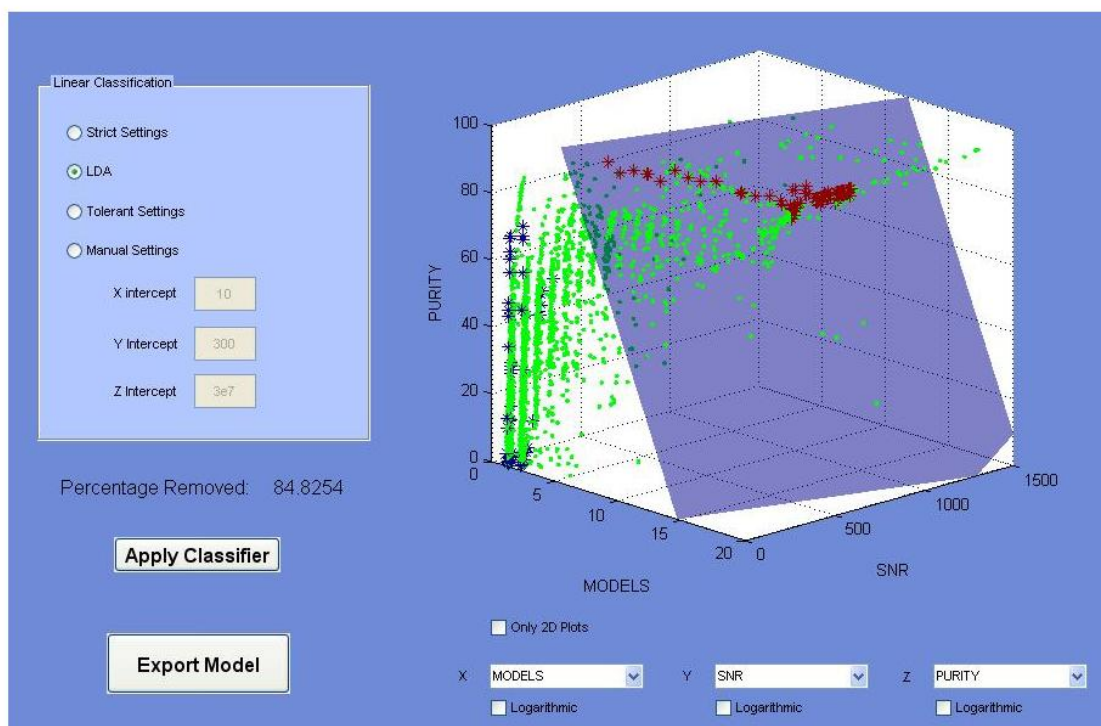


Figure I.3: Quality filtering tool

The user has the possibility of choosing 2 or 3 out of the 6 parameters available for the classification (models, SNR, abundance, purity, width and amount) and then determining a linear decision surface. This surface can be set manually, but may also be found with linear discriminant analysis. There is also the possibility of making the filtering very strict or very tolerant.

The resulting filter can be exported back into the main software tool. Within the main tool the quality filtering option can also be disabled completely, which may be useful if the user is willing to work with raw data.

### Linear Discriminant Analysis

In order to find the optimal decision surface for separating true and false positives, we use a linear discriminant analysis algorithm. Its objective is to find the plane that maximizes the separability of the two groups. Figure I.4 shows two plots that explain this. Plot a) shows a

poor choice of the separation direction. Plot b) instead shows the optimal choice of direction. If the points are projected onto this line, they are properly separable. LDA seeks directions that are efficient for discrimination.

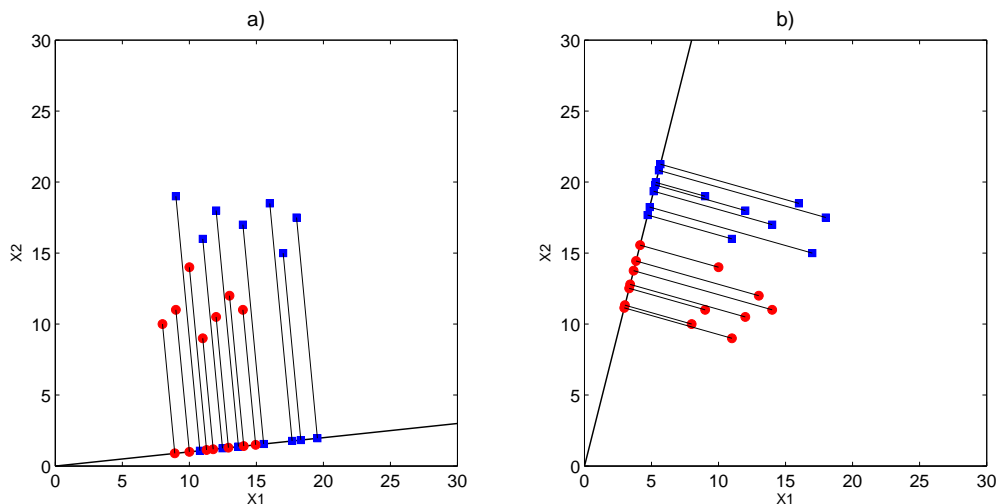


Figure I.4: Illustration of the concept of optimal separability

A brief explanation is as follows. Initially, the data is divided into several matrices, one per group, where the rows are different samples and the columns represent each attribute or biomarker. The program finds the mean of both groups and then calculates the covariance of each group matrix, shown in equations I.1 and I.2. The value  $p_j$  is proportion of samples of each group relative to the total. The sum of all covariances is computed and the between class scatter is calculated (see equation I.3). Finally, the direction of discriminatory plane (or line) is found by applying equation I.4.

$$S_w = \sum_j p_j \times cov_j \quad (\text{I.1})$$

$$cov_j = (x_j - \mu_j)(x_j - \mu_j)^T \quad (\text{I.2})$$

$$S_b = \sum_j (x_j - \mu_3) \times (x_j - \mu_3)^T \quad (\text{I.3})$$

$$v = inv(S_w) \times S_b \quad (\text{I.4})$$

### I.2.3 Alignment

The alignment stage of the processing is crucial for obtaining samples that can be compared between each other and to allow drawing statistical conclusions.

It is essential to define what is meant by alignment. Originally we start with a continuous chromatogram produced by the GC-MS instrument. After processing the chromatograms with AMDIS, we no longer have entire chromatograms but a list of discrete peaks found in the original data files. However, one difficulty arises when comparing different samples. Peaks have to be matched between different lists, because a certain compound may not always be found in the same position in different samples. Figure I.5 shows how the first component found in patient X may be the the second component found in patient Y, for instance, acetone. Other

components may be found in the same position, like the third and fourth, but others may be found in completely different locations. This differences between samples occur frequently, particularly for breath samples, where not all substances are found in all patients. In fact, there are about 3000 possible VOCs that may be found in breath, but only around 200 are found per patient, and less than 10% are shared by the majority of the population.

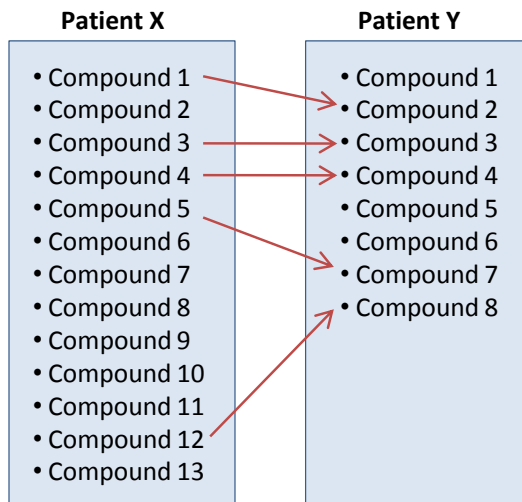


Figure I.5: Illustration of the concept of sample “Alignment”

If we intend to compare between a large number of samples, without identifying and giving a name to each component found first, then it is necessary to perform this type of alignment across all samples to find the abundance of every element in every file (zero if it is absent).

### Algorithm

Given a certain experimental configuration, the same chemical compound in different runs will have about the same retention time. Therefore, there is a window around a certain retention time where a compound is expected to be found. This fact is essential for speeding up the alignment process.

If in the alignment stage  $n$  different samples must be aligned, then there are  $n$  compound lists (of variable length). The algorithm starts by creating a new list (which we will call the “Total List”, and includes all of the components found in all samples). As a second step, it takes all of the compounds in Sample 1 and adds them to the list (see figure I.6), with their respective abundances.

Then it compares this partial total list with sample 2, starting by the first compound on the list. It looks for that compound in sample 2, though not anywhere: it searches for the compound in a time window around its retention time. If it is found, the abundance is added to the total list. All the compounds found in sample 2 that were not present in the total list are added at the end.

This process is repeated by comparing the incomplete total list against every sample available. With every comparison, the list may be come longer, to the point that at the end it will contain all the compounds found in all samples.



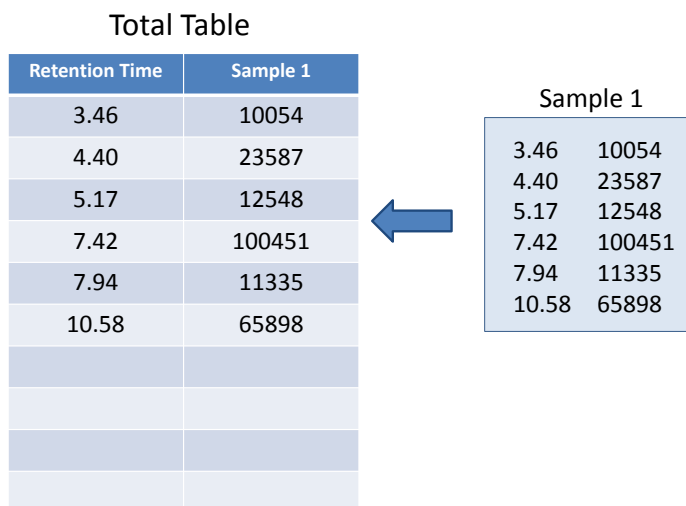


Figure I.6: First step of alignment process

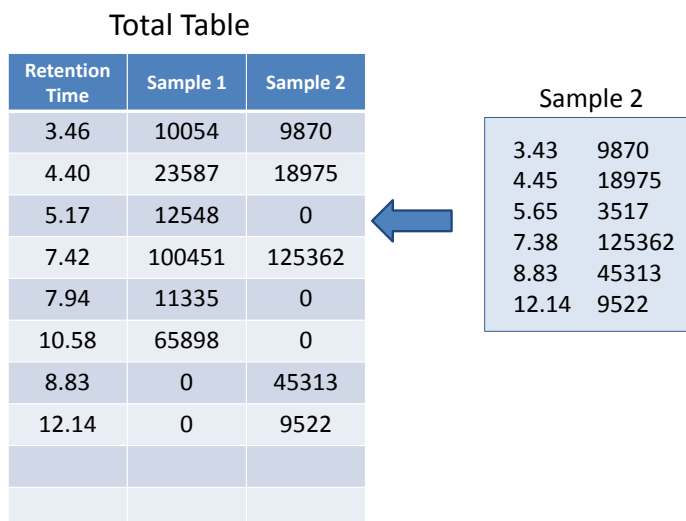


Figure I.7: Second step of alignment process

**Similarity scoring**

In order to find out whether two compounds are the same it is necessary to compare their mass spectra. Each spectrum is represented as row vector of order peak intensities. In this way, every mass spectrum is represented by a point in a multidimensional space defined by each of the masses. We choose to use the square of the cosine between two vectors as a measure of their similarity, following the conclusions by Stein and Scott.

The formula for the similarity S is the following:

**Total Table**

Retention Time	Sample 1	Sample 2	Sample 3	Sample 4	...	Sample n
3.46	10054	9870	0	11050	...	0
4.40	23587	18975	25564	0	...	19115
5.17	12548	0	11526	13554	...	0
7.42	100451	125362	0	116479	...	99891
7.94	11335	0	10989	11447	...	0
10.58	65898	0	59483	63659	...	0
8.83	0	45313	39795	43114	...	0
12.14	0	9522	0	0	...	0
3.30	0	0	1215	0	...	0
5.35	0	0	33456	0	...	0
4.58	0	0	0	12145	...	0
2.27	0	0	0	0	...	1561
11.11	0	0	0	0	...	9424

**Sample n**

2.27	1561
4.41	19115
7.39	99891
11.11	9424




Figure I.8: Final step of alignment process

$$S = \cos^2 \theta = \frac{\left( \sum_{i=1}^n A_i B_i \right)^2}{\left( \sum_{i=1}^n (A_i)^2 \right) \left( \sum_{i=1}^n (B_i)^2 \right)}$$

However, the higher masses of a spectrum have a higher diagnostic potential. Therefore weighting is performed on the data in order to give more importance to the higher masses. Therefore, the original abundances vector:

$$(A_1, A_2, \dots, A_n)$$

is weighted by the masses they correspond to, obtaining:

$$(\sqrt{M_1 A_1}, \sqrt{M_2 A_2}, \dots, \sqrt{M_n A_n})$$

Figure I.9 shows the effect of such weighting on a flat spectrum. The abundances of the fragments with a larger mass to charge ratio are amplified.

Finally, the complete similarity score formula is the following:

$$S = \frac{\left( \sum_{i=1}^n M_i \sqrt{A_i B_i} \right)^2}{\left( \sum_{i=1}^n M_i A_i \right) \left( \sum_{i=1}^n M_i B_i \right)}$$

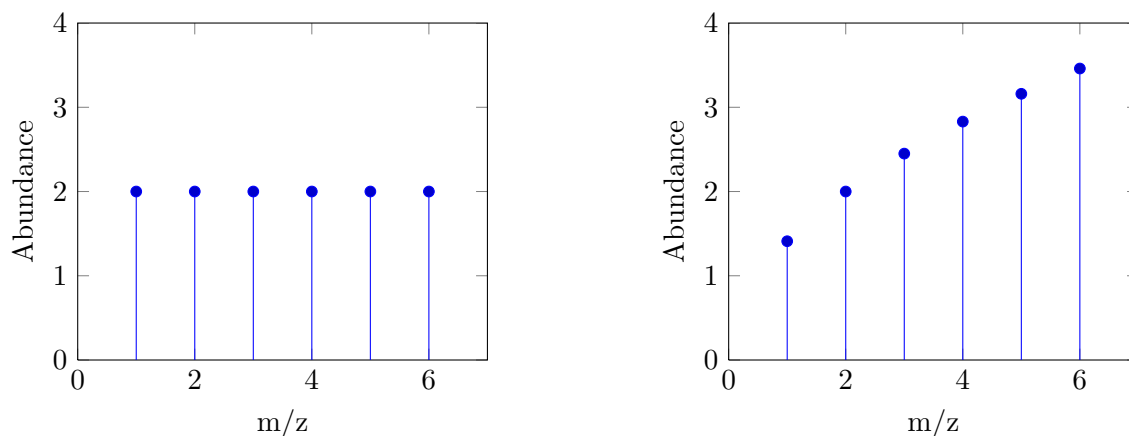


Figure I.9: Quality filtering tool

### I.2.4 Interface

Figure I.10 shows the graphical user interface. The “File Management” module takes AMDIS files as inputs and lists them. By default, all samples are considered to belong to the the same group (Group 1). The user can set a group label for each of the samples, which is necessary for proper plotting of the results.



Figure I.10: Graphical user interface for the aligner software

The “Quality Filtering” section offers the choice on whether to activate or disable the filter. If enabled, the user can choose to load a saved filter (designed with the designer tool in the past), to input the filter manually if the constants of the hyperplane are known, or to use the designer to create the filter wanted. Though the designer tool if the most clear option since it shows the effects on the data under study, having a predesigned filter is also useful when making repeated analysis.

The “Alignment” section allows the user to choose the two alignment parameters available, which were mentioned previously: retention time window and match factor. With this

information, the software is ready to fully process the data.

### I.2.5 Results

Results are provided in three different ways. Firstly, there is a visual output within Matlab. Here, compounds are plotted as bars according to their abundances, and are grouped depending on their labels. The user assigns a group label to each file at the beginning of the process, and this is used later for plotting purposes, since bars are shown clustered by group. An example of such output is shown in figure I.11. The colours show the two different groups.

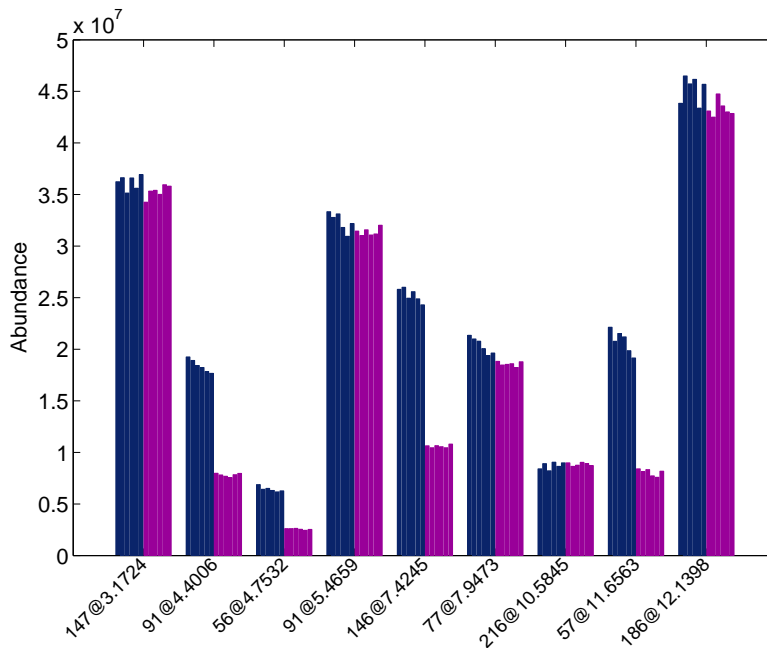


Figure I.11: Graphic output of the alignment process

Results are also stored as an excel file (see figure I.12), where the same information that is plotted is stored as an array. Furthermore, the array is also stored in Weka ARFF format for statistical processing. This is shown in figure I.13, for a small set of biomarkers (only four were considered) in a study of asthma data.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		S002.ELU	S003.ELU	S004.ELU	S005.ELU	S006.ELU	S007.ELU	S008.ELU	S009.ELU	S010.ELU	S011.ELU	S012.ELU	S013.ELU
2	147@3.1724	36235864	36637964	35145391	36603740	35604348	36940287	34258602	35330417	35404705	35014315	35951347	35805102
3	91@4.4006	19252502	18916004	18424430	18229633	17842600	17666168	7980697	7817627	7690220	7553586	7831152	7966993
4	56@4.7532	6878658	6422838	6513922	6305572	6174296	6264802	2610562	2609877	2628638	2542332	2440443	2528044
5	91@5.4659	33336618	32782286	33119767	31803450	30954019	32180927	31451225	31029486	31582187	31073309	31176652	32027311
6	146@7.4245	25812672	26016309	24954436	25575427	24892497	24305584	10643383	10450025	10652019	10550286	10450068	10804511
7	77@7.9473	21350717	20992463	20775759	20057437	19389743	19635614	18821035	18469104	18530872	18597733	18227162	18782058
8	216@10.5845	8400226	8910714	8209585	9044669	8641744	8990952	8980507	8649261	8761982	9035720	8917339	8725885
9	57@11.6563	22127731	20770632	21528092	21193561	19857550	19146705	8408267	8135425	8329093	7716992	7578973	8173924
10	186@12.1398	43846959	46496060	45723424	46169432	43380622	45687642	43092342	42503237	44754826	43591757	43015671	42856826

Figure I.12: Excel output of the alignment process

```

RELATION breath

@ATTRIBUTE A1 NUMERIC
@ATTRIBUTE A2 NUMERIC
@ATTRIBUTE A3 NUMERIC
@ATTRIBUTE A4 NUMERIC
@ATTRIBUTE class asthma,control

@DATA

600000,531000,1770000,266207,asthma
647000,76200,388000,783233,asthma
651000,758000,570000,647176,asthma
580000,200000,1750000,930896,asthma
686000,63900,447000,780405,asthma
503000,145000,1620000,185852,asthma
542000,1,410772,721292,asthma
541000,525074,587637,795470,asthma
519000,211000,2840000,681448,asthma
578000,98200,1250000,742391,control
635000,126470,1000000,680023,control
580000,140000,1630000,406390,control
601000,1,2460000,298126,control
652000,907000,767000,781246,control
599000,633000,715000,702737,control
400000,1,1500000,1,control
409000,112000,527000,61741,control
415000,1,946000,158206,control
350000,1,502000,1,control

```

Figure I.13: Reduced example of output in Weka's arff file format for statistical analysis of asthma data

### I.3 Conclusions

The software fulfilled the initial requirement of providing an accurate and reliable list of biomarkers present in a sample. In comparison to other commercial programs in the market, the reliability was improved through the development of a filtering system for poor quality compounds. This system allows the user to remove potential false positives or markers that do not meet quality criteria easily and with complete control over the process. The user is always aware of how much information is being lost because of the filtering, but also know that what remains is reliable enough to draw statistical conclusions.

The alignment stage also demonstrated to work properly, yielding adequate results for our only asthma dataset available.

In the future, the software should still be tested further with complex datasets, such as other breath samples.

# J User Guide for the complete processing workflow

## J.1 Introduction

This document summarizes the steps required to process a breath sample, from beginning to end. The original files are expected to be in Agilent's \*.d file format, though AMDIS is capable of working with many other source files. These files are initially processed with AMDIS, and are later imported into the Matlab tool. We describe all the steps required to obtain good results in detail.

## J.2 AMDIS Processing

### J.2.1 Description

AMDIS stands for “Automated Mass Spectral Deconvolution and Identification System”. It is a simple software running a complex algorithm for GC-MS data interpretation.

It can identify the peaks present in a certain GC-MS data file, separating even closely coeluting peaks thanks to its deconvolution algorithm. Later, it also allows the users to run a library search (if there is a NIST library present in the computer) and the peaks found may be identified. Though the Biomedical Sensor Systems department does not have this library, it may be used by people in MiPlaza.

### J.2.2 Loading Files

- Click **File** → **Open**
- Find the \*.d file to be loaded and click **Open**

### J.2.3 Configuration

- To analyze the file click **Analyze** → **Analyze GC-MS data**
- The following window will pop up:

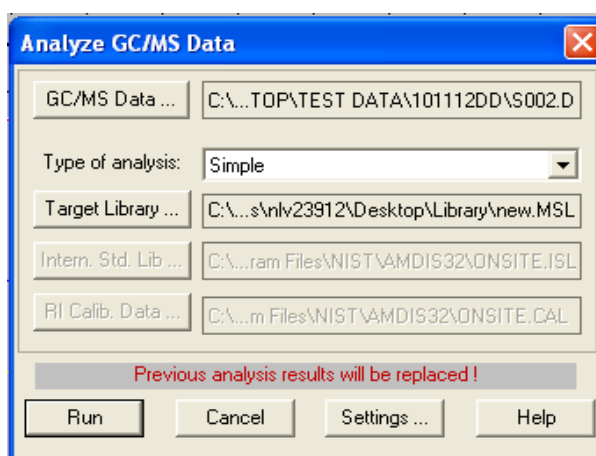


Figure J.1: Analyze pop-up menu

- **Type of analysis** should be set to **Simple**.
- Click on **Settings** to configure the analysis, and the next window will appear:

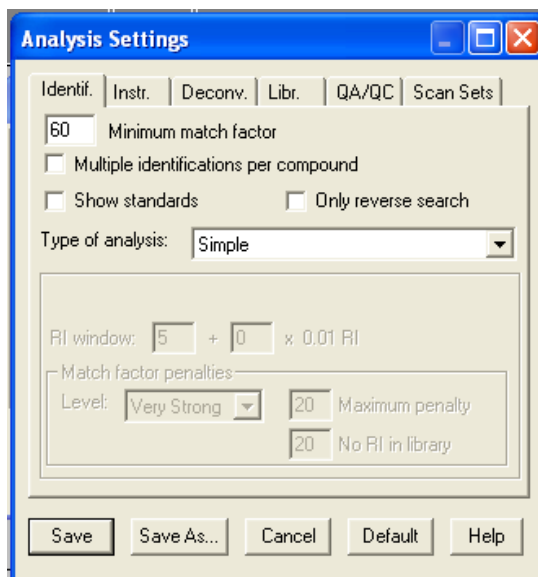


Figure J.2: Analysis settings menu

- The **Identification** tab in our case should remain as shown.
- The **Instrument** tab contains several parameters of interest. These are the setting required for the use with the Agilent GC-MS.

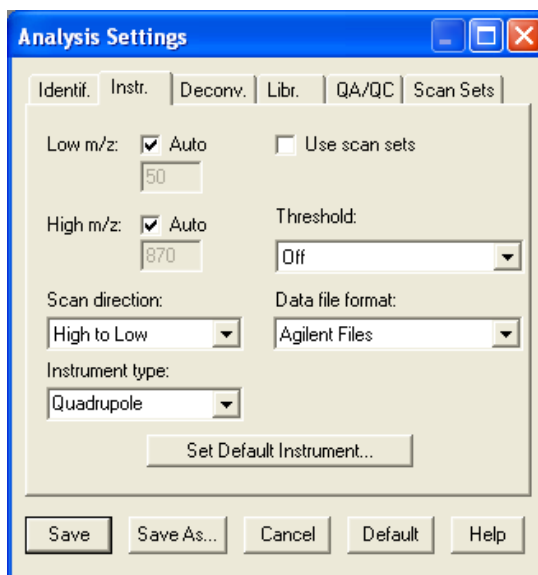


Figure J.3: Analysis settings menu

- The Deconvolution tab requires several things to be changed in order to work with our data.

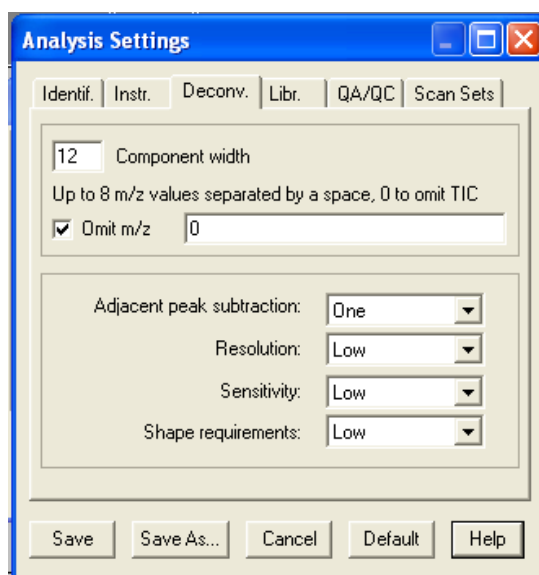


Figure J.4: Analysis settings menu

- The **Component Width** should be set (how many spectral scans are present across a peak).
- Tick **Omit m/z** and fill in 0 in order to omit the TIC as a model peak.
- The **Adjacent peak subtraction** is useful only when performing identification steps, if there is no library search performed it can be ignored. It is used to minimize the interference of adjacent peaks. Choose **none** if the spectrum is extremely clean, **two** if it is very congested.
- The choice of **Resolution** and **Sensitivity** will depend on the samples at hand. However, to minimize false positives, it is useful to start by setting both parameters to **Low**. If the sample is more complex, **Resolution** could be increased to **Medium** and so on.
- **Shape requirements** refer to how shape is taken into consideration in the deconvolution algorithm. Initially it is recommended to be set to **Low**. A higher value of **Shape requirements** implies stricter requirements on individual ions shape, and thus leads to less compound identifications.
- In the **Library**, **QA/QC** and **Scan sets** tabs nothing should be changed. If a special library should be used, then it can be selected in the **Library** tab.
- Finally, click on **Save** to store the new analysis settings.
- Back to the previous screen, click on **Run** to perform the analysis on the file opened at the beginning.
- In the following figure, a typical results window can be observed.



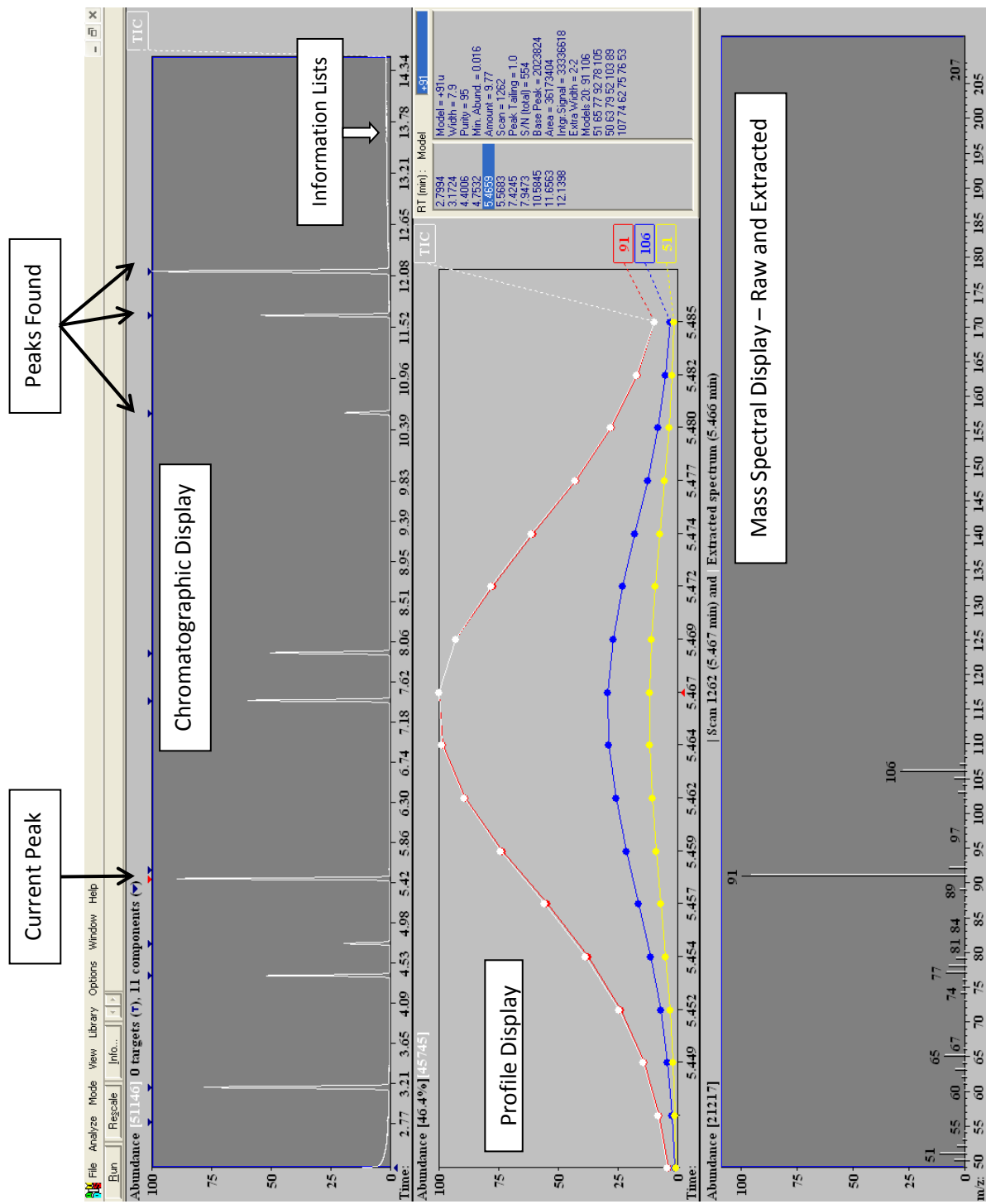


Figure J.5: AMDIS main window

- The **Chromatographic Display** shows the chromatogram extracted from the sample file, with small arrows above pointing at the peaks found by the deconvolution algorithm.
- The **Profile Display** shows the abundances of all the ions that are part of the currently selected component. When all of these curves have a similar shape, it usual suggests that this is a good quality peak and not likely to be a false positive.
- The **Information Lists** show all of the compounds found, with their retention times and by clicking on each, extra information is shown such as SNR, number of model ions, area, etc.

### J.2.4 Batch Processing

- The next step is to run a batch job on all the files we want to process. Click on **File** → **Batch Job** → **Create and Run Job**. The following screen pops up:

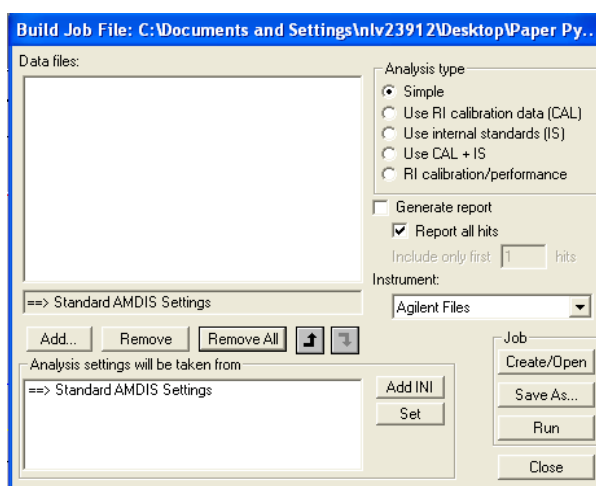


Figure J.6: Batch processing menu

- Click on **Add** and select all the files that require processing.
- **Analysis type** should be kept on **Simple**.
- Click on **Save As** in order to record the current batch job.
- Finally click on **Run** to perform the batch job.

### J.2.5 Export

- Output files \*.fin and \*.elu can be found in the same folder where each original file was found.

## J.3 Matlab Tool

The Matlab tool can be run from the command line in Matlab, by typing “Aligner”.

### J.3.1 Loading Files

- Once the application is open, click on **File** → **Open Files**
- When the file selection menu opens, choose all the files to be analyzed and click **Open**

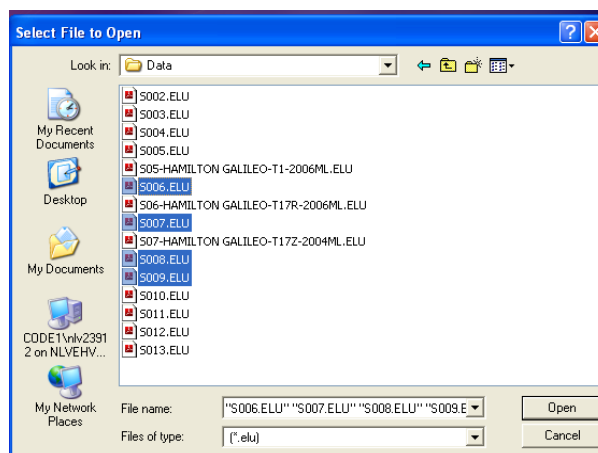


Figure J.7: File selection menu

- Set different group labels for each file by clicking on their respective label, typing another group name in the **New Label** box and clicking **Apply**.
- When all the labels are correctly set, click **Convert** to start processing the files.



Figure J.8: File management menu

### J.3.2 Quality Score Filtering

- Now that the **Quality Filtering** section is enabled, choose whether **No Filtering** will be applied or **With Quality Filtering**.
- If **No Filtering** was chosen, simply click **Filter** and move on to the **Alignment** section.
- If **With Quality Filtering** is chosen, there are three possibilities.
- Choose **Load Saved Filter** to open a \*.ftr file saved at the filter designer.
- Choose **Input Manually** to load the 7 constants for the 6 dimensional hyperplane dividing the data
- Or choose **Design Filter** to open the filter designer.
- At the filter designer, the data under analysis is loaded automatically in color green. To open a library of known true and false positives, click on **File** → **Load True/False Positives**
- In the pop-up menu, choose the \*.mat file that contains the library. In the default case, it is found in the Data folder and is called **scatter2.mat**

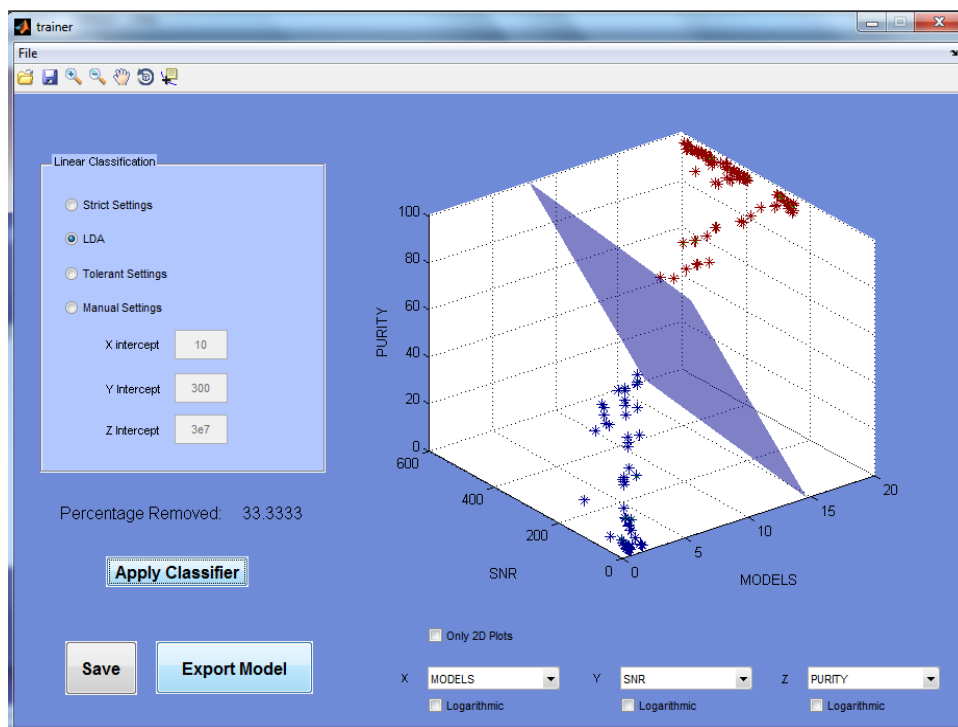


Figure J.9: Quality filtering tool

- Once the data is loaded, simply choose the parameters for plotting in the menus below the axes.
- Choose the linear classification technique from the possible options (LDA is the default) and click on **Apply Classifier**. This will show a plot of the decision surface.

- Once the desired filter is designed, the user may click on **Save** to save it as an \*.ftr file for future use
- When ready, click on **Export Model** to return to the main tool.
- Finally, click on **Filter** to move on to the next stage.

### J.3.3 Alignment

- When the alignment section is enabled, to cells appeared filled in by default.
- The **Retention Time Window** is the maximum expected time shift of peaks. In our case, 0.1 min is more than enough.
- The **Match Factor** sets the limits for detecting the similarity of two peaks. Spectra are never exactly the same, so 0.8 is a good starting value. If the program misses to align peaks, it may be lowered to a less strict condition.
- When the values are chose, click on **Align**.
- A plot with the results appears. If the results are not satisfactory and the user intends to reprocess, close the plot and click on **Reset** in the aligner.

### J.3.4 Other Output Files

- In the folder named **Output**, in the root directory, two more files are saved. A **data.arff** file for Weka, and a **data.xls** for use with Microsoft Excel. These files are overwritten with each run, so they should be renamed to preserve.

# Bibliography

- [1] Chromatography online. <http://www.chromatography-online.org/3/contents.html>.
- [2] *Handbook of Instrumental Techniques for Analytical Chemistry*. Prentice Hall, 1997.
- [3] Van den Velde S et al. Gcms analysis of breath odor compounds in liver patients. *Journal of Chromatography*, 2008.
- [4] Barker M et al. Volatile organic compounds in the exhaled breath of young patients with cystic fibrosis. *European Respiratory Journal*, 2006.
- [5] Dragonieri S et al. An electronic nose in the discrimination of patients with asthma and controls. *Journal of Allergy and Clinical Immunology*, 2007.
- [6] McGrath LT et al. Oxidative stress during acute respiratory exacerbations in cystic fibrosis. *Thorax - Journal of Respiratory Medicine*, 1998.
- [7] Montuschi P et al. Exhaled 8-isoprostane as an in vivo biomarker of lung oxidative stress in patients with copd and healthy smokers. *American Journal of Respiratory and Critical Care Medicine*, 2000.
- [8] Phillips M et al. Breath biomarkers of active pulmonary tuberculosis. *Journal of Tuberculosis*, 2009.
- [9] Studer SM et al. Patterns and significance of exhaled-breath biomarkers in lung transplant recipients with acute allograft rejection. *The Journal of Heart and Lung Transplantation*, 2001.
- [10] Van Berkel JJBN et al. A profile of volatile organic compounds in breath discriminates copd patients from controls. *Journal of Respiratory Medicine*, 2010.
- [11] University of Arizona. Introduction to mass spectrometry. [http://www.chem.arizona.edu/massspec/intro\\_html/intro.html](http://www.chem.arizona.edu/massspec/intro_html/intro.html).
- [12] Meryn S. Diagnosis of helicobacter pylori infection. 1994.