Non-targeted chromatographic high-resolution mass spectrometric analysis of biological and medical matrices with emphasis on the classification of fungal spores

Dissertation

for attaining the Academic Degree of
"Doktor rerum naturalium" (Dr. rer. nat.)
of the Departments
08 – Physics, Mathematics, and Computer Science,
09 – Chemistry, Pharmaceutics, Geography, and Geosciences,
10 – Biology,
and University Medicine
of the Johannes Gutenberg University

by

Regina Huesmann

born in Bielefeld, Germany

JG|U
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Max Planck Graduate Center mit der Johannes Gutenberg-Universität

Max Planck Graduate Center with the Johannes Gutenberg Universität Mainz

Mainz, April 2022

Faculty director: Prof. Dr. Tanja Schirmeister

1st supervisor: Prof. Dr. Thorsten Hoffmann

2nd supervisor: Prof. Dr. Eckhard Thines

Date of Examination: 23.06.2022

D77 – Dissertation of the Johannes Gutenberg University, Mainz

I hereby declare that I wrote the dissertation submitted without any unauthorized external assistance and used only sources acknowledged in the work. All textual passages which are appropriated verbatim or paraphrased from published and unpublished texts as well as all information obtained from oral sources are duly indicated and listed in accordance with bibliographical rules. In carrying out this research, I complied with the rules of standard scientific practice as formulated in the statutes of Johannes Gutenberg-University Mainz to insure standard scientific practice.

Mainz, April 2022

"I believe there is no philosophical high-road in science, with epistemological signposts. No, we are in a jungle and find our way by trial and error, building our road behind us as we proceed."

- Max Born-

Zusammenfassung

Primäre biologische Aerosolpartikel (PBAP) sind in der Erdatmosphäre allgegenwärtig und bestehen hauptsächlich aus Pilzsporen, Pollen, Bakterien und Viren. Pilzsporen sind die prominentesten PBAP. Sie können sowohl als Allergene als auch als Krankheitserreger wirken und damit die menschliche Gesundheit als auch landwirtschaftlich genutzte Pflanzen schädigen. Herkömmliche Methoden zur Untersuchung von Pilzsporen bestehen aus mikrobiologischen und mikroskopischen Methoden, oder aus der Analyse des Genoms. Diese Methoden sind zeitintensiv und entweder nicht in der Lage die Pilzspezies präzise zu bestimmen, oder, wie im Falle der Genomanalyse präzise, aber arbeits- und kostenintensiv. Daher werden zusätzliche Methoden zur schnellen und einfachen Klassifizierung von Pilzsporen benötigt.

Non-target Ultra-Hochleistungsflüssigkeitschromatographie gekoppelt mit hochauflösender Massenspektrometrie (UHPLC - HRMS) ermöglicht eine schnelle und umfassende Analyse des Metaboloms der Pilzsporen. Pilze produzieren eine Vielzahl von Verbindungen, darunter potenziell klassenunterscheidende Sekundärmetaboliten. Ziel dieser Arbeit war die Bewertung, ob eine Klassifizierung von Pilzsporen in die entsprechenden Klassen oder Spezies anhand ihres Metaboloms möglich ist, und wenn ja, die Entwicklung einer geeigneten Methode. Verschiedenste Algorithmen des maschinellen Lernens wurden untersucht, unter anderem Dimensionsreduktion, unüberwachtes Clustering und überwachte Klassifikation, um geeignete Methoden zu identifizieren. Die entwickelte Methode wurde für die Klassifizierung sowohl auf Klassen- als auch auf Spezies-Ebene angewendet.

Der Methode beginnt mit der Extraktion der Pilzsporen durch Methanol, gefolgt von einer Normalisierung anhand der Anzahl und Größe der Pilzsporen. Die UHPLC-HRMS Messungen wurde auf einer C18-Säule durchgeführt, anschließend folgte die Massenanalyse mittels Orbitrap-Massenspektrometer. Die Ionisierung wurde mit Elektrospray (ESI) und chemischer Ionisierung bei Atmosphärendruck (APCI) durchgeführt, sowohl im positiven als auch im negativen Modus. ESI im positiven Modus erwies sich als die am besten geeignete Ionisierungsmethode. Die umfassende Datenanalyse enthält eine Log Transformation, gefolgt von einer z-Score Standardisierung und einer Dimensionalitätsreduktion durch Hauptkomponentenanalyse (PCA). Die überwachte Klassifizierung zeigte eine höhere Genauigkeit als das unüberwachte Clustering, wobei die Support Vector Machine mit linearem Kernel (SVM) die besten Ergebnisse lieferte. Die Unterscheidung verschiedener Pilzporen Klassen erreichte eine Genauigkeit von 99 % mit

einer Standardabweichung von 3 %. Die Proben stammten aus vier Klassen respektive 5 Familien (*Aspergillus, Botrytis, Cladosporium*, *Verticillium* und *Trichoderma* spp.), mit 75 biologischen und zusätzlichen 20 technischen Replikaten. Für die Klassifizierung von verschiedenen Spezies wurden eine Genauigkeit von 95 % mit einer Standardabweichung von 5 % erreicht. Bei den Proben handelte es sich um 5 verschieden Spezies respektive 6 Stämmen der Gattung *Trichoderma* mit 42 biologischen und zusätzlichen 14 technischen Replikaten. Die Ergebnisse wurden durch 10-fache stratifizierte Kreuzvalidierung ermittelt und mittels Validierungsproben überprüft. Darüber hinaus wurde untersucht, ob klassen- oder spezies-spezifische Merkmale identifiziert wurden. Die hierarchische Clustering-Analyse zeigte einige spezies-spezifische Merkmale. Jedoch konnten aufgrund der hohen Variabilität zwischen den Spezies keine spezifischen Substanzen identifiziert werden, die als Marker verwendet werden könnten. Eine weitere Anwendung von non-target-UHPLC-HRMS umfasste ein kleines Probenset von Basidiomyceten Sporen auf Filtern aus dem Amazonas-Regenwald. Das hierarchische Clustering zeigte speziesspezifische Merkmalsregionen, die darauf hindeuten, dass eine Klassifizierung möglich ist. Maschinellen Lernalgorithmen für Non-Target-HRMS-Daten wurden nicht nur zur Analyse biologischer Matrices angewendet. Die hierarchische Clusteranalyse wurde in dieser Arbeit zur Untersuchung von E-Zigaretten Liquids und Kondensaten verwendet.

Abschließend wurde gezeigt, dass eine Differenzierung von Pilzsporenklassen und -spezies auf der Grundlage der non-target UHPLC-HRMS-Analyse mittels Algorithmen des maschinellen Lernens möglich ist. Die Anwendung von maschinellem Lernen ermöglichte Einblicke in komplexe biologische und medizinische Matrizes, die sonst nicht möglich gewesen wären.

Abstract

Primary biological aerosol particles (PBAP) are ubiquitous in the earth's atmosphere and consist of fungal spores, pollen, viruses, bacteria, and debris of such. Fungal spores are the most prominent PBAP and can negatively impact human health and agricultural crops, as they can act as allergens and pathogens. Traditional methods for the investigation of the fungal bioaerosol consist of microbiological cultivation and microscopy techniques, or analysis of the fungal genome. These methods are time-consuming and either not able to precisely determine the fungal species, or in the case of genome analysis precise but labor and cost intensive. Therefore, additional methods for rapid and easy classification of fungal spores in environmental samples are needed.

Non-target ultra-high-performance liquid chromatography high-resolution mass spectrometry (UHPLC – HRMS) allow a fast and comprehensive analysis of the fungal spores' metabolome. Fungi produce a variety of compounds, including potentially class- or species-distinguishing secondary metabolites. The aim of this work was the evaluation whether it is possible to differentiate fungal spore classes or species based on their metabolome and if so, to develop a suitable workflow. Therefore, various machine learning algorithms, including dimensionality reduction, unsupervised clustering, and supervised classification methods, were investigated to find a suitable classification method. The developed workflow was applied to both class- and species differentiation.

The developed workflow starts with the extraction of the fungal spores by methanol and is followed by sample normalization based on fungal spore count and size. UHPLC-HRMS measurements were performed on a C18 column followed by mass analysis with an orbitrap mass spectrometer. Ionization was carried out with electrospray- (ESI) and atmospheric pressure chemical ionization (APCI) in both positive and negative modes. ESI in positive mode was found to be the most suitable ionization method. The comprehensive data analysis included a log transformation followed by a z-score standardization and a dimensionality reduction by principal component analysis (PCA). Supervised classification showed higher accuracies than unsupervised clustering, whereby the support vector machine with linear kernel (SVM) producing the best results. The class differentiation of fungal spores resulted in classification accuracies of 99 % with a standard deviation of 3 %. Samples consisted of four classes from five different families (*Aspergillus, Botrytis, Cladosporium*, *Verticillium* and *Trichoderma* spp.) with a total of 75 biological and additional 20 technical replicates. Species differentiation resulted in classification accuracies of 95 % with standard deviations of 5 %. Samples belong to the genus

Trichoderma and contained 5 species from six strains with a total of 42 biological and additional 15 technical replicates. Results were obtained by 10-fold stratified cross-validation and verified with validation samples. In addition, the classification of mixed-species samples was tested, resulting in correct classification according to the prevailing species in the sample. Furthermore, it was studied if features were detected that are class- or species-specific. Hierarchical clustering analysis revealed some species-specific feature spaces but due to high inter-species variability, no specific features which could be used as chemical tracers were detected. Additional applications of non-target UHPLC-HRMS included a provisional sample set of basidiomycetes spores on filters from the Amazonian rainforest. Hierarchical clustering showed species-specific feature regions, suggesting that classification by is possible. Data analysis with machine learning algorithms for non-target HRMS data was not only applied to biological matrices. Hierarchical clustering analysis was used in this work to study e-cigarette liquids and condensates.

Concluding, it has been shown, that differentiation of fungal spore classes and species is possible based on non-target UHPLC-HRMS analysis using machine learning algorithms. The application of machine learning provided insights into complex biological and medical matrices, that would not have been possible otherwise.

Table of Contents

Zusammenfassung **I**

Abstract **III**

1. Introduction **1**

1.1. Bioaerosols **1**

1.1.1. Aerosols 1

1.1.2. Bioaerosols 3

1.1.3. Contributions of bioaerosol to climate and health 4

1.1.4. Fungal bioaerosol 5

1.2. Fungi **6**

1.2.1. General information about fungi 6

1.2.2. Fungal metabolism 8

1.2.3. Biology of fungal spores and their implication for health and environment 10

1.2.4. Measurement methods 14

2. Analytical Methods and Instruments **18**

2.1. Instruments **18**

2.1.1. High-Performance Liquid Chromatography 18

2.1.2. Gas chromatography 19

2.1.3. Mass Spectrometry 21

2.1.4. Ionization Sources 21

2.1.5. Mass spectrometers 26

2.2. Data analysis **30**

2.2.1. Raw data processing 30

2.2.2. Normalization 31

2.2.3. Machine learning algorithms 34

2.2.4. Dimensionality reduction 35

2.2.5. Unsupervised learning methods 39

2.2.6. Supervised learning methods 45

3. Thesis Motivation **50**

4. Experimental Setup **52**

4.1. Fungal spore samples **52**

4.2. Workflow for extraction, measurement, and data processing **58**

5. Method Development **63**

5.1. Sample preparation and measurement **63**

5.1.1. Fungal spore cultivation and extraction 63

5.1.2. Sample-based normalization 66

5.1.3. Ionization source and polarity, chromatography, and quality control 67

5.1.4. Raw data processing, blank subtraction, and alignment 69

5.1.5. Data evaluation – Semi target approach 70

5.2. Data analysis with machine learning algorithms **71**

5.2.1. Data-based normalization 71

5.2.2. Visualization and dimensionality reduction 74

5.2.3. Unsupervised Machine Learning: Clustering 78

5.2.4. Supervised Machine Learning: Classification 83

6. Results and Discussion **91**

6.1. Differentiation between fungal classes and families **91**

6.1.1. Comparison of SVM results for different ionization methods 91

6.1.2. Validation of classification 93

6.1.3. Genus's differentiation in mixed-species samples 95

6.1.4. Evaluation of feature space 97

6.1.5. Hierarchical clustering of features 98

6.1.6. Evaluation of possible specific fingerprints 103

6.1.7. Summary of fungal class/family differentiation 104

6.2. Differentiation between fungal species **106**

6.2.1. Supervised classification 106

6.2.2. Hierarchical clustering analysis 109

6.2.3. Evaluation of possible species-specific fingerprints 111

6.2.4. Species differentiation including *T. harzianum* strains 115

6.2.5. Summary of fungal species differentation 117

6.3. Additional fungal samples **118**

6.3.1. Basidiomycetes spores from the Amazonian rainforest 118

6.3.2. Fungal volatile organic compounds 121

6.4. Non-target LC-MS analysis of electronic cigarettes **124**

6.4.1. Introduction 124

6.4.2. Experimental work 125

6.4.3. Results 126

7. Conclusions and Outlook **130**

8. Appendix **133**

8.1. Supporting information **133**

8.1.1. Instruments, chemicals, and programs 133

8.1.2. Supporting information for fungal class differentiation 151

8.1.3. Supporting information for fungal species differentiation 153

8.1.4. Further supporting information 157

8.2. List of abbreviations **160**

8.3. List of Figures **162**

8.4. List of Tables **164**

9. References **167**

10. List of related poster presentations and publications **187**

11. Acknowledgements **188**

12. Curriculum Vitae **189**

1. Introduction

Fungal spores are ubiquitous in the air. They influence our lives in many ways, from enhancing cloud formation over acting as allergens to either destroying crops as plant pathogens or acting beneficially as biological plant protectants. Fungal spores are primary biological aerosol particles which are also called bioaerosol.

1.1. Bioaerosols

1.1.1. Aerosols

Aerosols generally are described as a heterogeneous suspension of solid or liquid particles in a gas phase. Sources of atmospheric aerosols can vary between natural and anthropogenic sources and primary and secondary particle formation. Primary particles are directly emitted from their source, natural aerosols like mineral dust (e.g., Sahara dust events), sea spray, but also anthropogenic like industrial exhaust or soot. Secondary particles are formed in the atmosphere by gas to particle conversion, precursors can be plant emissions like isoprene or alpha-pinene, or anthropogenic gases from e.g., car engines like NO_x. In total 12800 teragram (Tg) of particles are emitted or formed in the troposphere (2 – 20 km altitude) with some particles even reaching the stratosphere (16 – 25 km altitude). Aerosols from anthropogenic sources contribute less to the global aerosol load than those from natural sources but are prone to negatively influencing climate and health (Pöschl, 2005, Schnelle-Kreis et al., 2007, Seinfeld, Pandis, 2012). An overview of the source's contribution to the aerosol load is given in Table 1.1.

Table 1.1: Estimated aerosol emissions in teragrams per year according to their origins (Seinfeld, Pandis, 2012).

Source	Estimated Tg/a
Natural - Primary	
Mineral dust	2407
Sea salt	10000
Volcanic ash	30
Bioaerosol	50
Natural – Secondary	
Sulfates	32
Organic aerosol, from biogenic precursors	11
Total natural	**12530**
Anthropogenic- Primary	
Industrial dust (without soot)	100
Soot	12
Organic aerosol	81
Anthropogenic - Secondary	
Sulfates	49
Nitrates	21
Total anthropogenic	**263**

Mineral dust and Sea salt are the dominant sources of aerosol particles. However, anthropogenic particles like industrial dust contribute to the aerosol load as well, especially in densely populated areas. The effects of aerosols on climate and health depend on the nature of the particle.

Aerosols can influence the earth's climate directly and indirectly. They influence directly by scattering and absorbing light from solar and terrestrial sources. Scattering solar radiation has a cooling effect, whereas black carbon absorbs solar radiation and has a warming effect (Pöschl, Shiraiwa, 2015). They influence indirectly by contributing to cloud formation, as cloud condensation nuclei (CCN) and ice nuclei (IN). Without aerosol particles to enable condensation or freezing, water would need to be supercooled for hours, or in the case of homogenous freezing, temperatures need to fall to -36.5 °C or lower (Henderson-Begg et al., 2009). Ice nuclei enable freezing processes at temperatures between – 30 °C (mineral dust) and 1.5. °C (biological particles, see also in chapter 1.1.3). By enabling cloud formation, thus increasing radiation reflection aerosols contribute to negative forcing, as well as influencing the earth's hydrological cycle. The total feedback of aerosols to the global climate is still not completely understood, a total positive radiative forcing is estimated (Lohmann, Feichter, 2005).

Aerosols can also influence human health. Air Quality with a high aerosol load of fine particulate matter (PM 2.5.) is known to correlate with higher mortality and increased risk for cardiovascular disease. The finer the particle, the deeper they can be inhaled into the lung. Bigger particles (> 5 µm) deposit in the larger bronchial areas but smaller particles (< 0.5 µm) reach the alveoli where they can cause chronic inflammatory diseases or even cancer. (Krug, Wick, 2011, Pöschl, 2005, Pöschl, Shiraiwa, 2015). The health effects of biological aerosols are described in chapters 1.1.3 and 1.2.3.

1.1.2. Bioaerosols

Bioaerosols or more exactly primary biological aerosol particles (PBAP) are natural, primary aerosols from living or dead biological organisms, such as fungal spores, bacteria, viruses, and pollen but also from plant debris and animal dander. In subordinate quantities also archaea, fern spores, and lichen, cyanobacteria are part of the biological aerosol. The estimated total global emissions of PBAP lay in a wide range between 10 and 1000 Tg/a, depending on which size range and if cellular fragments are included in the estimation (Després et al., 2012, Fröhlich-Nowoisky et al., 2016). PBAP are between 1 nm and 100 µm in size, an overview is given in Figure 1.1.

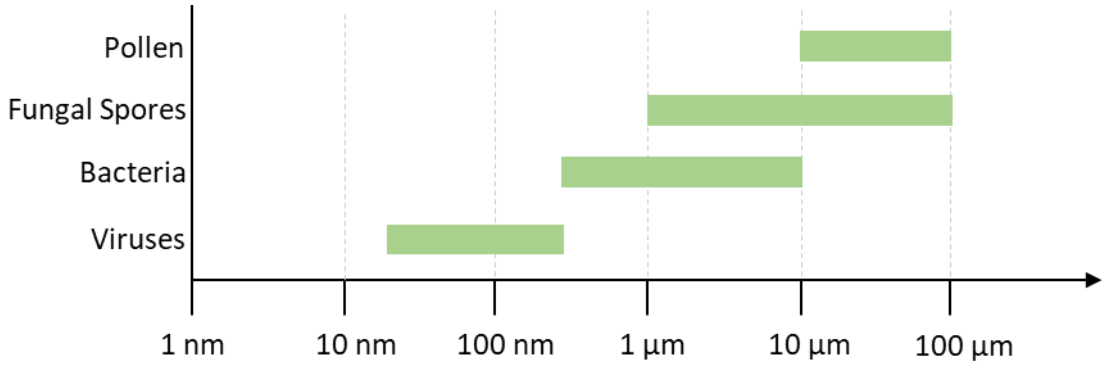


Figure 1.1: Size ranges of major biological aerosols. After (Fröhlich-Nowoisky et al., 2016).

Viruses are the smallest PBAP, followed by bacteria (<1 – 4 µm). Spores are usually bigger, with sizes between 1 and 50 µm. The largest PBAP type is pollen with sizes between 9 to over 40 µm. Fungal or plant fragments can reach sizes up to 100 µm. Larger particles sediment quickly and therefore do not travel far, but smaller particles can travel far distances either by themselves or attached to dust and even cross oceans. Also, high altitudes are possible with metabolically active PBAB (fungi and bacteria) found at 20 km height (Bowers et al., 2009, Després et al., 2012, Fröhlich-Nowoisky et al., 2016).

Biological particles occur in high number concentrations of up to 10^4 particles per m^3, depending on the particles' size they contribute to the mass load as well. An overview of the aerosol load by different PBAP classes is given in Table 1.2.

Table 1.2: Estimated PBAP emissions (Després et al., 2012, Fröhlich-Nowoisky et al., 2016).

Source	Global emission Tg/a	Number concentration [cells m^{-3}]	Mass concentration [µg m^{-3}]
Bacteria	0.4 - 28	10^4	0.1
Fungal spores (and fragments)	8 - 190	10^3 - 10^4	0.1 - 1
Pollen	47 - 84	10 - 10^3	1
Plant debris	-	-	0.1 - 1
Algae	-	100 - 10^3	10^{-3}
Viruses	-	10^4	10^{-3}

Bigger particles like plant debris have a high mass, but a low number concentration, whereas very small particles like viruses have a high number concentration but a small mass concentration. Interesting are PBAP with a high global emission, like bacteria, pollen, and fungi. Fungal spores and fragments are one of the biggest contributors to the global bioaerosol load and one of the most prevalent classes of PBAP. General implications of biological particles for health and the environment are given in the following chapter, with more in-depth information on fungal bioaerosols in chapters 1.1.4 and 1.2.3.

1.1.3. Contributions of bioaerosol to climate and health

As mentioned in chapter 1.1.1 aerosols can act as cloud condensation nuclei (CCN) and ice nuclei (IN). PBAP are known to act as very effective ice nuclei, enabling freezing in clouds at much higher temperatures than mineral ice nuclei. Biological particles, mostly fungi, and bacteria can induce freezing at temperatures above -10 °C and even at temperatures as high as -1.8 °C (*Pseudomonas syringae*). *P. syringae* for example, has ladder-like proteins on its surface, which order water molecules, thus supporting ice formation. (Henderson-Begg et al., 2009, Pandey et al., 2016). Acting as IN/CCN enables the deposition and distribution of biological particles. It is estimated, that PBAP are important for ice formation in warm clouds (temperature > -15 °C) and even dominate cloud formation in warm, humid regions with high biological aerosol load, like rainforests (Andreae, Rosenfeld, 2008, Spracklen, Heald, 2014). Especially over pristine regions, PBAPs can influence the hydrological cycle. Over the Amazonian rainforest, PBAB contribute up to 67 % of the total particulate matter

and absorb up to 47 % of the radiation, which can therefore influence the climate on a local scale (Fröhlich-Nowoisky et al., 2016, Spracklen, Heald, 2014).

Biological aerosols can be harmful to humans' health, not only by the effects described in chapter 1.1.1 but also because of their unique biological properties. The most prominent example right now is the COVID-19 (coronavirus disease 2019) pandemic, with SARS-CoV-2 (severe acute respiratory syndrome coronavirus type 2) viral particles being transmitted by air (Robert Koch-Institut, 2021). Many diseases caused by airborne biological particles are known, from severe sicknesses like tuberculosis (bacteria) or diphtheria (bacteria) to milder illnesses like the common cold (virus). PBAP can also cause allergies and asthma, or chronic inflammatory diseases when biological particles are inhaled in high concentrations over a long period. Biological aerosols can also produce endotoxins (bacteria) and mycotoxins (fungi). (Fröhlich-Nowoisky et al., 2016, Kim et al., 2018). Further elaboration on the health impact of fungal spores is given in chapter 1.2.3.

1.1.4. Fungal bioaerosol

One of the most prevalent classes of PBAP, and a large source of organic aerosol, are fungal spores. Global estimated emissions for fungal spores vary greatly between 50 and 186 Tg/year (Elbert et al., 2007, Jacobson, Streets, 2009, Janssen et al., 2021a, 2021b). Modelling estimated high fungal spore abundancy over tropical rainforest., accounting for up to 45 % of the particle mass (1 – 10 µm) (Elbert et al., 2007). Figure 1.2 shows the estimated mean number concentration of fungal spores in the boundary layer. Notable is the intensity over rainforests, especially the Amazonian rainforest, where an influence on the local hydrological cycle by PBAP is proposed (Elbert et al., 2007, Heald, Spracklen, 2009).

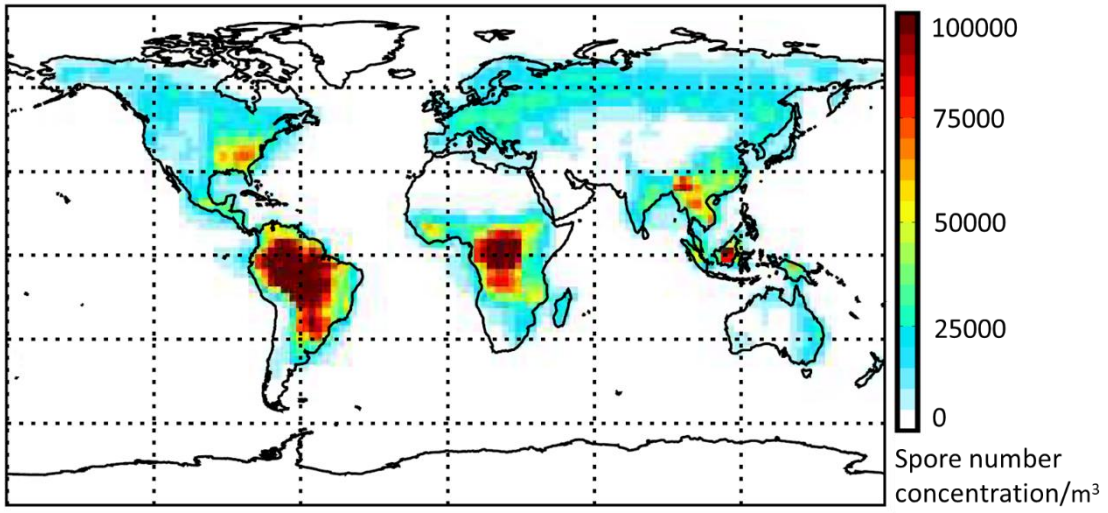


Figure 1.2: Simulated surface annual mean of fungal spore number concentration per cubic meter in the boundary layer, dark red represents the highest concentration, mainly in regions with rainforests (Amazonian, central Africa, and southeast Asia) (Spracklen, Heald, 2014).

In urban and suburban areas fungal spore contribution is smaller but still essential at up to 11 % in fine particle mass (< 2.5 μm). Fungal spores contribute to 8 % of the continental supermicron (> 1 μm) number concentration. (Spracklen, Heald, 2014). To gain further understanding of the fungal spore composition some basics about fungi are needed which are discussed in chapter 1.2.

1.2. Fungi

1.2.1. General information about fungi

Fungi are important for our everyday life, from food like mushrooms, over symbiotic relationships with plants to the production of life-saving medications (penicillin, cyclosporine). They are essential for ecosystems as they act as decomposers and symbionts, as well as pathogens. The symbiosis with plants, e.g., trees or photobionts like algae improves plant growth by absorbing soil nutrients. Fungi are heterotroph organisms, meaning they obtain carbon and energy by absorption, and not by photosynthesis (Bonfante, Genre, 2010, Webster, Weber, 2007, Wu et al., 2019).

As eukaryotes, fungi are, contrary to beliefs in former times, more closely related to animals than to plants. Fungi separation from animalia is suggested to have taken place 1.5 billion years ago, about 9 million after separation from plantae, with fungal fossils found to be 550

to 350 million years old (Hibbett et al., 2007, James et al., 2006, McLaughlin et al., 2009, Wang et al., 1999). Fungi are the third-largest eukaryotic kingdom besides animalia and plantae, with estimated species numbers between 2.2 – 3.8 million and 11.7 – 13.2 million. (Hawksworth, Lücking, 2017, Wu et al., 2019). An overview of the evolutionary separation is given in Figure 1.3. The taxonomic kingdoms, e.g., bacteria, animalia, or fungi are further taxonomically separated into phyla, classes, orders, families, genera, and finally species, reflecting evolutionary relationships between organisms. An overview of the separation of the kingdom fungi into phyla is also given in Figure 1.3.

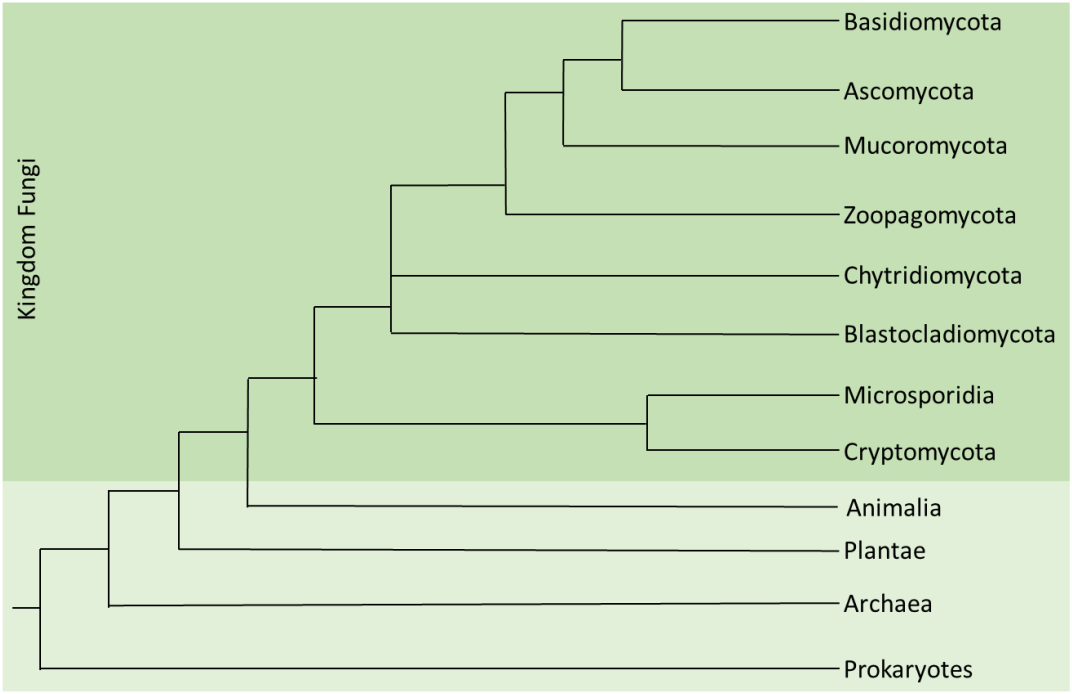


Figure 1.3: Taxonomy of the kingdom true fungi, starting from the proposed last universal common ancestor separation. The separation of the kingdom fungi from the kingdom prokaryotes, archaea, plantae, and animalia is proposed to have taken place in the shown order (light green area). The kingdom fungi is separated into several phyla, shown in the dark green area. The length of the branches is not proportional to evolutionary distances. Drawing after: (Moore et al., 2020, Spatafora et al., 2017).

Classification of fungi remains a dynamically changing research field, with phylogenetic analysis bringing new information to light which leads to regular alterations of taxonomic trees. Historically fungi were separated into the kingdom fungi, the oomycetes, and the slimes molds, e.g., myxomeycetes. Nowadays it is clear, that oomycetes are closer related to algae than to the rest of the fungi and slime molds are not classified as fungi anymore. Later the kingdom fungi was divided into four phyla: Ascomycota, Basidiomycota, Chytridiomycota, and Zygomycota. Nowadays phylogenetic analysis suggests seven to eight different phyla: Basidiomycota and Ascomycota, Mucoromycota, Zoopagomycota,

Chytridiomycota, Blastocladiomycota, Microsporidia, and Crypotmycota, the last two were suggested to belong to the phyla Opisthosporidia (Adl et al., 2012, Li et al., 2021, Moore et al., 2020, Naranjo-Ortiz, Gabaldón, 2019, Tedersoo et al., 2018, Webster, Weber, 2007).

The most important lineages, Ascomycota and Basidiomycota together form the clade Dikarya (Hibbett 2007). Ascomycota has 64000 described species and Basidiomycota 31500 described species. Most edible fungi are Basidiomycota, like the button mushroom *Agaricus* or the *Boletus edulis* . Commonly known Basidiomycota have a stem and a fleshy cap, but there are also basidiomycetes without a stem, like rusts (fungi from the order *Pucciniales*) and smuts (fungi from the order *Ustilagnales*). Typical examples of ascomycetes are brewer's/baker's yeast, *Penicillium* of which the first antibiotic was detected, or *Aspergillus* which is commonly known as black mold (*Aspergillus niger*). Ascomycetes can act as pathogens for either human/animals (*Candida alibans, Aspergillus niger*) and plants (mildews, etc.) but also as beneficial organisms. For further information see chapter 1.2.3 (Hibbett et al., 2007, Moore et al., 2020, Spatafora et al., 2017, Webster, Weber, 2007).

1.2.2. Fungal metabolism

The primary metabolism describes the processes and molecules needed for the homeostasis, like growth, respiration, and reproduction. The primary metabolism is evolutionary conserved and consists of compounds like sugars, amino acids, fatty acids, and nucleosides. Out of those precursors, more complex molecules are formed: polysaccharides, proteins, lipids, and nucleic acids. All other metabolites are described as secondary metabolites, meaning they are not immediately necessary for an organism's life, but increase the fitness and chance of survival immensely. Some secondary metabolites can specifically occur in certain steps of the fungi's life cycle, like reproduction or when influenced by the environment. Fungi produce a broad variety of metabolites, from highly sought metabolites like penicillin to toxic compounds like aflatoxin (Bayram, Braus, 2012, Boruta, 2018).

The primary metabolism of fungi is very similar to other eukaryotes. Fungal cell walls consist of 80 – 90 % polysaccharides, with lipids and proteins as the remainder. Also dominant in fungal cell walls, at least in Asco- and Basidiomycota, is the sterol ergosterol, in its function comparable to animals' cholesterol. As well as ergosterol, fungi use chitin in their cell walls, comparable to plants, which use cellulose. Further compounds are β 1,3-

glucan and on the outside of the cell wall a layer of mannoproteins (Moore et al., 2020, Weete et al., 2010).

The secondary metabolism varies immensely between fungal species (Zeilinger et al., 2015b). Fungal secondary metabolites are diverse and complex classes of molecules, like polyketides, terpenoids, alkaloids or non-ribosomal peptides. They can be used for promoting growth, defense, or even communication between microorganisms. The occurrence or concentration of secondary metabolites depends on the life cycle, the environment, and the growth condition of the fungi (Boruta, 2018, Keller, 2019, Zeilinger et al., 2015b, Zeilinger et al., 2016). To highlight the variety of fungal secondary metabolisms a few structures are presented in Figure 1.4.

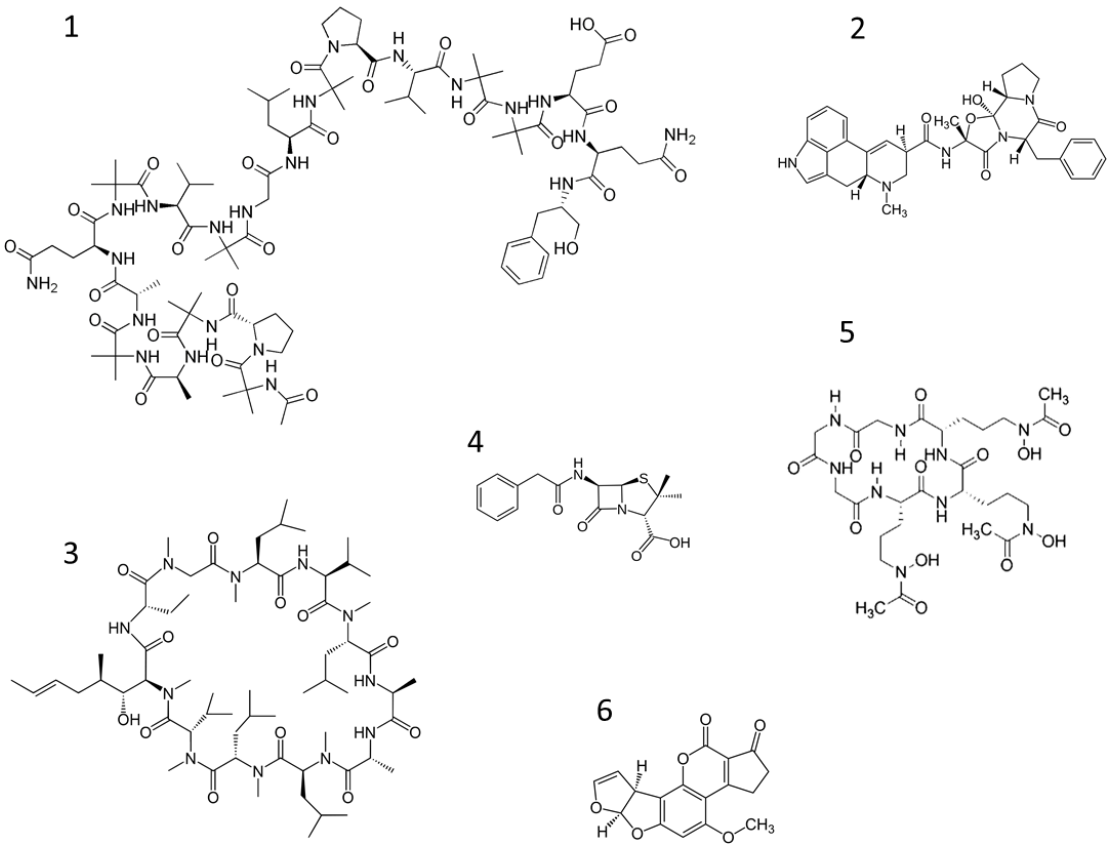


Figure 1.4: Examples of secondary metabolites. 1) Alamethicin, peptaibol from *Trichoderma* spp. (Payne et al., 1970). 2) Ergotamin, alkaloid from *Claviceps purpurea* (german: Mutterkorn) (O'Neil, 2006). 3) Cyclosporin A, first detected in *Tolypocladium inflatum* (Merluzzi, 1995). 4) Penicillin G, beta-lactam antibiotic from *Penicillium notatum* (Dexter, van der Veen, 1978). 5) Ferrichrome, siderophore first detected in *Ustilago sphaergoena* (van der Helm et al., 1980). 6) Aflatoxin B1, mycotoxin from *Aspergillus* spp. (Do, Choi, 2007).

Shown are diverse secondary metabolites, with different heteroatoms and functional groups, some are peptides, and some are small molecules. The secondary metabolites have different

functions, e.g., Ferrichrome (5) helps with the fungi's iron uptake whereas Alamethicin (1) or Penicillin (4) are part of the fungis' defense mechanisms. Secondary metabolites of fungi can be used as drugs, e.g., Cyclosporin (3). Many antibiotics, like Penicillin (4) originate from fungi. (Keller, Turner, 2012, Moore et al., 2020, Webster, Weber, 2007).

1.2.3. Biology of fungal spores and their implication for health and environment

Spores are the reproductive units of fungi. The size ranges from 1 up to 50 µm, with most spores varying between 2 and 10 µm. Spores can be dispersed by wind and over wide distances, e.g., in the case of dust events across oceans. They withstand extreme temperatures, UV radiation, and dryness and still germinate after years. Viable forms of fungal spores were found in deserts, ice from glaciers, tundra, hailstones, the artic, and even on the international space station (ISS) (Cortesão et al., 2020, Feofilova et al., 2012, Fröhlich-Nowoisky et al., 2016, Griffin, 2007).

Spores have a reduced metabolism, thick cell walls, and usually some form of pigmentation to withstand UV radiation. They are rich in carbohydrates like glycerol, mannitol, arabitol, erythritol, and trehalose increasing their heat resistance. Other components increasing the spore's survival in unfavorable conditions are e.g., glycolipids, lysophosphatidic acid, chitin, chitosan, glucans, sporopollenin, hydrophobins, and melanin. Spores remain dormant, meaning with a reduced metabolism (maximum 50% metabolic activity) until a suitable environment is reached. Water alone is not sufficient to "awake" any spore, usually, carbon and/or nitrogen sources are necessary. Secondary metabolites can control the dormancy, with several substances suppressing or inducing germination (Dijksterhuis, 2019, Feofilova et al., 2012, Keller, 2019, Madelin, 1994, Moore et al., 2020, Thines et al., 2004).

Spores can result from sexual or asexual reproduction and are mostly haploid. Most fungi can reproduce sexually (teleomorph) and asexually (anamorph), but some are only known with asexual reproduction (imperfect fungi). Some fungal spore types are characteristic for their clade, like basidiospores (Basidiomycetes) or ascospores (sexual reproduction of Ascomycetes). Conidia are produced by mitosis and are typical for the asexual reproduction of Ascomycetes. Some spore types are more specialized, like teliospores or chlamydospores which are thick-walled resting spores to survive unfavorable conditions (Elbert et al., 2007, Janssen et al., 2021a, Moore et al., 2020).

Spores can be dispersed actively, e.g., by osmotic pressure or surface tension and other mechanisms, leading to the accelerated discharge of spores. Passive, "dry" discharge is

achieved by wind speeds at 1 $m \cdot s^{-1}$. Spore discharge is connected to meteorological parameters like temperature, humidity, and wind speed, also day-night rhythms were observed with spore discharge higher at night when the relative humidity was increased and can vary depending on the season, especially regarding different species. Additionally to metrological conditions, location can influence the presence of certain spore species, with rural locations showing higher fungal diversities and abundances than urban locations. (Abrego et al., 2020, Oliveira et al., 2010). The highest fungal spore concentrations were found over tropical rainforests (see chapter 1.1.4) (Elbert et al., 2007, Fröhlich-Nowoisky et al., 2012, Fröhlich-Nowoisky et al., 2016).

Fungal spores are the most genetically diverse group of PBAP. The composition of the fungal bioaerosol is complex and mostly unknown. Besides the unknown diversity of airborne fungal spores, their contribution to the global aerosol load remains unclear. Previous studies examined fungal spores sampled in outdoor air, with conidia compromising 30 to 60 % of the total amount, with the remaining percent being sexual spores. Cladosporium seems to be the most predominant genus found, present in almost all samples on every continent (Ovaskainen et al., 2020, Pace et al., 2019). Fungal spores were present in the fine (< 2.5 µm) and the coarse (> 2.5 µm) mode with similar species richness in both fractions (Fröhlich-Nowoisky et al., 2016, Janssen et al., 2021b). Most detected airborne fungal spores belong to the clade Dikarya, with one study estimating Basidiomycetes at 64 % and Ascomycota at 34 %, as well as examining the classes of Ascomycetes and Basidiomycetes, see Figure 1.5 (Fröhlich-Nowoisky et al., 2009).

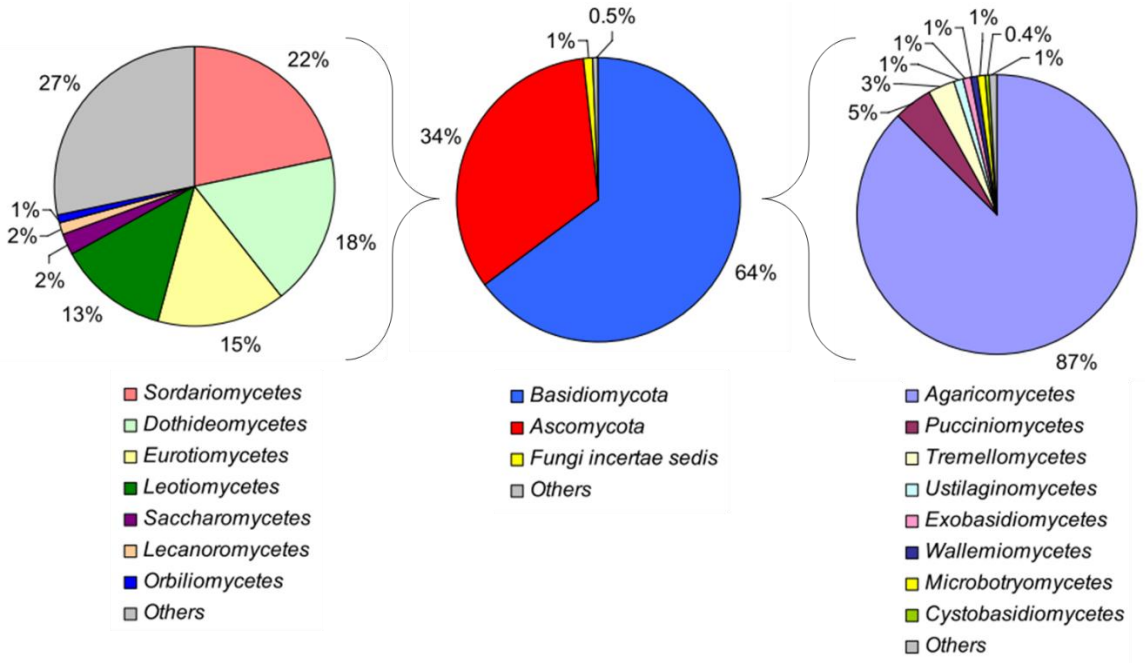


Figure 1.5: Diversity of airborne fungal spores, as examined by DNA analysis. Left: Class diversity of Ascomycetes. Middle: Diversity of phyla, Ascomycota (red), and Basidiomycota (blue) are dominant. Right: Class diversity of Basidiomycetes (Fröhlich-Nowoisky et al., 2009).

For Basidiomycetes, the class Agaricomycetes is dominant with 87 %. Among the Ascomycetes the diversity between classes is more distributed, whereby the composition of the different classes can change with the season, e.g., the genera *Cladosporium* spp. and *Alternaria* spp. occurring more in summer and *Penicillium* spp. or *Blumeria graminis* in winter (Fröhlich-Nowoisky et al., 2009). There are several analytical methods to examine airborne fungal spores, an overview is given in chapter 1.2.4.

Implications of fungal spores on health and environment

Fungal spores in the air can have diverse influences on ecosystems and their inhabitants. Additionally, to the already mentioned climate effects, fungi can act as major pathogens for humans (see also chapter 1.1.3). They can cause allergic reactions or infections. The impact on human health is significant, 20 to 30 % of all allergic asthma diseases are caused by ~~mildew~~ mold allergies, and 30 % of the total population reacts sensitively to fungal allergens. Common fungal allergens are known from *Alternaria*, *Cladosporium*, *Aspergillus*, and *Penicillium*. Fungal spores are small enough to be inhaled deep into the lung, reaching bronchia and causing inflammations. Serious or life-threatening fungal infections usually only affect immuno-compromised persons. (Dales, Munt, 1982, Fröhlich-Nowoisky et al., 2016), (Fischer, Dott, 2003, Rivera-Mariani, Bolaños-Rosero, 2012). Fungal spores can also infect crops and in severe cases, fungal diseases destroy them completely, which results in

famine and huge economical damage to agriculture. A prominent example is the Irish potato famine from 1845 to 1852, where the potato harvest was destroyed by the fungal-like microorganism oomycetes *Phytophthora infestans* causing the potato blight (Haas et al., 2009). Nowadays fungal diseases are on the rise, possibly due to climate change, in some cases with extensive resistance to fungicides. The most prevalent fungal plant pathogens now are *Magnaporthe oryzae*, *Botrytis cinerea.* and *Puccinia* spp. Fungal diseases will remain a threat to food security (Fisher et al., 2012, Ghosh et al., 2018, (Almeida et al., 2019, Kim et al., 2018).

But besides the negative impact fungi also act as symbionts and can enhance plant growth and can improve plant stress tolerance (Singh et al., 2011). In some cases, fungi can be used as a biological plant protectant, meaning a suspension of suitable fungal spores is applied on the field, introducing the beneficial organism onto the plant. One of the most prominent biological plant protectants is the genus *Trichoderma*. It emits several secondary metabolites which can promote plant growth and influences root formation. *Trichoderma* spp. are used in this work. A more detailed description of the fungal species used in can be found in chapter 4.1. As one example *Trichoderma atroviride Sc1* is sold under the name "Vintec" as a biological control acent against the grapevine disease "Esca" which is caused by pathogenic fungi (Zin, Badaluddin, 2020). Biological plant protectants are seen critically by some, as it can't be excluded that toxic secondary metabolites are formed, which do not form under artificial conditions in a laboratory. A need for modern tools which could control whether toxic compounds are formed *in situ* is needed. Also, the influence of biological plant protectants on biological diversity is not well examined. (Abdelfattah et al., 2018, Deising et al., 2017) Nonetheless, biological plant protectants might play an important role in the future, as the global population is increasing and the need for food production with it. Chemical pesticides/fungicides can be problematic as they can pose risk to health, environment, and biodiversity (Almeida et al., 2019, Ghosh et al., 2018, Kvakkestad et al., 2020).

Biological plant protectants might present an alternative to conventional plant protectants, but further research is needed. Overall, there are many uncertainties about the occurrence of fungal spores in the air, whether there are beneficial or harmful. Several measurement methods are available, which are presented in the next chapter with established and possible novel methods discussed.

1.2.4. Measurement methods

The detection and analysis of fungal spores in the air can be accomplished in different ways. Either online/real-time measurements are employed, or, more often, samples are collected by filters, spore traps, impactors, cyclones, electrostatic precipitation, or impingers (Caruana, 2011, Després et al., 2012). In the following an overview of established and novel methods is given.

Established methods

Real-time sensing uses fluorescence spectroscopy, Raman spectroscopy, light scattering, or online mass spectrometry. Current research examines the real-time determination of which PBAP is present, e.g., if it is pollen or bacteria. The identification on a more in-depth taxonomic level e.g. if a fungal spore is basidiomycetes or ascomycetes is not possible. (D Strycker et al., 2019, Gosselin et al., 2016, Huffman et al., 2019). For in-depth identification, offline methods are used, like cultivation, microscopy, DNA/RNA sequencing, or chemical tracers. With cultivation living spores from viable samplers can be grown on a suitable medium and further analyzed. The back draw is that only a very small fraction of fungal spores is cultivable (~17 %). To examine all airborne fungal spores other, more elaborate methods are needed (Fröhlich-Nowoisky et al., 2016, Gosselin et al., 2016, Rivera-Mariani, Bolaños-Rosero, 2012).

The easiest offline/sample-based method would be light microscopy, where a trained person distinguishes spore types, some on the genus level, but others only on the family or class level, by their image. Manual counting can give the total spore count but is time and labor-intensive and prone to mistakes. With fluorescence microscopy and elaborate data analysis, current research develops methods to identify biological particles from non-biological. With flow cytometry counting of labeled cells is possible (Després et al., 2012, Kumar, Attri, 2016).

The most accurate method is by sequencing the fungi's DNA or RNA. This approach is very exact but cost-intensive and needs highly trained personnel and laboratory equipment. Some regions in an organisms' genome are very specific for the genus or even species and can be used for identification. In the case of fungi, the ribosomal DNA internal transcribed spacer (*ITS*)-region is the primary fungal barcode. As not all fungal species are accurately identified by *ITS*-sequencing, a secondary barcode, the translational elongation factor 1α (*TEF1α*) was introduced but is not commonly used yet. For sequencing, either Sanger sequencing or next-generation sequencing is used. Quantitative PCR can be used to estimate

the DNA concentration, thus allowing conclusions about the initial sample concentration. With a BLAST approach also unidentified samples can be compared to find matching sequences in a data bank. Another DNA-based approach is the terminal restriction fragment length polymorphism technique (TRFLP) where target genes are fluorescence marked. The fluorescence dyed fragments give a TRFLP profile that can be compared to other samples or a data bank. In contrast to barcoding, TRFLP can't be used for taxonomic studies. (Després et al., 2012, Meyer et al., 2019, Ovaskainen et al., 2020, Womack et al., 2015).

The third major approach besides microscopy and DNA analysis is the chemical tracer analysis. Chemical tracers, also called biomarkers are molecules that are characteristic of the sample. There are tracers available for all major PBAP groups, with ergosterol, mannitol, and arabitol used for fungi. As mannitol and arabitol also occur in plants, only ergosterol is a reliable marker for fungal spores. The quantitative abundance of fungal spores and PBAP can be determined by this approach, but not the spore's species. (Buiarelli et al., 2013, Buiarelli et al., 2019, Di Filippo et al., 2013).

Overall, all current methods have their limitations, either in the accuracy or in costs and time. Still, the airborne fungal diversity is not well examined, and in times of growing health threats and food insecurities more knowledge, about when, where, which, and how many fungal spores are present, especially regarding pathogens and agricultural pests, is needed.

Novel methods

A novel approach is using chemotaxonomy combined with metabolomics. Fungal secondary metabolites can be species or even strain-specific and present a metabolic fingerprint. In some cases, DNA analysis, especially when just analyzing the ITS sequence cannot differ between more closely related species, e.g. *Aspergillus* and *Penicillium*, or within fungal species like *Trichoderma* (Kang, 2011, Lücking et al., 2020). Analysis of the metabolome can bring insights into relationships between fungi (Aliferis et al., 2013, Kang, 2011, Lücking et al., 2020, Maciá-Vicente et al., 2018, Zwickel et al., 2018).

For the chemotaxonomic approach non-target analysis, where all features of a sample are detected, is needed (Kluger et al., 2015). Non-target analysis usually requires high-resolution mass spectrometry, often coupled to chromatography (for further information see chapter 2.1). As non-target high-resolution mass spectrometry data produces a considerable feature list, the application of machine learning algorithms (see chapter 2.2.3) can help obtain sensible information out of non-target data. Many of these techniques were first used in bioinformatics/DNA/RNA analysis, but with the further spread of high-

resolution mass spectrometers are made usable in the field of e.g., metabolomics. An explanatory workflow for non-target metabolomics is shown in Figure 1.6, with further explanations in the following chapter.

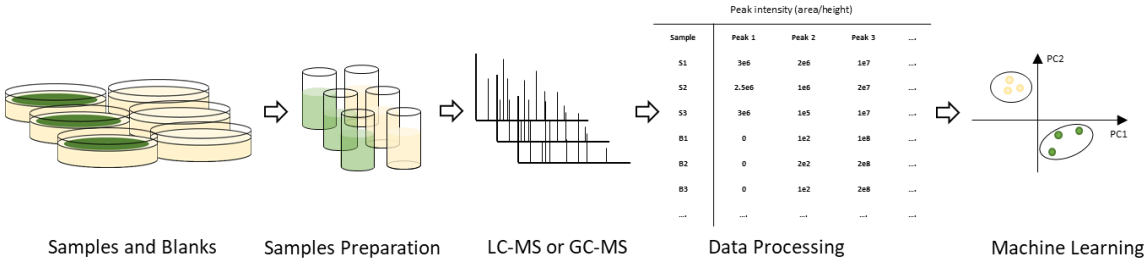


Figure 1.6: Explanatory workflow for metabolic profiling and fingerprinting. Figure after (Zeilinger et al., 2015b).

Samples need to be carefully chosen, as well as conditions in which samples are grown to account for biological variation. Blank samples are needed to make sure detected substances originate from the sample and not e.g., the growth media. Data needs to undergo extensive pre-processing like blank subtraction, filtering steps, etc., resulting in an extensive list of "features" that might be specific for the fungal species or strain. As manual control and comparison between samples are not possible anymore, as feature lists easily have thousands of entries, several algorithms and techniques from machine learning are used to identify which compounds are of interest. Dimensionality reduction and clustering are possibilities to get trends out of non-target data (Aliferis et al., 2013, Kluger et al., 2015, Smedsgaard, Nielsen, 2005, Zeilinger et al., 2015b).

Non-target analysis in general and of biological samples brings some challenges. Metabolite concentration between primary and secondary can differ in orders of magnitude, with primary metabolites like e.g., ergosterol being much more abundant than secondary ones. Most primary metabolites are evolutionary conserved and have relatively constant concentration therefore won't help much in the differentiation between species. Secondary metabolites on the other hand show much larger differences and can be influenced by the environment. Environmental factors can influence which secondary metabolites are expressed, possibly leading to different metabolic profiles of the same species, if grown under different environmental influences (González-Riano et al., 2020, Müller et al., 2013, Smedsgaard, Nielsen, 2005).

In recent years metabolic profiling was performed on several fungal species, but to our knowledge always on filamentous fungi and not on the spores. Metabolic profiling was used

in chemotaxonomy and enabled differentiation between species and strains that would not have been possible by traditional phenotyping (Smedsgaard, Nielsen, 2005). In some cases, chemotaxonomy was consistent with results from ITS-sequencing. Several groups used LC-MS measurements to examine differences in the metabolome. Separation of *Alternaria* species was performed by (Gotthardt et al., 2020), evaluation of root endophytic fungi belonging to different orders by (Maciá-Vicente et al., 2018), of the species *Ascochyta* and *Phoma* which belong to the same order by (Kim et al., 2016), of *Trichoderma* species by (Kang, 2011) and the differentiation of *Alternaria* species by their toxin profile by (Zwickel et al., 2018). Also, the detection of fungal genera, in general, is examined by trying to find genera-specific biomarkers to determine if said fungi are present. (Xie et al., 2022) Also possible is the use of GC-MS to either find differences in volatile compounds of a fungal species or just look at the volatile organic compounds themselves finding an "odor" profile (Aliferis et al., 2013, Guo et al., 2020a, Müller et al., 2013).

Furthermore, MALDI-TOF is used for fungal chemotaxonomy, but whereas GC-MS and LC-MS look at the metabolome with MALDI-TOF the proteome is examined as bigger molecules like proteins in the 2- 20 Dalton range are checked. In some studies, fungal spores were evaluated, using proteins on the spores' surface, but mostly protein extracts from the mycelium are used (Becker et al., 2014, Chalupová et al., 2014, Lau et al., 2013, Li et al., 2000, Ulrich et al., 2016).

Chemotaxonomy is a promising approach, but it must be kept in mind that biological variation in a species can be high, as different strains can express different secondary metabolites. Therefore, enough samples need to be processed to take inter-species variability into account (Becker et al., 2014). Overall application of machine learning algorithms can enable novel insight not only in the field of chemotaxonomy or metabolomics but also in other research fields using non-target mass spectrometric data, like aerosol research. Examination of fungal spores combines chemotaxonomy and bioaerosol research. Required methods are described in the following chapter 2.

2. Analytical Methods and Instruments

For the non-target analysis of complex biological samples, the separation of analytes by chromatography with subsequent exact mass measurements is very important. In this work, this is reached by coupling liquid chromatography to an Orbitrap ultra-high resolution mass spectrometer. Additional information about volatile compounds can be obtained using gas-chromatography coupled to mass spectrometry. In the following chapter, the instruments (chapter 2.1.) and techniques for data exploration (chapter 2.2.) are described.

2.1. Instruments

2.1.1. High-Performance Liquid Chromatography

High-Performance Liquid Chromatography (HPLC) enables the separation of analytes based on their affinity for the solid and liquid phase. The analytes are separated from one another by interacting with the stationary and the mobile phase, leading to an equilibrium of distribution between the two phases. The set-up is the following: The mobile phase is stored in a solvent reservoir (also called eluent reservoir) and is pumped under high pressure. A static mixer ensures that eluents are well mixed before reaching the 6-way valve, in which the sample is injected into the sample loop and waits for entering the mobile phase flow. The sample is then transported by the mobile phase into the HPLC column which contains the solid phase. This is where the separation takes place. The separated analytes then reach the detector one after another.

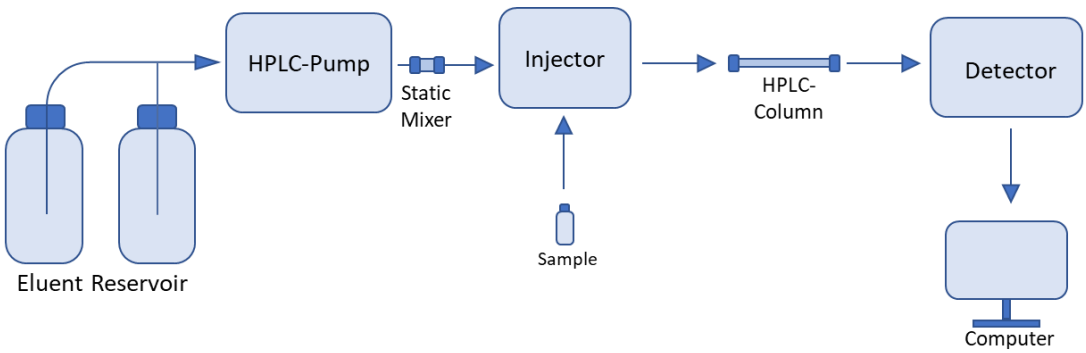


Figure 2.1: Explanatory construction of an HPLC. The eluent is transported by the pump onto the HPLC column, where the analytes are separated. Figure after (Gey, 2015).

The stationary phase consists of a column (5 to 20 cm) filled with spherical particles with a modified surface. The particles are as small as 1 µm in ultra-high performance liquid chromatography (UHPLC) resulting in a high packing density and high back pressures of up to 1000 bar. Usually, reverse phase (RP) columns are used, where the column material is functionalized with a nonpolar component, like C_{18} (n-Octadecyl) residues. Different functionalization and functional groups can influence the polarity of the stationary phase. The mobile phase consists of organic solvents mixtures of different eluents usually a watery phase as eluent A and acetonitrile or methanol as eluent B. To improve the speed and quality of the separation a gradient is used, changing the composition of the eluent over time to increase its eluting force. In RP chromatography the eluent starts with a high percentage of water and ends with a high percentage of an organic eluent like methanol or acetonitrile. Possible detectors for an HPLC system are e.g. a UV/VIS detector or a mass spectrometer (Gey, 2015, Harris, 2014).

2.1.2. Gas chromatography

In gas chromatography (GC) analytes are separated based on their vapor pressure and polarity. The gas chromatograph consists of an injector, a column inside an oven, and a detector. The stationary phase is usually inside of a capillary column, the mobile phase is a gas. Analytes need to be thermally stable as high temperatures of 400 °C can be reached.

Injection of the analyte usually is performed by solubilizing the analyte and transferring a small quantity into an injector where it is vaporized and transferred onto the column, either completely (splitless) or at a percentage (split), meaning only e.g., 10 % of the sample is transferred onto the column and 90 % discarded, to avoid column overload. Another method is thermal-desorption gas chromatography, which is shown in Figure 2.2. The volatile analyte is trapped during sampling on an adsorption material like Tenax, molecular sieve, or active charcoal inside of a tube. By this approach, the sample is pre-concentrated on the tube. After sampling the tube is capped and stored until analysis. The injection takes place by heating the tube, thus desorbing the sample. The sample is then transported by the mobile phase into a trap, where it is focused before injection. The trap works by cooling to less than -180 °C, condensing the analyte. The injection takes place by rapidly heating the trap, transferring the analytes into the gas phase, and transporting them onto the column (Gerstel, 2021, Gey, 2015, Harris, 2014).

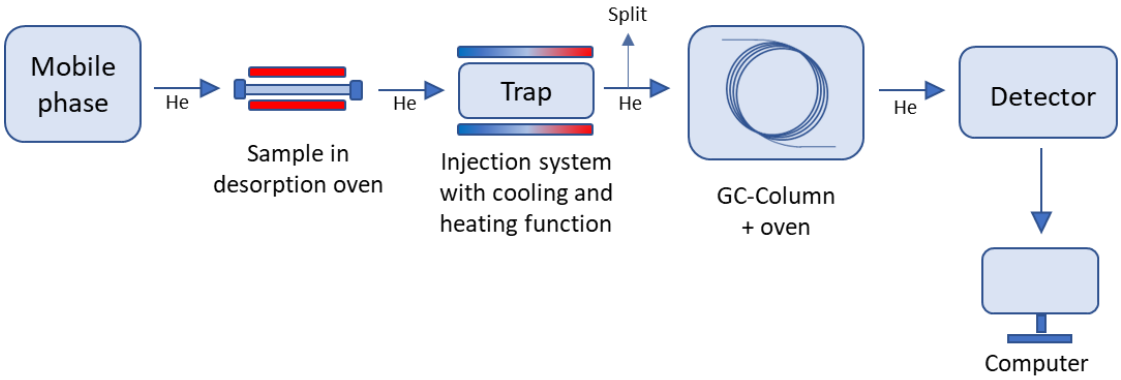


Figure 2.2: Explanatory scheme of gas chromatography with thermal desorption system and helium (He) as the mobile phase. The sample is desorbed from a sampling tube, transferred into the cryofocussing trap, and injected from there. Figure after (Gey, 2015).

Modern columns consist of a thin film on the inside of the capillary column with lengths of usually 30 or 60 meters. The thin film consists of Polyphenylmethylsiloxane or other siloxane derivates, depending on the required polarity of the column. For the mobile phase gases like nitrogen, helium or hydrogen can be used. Helium is the most common one as it has better properties when compared to nitrogen and is not explosive as hydrogen (Gey, 2015, Harris, 2014).

The column is inside an oven, where the temperature can be controlled precisely. Gas chromatography usually is performed with a temperature gradient, starting at low temperatures, e.g., 30 °C rising to high temperatures of over 200 °C. This enables a good separation of analytes with different vapor pressures and boiling points. The analyte is transported by the mobile phase and adsorbs and desorbs on the stationary phase several thousand times. The higher the vapor pressure the shorter the analyte will adsorb on the stationary phase resulting in shorter retention times. Polarity can play a role in the interaction between the analyte and the stationary phase, influencing separation and retention times. For detection, several detectors are possible with flame ionization detectors and mass spectrometers being the most common ones (Gey, 2015, Harris, 2014).

2.1.3. Mass Spectrometry

In a mass spectrometer ionized atoms and molecules are separated according to their mass-to-charge ratio (*m/z* ratio) and detected qualitatively as well as quantitatively. A mass spectrometer consists of different parts: An ionization source one or more mass analyzers, and a detector.

The whole mass spectrometer is operated under vacuum, reaching pressures lower than $1 \cdot 10^{-9}$ millibar (mbar) in the case of the orbitrap mass spectrometers. Other mass spectrometers work at pressures under $1 \cdot 10^{-7}$ mbar. The pressures are reached by operating several (turbo) vacuum pumps.

The ionization source transfers non-charged analytes into charged species. The introduction of the sample into the ionization source can be done directly or by coupling to a chromatographic system. There are different sources available for the ionization step, three of which are introduced in the following sub-chapter.

The analyte ions reach the mass spectrometer, are focused into an ion beam, and the pressure is reduced to a high vacuum ($<10^{-5}$ mbar). The separation takes place in the mass analyzer, of which there are different types (Quadrupole, Ion trap, Time of Flight, Orbitrap, etc.). The detector follows the mass analyzer, but in some cases, the detection occurs in the mass analyzer itself by *ion current imaging* (Gross, 2013). For further structure clarification, MS^n experiments are possible, where single *m/z*-ratios are chosen and fragmented inside the mass spectrometer. Those fragments can be analyzed again, providing insights into the structure of the analyte (Gross, 2013, (Harris, 2014)).

2.1.4. Ionization Sources

Several ionization sources are available for introducing the analyte into the mass spectrometer. For the combination of gas chromatography with a mass spectrometer, the electron ionization (EI) is usually used, description is on page 25. With HPLC the electrospray ionization (ESI), the atmospheric pressure chemical ionization (APCI), or the atmospheric pressure photoionization (APPI) are commonly used.

With ion sources coupled to an HPLC, ionization happens at atmospheric pressure, and resulting ions are introduced into the mass analyzer where the pressure is stepwise reduced until a sufficient vacuum is reached. All three ionization types have in common, that they are

"soft" ionization methods, meaning compounds are ionized almost without fragmentation, leading to molecular ions, which maintain the analytes' structural information.

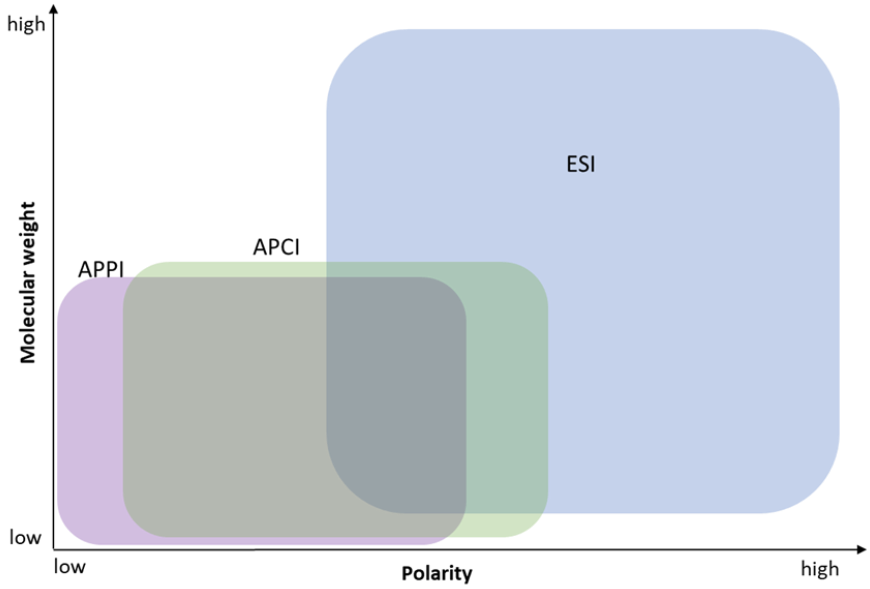


Figure 2.3: Suitable analyte polarity and molecular weight for different ionization sources (ESI, APCI, APPI). Figure after (Riches et al., 2017).

As shown in Figure 2.3 the different ionization sources have different strengths. ESI is extensively used, as a wide range of mid- to non-polar compounds can be easily ionized. The formation of multiply charged ions is especially useful for the detection of high molecular weight compounds like proteins. With charges of more than one, the *m/z*-ratio stays small even at high molecular weights. This enables detection by common mass spectrometers, e.g., the orbitrap can detect ions up to a *m/z* ratio of 6000. Compounds that are not well ionizable by ESI are non-polar compounds, like sterols, which need other ionization sources like the APCI or APPI. As seen in Figure 2.3 APCI and ESI complement each other (Rosenberg, 2003).

<u>Electrospray ionization</u>

The analyte is introduced into the source at atmospheric pressure, dissolved in a liquid, e.g. the HPLC eluent. The sample is sprayed through the capillary, at the tip of which a potential is applied, forming an electrically charged aerosol. Heated sheath and auxiliary gas enhance the formation of the aerosol and the evaporation of the solvent. Ion formation in positive mode is shown in Figure 2.4.

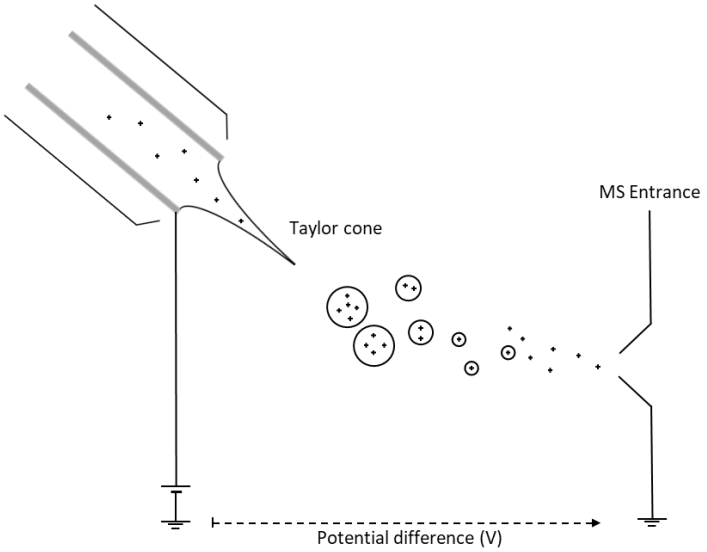


Figure 2.4: Construction of an electrospray ionization source. The analytes are solubilized and enter the ionization area in a fine spray. Figure after (Riches et al., 2017).

When the solved sample leaves the capillary, it is exposed to an electric field of 3-4 kV, leading to charge separation and the formation of the so-called Taylor cone. From the tip of this cone small, highly charged droplets are emitted. The droplet's liquid evaporates and when the Rayleigh Limit is reached, smaller highly charged droplets are formed from the bigger ones. The final ion release is discussed in two models. The charged residue model (CRM) suggests smaller and smaller droplets until the final droplet contains only one analyte molecule. Finally, evaporation leads to ion formation by charge transfer from e.g., the solvents protons. The ion evaporation model suggests that ions leave highly charged droplets when the field strength on the droplet's surface enables field desorption. As a third option, proton transfer in the gas phase is discussed (Gross, 2013, Ho et al., 2003, Sleighter, Hatcher, 2007). The potential difference between the capillary tip and the MS entrance leads the freshly formed ions into the MS.

Atmospheric Pressure Chemical Ionization

Atmospheric pressure chemical ionization, short APCI is a technique in which molecules in the gas phase are ionized under ambient pressure. Analytes need to be thermally stable as temperatures in the ion source reach up to 400 °C. APCI can be used for direct input by a gas stream or coupled to liquid chromatography systems. (Hoffmann, Stroobant, 2011).

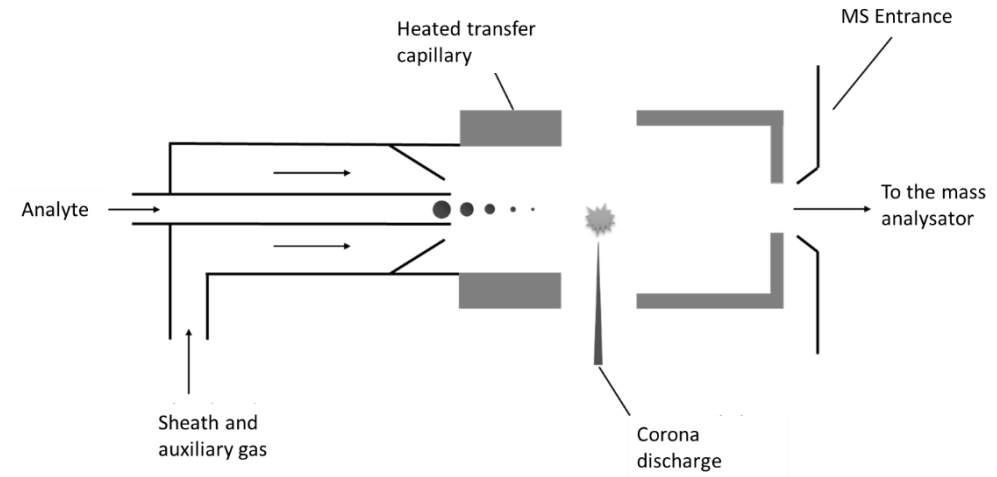


Figure 2.5: Schematics of an APCI ionization source. The analytes evaporate in the transfer capillary and reach the corona discharge area, where the ions are generated. Figure after (Hoffmann, 2013).

The analytes are introduced into the ion source via gaseous or liquid phase and enter through an up to 400 °C heated transfer capillary in which solvents are evaporated. Heated sheath and auxiliary gas (N_2) support the evaporation. After evaporation, the analyte-molecules reach the plasma which is produced by a corona discharge (up to 5 kV). Here a row of ion-molecule reactions takes place. At first, primary ions are formed out of the reactant gas, e.g., evaporated solvent molecules, nitrogen, or oxygen. Those primary ions transfer their charge onto the analytes. Positive as well as negative ion formation is possible, with the reaction mechanism depending on the applied voltage (Gross, 2013). For the reaction mechanism see Figure 2.6.

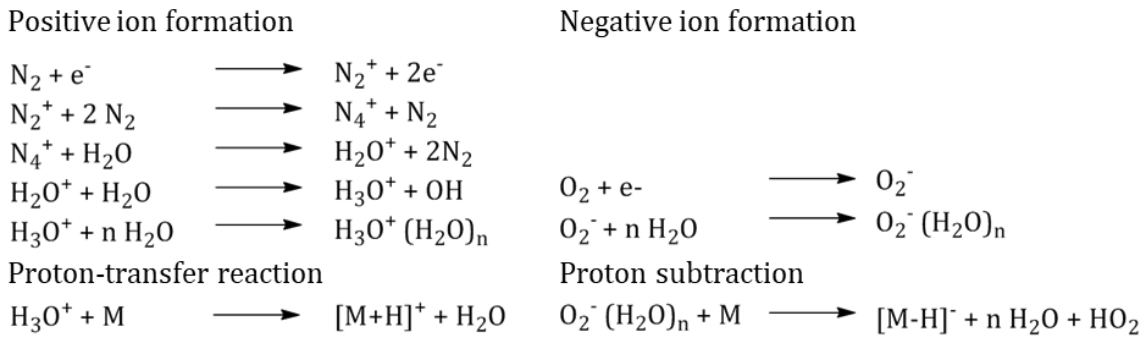


Figure 2.6: Positive and negative ionization formation in the APCI. Sheath and auxiliar gas, containing water, nitrogen and oxygen are crucial for the final ionization of the analyte molecule M (Gross, 2013).

Analyte ions are formed by a chain reaction with proton transfer or abstraction as the last step. The addition or loss of *n* water molecules, resulting in adducts is also possible. For protonation, the proton affinity of the analyte needs to be higher than the one of water,

which is the case for many organic molecules. In negative mode especially acid groups are easily deprotonated. After ionization the ions are guided into the mass spectrometer, reaching the vacuum area ((Gross, 2013, Harris, 2014, Hoffmann, 2013).

Electron ionization

With electron ionization, the analyte is ionized with energetic electrons. In contrast to ESI and APCI, it is a hard ionization method, meaning the analyte is fragmented into several pieces. This can enable structure determination. A scheme of an electron ionization source is shown in Figure 2.7.

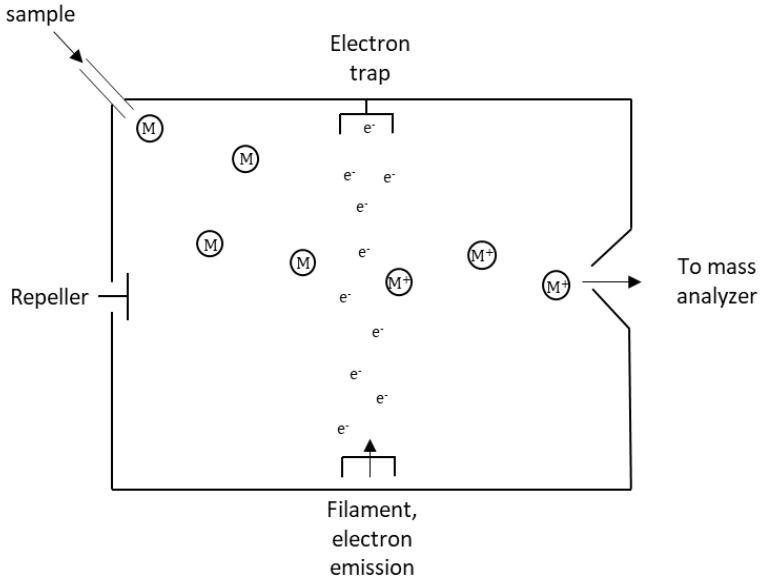


Figure 2.7: Schematics of an EI ionization source. The analyte is ionized by electrons which are emitted and accelerated to reach a kinetic energy of 70 eV. Figure after (Hoffmann, 2013).

The source consists of a heated filament emitting electrons. The electrons are accelerated in the direction of the anode/electron trap. The electrons' velocity influences the wavelength and usually, a kinetic energy of 70 electron volt (eV) is chosen, where the electrons have a wavelength of 1.4 Å, which is roughly the length of a C-C bond. The analyte, which is in the gas phase and the electron collide, forming a positive ion with an odd electron number $M + e^- \rightarrow M^{+\cdot} + 2e^-$. The resulting radical cation is transferred into the direction of the mass analyzer, usually fragmenting. Neutral fragments won't reach the analyzer and are removed by the vacuum system. Fragment patterns can be replicable at 70 eV and be used to compare with databases.

2.1.5. Mass spectrometers

In the following chapter, the Orbitrap mass analyzer for high-resolution mass measurements and the ion trap analyzer with a nominal mass resolution are described.

Ion-trap mass spectrometer

An ion trap is a mass analyzer consisting of a ring-shaped electrode (cathode) and two end electrodes (anode). A three-dimensional high-frequency field is formed in which ions can be "stored". The ions move in stable, saddle-like trajectories according to their mass-to-charge ratio. The corresponding voltages and frequencies to maintain stable trajectories can be described by the Mathieu formulas which give two parameters for a_z and q_z. These two parameters can be varied, destabilizing the trajectories of chosen ions, which then leave the analyzer in direction of the detector. This procedure is called mass selective instability (Gross, 2013, Hoffmann, 2013).

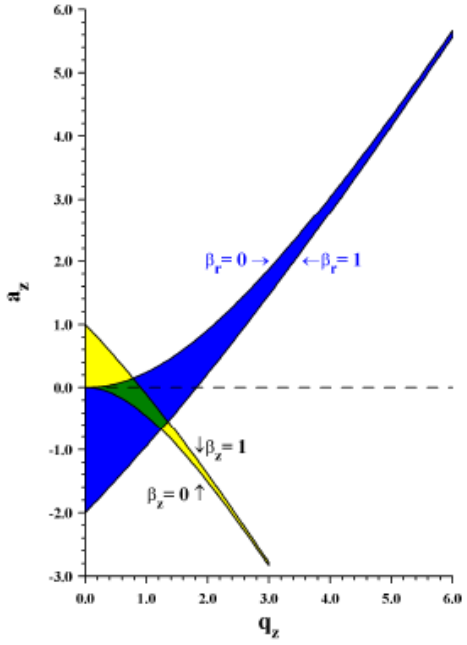


Figure 2.8: Stability diagram for a 3D ion trap, green represents the area where ion trajectories are stable. Adapted after (Hoffmann, 2013).

Figure 2.8 describes at which AC voltage (q_z term) and at which DC voltage (a_z term) the ions move on stable trajectories. Usually, the DC voltage is not changed, resulting in a working line at $a_z = 0$. The ions are in different positions on the working line depending on their *m/z* ratio.

A small amount of helium is present in the ion trap; it has two functions. On the one hand, it serves as a buffer gas, the kinetic energy of the ions is reduced by collisions and the ions are focused in the ion trap. They all have the same kinetic energy, which increases the resolution. On the other hand, in MS/MS experiments, the helium fragments selected ions by collisions (Gross, 2013, Hoffmann, 2013).

As a detector, a secondary electron multiplier is used. Ions leave the ion trap and impinge onto a diode, knocking electrons out of this diode. The electrons are accelerated by applying a voltage, hitting other dynodes, from which they knock out several secondary electrons, amplifying the signal. In the end, the electrons are measured by an electrode that detects the voltage change (Gross, 2013, Hoffmann, 2013).

Orbitrap mass spectrometer

The Orbitrap mass analyzer enables ultra-high resolution mass spectrometry with a mass resolution of R = 140,000. The mass accuracy can achieve under 1 ppm with internal and under 5 ppm with external calibration. A mass range of m/z 6000 is covered (Thermo Fisher Scientific Inc., 2012) The only other mass spectrometers reaching resolutions this high or higher are FTICR (Fourier transforming ion cyclotron resonance) devices which are much more expensive. The mass-to-charge ratio is determined by measuring the oscillation frequency of the respective ions (Gross, 2013, Hoffmann, 2013). The scheme of an Orbitrap Q Exactive is shown in Figure 2.9.

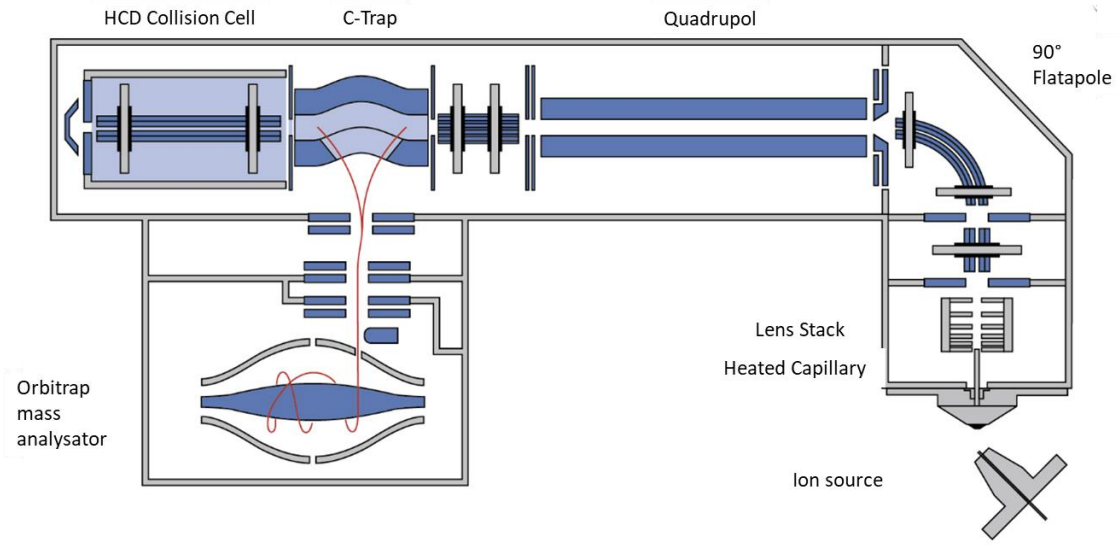


Figure 2.9: Schematics of the Orbitrap Q Exactive Hybrid Quadrupol Orbitrap mass spectrometer. Adapted from Thermo Fisher Scientific Inc., 2012.

First, the analytes are ionized in the ion source (APCI or ESI) and enter the orbitrap's high vacuum area through the ion entry system. A lens system focuses the ions and a 90° bend flatapole filters neutral particles by only allowing charged particles to enter the quadrupole mass analyzer, which works as a pre-filter for the orbitrap mass analyzer.

The quadrupole mass analyzer consists of four hyperbolical stab rods, arranged parallel to one another. On two opposite rods, a radio frequency (RF) or respectively a DC offset voltage is applied. A periodical variation of the voltage attracts or repels the passing ions alternately. Only ions of certain *m/z* ratios can pass the quadrupole on stable trajectories. All other ions either impact on the electrodes or are removed by the vacuum system. Using this technique single nominal masses can be filtered and chosen for MS^n.

After passing the quadrupole the ions reach the C-trap (*curved linear trap*), a bend RF-Quadrupole. Ions are stored and focused to a package by collisions with a cooling gas, which reduces kinetic energy. The ion package is then injected into the orbitrap mass analyzer. Behind the C-Trap is the HCD Collision Chamber (*higher-energy collisional dissociation* HCD) in which chosen mass-to-charge ratios can be fragmented for MS^2 experiments. Fragmentation is performed by the acceleration of the analyte ions followed by collision with nitrogen. After fragmentation, the ions are led through the C-trap into the Orbitrap mass analyzer (Hu et al., 2005, Thermo Fisher Scientific Inc., 2017, Zubarev, Makarov, 2013)

The Orbitrap mass analyzer itself consists of a spindle-like electrode surrounded by a barrel-like electrode. The outer electrode is separated into two pieces. Through a split ions are inserted decentral axial from the C-trap. The ions start oscillating around the inner electrode on stable trajectories. Attraction by the inner electron is in balance with the oppositional acting centrifugal force.

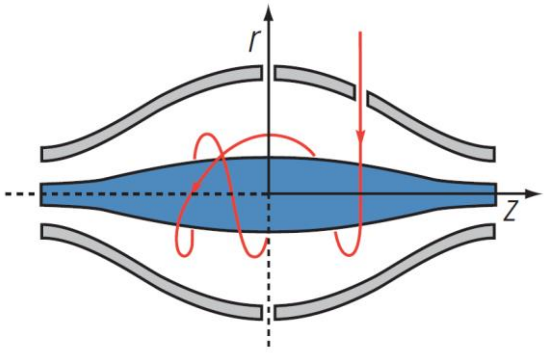


Figure 2.10: Schematics of the Orbitrap mass analyzer and movement of ions around the inner electrode (Thermo Fisher Scientific Inc., 2017).

The ions move periodically along the z-axis (see Figure 2.10). The movement is harmonic and independent of all parameters except the mass-to-charge ratio. The only dependent measurand is the frequency which can be measured very precisely, the corresponding formula is shown in Formula (1). This enables the very high mass resolution of the Orbitrap (Hu et al., 2005).

$$\omega = \sqrt{\frac{z}{m} \times k} \tag{1}$$

ω = frequency of the harmonic oscillation, m/z: mass-to-charge ratio, k: Instrumental constant

The harmonic axial motion induces a current on each half of the electrode. The current is determined by a differential amplifier on the respective half of the outer electrode. The resulting frequency out of this ion current image is Fourier transformed into the mass-to-charge ratio of the ions (Hu et al., 2005, Zubarev, Makarov, 2013).

High-resolution mass spectrometry

The mass resolution R describes the difference between two m/z -ratios which can just be separated. Often the resolution is indicated for the full width at half maximum (FWHM) (Gross, 2013). The formula for the mass resolution is shown in (2).

$$R = \frac{m}{\Delta m} \tag{2}$$

The influence of the mass resolution is shown in Figure 2.11:

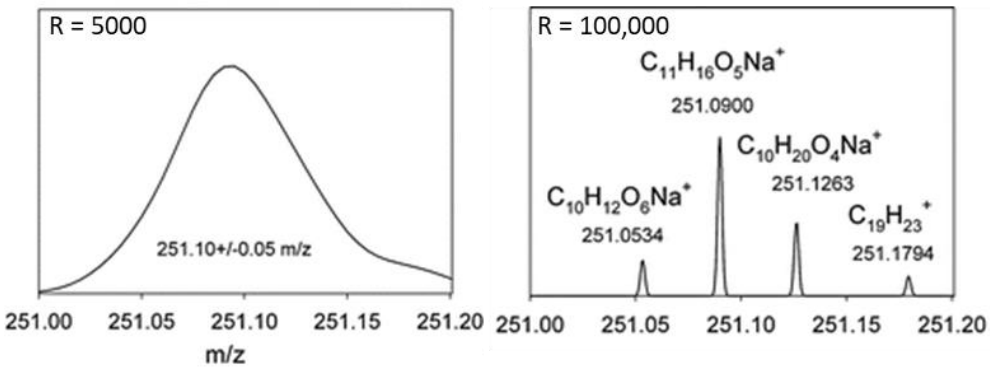


Figure 2.11: Zoom on nominal mass 251.7 in low (left) and high (right) mass resolution (Nizkorodov et al., 2011).

At a resolution power of $R = 5000$ isobaric peaks of a nominal mass can't be determined from one another. At R =100,000 four different peaks can be determined. (Nizkorodov et al., 2011, Nozière et al., 2015).

Also important is the determination of the exact mass. The calculated exact mass and the mass determined by the mass analyzer should be as close as possible to each other, desirable are relative mass accuracy's lower than 5 ppm (Nozière et al., 2015).

$$\Delta \frac{m}{z} = \frac{\frac{m}{z}_{experimental} - \frac{m}{z}_{theoretical}}{\frac{m}{z}_{theoretical}} \quad (3)$$

A high mass accuracy enables the calculation of the molecular formula. To improve the reliability of the molecular formula's calculation, the Seven Golden Rules can be applied, as "older" rules like the nitrogen rule or the RDBE are insufficient at masses higher than 500 Da (Kind, Fiehn, 2007). The seven golden rules give restrictions for element numbers during formula generation for small molecules, depending on the mass. For molecules smaller than 1000 Da, e.g., carbon numbers of more than 78 are unlikely. Also, the LEWIS rules, the isotopic pattern filter, the element ratios, and an element probability check are used by most programs to increase the reliability of the molecular formula calculation (Kind, Fiehn, 2007).

2.2. Data analysis

Obstacle in the analysis of non-target high-resolution mass spectrometry data is the sheer abundance of detected compounds, resulting in compound lists of thousands or even tens of thousands of possible features. Data analysis by the human eye is not possible. To enable data analysis a data processing workflow consisting of several steps is needed, including raw data processing, spectral deconvolution, component detection, data normalization, and multivariate statistical analysis (Blekherman et al., 2011).

2.2.1. Raw data processing

Raw data from mass spectrometry measurements need to be processed. Open-source and commercial options are available, to perform peak detection and deconvolution, extracting mass-to-charge ratios and retention times of the sample features, filtering for adducts and complexes, and calculating molecular formulas. Pre-processed data is further filtered, blank

subtracted, and aligned, resulting in a feature list ready for further data analysis by e.g., *van Krevelen* plots or machine learning methods like Principal Component Analysis or clustering. Those methods are described in the following chapters. (Blekherman et al., 2011, Enot et al., 2008, Yi et al., 2016).

2.2.2. Normalization

Before clustering or classification, the data must be normalized to take differences in metabolite recoveries, concentrations, or instrumental electronic noise into account. Different samples concentrations can influence the clustering result, as well as fluctuations in the experimental setup. With LC-MS not only mass accuracy drift or variability in retention times need to be controlled, but also ion suppression poses a problem, with biological samples there is high natural variability in samples. Also, metabolite concentration varies immensely, with the highest concentration not necessarily belonging to metabolites of interest. Metabolites like ergosterol are highly abundant and present in all samples whereas secondary metabolites, which make the difference between samples/species might have low concentrations. Normalization ensures that the clustering depends on relative intensities and not on absolute values with large differences. (Bouguettaya et al., 2015, Filzmoser, Walczak, 2014, Frochte, 2019, Meinicke et al., 2008).

Normalization can be done sample- or data-based and performed pre- or post-acquisition. Figure 2.12 gives an overview of the normalization techniques. There's a wide range of possible normalization methods, which are very controversially discussed and used (Blekherman et al., 2011, Bouguettaya et al., 2015, Daellenbach et al., 2019, Enot et al., 2008, Filzmoser, Walczak, 2014, Forsberg et al., 2018, Livera et al., 2012, Misra, 2020, Sysi-Aho et al., 2007, van den Berg et al., 2006, Veselkov et al., 2011, Winkler, 2020, Wu, Li, 2016, Yi et al., 2016).

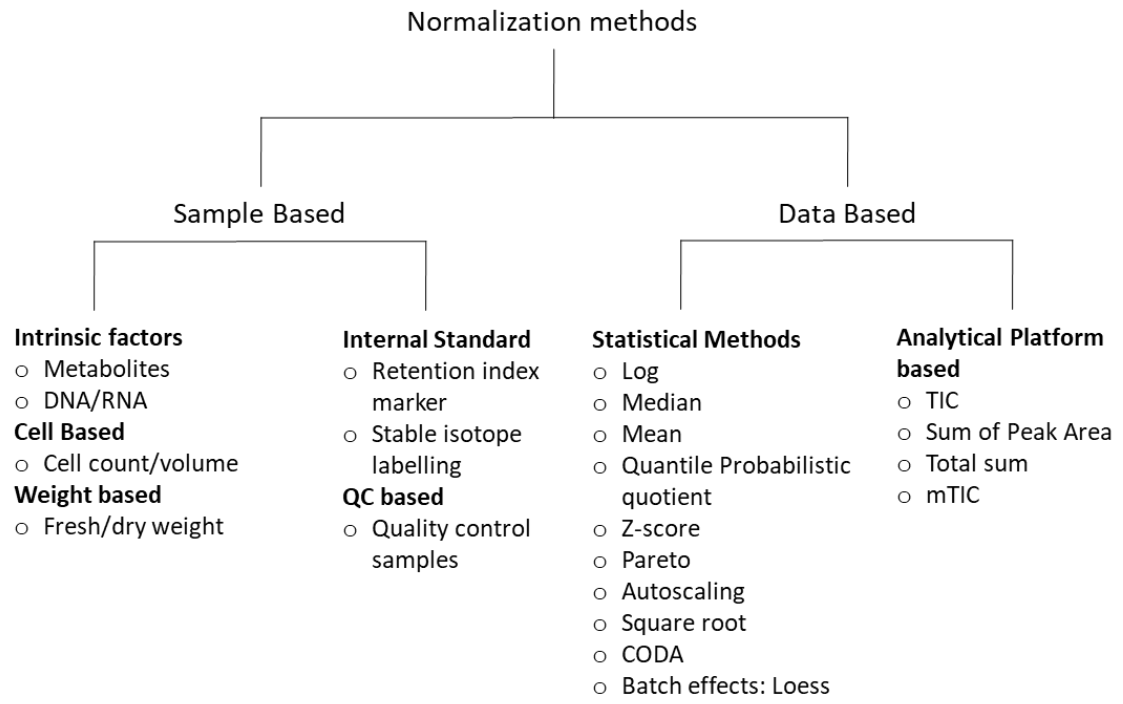


Figure 2.12: Normalization strategies divided into sample-based and data-based approaches. Adapted after (Misra, 2020).

Sample-based methods would be to normalize data based on properties like cell count, the concentration of a marker (e.g., ergosterol), weight, or the use of different internal or external standards. Cell count or weight can be used pre-acquisition by diluting the samples to their respective factors, approaches like standards can be used post-acquisition. Data-based approaches are used post-acquisition. In data-based approaches the terms Scaling, Centering, Standardizing, Normalizing, and Transforming are sometimes used interchangeably.

Transformation

Often the first normalization step is a transformation to reduce data heteroscedasticity and skewness. Several transformation methods are available: log (base 2 or 10), BoxCox, square-root, cube-root, CODA (centered log-ratio), arcsine, and other less common transformations. Log transformation is the most common one. Negative or zero values need to be adjusted with a factor, shifting the complete dataset. By a transformation, larger values are relatively more reduced than smaller ones. (Enot et al., 2008, Livera et al., 2012, Misra, 2020, van den Berg et al., 2006).

Standardization and centering

After transformation normalization follows for which the terms centering/standardizing /scaling/normalizing are used interchangeably. Centering adjusts the data to a zero mean, or a zero median e.g., by subtracting the mean. Standardization adjusts to a zero mean and a variance of 1. The most common method is the z-score standardization, which is also called autoscaling (see Formula (4)) with μ being the mean and σ the standard deviation,

$$x' = \frac{(x - \mu)}{\sigma} \tag{4}$$

The prerequisite is that the data follows near Gaussian distribution, e.g., by previous log transformation. Pareto scaling is similar to autoscaling but used the square root of the standard deviation instead of the standard deviation as the denominator, being a less "invasive" normalization method than autoscaling (van den Berg et al., 2006).

Another popular normalization method is global normalization where the same scaling factor for all features is used. In mass spectrometry, one would be using the total ion count (TIC) as the scaling factor. Another method would be using level-scaling with the mean or median:

$$x' = \frac{(x - \mu)}{\mu} \tag{5}$$

Other methods are min-max normalization, min-max scaling, probabilistic quotient normalization (PQN), median fold change normalization, LOESS normalization (Locally weighted scatter plot smoothing), normalization to Euclidean unit length, quantile normalization, and more, shown in table Table 2.1.

Table 2.1: Normalization methods by formula.

$q_{ij} = x_{ij}^{TIC} / x_{control,j}^{TIC}$	TIC normalization (Wulff, Mitchell, 2018)
$X_i = \{x_{i[1]} \dots x_{i[m]}\}, \quad \bar{X} = \{\bar{x}_{[1]} \dots \bar{x}_{[m]}\} = \left\{\frac{\sum_{i=1}^{n} x_{i[1]}}{n} \dots \frac{\sum_{i=1}^{n} x_{i[m]}}{n}\right\}$ With X_i ordered set of intensities for sample i $X_i^N = \{\bar{x}_{[rank(x_{i1})]} \dots \bar{x}_{[rank(x_{im})]}\}$	Quantile normalization (Wulff, Mitchell, 2018)
$n_k^{\mathrm{MFC}} = median\left(\frac{x_{ik}}{x_{ir}}\right)$	Median fold change (MFC) (Veselkov et al., 2011)
$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{(x_{imax} - x_{imin})}, \bar{x}_i = \mathrm{mean}$	Range scaling (van den Berg et al., 2006)

Normalization methods are very controversially discussed having their advantages and disadvantages and should be chosen carefully. Nonetheless, normalization is needed for methods like principal component analysis, as variables need to have the same standard deviation (for detail see chapter 2.2.4) (Blekherman et al., 2011, Enot et al., 2008, Filzmoser, Walczak, 2014, Sysi-Aho et al., 2007, van den Berg et al., 2006, Veselkov et al., 2011, Wu, Li, 2016).

2.2.3. Machine learning algorithms

Machine learning algorithms enable determining trends in non-target data and detecting features that differ samples from one another. As the number of features and/or samples can be very high dimensionality reduction techniques like Principal component analysis (PCA), linear discriminant analysis (LDA) or t-distributed stochastic neighbor embedding (t-SNE) can help. PCA can be used to reduce the data set to a smaller size. This smaller dataset is then used as input for further algorithms from the areas of unsupervised- and supervised learning. Unsupervised means, that no knowledge of which sample is which class is needed, whereas in supervised methods a training set with known classification is needed. Unsupervised learning includes clustering techniques like hierarchical clustering (HCA), k-means clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), or self-organizing maps (SOM). Clustering techniques can be used as a first step in determining

interesting features. Supervised learning requires some input in form of a training set (see chapter 2.2.5) and includes classification methods like the k-nearest neighbor, random forest (RF), support vector machine (SVM), partial least squares discriminant analysis (PLS-DA), or linear discriminant analysis (LDA). To determine how well a classification performs cross-validation is used. The variance-bias trade-off and the problem of over-or underfitting can be handled by choosing the right parameters for the algorithms and performing cross-validation (see chapter 2.2.6) (Blekherman et al., 2011, Bouguettaya et al., 2015, Frochte, 2019, Yi et al., 2016). The methods used in this work are explained in the following chapters. In-depth mathematical explanations were omitted as they would go beyond the scope of this work, but can be found in (Frochte, 2019, Merkl, 2015).

2.2.4. Dimensionality reduction

Dimensionality reduction is a necessary step before applying unsupervised or supervised machine learning algorithms because of the so-called "curse of dimensionality". Especially non-target LC-MS data can be very high dimensionally with several hundred to thousands of features per sample. With high-dimensional data, the distance measurements in a Euclidean space become less meaningful the more dimensions are included. This is also called distance concentration and is explained in literature (Zimek et al., 2012). The curse of dimensionality leads to less meaningful clustering and classification results with high dimensionality data. Several methods are available to reduce dimensionality with principal component analysis being the most prominent one. A rule of thumb says that at least five samples should be available per dimension (Koutroumbas, Theodoridis, 2010). This rule of thumb should be carefully examined for each data set, as also more dimensions can be useful as long as they contain relevant information.

Principal Component Analysis

Principal component analysis is a multivariate statistical method that uses linear combination to reduce dimensionality in a data set and helps to get a first look at the data. A dataset with 10,000 dimensions can be reduced to only two or three, enabling graphic representation. PCA also gives information on which variable is the most valuable for clustering the data.

Principal components are eigenvectors of the data's covariance matrix. As a first step, the data imperatively need to be standardized, so that features have the same weight independent of their unit of measure. Secondly, the covariance matrix is calculated as well

as the eigenvalues and eigenvectors for the covariance matrix. eigenvalues and their corresponding eigenvectors are sorted, depending on how much of the data's variance is explained by the eigenvalue. The eigenvector or principal component which explains the highest amount of variance is the first, the one which explains the second most of the variance the seconds, and so forth. A scree plot can help to see how much of the data's variance is explained by the respective principal component, an example is shown in Figure 2.13.

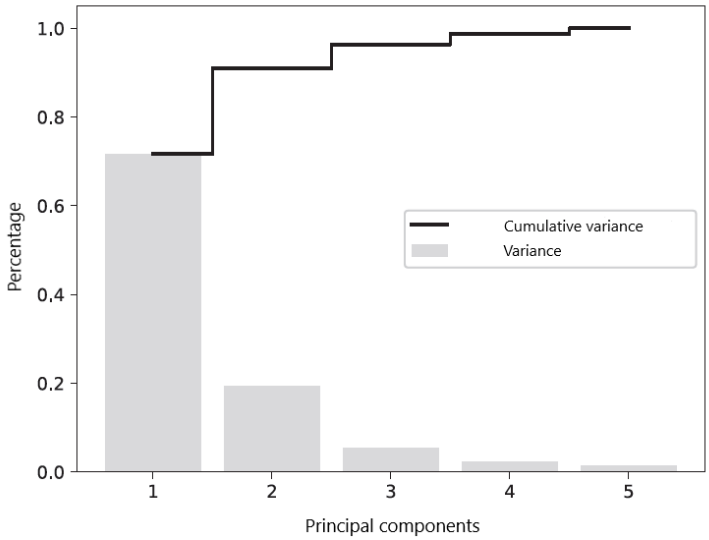


Figure 2.13: Explanatory scree plot showing how much of the variance is explained by the principal components.

Depending on the user a certain number k of principal components is chosen, e.g., the number of PCs which explain 90 % of the data's variance. In the case of Figure 2.13, these would be the first two principal components. The chosen k eigenvalues are used to form a matrix of corresponding eigenvectors, which are then used to transform the original matrix (feature matrix × chosen number of eigenvectors = transformed, dimensionality reduced data). Results can be visualized by making a scatter plot representing the samples according to the first two principal components, see the example in Figure 2.14. The data represented is the Fisher iris data set concerning three species of the flower genus *Iris*, which is commonly used in machine learning contexts (FISHER, 1936).

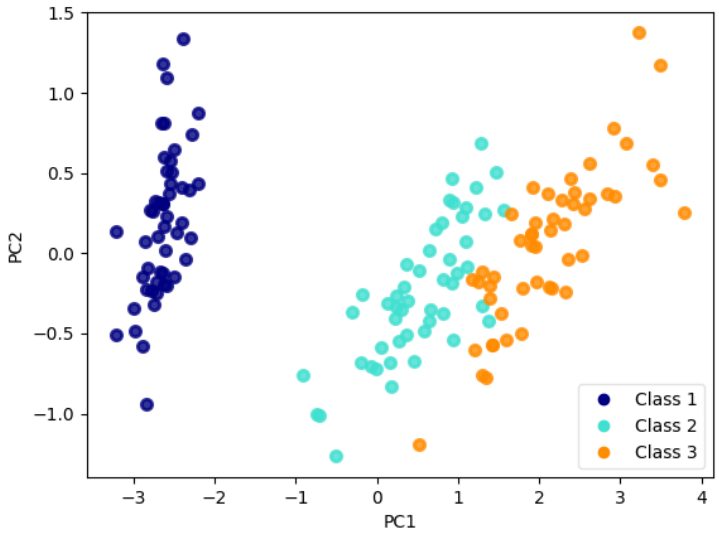


Figure 2.14: Exemplary score plot on the example of the Fisher Iris Data set. Even with only the information of two principal components a difference between the classes, especially between class 1 and the other two classes can be visualized.

In some cases, differences between classes can be visualized by a PCA and allow a first visual inspection of the data (see Figure 2.14). Additional plots like loading plots where the eigenvectors are visualized to see which variables correlate with which PC or a biplot, where score and loading plot are superimposed can give additional information, e.g., which variable is important for differences between samples. An example of a loading plot is shown in Figure 2.15. Features that correlate (Feature 1 and 2) are very close to one another.

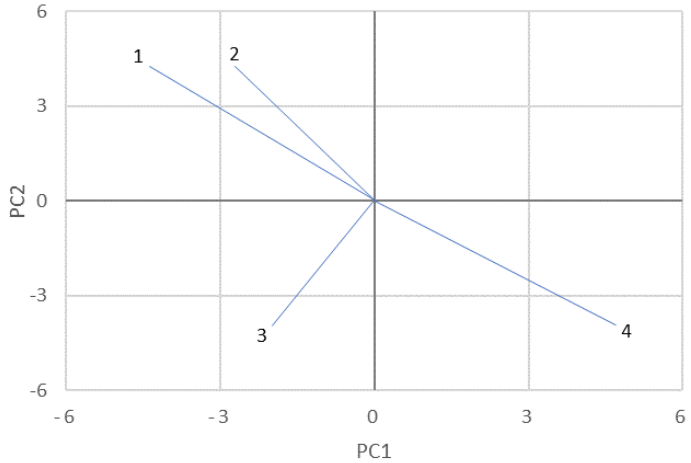


Figure 2.15: Explanatory loading plot with four features with different correlations.

No correlation would mean that features are at a 90 ° angle as shown with features 2 and 3. Anti-correlation is shown with features 1 and 4 which are placed at a 180 ° angle.

Score, loading, and biplot have to be carefully interpreted as they show only the variance that the first two principal components can explain (Frochte, 2019, Merkl, 2015, Wentura, Pospeschill, 2015).

t-distributed stochastic neighbor embedding

t-SNE is a dimensionality reduction technique that is used for visualization. In contrast to PCA it is not used for feature extraction, and it is a non-linear dimensionality reduction technique with a focus on keeping similar data points close together in low dimensional space. t-SNE briefly explained uses two steps, at first, a probability distribution is constructed with high probabilities for similar objects and low probabilities for less similar objects. Then the probability distribution is performed in low dimensions and compared with the one in the high dimensional space while minimizing the Kullback-Leibler (KL) divergence. The KL divergence is a measure of the difference between two probability distributions, in this case between the one in the high- and the one in the low dimensionality space (van der Maaten, Hinton, 2008). The Python implementation of t-SNE needs several parameters to be set by the user, mainly the perplexity, with values between 5 and 50. It is equivalent to the number of neighbors used in other machine learning algorithms, the bigger the dataset the higher the perplexity should be. An explanatory picture is shown in Figure 2.16 with the Fisher iris data set. T-SNE performed similarly to PCA as in the iris data set mostly linear relationships exist.

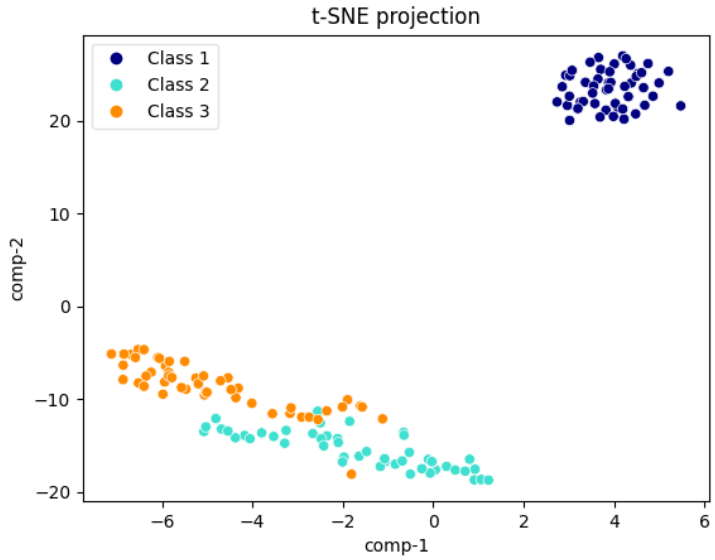


Figure 2.16: t-SNE protection of Fisher iris data set. Class 1 is separated from classes two and three. t-SNE component 1 on x-axis, t-SNE component 2 on y-axis.

2.2.5. Unsupervised learning methods

With unsupervised learning, no previous input before clustering is needed. The information on which sample belongs to which class is not necessary.

Hierarchical Clustering

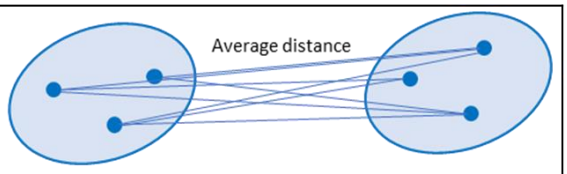
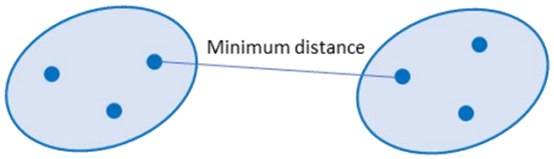
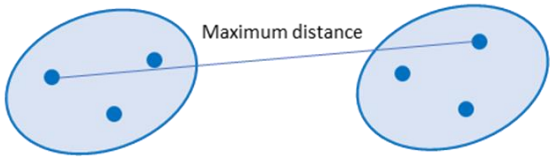
Hierarchical clustering can be performed divisive or agglomerative ("bottom-up"). Here the agglomerative approach is described. Results can be visualized as dendrograms and as heat maps. Agglomerative hierarchical clustering consists of several steps. Firstly, the normalized data is saved in an m × n matrix (m: feature, n: sample), and the distance matrix between two objects i and j is calculated with the Euclidean Distance (Formula (6)). Other distances like Manhattan/City block, Minkowski's, Mahalanobis, cosine, hamming, etc. can be used.

$$e_{jk} = \sqrt{\sum_{i=1}^{n} (X_{ij} - X_{ik})^2} \qquad (6)$$

With j and k being two objects, n the number of attributes, X_{ij} and X_{ik} the coordinates for the two objects. Often also the quadrated Euclidean distance is used (Wentura 2015).

After distance calculation, the proximity of two objects i and j are used to group them into clusters. The two objects with the minimal distance are combined into a cluster. Subsequently, the distances between this newly formed cluster and the other clusters are calculated to find the next clusters to be joined together. Several linkage algorithms are available: Unweighted average distance (UPGMA), centroid (UPGMC), complete, single, median, ward (minimum variance algorithm), etc. Table 2.2 gives an impression.

Table 2.2: Typical linkage algorithms with an explanatory picture on the right (The MathWorks, 2021b).

Average	$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj})$	
Single	$d(r,s) = \min\left(dist(x_{ri}, x_{si})\right), i \in (i, \dots, n_r), j \in (1, \dots, n_s)$	
Complete	$d(r,s) = \max\left(dist(x_{ri}, x_{si})\right), i \in (i, \dots, n_r), j \in (1, \dots, n_s)$	

Average linkage is considered robust compared to other linkage methods (Everitt 1993, Kalkstein 1987, Robinson 2013). A cophenetic correlation coefficient can be used to determine which linkage method is best suited to find the closest relationship between the two variables X and Y. The cophenetic correlation coefficient measures how accurately the linkage function determined the distance of the original data. Ideally, the correlation coefficient should be close to one.

$$c = \frac{\sum_{i<j}(Y_{ij} - y)(Z_{ij} - z)}{\sqrt{\sum_{i<j}(Y_{ij} - y)^2 \sum_{i<j}(Z_{ij} - z)^2}} \tag{7}$$

Z: output of linkage function, Y: distance, c: cophenetic distance. Y_{ij}: distance between objects I and j in Y. Z_{ij}: cophenetic distance objects I and j in Z. y and z: average of Y and Z (The MathWorks, 2021a).

After completion, a clustering tree is formed, where the height of each arm shows the "similarity" of the clusters to one another. Features are then sorted following the clusters' order. The whole process is repeated sample-wise. After sorting, a heat map is used to visualize the results, enabling fast detection of similar and distinctive features (Bouguettaya et al., 2015, Frochte, 2019, Murtagh, Contreras, 2012). An explanatory heatmap is shown in Figure 2.17 with tree diagrams visualizing the clustering for samples (horizontal) and features (vertical). Dark red implicates high intensities, and dark blue low intensities of the features in the heat map.

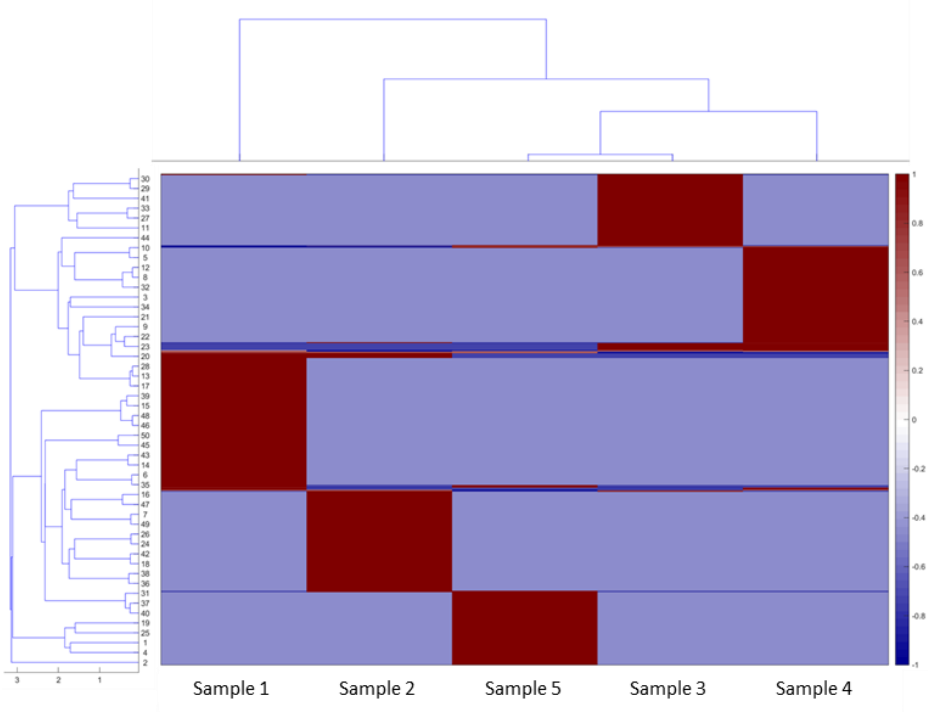


Figure 2.17: Explanatory heat map with tree diagrams to visualize hierarchical clustering results. The vertical tree diagram represents features and the horizontal tree diagram the samples. The heatmap shows the intensities of the different features, sorted according to the clusters. Own work, LC-MS analysis of fungal spore samples.

The heatmap gives information about the relationship between samples and the relationship between features. Additionally, the heatmap can show if certain features are intense for a certain sample group. However, it doesn't give information on which kind of molecule the features of interest are. A *van Krevelen* plot can help visualize features and get information about the feature's nature. In a *van Krevelen* plot, the hydrogen (H) to carbon (C) ratio is plotted against the oxygen (O) to carbon ratio. As biological molecule groups have definitive ranges in which the O/C and the H/C ratios lay, regions can be defined, as seen in Figure 2.18 (Brockman et al., 2018, Kew et al., 2017, Kim et al., 2003, Rivas-Ubach et al., 2018).

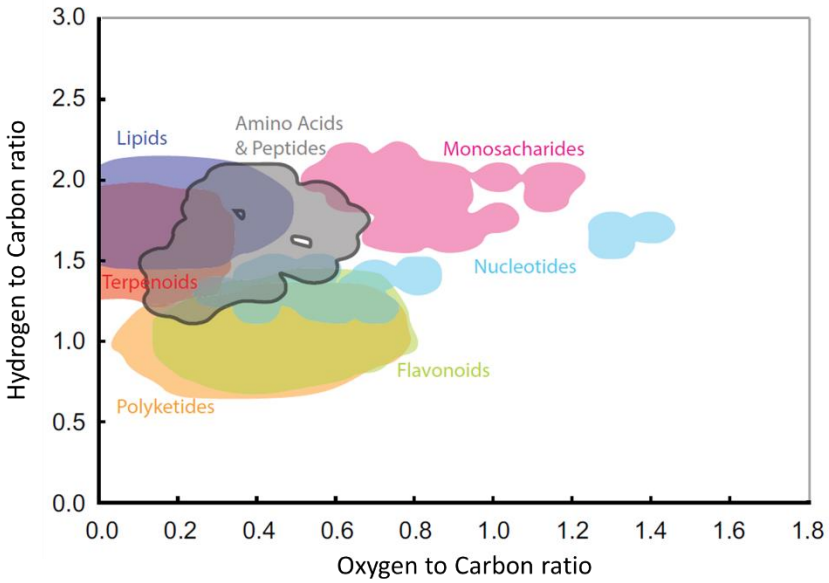


Figure 2.18: *Van Krevelen* plot with regions for biomolecules marked in color. Adapted from (Brockman et al., 2018).

The *van Krevelen* plot gives information if a molecule is belonging to a certain group of biomolecules. Areas can overlap as molecules can have the same stoichiometric constraints but different structures.

k-means clustering

With k-means clustering, a dataset is grouped into k clusters. The number of clusters k is fixed and determined by the user, depending on the data. An elbow plot helps find a good value for k.

k gives the number of centroids whose position is selected randomly. Afterward, the data points are ordered to their nearest centroid. The position of the centroid is recalculated, and the ordering is repeated until the sum of squares within the cluster is minimal and the clustering centroids have their final position (Hartigan, Wong, 1979) (Blekherman et al., 2011, Frochte, 2019, Merkl, 2015).

$$J = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2 \tag{8}$$

With x_j as the data points, μ_i as the centroids of the clusters S_i and $\left\| x_j - \mu_i \right\|^2$ as the quadrated Euclidean distance.

For choosing the right value k an elbow plot, using WCSS is calculated. WCSS (Within-Cluster Sum of Square) is the sum of squared distances between each point and the cluster's centroid. The optimum k value is at the “elbow”, the number k where the slope of the graph changes. For clustering, the fisher iris data set the elbow plot shows the optimal number of clusters at 3, see Figure 2.19.

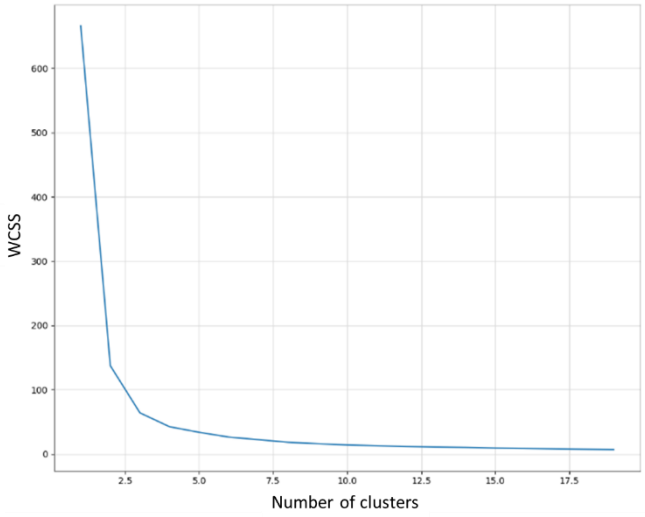


Figure 2.19: Elbow plot on the example of the Fisher iris data set. The optimal number of clusters is three.

Clustering with 3 centroids results in the picture shown in Figure 2.20. Class 1 is determined correctly. Class 2 and 3 are closer together (see PCA results Figure 2.14) and not all data points were correctly determined as k-means can’t separate correctly between the two classes.

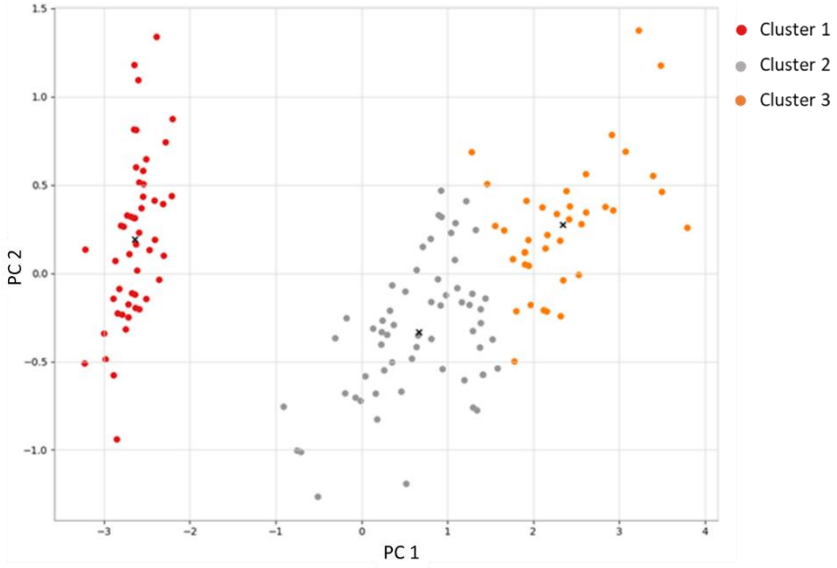


Figure 2.20: k-means clustering for the Fisher iris data set. The black x represents the centroid. Red = cluster 1; Grey = cluster 2; Orange = cluster 3.

The k-means algorithm is susceptible to noise and outliers. Density-based clustering (DBSCAN) is an alternative unsupervised clustering method that performs better when noise is present.

DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) uses the density of point groups to determine clusters and noise. The parameter epsilon (ε) determines a radius and another parameter the minimum number of points required to form a dense region. DBSCAN works by marking points as core points, edge points, or outliers. Core points are in a certain environment ε to other points and have a certain number of neighbor points belonging to the same cluster. The cluster can grow from these points. In Figure 2.21 they are marked red, each point has at least two neighbor points which are also core points. Edge points marked yellow in the figure are in the correct radius to another point, but only have one neighbor. The cluster can't grow any further from an edge point. Some points (blue color in the figure) are neither in the distance ε nor have any neighbors, determining them as outliers (Frochte, 2019, Merkl, 2015).

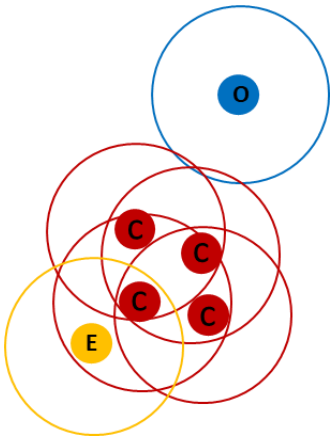


Figure 2.21: Exemplary DBSCAN with core points (red), edge points (yellow), and outlier points (blue). Modified after (Frochte, 2019).

DBSCAN is commonly used but doesn't work in high-dimensionality datasets. Data dimensionality is needed before performing DBSCAN. Overall is unsupervised clustering able to give an idea of how many different classes are in a dataset, but the results aren't as accurate as results from supervised learning.

2.2.6. Supervised learning methods

Supervised learning methods need a training data set, where classes are known. The algorithm is trained, and parameters are optimized. After training a test data set is used to determine how well the classifier performed. Performance and robustness testing of the classification is often done by cross-validation, which is described in the following chapter.

Cross-Validation

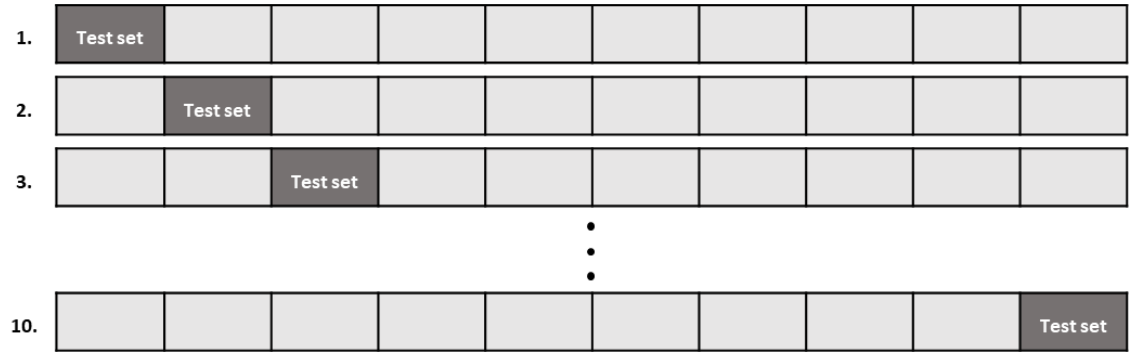
Cross-validation makes it possible to estimate the best parameters for an algorithm and to estimate the accuracy of a predictive classification. Usage reduces overfitting and helps with the bias/variance trade-off. Bias and variance are two error sources that need to be controlled so that the algorithm performs correctly when faced with unknown samples. A high bias error would result in underfitting, meaning the assumptions from the training data do not represent the data. A high variance error would result in overfitting, meaning the algorithm is perfectly trained for the training data, but as soon as an outlier or novel (unknown) data is presented the performance of the algorithm is not accurate.

Cross-validation is used by splitting a dataset with known classes into a test and a train set. The train set is introduced into the algorithm after training, and it can be determined how many of the samples were correctly classified. Generally, 70 – 85 % of the data set is used for training and 15 – 30 % for testing. There are different forms of cross-validation, depending on the size of the data set. Typically, n-fold cross-validation (usually n=10) is used, whereas the "leave-one-out" approach is used for small data sets.

n-fold cross-validation

The dataset is separated into *n* randomly chosen subsets of similar size. $\frac{n-1}{n}$ of the data set is used for training and $\frac{1}{n}$ for testing. This is repeated *n*-times while variating which subset is used for training and which for testing (see Table 2.3).

Table 2.3: Schematics of n-fold cross-validation, adapted from (Frochte, 2019).



The dataset usually is randomly partitioned into subsets. As shown in Table 2.3 90 % of the data set is used for training and 10 % for testing. This is repeated 10 times with a different subset for training and testing in each round. 10 repetitions are the usual number for n-fold cross-validation.

Leave one out cross-validation

With small datasets, a leave-one-out approach can be chosen, where n-1 objects are used for training and just 1 for testing. This can be repeated for each object of the data set.

In the end, for all cross-validation methods, the accuracy, see formula (9) can be calculated (Frochte, 2019):

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{9}$$

The accuracy itself doesn't give more information about the false predictions. It doesn't show if a sample from the wrong class was classified as X or if a sample from class X was classified wrongly. To visualize this a confusion matrix is helpful (see Figure 2.22).

		Actual class	
		species X	not species X
Predicted class	species X	7	0
	not species X	5	12

Figure 2.22: Example of a confusion matrix. Adapted after (Frochte, 2019).

In the example from Figure 2.22, one can find out if a sample is from species X or not. The classifier determined 7 samples as true positive and 12 samples as true negative. However,

5 samples were determined as not belonging to species X although they belong to species X, making those false-negative results. There were no false-positive results from the classifier. By this, the specificity (true negative (TN) rate), the sensitivity (true positive (TP) rate), and the precision (positive predictive value) can be determined. Confusion matrices can also be used with multi-level classifiers (Frochte, 2019, Merkl, 2015).

Support Vector Machine

A support vector machine (SVM) is a supervised learning method that performs classification and regression analysis. SVM is robust, flexible, and shows a high classification performance. SVM can also classify if many features are present, as they can work in higher dimensions. With a support vector machine, the margin between classes is maximized. SVM uses hyperplanes to separate classes from one another (see Figure 2.23), with the hyperplane having the maximum distance between the classes.

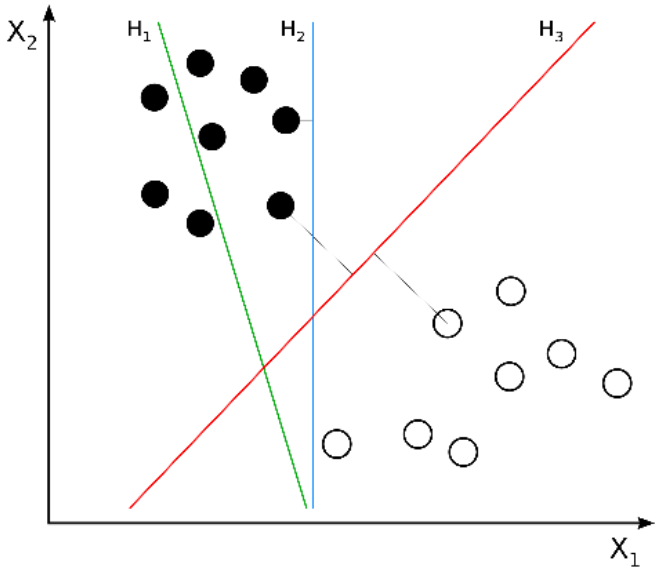


Figure 2.23: Hyperplanes separating two classes. H1 doesn't separate, H2 separates at a small margin and H3 separates at the maximum margin. Figure after (Weinberg, 2012).

In Figure 2.23 hyperplane H1 doesn't separate the classes at all, hyperplane 2 does, but only by a small margin, and hyperplane 3 gives the maximum margin. The figure shows the separation in two dimensions, but SVM performs the classification in higher dimensions using the so-called "kernel trick". The hyperplanes are determined in a high-dimensionality space and then transferred back into lower dimensions. Several kernels are available, the most common ones being linear, polynomial, and radial basis function (rbf) (Frochte, 2019, Merkl, 2015). When running SVM several parameters need to be chosen, the kind of kernel,

the value C, and gamma being the main ones. C is a regularization parameter, penalizing wrong classifications during training, high C values can lead to overfitting. Gamma is a kernel coefficient, with $1/n$ features being a common setting. Cross-validation can help find the optimal parameters.

k-nearest neighbors algorithm

k-nearest neighbor (kNN) is a supervised non-parametric classifier used for classification and regression. Data should be normalized as a distance metric is used and a dimensionality reduction should be performed before applying the algorithm. The advantage is, that there are no parameters besides the number of *k* neighbors that need to be chosen. Objects are classified by looking at the *k* nearest neighbors. E.g., with $k=1$ only the nearest neighbor is used to define into which class the new sample belongs. With k=3 the three nearest neighbors are used and so on. Cross-validation can be used to determine the best number for *k*. (Frochte, 2019, Merkl, 2015).

Machine learning for non-target MS analysis

Overall, supervised learning methods can be a powerful tool to determine which class a new, unknown sample belongs to. A prerequisite is that there is a sufficient training data set. Generally, in machine learning contexts several thousand samples are used as training data, which is usually not possible with biological or environmental sample sets. Studies on biological samples like fungi differentiation or environmental samples like water usually include less than 100 samples (Aliferis et al., 2013, Erler et al., 2020, Gotthardt et al., 2020, Guo et al., 2020b, Kang, 2011, Kim et al., 2016, Kruve, 2019, Maciá-Vicente et al., 2018, Müller et al., 2013, Samanipour et al., 2019, Zwickel et al., 2018).

Biological samples show intrinsic biological variation, which makes classifications and unsupervised clustering approaches a difficult task. Biological or environmental variation leads to dynamic feature spaces. Especially with high-resolution mass spectrometry features spaces become more and more complex, consisting of several thousand features. Nonetheless, machine learning can do what a human cannot, extracting information out of complex datasets with several thousand dimensions. For biological samples, relative abundances between features can be more meaningful than absolute (yes/no) abundancies. It can't be expected that all samples from class X will express a certain feature. Some samples might express the feature in very low abundances, below the measurement threshold, or not at all. But machine learning algorithms evaluate hundreds or thousands of

features, detecting patterns within complex feature spaces, which enables the determination of classes. Having a multiclass problem when classifying different fungal species for example enhances the difficulty of the classification. Careful method development including cross-validation is necessary (Liebal et al., 2020, Samanipour et al., 2019).

In conclusion, machine learning is a valuable technique for various analytical questions including the analysis of non-target LC-MS data. Applying machine learning algorithms on LC-MS data is a relatively novel approach, first applied in metabolomic studies in medical fields. First publications emerged in the early 2000s (Liebal et al., 2020). Together with high-resolution mass spectrometry, machine learning algorithms show high potential for the analysis of environmental and biological samples.

3. Thesis Motivation

Fungal spores are part of primary biological aerosol particles and influence our lives every day. They can act as allergens and human or plant pathogens, but also as biological plant protectants. Monitoring methods are either not precise, like microscopy, or time-consuming and expensive, like DNA analysis. Additional methods for fast and easy classification of fungal spores in environmental samples are needed.

Direct analysis of fungal spores enables filter sampling from the air and passes over the cultivation step. This is especially useful as less than 20 % of fungal species can be cultivated. Results from fungal spore analysis might be better generalizable and therefore transferable to real-world samples than results from axenic, single-species cultures. In axenic cultures, the fungi's metabolome usually does not represent the metabolome under "real-world" environmental conditions, as the expression of metabolites is highly dependent on growth conditions and the presence of other organisms (Begley, 2020, Overy et al., 2014, Rämä, Quandt, 2021). Because fungal spores are metabolically dormant and need to survive in many environments, they might not be as adapted to a certain environment and results might therefore be generally applicable.

Non-target high-resolution mass spectrometry can give a picture of the fungal spores' metabolome, including potentially class-differentiating secondary metabolites. High-resolution mass spectrometry allows a fast and comprehensive analysis. Even for metabolically dormant fungal spores the features detected by non-target LC-HRMS can be highly complex. Biological variation in-between species can be high, which hinders the detection of class-specific features. This makes novel data analysis methods including machine learning algorithms, particularly useful. The data analysis of biological samples with machine learning algorithms is demanding. Biological sample sets usually provide only a limited number of samples whereas high-resolution mass spectrometry data provides high dimensionally feature spaces. Data pre-processing, machine learning algorithms, and parameters need to be carefully chosen and validated, to obtain a sensible and robust method.

Fungal classification based on LC-or GC-MS analysis of fungal axenic cultures has been reported, but not based on fungal spores. The main aim of this doctoral work was the development of the first classification method of fungal species and classes based on non-target LC-HRMS data of fungal spores.

During this work, several machine learning algorithms, including unsupervised clustering and supervised classification methods, are investigated to find a suitable classification method. The developed method is applied for the class differentiation of fungal spores from four fungal classes from five different families. Additionally, it is examined if fungal spores of different species, but the same genus can be classified. This is evaluated on five fungal species from six different strains. The mass spectrometric analysis is performed with ESI and APCI ionization in both positive and negative modes to determine the most suitable ionization methods which would give a comprehensive picture of the fungal spores' metabolome. In addition, it is studied if features are detected that are class- or species-specific. Furthermore, basidiomycetes spores from the Amazonian rainforest are examined as well as a thermal desorption GC-MS system for the volatile organic compound profile of fungi. Additional application of non-target LC-HRMS data analysis for the evaluation of e-cigarette liquids and condensates is investigated.

4. Experimental Setup

In the following chapters, the experimental setup is described. First, the fungal spore samples are explained, then the experimental workflow. Instruments, chemicals, and computer programs used in this work are shown in the supplementary information, see appendix chapter 8.1.1.

4.1. Fungal spore samples

The main sample set consists of different ascomycetes, *Aspergillus versicolor*, *Cladosporium cladosporioides*, *Botrytis cinerea*, *Verticillium dahlia,* and several *Trichoderma* spp. (*T. longibrachiatum*, *T. fasciculatum*, *T. minutisporum*, *T. atroviride* and *T. harzianum* with two different strains). The taxonomic classification is shown in Table 4.1.

Table 4.1: Table of ascomycetes used in this work. Taxonomic classification from (Index Fungorum, 2020).

Genus	Family	Order	Class
Aspergillus	Trichocomaceae	Eurotiales	Eurotiomycetes
Cladosporium	Davidiellaceae	Capnodiales	Dothideomycetes
Botrytis	Sclerotiniaceae	Helotiales	Leotiomycetes
Trichoderma	Hypocreaceae	Hypocreales	Sordariomycetes
Verticillium	Plectosphaerellaceae	Glomerellales	Sordariomycetes

The samples belong to different classes, with exception of *Verticillium* and *Trichoderma*, which both belong to the class Sodariomycetes. An overview of different fungal classes and their connection with one another is given in Figure 4.1.

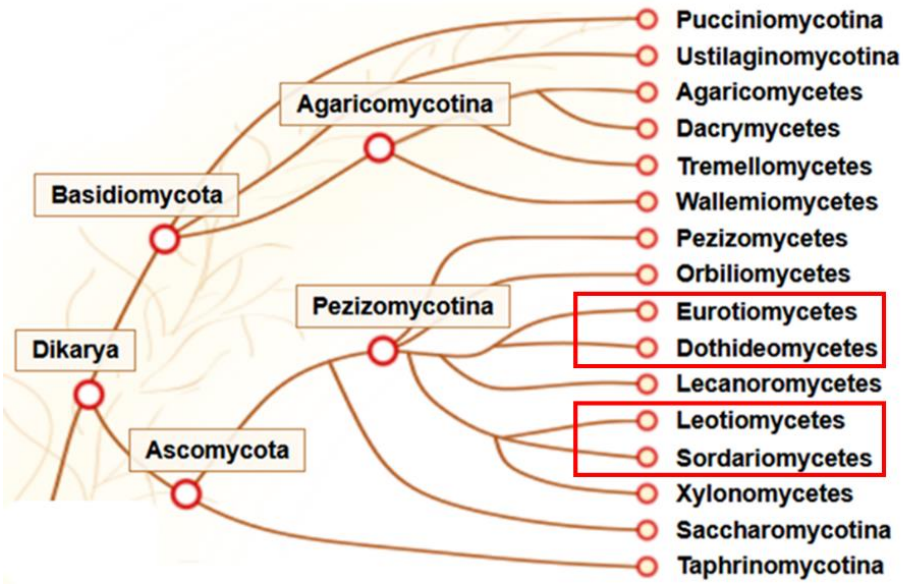


Figure 4.1: Tree illustrating relationships between groups of fungi. Fungal classes of Ascomycetes used in this work are marked in red. Adapted from (Grigoriev et al., 2014).

All classes belong to the subclade Pezizomycotina with Eurotiomycetes and Dothideomycetes_being more closely related on the one hand and Leotiomycetes and Sodariomycetes on the other hand.

Aspergillus versicolor is ubiquitously occurring in different environments like soil, marine, and indoor. It is commonly associated with indoor molds. The genus *Aspergillus* contains more than 350 species and is known to produce toxic or allergenic secondary metabolites (see chapter 1.2.3). Aspergillosis affects 14 Mio. People yearly, causing death to 600,000 individuals (Powell, 1994).

Cladosporium cladosporioides occur worldwide, outdoor as well as indoors. The genus *Cladosporium* contains over 700 species and is often found in bioaerosol samples. Cladosporium species produce no major mycotoxins of concern but are often found in very high concentrations outdoors, especially in summer, and can cause allergies (Grinn-Gofroń et al., 2019).

Botrytis cinerea is a plant pathogen known to infect more than 200 different kinds of plants. It lives parasitic and induces apoptosis in plants. If *Botrytis* infects grapes in spring it leads to a loss of harvest, but if it infects ripe grapes in fall, it can induce the so-called noble rot (german: *Edelfäule*), resulting in sweeter and more complex wines. Control is possible with chemical fungicides, although multiple fungicide resistances were reported (Rupp et al., 2016).

Verticillium dahlia is a plant pathogen with more than 50 species known to the genus *Verticillium*. Over 300 different kinds of plants can be infected and control by fungicides is difficult as *Verticillium* can persist in the soil (Barbara, Clewes, 2003).

Trichoderma has a teleomorph form called *Hyprocrea*. Nevertheless, the genus is called *Trichoderma*. *Trichoderma* consists of more than 200 species and is usually found on plant material or in soil, in some cases also as house mold. *Trichoderma* produces a wide range of secondary metabolites, peptaibols, siderophores, and diketopiperazines like gliovirin, polyketides, terpenes, pyrones, and isocyane metabolites. Some have antibiotic activity and/or improve plant and root growth and act against phytopathogenic fungi. Some *Trichoderma* strains are used as biological plant protectants, improving crop productivity. The strain *Trichoderma atroviride* , which was used in this work is used as a biological plant protectant in vineyards (tradename "Vintec" by Belchim Crop Protection). *Trichoderma* also produces VOCs some of them with a characteristic flavor ("coconut") which can be used for species or strain determination (Almeida et al., 2019, Bissett et al., 2003, Harman et al., 2004, Harman, 2006, Reino et al., 2007, Sood et al., 2020, Stoppacher et al., 2007, Zeilinger et al., 2015a, Zeilinger et al., 2016).

The CBS (*Centraalbureau voor Schimmelcultures*- central bureau of fungal cultures) or DSM (*Deutsche Sammlung von Mikroorganismen und Zellkulturen* German Collection of microorganisms and cell cultures) number of the used strains is given in Table 4.2.

Table 4.2: Table of species used in this work. Corresponding numbers are from the Centraalbureau voor Schimmelcultures (CBS) and Deutsche Sammlung von Mikroorganismen (DSM).

Species/ Strain	Number
Botrytis cinerea	DSM 877
Aspergillus versicolor	DSM 19652
Cladosporium cladosporioides	CBS 109.21
Verticillium dahliae	DSM 11938
Trichoderma longibrachiatum	CBS 488.78
Trichoderma fasciculatum	CBS 118.72
Trichoderma minutisporum	CBS 584.95
Trichoderma harzianum strain B	CBS 608.89
Trichoderma harzianum strain A	CBS 348.96
Trichoderma atroviride	CBS 122089

Fungal spore samples were grown under different conditions to induce phenotypical plasticity. Differences were growth media, temperature, day/night rhythm, and partly also season. Samples from *Trichoderma harzianum* strain A and B and *Trichoderma atroviride* were examined for three years and stored in between in the "*Stammsammlung*" at 4 °C. Fungi that are reactivated from a dormant state can change their phenotype after each activation. The media used were PDA (potato dextrose agar) and HMG (yeast malt agar) at 26 °C (darkness) and 20 °C (day-night light rhythm) with incubation periods between 14 and 30 days. Both T. *harzianum*. and *T. atroviride.* were also subjected to growth at 20 °C in darkness and 26 °C with day/night rhythm in another laboratory building. In Figure 4.2 it is shown that fungal samples of the same species, grown under different conditions, have different appearances. Samples shown in the upper right corner (samples 5-10), which were grown in darkness show noticeably less pigmentation than samples shown in the lower right corner, which were grown in daylight.

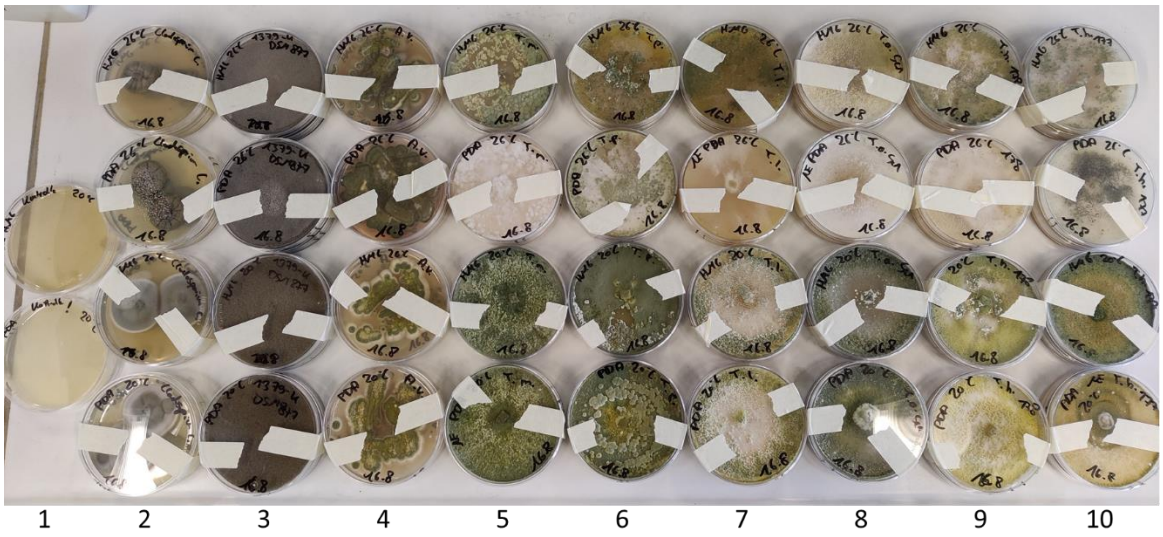


Figure 4.2: Photograph of fungal cultures before harvest. Left to right. Column 1: Blank samples, 2: *Cladosporium cladosporioides*, 3: *Botrytis cinerea*, 4: *Aspergillus versicolor*, 5: *Trichoderma minutisporum*. 6: *Trichoderma fasciculatum*, 7: *Trichoderma longibrachiatum*, 8: *Trichoderma atroviride*, 9: *Trichoderma harzianum* strain B, 10: *Trichoderma harzianum* strain A.

An overview of the biological replicates is given in Table 4.3.

Table 4.3: Overview of biological replicates of fungal spore samples.

Genera/Species/ Strain	Biological replicates parameters	Biological replicates
Aspergillus versicolor	PDA and HMG 20°C/26°C each. Repetition after 1 month.	8
Botrytis cinerea	PDA and HMG 20°C/26°C each. Repetition after 3 months.	8
Cladosporium cladosporioides	PDA and HMG 20°C/26°C each. Repetition after 3 months.	8
Verticillium dahliae	PDA and HMG 20°C/26°C each. Repetition after 1 month.	9
Trichoderma longibrachiatum	PDA and HMG 20°C/26°C each. No repetition.	4
Trichoderma fasciculatum	PDA and HMG 20°C/26°C each. No repetition.	4
Trichoderma minutisporum	PDA and HMG 20°C/26°C each. No repetition.	4
Trichoderma harzianum strain B	PDA and HMG 20°C/26°C, grown in fall 2019 (old laboratory), spring 2021, and fall 2021.	8
Trichoderma harzianum strain A	PDA and HMG 20°C/26°C, grown in fall 2019 (old laboratory), spring 2021, and fall 2021.	9
Trichoderma atroviride	PDA and HMG 20°C/26°C, grown in fall 2019 (old laboratory), spring 2021, and fall 2021.	13
Trichoderma samples biological replicates total	Total of *Trichoderma* biological replicates, see above.	42
Biological replicates	Total of all biological replicates, see above	75

At least 8 samples which are biological replicates are available per genera except for *Trichoderma* where 5 different species with a total of 6 strains are available, totaling 42 *Trichoderma* samples. All together 75 biological replicates are available.

Basidiomycetes samples from the Amazonian rainforest

Additionally, fungal spore samples from Brazil were examined, which were sampled on filters. They were collected in the Amazonian Rainforest at the Amazonian Tall Tower Observatory (ATTO), close to São Sebastião do Uatumã, Brazil. Sampling and taxonomical classification were performed by Cybelli Barbosa (C.Barbosa@mpic.de). The samples are four single spore samples and one mixed sample. The single spore samples were taken very close to the fungi's fruiting body, by a pump with sucked the spores onto a filter. Taxonomic classification was based on the fruiting body. The mixed sample was taken on ground level at the research site, close to the trees where the fruiting bodies occur. The samples are basidiomycetes of the genera *Trametes, Picnoporus,* and *Ganoderma*. One sample was classified on the family level as Polyporaceae. The sample set is very small. Planned further samples couldn't be collected due to COVID-19 related closures in 2020 and 2021.

Table 4.4: List of basidiomycetes used in this work.

Sampling Period	Genus	Family	Order	Class
Wet 2019	*Trametes*	Polyporaceae	Polyporales	Agaricomycetes
Wet 2019	*Picnoporus*			
Wet 2019	*Polyporaceae*			
Dry 2018	*Ganoderma*	Ganodermataceae		
Dry 2018	Field Mix	mixed	mixed	mixed

4.2. Workflow for extraction, measurement, and data processing

The workflow is shown in Figure 4.3. In short, samples were harvested, followed by extraction and spore counting. The resulting raw extracts were resolubilized and diluted, measured by LC-MS, and pre-processed by MZmine. Afterward, blank subtraction and alignment in MATLAB finalized the pre-processing step. Machine learning, including normalization, dimensionality reduction, unsupervised and supervised learning was performed with Python.

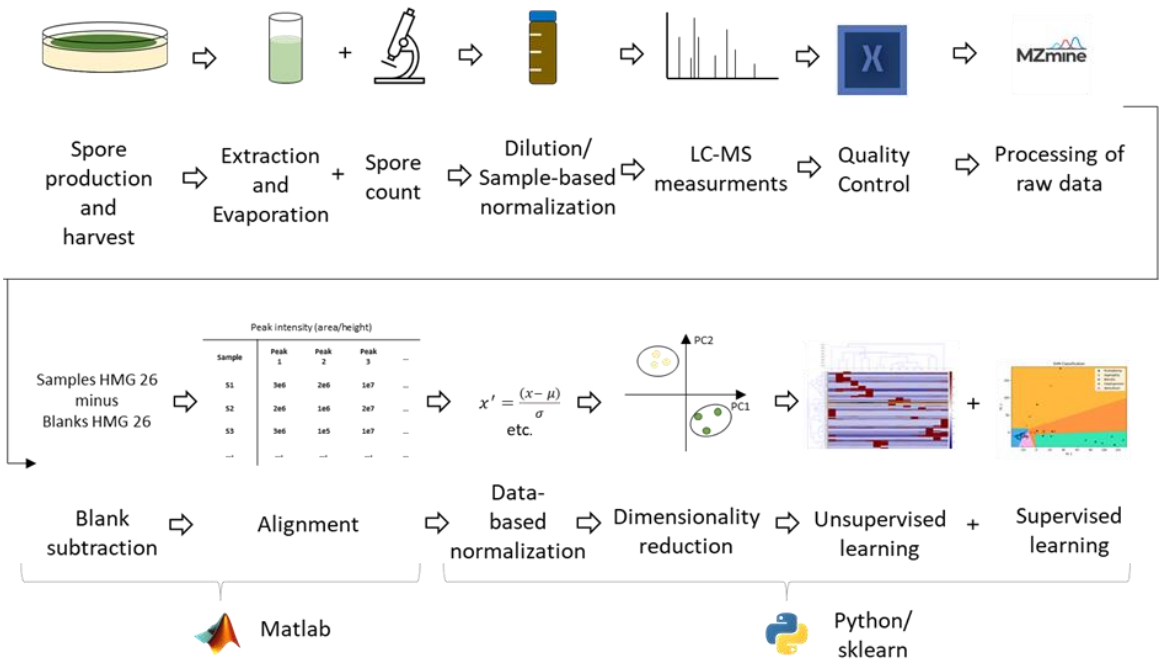


Figure 4.3: Workflow for sample preparation, extraction, measurement, and data processing of fungal spore samples.

Fungal spore samples were grown by Hendrik Neumann and Andrea Ebert-Jung, Thines Group at the Biology Department, University Mainz. PDA consists of 26.5 grams of Potato-Extract-Glucose Bouillon (Carl Roth GmbH) on 1 L of water with 2 % agar. HMG consists of 10 grams glucose, 4 grams yeast extract, and 10 grams barley malt on 1 L water with 2 % agar at pH 5.5. Blank medium samples were incubated at the same time as fungal samples. The incubation varied between 14 and 30 days, depending on the individual growth progression of the samples.

The fungal spores were harvested with Milli-Q water and a scraper and filtered through Miracloth to separate the spores from the mycelium. The absence of mycelium was controlled by light microscopy. The filtrated spores in water were filled to the maximum in a 50 mL Grainer tube, homogenized, and a 1 mL aliquot was taken for counting. The

suspension was centrifuged (5 min, 3500 rpm) and the aqueous supernatant was discarded. The spore pellet was frozen at – 25 °C until further use. Samples were frozen at least once before extraction as freezing improves cell wall breakage. Counting was performed on a Neubauer improved counting plate, using double sampling. Samples were diluted until a suitably countable dose, equaling around 150 cells per middle square, was reached.

Extraction of the metabolites was performed as follows: 15 mL of an organic solvent were given onto the pellet, followed by homogenization. Extraction was performed in an ultrasonic bath at ambient temperatures for 15 minutes. Centrifugation (3500 rpm, 5 min) separated the spores from the solvent which was removed by a syringe. The extract was filtered through a 0.2 µm PTFE syringe filter into a brown glass vial. The solvent was evaporated at 30 °C under a gentle nitrogen stream. This procedure was repeated twice.

Fungal spores from filters were prepared by taking 1/8 of the filter (diameter 47 mm) and cutting it into small pieces. The filter pieces were placed into a brown glass vial, 15 mL of methanol was added, and extraction was performed by ultrasonication. The liquid extract was removed by a syringe and filtered through a 0.2 µm PTFE syringe filter into a brown glass vial. The solvent was evaporated at 30 °C under a gentle nitrogen stream. This procedure was repeated twice.

The solvents were evaporated until dry, and the extract was taken up in 2 mL methanol. The solution was transferred into an HPLC vial through a syringe filter (PTFE, 0.2 µm). Dilution of samples was performed according to their spore count and a factor concerning the spore size (see page 66), reaching $1 \cdot 10^6$ spore equivalents per µL.

The UHPLC was operated with water/acetonitrile (98:2 % with 400 µL/L formic acid) as eluent A and methanol as eluent B. As a column, the Hypersil Gold C18 with 50 mm length and a particle size of 1.9 µm was chosen. The injection volume was 5 µL and the following gradient was used:

Table 4.5: Eluent gradient for UHPLC measurements. Eluent A; Water/acetonitrile, eluent B: Methanol.

Minute	Eluent A [%]	Eluent B [%]	Flow [µL/min]
0	95	5	500
5	50	50	500
10	0	100	500
15	0	100	500
16	95	0	500

Samples were measured thrice, in positive and negative mode, with HESI and APCI ionization. Between samples of a different species or growing condition, a blank sample of pure methanol was used. Every 20 samples a quality control sample was used with the following compounds (500 ng/mL), measured in positive and negative mode:

Table 4.6: Composition of the quality and retention time control sample.

Substance	Corresponding ion	Retention time
Ergosterol	379.3359 [M+H-H_2O^+]	10.57
Reserpin	609.2806 [M+H^+]	5.81
Caffeine	195.0876 [M+H^+]	2.07
Syringealdehyde	183.0651 [M+H^+]	2.76
Vanillin	153.0546 [M+H^+]	2.53
Gluconic acid	195.0510 [M-H^-]	0.32
Camphersulfonic acid	231.0696 [M-H^-]	2.38
Lauric acid	199.1703 [M-H^-]	8.93
Mannitol	181.0717 [M-H^-]	0.31
Xylitol	151.0611 [M-H-]	0.33

The Orbitrap was calibrated daily with a customized sodium acetate (2 mM/mL) solution, where the acetate clusters are used for calibration. Calibration with commercially available CalMix was performed approximately every 3 months and after each bake-out. The Orbitrap and ion sources were operated under the following parameters:

Table 4.7: Parameters for Orbitrap and ionization sources.

Parameter	HESI positive	APCI positive	HESI negative	APCI negative
Scan Range (*m/z*)	50 - 750	50 - 750	50 - 750	50 - 750
Polarity	positive	positive	negative	negative
Resolution	140,000	140,000	140,000	140,000
Microscans	1	1	1	1
Lock masses	Off	Off	Off	Off
AGC target	1e6	1e6	1e6	1e6
Maximum inject time (ms)	50	200	200	200
Sheath gas flow rate (a.u.)	53	10	53	24
Aux gas flow rate (a.u.)	14	0	14	5
Sweep gas flow rate (a.u.)	3	0	3	0
Spray voltage (kV)	3.5	4.0	2.5	3.3
Capillary temp. (°C)	269	320	269	250
Vaporizer temp (°C)	438	388	438	388

The resulting raw files were checked if their ergosterol peak showed mass accuracy or signal intensity deviation to detect outliers early. Samples with low ergosterol content were usually too diluted. If possible, measurements were repeated at a higher sample concentration. The quality control samples were checked if the retention times were stable and if signal intensities and mass accuracy were as expected.

The raw data was processed by MZMine 2.51, using the ADAP algorithm. For further information on the ADAP, algorithm see (Du et al., 2020) One sample, including its triple technical replicates and the corresponding blank measurement, was processed in one batch. Parameters for the data processing in MZMine are shown in the supporting information, see Table 8.4. Molecular formulas were calculated by MZMine. The raw data processing resulted in a feature list with *m/z* ratio, retention time, peak area, and predicted molecular formula.

Subsequently, the respective blank sample is subtracted by an in-house build MATLAB script (adapted from Martin Brüggemann, brueggemann@tropos.de see supporting information page 136). Blank values were multiplied by the factor 3 to ensure complete subtraction. After subtraction, only signals which were present in all three technical replicates were kept and the resulting values were averaged. After blank removal the samples were aligned by an in-house MATLAB script, at retention times and mass accuracy

ranges adapted to the samples, usually at around 1 minute retention time range and 5 ppm mass accuracy range (see also chapter 5.1.4). After alignment, data was checked, if the ergosterol peak was aligned correctly, and if there were deviations in the mass accuracy, which would lead to incorrect alignment.

Afterward, a dimension reduction by PCA was performed using the Python 3 scikit-learn package. Before performing the PCA, data were log-transformed followed by a z-score standardization. t-SNE visualization, k-means, DBSCAN clustering, and supervised learning classification (Support Vector Machine and k.nearest neighbors) were performed in Python, using the respective functions implemented in Scikit-learn. More information on the functions can be found in the documentation of each function online (scikit-learn developers, 2021). Parameters will be discussed in method development. Visualization was performed with Pythons Matplotlib and Seaborn. Hierarchical clustering analysis was performed with a MATLAB script adapted from Denis Leppla (Hoffmann group). Parameters were adjusted to the respective problem as part of the method development and are discussed in chapters 5.2.2 to 5.2.4. The MATLAB and Python scripts are available in the supporting information see page 136 et seq.

5. Method Development

5.1. Sample preparation and measurement

In the following chapter the method development of the extraction and measurement procedure is presented in chronological order of the workflow, see Figure 4.3. Data processing and machine learning development will be presented in chapter 5.2.

5.1.1. Fungal spore cultivation and extraction

Cultivation

Fungal spore cultivation was tested on Petri dishes (diameter 94 mm), Fernbach flasks, and canisters. As growth was good and harvest was easier from Petri dishes those were chosen for further use. Different media for growth were tested, small animal litter (wood chips), oatmeal agar (OM), yeast-glucose agar (CM), yeast-malt extract (HMG), and potato-dextrose extract agar (PDA). HMG and PDA showed good sporulation and were therefore chosen for further experiments. PDA consists of potato starch, glucose, and agar, whereas HMG consists of dextrose, malt extract, yeast extract, and agar, showing different nutritional profiles, resulting in different environments for the fungi to grow on. Together with different temperatures and day-light cycles, this can induce phenotypical plasticity. Additional to the fungal spores described in the experimental two more samples from the *Trichoderma* genus, *Trichoderma hamatum,* and *Trichoderma viride* were tested, but they showed very minimal sporulation and weren't used further.

Extraction

Fungal spores have a sturdy membrane that needs to be penetrated to extract the metabolites of interest. Extraction was tested under different conditions, from protocols from IBWF (Institut für Biotechnologie und Wirkstoff-Forschung, Mainz, Germany) and literature (Castrillo, Oliver, 2011, Feussner, Feussner, 2019, Gummer et al., 2012, Madla et al., 2012, Winder, Dunn, 2011). Extraction was performed with ultrasonication to break cell walls. Additionally, samples were frozen before extraction and in between extractions as ice crystals can penetrate the cell wall as well.

For comparison of solvents, methanol (MeOH), ethyl acetate (EtOAc), and a methanol: water Mixture (60:40) were tested, all three commonly used in the extraction of fungi. The methanol: water mixture (polarity index ~7.2) is the most polar, methanol is polar (polarity index 5.1), and ethyl acetate is semi-polar, with a polarity index of 4.4 (Snyder, 1974). As test samples, two *Trichoderma harzianum* strain A samples were aliquoted into three subsamples and extracted with the three solvents. Samples were measured with APCI and ESI to check if any major differences are detected by the methods.

The ergosterol value was examined because detection of ergosterol as a membrane compound implies breakage of the membrane. The methanol extract showed the highest ergosterol peak areas with ~$3 \cdot 10^7$ arbitrary units (a.u.), followed by ethyl acetate with ~$9 \cdot 10^6$ a.u. and water: methanol extract with the lowest ~$1 \cdot 10^6$ a.u. As ergosterol is a non-polar molecule with only one hydroxy group it is expected that very polar solvents wouldn't work well, however, methanol seemed to have extracted/solubilized ergosterol the best.

More importantly, it was tested how many different substances were extracted as in non-target analysis a compound profile as complete as possible is desirable. With ESI ionization in positive mode, a total of 761 compounds could be detected. On average between the *T. harzianum* replicates, 368 compounds were detected in the methanol extract, 361 in the methanol: water extract, and 170 compounds in the ethyl acetate extract. 82 compounds were exclusively found in methanol: water extract, 60 exclusively in the methanol extract, and 30 compounds exclusively in the ethyl acetate extract. Especially with methanol and methanol: water some compounds were present in both extracts. To check if extracts of a certain solvent solvated groups of biomolecules specifically a *van Krevelen* plot were calculated for the three extracts, see Figure 5.1.

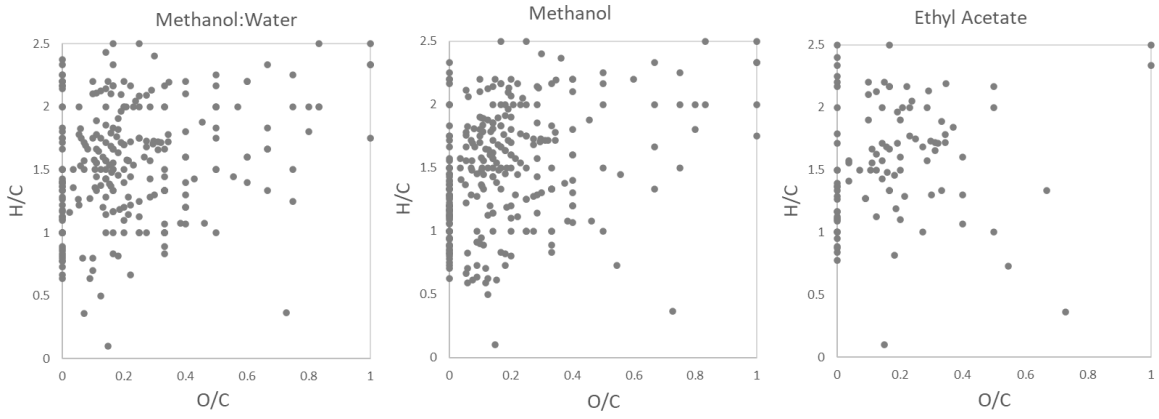


Figure 5.1: *van Krevelen* plot for *Trichoderma harzianum* extracts with different solvents. Ionization with ESI in positive mode. Left: Methanol: water extract, middle: Methanol extract, right: Ethyl acetate extract.

Methanol: Water and Methanol extracts show similar *van Krevelen* profiles, with a higher density in the area where peptides are found (see page 42). The ethyl acetate extract contains fewer compounds and doesn't show an emphasis on a certain biomolecule group. Methanol and methanol:water show some compounds in the carbohydrate group where ethyl acetate is not. This is presumably due to the polar nature of carbohydrates. Methanol shows some compounds in the polyketide region (O/C ratio: 0-0.4; H/C ratio: 0.5 – 1) which might be interesting for metabolic profiling.

It was checked if results would be different when ionized with APCI, which is more likely to also ionize non-polar compounds. With APCI 508 compounds were detected, with 343 in the methanol extract, 314 in the methanol: water extract, and 127 in the ethyl acetate extract. 119 compounds were present in both methanol and methanol: water, of which none was present in the ethyl acetate sample. 38 compounds were only available in the ethyl acetate extract, 139 in only the methanol extract, and 120 only in the methanol: water extract. 67 compounds were present in all three extracts.

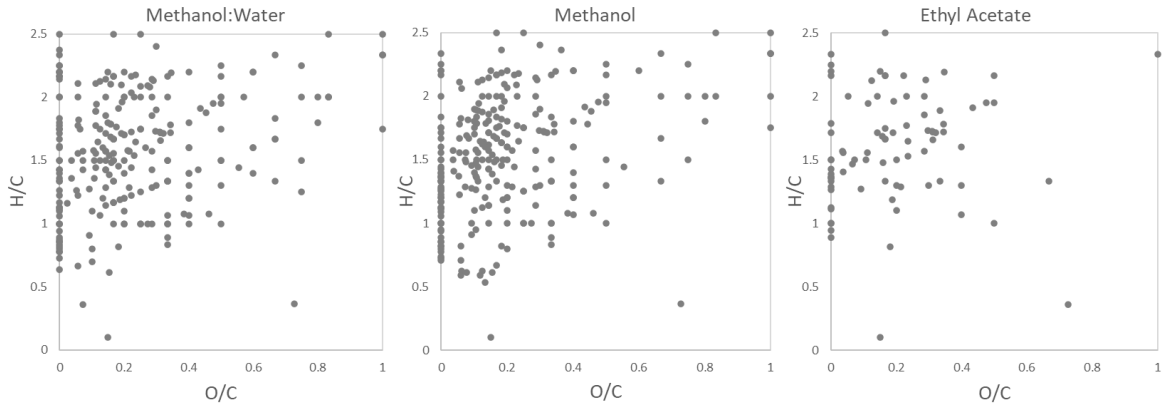


Figure 5.2: *van Krevelen* plot for *Trichoderma harzianum* extracts with different solvents. Ionization with ACPI in positive mode. Left: Methanol: Water extract, middle: Methanol extract, right: Ethyl acetate extract.

The *van Krevelen* plots are very similar to the ones measured with ESI. As differences are minor it can't be determined which ionization method is better for metabolite profiling. Ethyl acetate which is the least polar compound was expected to extract compounds that are more likely to be ionized with APCI but that doesn't seem to be the case.

Overall, ethyl acetate showed the least number of compounds, and very few compounds that could not have been extracted by methanol or methanol:water. Methanol and methanol:water show very similar results with ergosterol being better soluble in pure methanol. As the evaporation times of pure methanol extracts are distinctively shorter,

Methanol was chosen to extract the fungal spore samples. The ionization methods didn't give majorly different results; therefore, ionization will be performed by both methods to determine which is better for non-target profiling on the full data set.

5.1.2. Sample-based normalization

To compare profiles of different spore genera and for performing analysis by clustering/classifications, roughly the same concentration of all samples is needed, otherwise, differences between samples are based on concentration and not on the difference of the actual sample. Sample-based normalization was tested with spore count and spore weight. As a mix of data- and sample-based normalization the relationship between spore count/weight and content was examined.

Samples were counted with a Neubauer counting chamber and diluted to the same spore number in each sample (about $1 \cdot 10^6$ spore equivalents per microliter). With spores of the same genera/type, this worked smoothly, whereas comparing spores of different genera showed difficulties because of different spore sizes. The spore volume was calculated from the reported spore sizes (see Table 5.1), to calculate a factor that compensates for the different sizes.

Table 5.1: Spore sizes and calculated approximate volumes for spores of different genera.

Genera	Mean size (µm)	~Volume (µm³)	Source
Trichoderma	3.3	19	(Di Filippo et al., 2013, Harman et al., 2004)
Aspergillus	2.9	12	(Lau et al., 2006, Miller, Young, 1997, Pasanen et al., 1999)
Cladosporium	3.7	27	(Buiarelli et al., 2013, Lee et al., 2007, Pasanen et al., 1999, Yamamoto et al., 2012)
Botrytis	8.8	350	(C.H. Chen et al., 2009)
Verticillium	3.9	31	(Feng et al., 2002, Qin et al., 2008)

Spore size and volume and the respective dilution factor were compared with *Trichoderma*, which has a spore size of around 3.3. µm and a volume of approximately 19 µm³. *Aspergillus* spores are smaller and were therefore diluted less by a factor of 0.5 than the other samples. *Botrytis* was diluted 10-fold more than the other samples, as the spores were the biggest of the data set. Overall, the dilution factors were not always reliable, and samples needed to be

either diluted further or concentrated after a first preliminary measurement. The counting is a source of error, as sometimes spores clumped together despite homogenization, or the necessary spore count on a middle field wasn't reached at lower concentrated samples. Additionally, the calculated spore volume is only a rough estimate and therefore also the dilution factor.

To enhance sample base normalization, it was tested if the samples could be diluted according to the weight of the raw extract after evaporation. Raw extract weights varied between 2 and 20 mg, depending on how well the samples sporulated. The blank weight showed values between 0.5 and 3.8 mg with a standard deviation of 1.5 mg. This might be because the media was sometimes very solid and sometimes quite fragile, leading to different amounts of media being incorporated into the samples. This makes it impossible to determine the actual weight of the spore's raw extract.

To have additional control over the concentration of the sample it was checked if the ergosterol content or the total ion count could be used for normalization respectively for validation if a sample needed to be measured again at a different concentration.

5.1.3. Ionization source and polarity, chromatography, and quality control

All samples were measured in positive and negative modes with the ionization sources HESI and APCI. This should enable a comprehensive picture of the fungal spores' composition and determine which method is most suitable for the differentiation of genera and species.

Chromatography

Chromatography was performed on a standard C18 column. C18 is a universal reverse phase column that is suitable for most analytes. Polar components of fungal spores like sugars won't be separated, but secondary metabolites, which often have a non-polar part will be separated on the column. Several column lengths were tested, but as methanol was to be found the most suitable eluent only the shorter column of 50 mm could be used due to the upper-pressure limit of the UHPLC. For eluents, acetonitrile and methanol as the organic solvent were tested. Previous work in the Hoffmann group showed a better separation and peak shape for ergosterol when using methanol as the organic eluent (Martin Müller, 2016). Acetonitrile has a higher elution strength than methanol, leading to shorter retention times, which is a good choice for target analysis but not necessarily for non-target analysis, which is why methanol was chosen.

To control the quality of the separation and the stability of retention times a quality control standard was chosen, of which ergosterol is an ingredient. ergosterol will also be important for the control of the concentration. For the quality control sample, a wide range of mass-to-charge ratios and retention times was preferred, ionizable in positive and/or negative mode to control retention time and mass accuracy as well as overall instrument performance for the complete run. Several compounds were tested, e.g. glutamine, glycerine, asparagine, bilirubin, cholesterol, palmitic acid, cortisone, glucose, urea, beta-carotene with the final mixture containing the following compounds: ergosterol (fungal membrane compound), reserpine, caffeine, syringaldehyde, vanillin, gluconic acid, camphor sulfonic acid, lauric acid, mannitol (fungal compound) and xylitol, which cover as a mass-to-charge range from *m/z* 151 to *m/z* 609 and a retention time range from 0.3 min (no retention) to 10.5 min.

Ionization

Ionization was performed in positive and negative modes to evaluate which method would give the most comprehensive picture and if certain substances are unique for a fungal genus.

As ergosterol is a biomarker for fungal spores (see chapter 1.2.4) its ionization was evaluated. Usually, ergosterol measurements are performed with GC-MS or LC-MS (Headley et al., 2002, Lau et al., 2006, Miller, Young, 1997, Srzednicki et al., 2004). Previous methods for the quantitation of Ergosterol in fungal spores included GC-MS, Limit of detection (LOD) 0.22 – 2.1 ng/mL (Miller, Young, 1997); HPLC-UV, LOD 20-80 ng/mL (Beni et al., 2014, Miller, Young, 1997), and HPLC-MS; LOD: 8.6 ng/mL (Headley et al., 2002).

The UHPLC-Orbitrap was tested for its LOD and Limit of Quantitation (LOQ) of ergosterol, after DIN 32465 from the slope of the calibration line. Two ions were found corresponding to ergosterol: *m/z* 397.3465 $[M+H]^+$ and *m/z* 379.3357 $[M+H-H_2O]^+$. The intensity of the $[M+H-H_2O]^+$ adduct is considerably higher, therefore the adduct will be used for quantitation and the $[M+H]^+$ ion for additional identification. The Orbitrap and ionization source settings are described in the experimental information, chapter 4. The Limit of Detection and Quantitation of ergosterol ([M+H-H2O]+) with two different ionization sources are presented in Table 5.2.

Table 5.2: Limit of Detection and limit of quantitation for ergosterol with HESI and APCI ionization for m/z $[M+H-H2O]^+$.

	HESI	APCI
LOD [ng/mL]	20.5	0.9
LOQ [ng/mL]	76.8	4.8

The LOD with the electrospray ionization source was in the range of the reported values in Literature for HPLC-MS measurements. The APCI LOD is at 0.9 ng/mL significantly lower than the HESI LOD.

5.1.4. Raw data processing, blank subtraction, and alignment

For the processing of raw data several commercial and open-source programs are available, like XCMS, OpenMS, Thermo Fishers SIEVE, or MzMine (Domingo-Almenara, Siuzdak, 2020, Du et al., 2020). In this work open-source programs were tested, as support for the available program SIEVE has been discontinued; novel commercial programs are expensive and not as versatile as open-source programs. MzMine was chosen as it implemented the ADAP algorithm which is favorable compared to XCMS due to fewer false-positive peaks (Myers et al., 2017).

Data processing in MzMine consists of several steps; extracting the mass-to-charge ratios, forming extracted ion chromatograms, detection of chromatographic peaks and differentiation from noise, deconvolution, several alignments, and filtering steps, and calculation of molecular formulas. In total 11 steps with 50 settings need to be chosen. Settings were evaluated on several samples of *Trichoderma harzianum* and kept for the whole data set to maintain reproducible results. The three technical replicates of a sample were processed together with the technical replicates of the corresponding blank sample, resulting in a list with mass-to-charge ratio, retention time, predicted molecular formula, and intensity.

For background subtraction, the blank intensities were multiplied by the factor 3 to ensure complete blank subtraction. Some features, like signals for sugars like mannose, etc. tend to be subtracted as well as the blanks are rich in sugars as well. This loss of information is necessary as the blank information should not influence the final feature list.

Alignment was first tested based on features having the same molecular formula and the same retention time. It has turned out that this approach is not tenable, as not all features of

the same mass-to-charge ratio will be calculated with the same molecular formula. Small deviations in the fourth digit after the comma will lead to different molecular formulas or no molecular formula at all. Therefore, the alignment was changed to align according to the mass-to-charge ratio and retention time. Tolerance of mass-to-charge value corresponds to the mass accuracy of the measurements which was ideally below 2 ppm. But as instrument performance varies higher mass tolerances need to be chosen to enable alignment of samples. Usually, a mass tolerance of 5 ppm is sufficient but in the case of alignment of samples that are one or more years apart, with several instrument maintenance and technical procedures in between the mass tolerance needed to be risen to 7 ppm to ensure correct alignment. Alignment correctness was mainly controlled by the known mass-to-charge ratios for the ergosterol $[M+H^+]$ and $[M+H^+-H_2O^+]$ ion as well as with randomly chosen features. Retention time tolerance was usually chosen to be around one minute to account for all peaks. In the case of alignment of samples which were measured with an old column and different capillaries retention time tolerance needed to be risen to up to 2 minutes. This was especially important when samples that were measured on an old C18 column were aligned with samples measured on a new C18 column. Despite being the same model and manufacturer retention times varied up to a minute between measurements. After alignment samples were ready for further data analysis.

5.1.5. Data evaluation – Semi target approach

In non-target high-resolution mass spectrometry, a wide range of compounds is available for analysis. Depending on the parameters chosen for analysis, like peak intensity threshold or filtering steps, several thousand compounds are measured per sample. Many of those compounds are not sample-specific, they can originate, e.g., from the fungi's primary metabolism, meaning they occur in all fungal samples. However, secondary metabolites can be strain-, species- or genera-specific and are proposed to differentiate fungal spore samples from one another.

The easiest approach to finding those specific compounds would be a filtering step if certain compounds are only present in samples of a specific species. Evaluating this approach for fungal spores showed, that the phenotypical plasticity impedes this approach. Supposedly species-specific metabolites are also present in species of the same genera. Finding genera-specific compounds showed to be unsuccessful as well. When evaluating all five species of *Trichoderma* no compounds were found that are present in all samples. Detailed results are shown in chapter 6.2.3.

Overall, finding species or genera-specific markers in fungal spore samples was not successful. Even if analyzed fungal spores were metabolically dormant, they showed a wide variety of phenotypes. The human eye can't make sense of a fungal spores' profile if compounds are only present in 80 or 90 % of samples, especially when confronted with thousands of compounds. However, computer-aided methods can find patterns in the sample's profiles and find similarities or differences where a human can't. The application of machine learning in non-target analysis gets more and more important, especially with the wider availability of high-resolution mass spectrometers. In the following chapter, the method development for the application of machine learning on fungal spore samples is described.

5.2. Data analysis with machine learning algorithms

All method development was performed on the smaller and the larger datasets (see Table 4.3 and Table 5.5) and all ionization methods, Electrospray ionization (ESI) and Atmospheric Pressure chemical ionization (APCI) in both positive and negative modes. As results are similar, the method development results are discussed for the larger dataset, with all samples, including samples from 2020 (see Table 4.3), ionized by ESI in the positive mode if not stated otherwise. Results from method development for the other ionization methods are shown in the attachment.

5.2.1. Data-based normalization

Data-based normalization is an important step before applying machine learning algorithms. Sample-wise normalization is performed to correct different sample concentrations so that distances between features of different samples are due to biological differences and not due to concentrations. Additionally, feature-wise normalization ensures that highly abundant metabolites don't overpower the analysis as absolute distances between highly concentrated metabolites could be larger than smaller distances between less abundant metabolites. Sample-wise normalization was tested with ergosterol as a marker, with total ion count (TIC), mean and median. Feature-wise normalization by mean, median, and z-score was tested. Additionally, a log transformation was performed. The theoretical background is described in chapter 2.2.2.

As a first step, the data were log2 transformed to remove heteroscedasticity. Heteroscedasticity is typical for MS data as the distribution of the signal intensities does not correspond to a normal distribution. As log2 can't be calculated from zero values all feature intensity values were shifted by 1 before the transformation. (Blekherman et al., 2011, Enot et al., 2008, Filzmoser, Walczak, 2014, Sysi-Aho et al., 2007, van den Berg et al., 2006, Veselkov et al., 2011, Wu, Li, 2016, Yi et al., 2016)

Sample-wise normalization with a marker compound is a typical approach. Ergosterol would be a suitable marker compound as it is available in all samples and measurable at least in positive mode. However, the ergosterol content of fungal spores differs between species with values ranging between 0.68 and 5.11 picogram(pg)/spore depending on the publication (Lau et al., 2006, Miller, Young, 1997). *Cladosporium* shows ergosterol concentrations between 1.9 and 3.1 pg/spore with *Trichoderma* in the same range, *Aspergillus* shows lower values between 1.3 and 2.5 pg/spore (Di Filippo et al., 2013, Miller, Young, 1997, Pasanen et al., 1999). Overall ergosterol concentrations vary widely between species. Most of these calculations assume the same size/weight of the spores which doesn't reflect reality. Spore sizes and morphology vary greatly with spores in this work being either spherical or elongated (*Cladosporium*) (Buiarelli et al., 2013, Di Filippo et al., 2013, Lau et al., 2006, Lee et al., 2007, Miller, Young, 1997). Therefore the ergosterol concentration isn't necessarily linearly correlated with spore size and spore content. (Gutarowska et al., 2015, Pasanen et al., 1999). Additionally, the extraction efficiency can vary. Normalizing by ergosterol can introduce bias and results were not sensible.

Another common normalization approach for LC-MS analysis is by total ion count (TIC). This presumes, that the total sum of the TIC reflects the sample's total concentration. Each feature value would be divided by the sum of the respective TIC so, resulting in all samples having the same sum of total ion count. This should theoretically adjust the concentrations of the different samples. But this representation of a sample's concentration by TIC is not always the case for biological samples like fungal spores. Different environmental factors can induce or suppress the presence of secondary metabolites resulting in different metabolite abundances. One well ionizable metabolite may be expressed intensely leading to a very large signal intensity. The resulting total sum of ion counts is also very large, even if the sample's fungal spore concentration might not be exceptionally large. An example to further visualize the problem is shown in Figure 5.3:

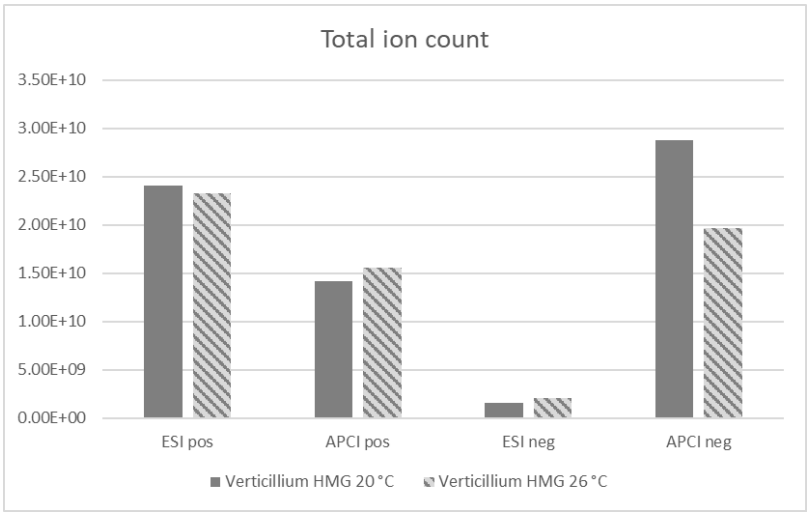


Figure 5.3: Total ion counts of two samples (*Verticillium* on HMG at 20°C and 26°C) with different ionization methods.

Two *Verticillium* samples were measured with each method. Spore number concentrations for the samples were the same in every ionization method, therefore the concentration ratio should be the same. If the TIC would represent the concentration correctly, it should present a very similar TIC ratio for the samples. . However, the TIC of sample 1 (*Verticillium* at 20 °C) with ESI positive ionization is 1.03 times the TIC of sample 2, with APCI positive just 0.91 times, with ESI negative just 0.77 and with APCI negative 1.46 times the TIC of sample 2. This shows that the sum of the total ion count represents the abundancies of ionizable compounds in a sample but not necessarily the sample's actual concentration. If the TIC can't represent the concentration of a sample that was grown at the same time on the same media just at different temperatures, then it is not suitable for normalization. Other sample-wise normalizations like median or mean normalization suffer from the same problem.

Additionally, is the use of markers/ions problematic as ionization efficiencies vary and ion suppression can alter actual concentrations. High analyte signals don't necessarily mean high metabolite concentration, especially in complex biological matrices (Bouguettaya et al., 2015, Wu, Li, 2016).

Feature-wise scaling is necessary for the PCA and subsequent clustering and classifications. Feature-wise normalization ensures, that largely abundant metabolites don't overpower smaller metabolites, especially if those metabolites are of interest. The most common and robust scaling method is z-score scaling. Adaptation of z-score scaling for sparse matrices didn't show an influence, leading to the use of the "classic" z-score scaling (see chapter 2.2.2).

5.2.2. Visualization and dimensionality reduction

Principal component analysis was used for dimensionality reduction. For visualization PCA and t-SNE were tested. The theoretical background is described in chapter 2.2.4.

PCA

As machine learning algorithms are sensitive toward high-dimensional feature spaces usually a dimension reducing step is performed. High resolution mass spectrometric data is very high dimensional with 26,047 features and 89 samples for the ESI positive mode full data set. Those 26,047 dimensions are reduced by PCA to a maximum of 89 dimensions/principal components (PC) which explains the full variance of the data set. By choosing only a subset of PCs the information to differentiate between different classes/fungal species is kept while information explaining noise is discarded.

To determine how many principal components are used as input for further machine learning the explained variance per principal component was evaluated for each dataset. A scree plot that shows the variance explained by the principal components helps to choose the right number of PCs (Figure 5.4). The first PCs don't explain high amounts of variance with the first PC explaining only 5.8 %. This means taking only the first two or three principal components wouldn't explain the data set well.

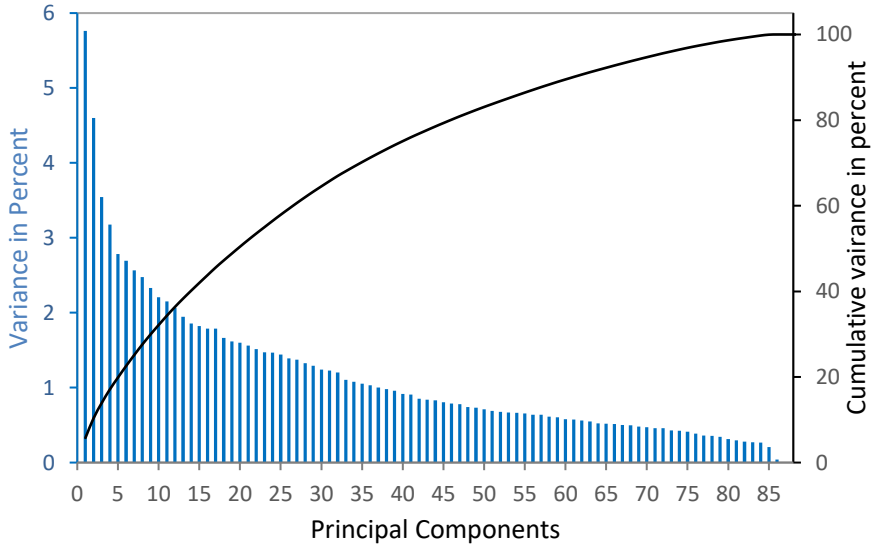


Figure 5.4: Scree plot showing the variance explained by principal components, shown for the full data set B with ESI positive mode ionization. The blue bar chart represents the variances explained by each component and the black line diagram the cumulative variance.

A commonly used threshold is the number of PCs that are needed to explain 80 % of the dataset's variance (Jolliffe, 2002). In the example shown in Figure 5.4, one would need 45 PCs to explain 80 % of the variance. For some machine learning algorithms, this is still too high-dimensional, therefore further method development needs to focus on determining how few principal components can be used as input without losing too much information. This is discussed in the following chapters for each algorithm.

Visualization of the principal components was performed by a scatter plot, showing the sample's distribution as explained by the first two principal components.

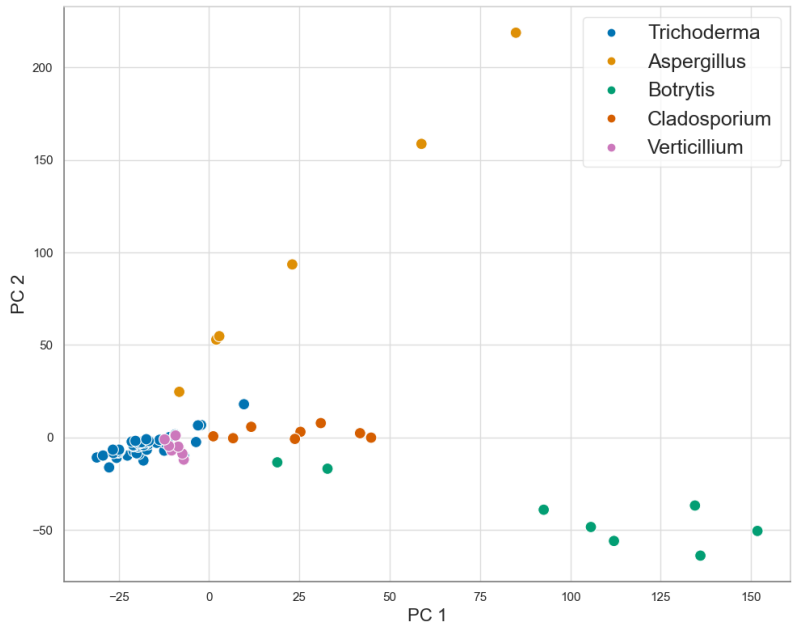


Figure 5.5: Visualization of the data as explained by the first two principal components. Different fungal genera are marked in color.

In Figure 5.5 one can see that samples of different genera are scattered differently. Most *Botrytis* samples are scattered along the x-axis (PC1), meaning PC1 contains information that differentiates *Botrytis* from other genera. *Aspergillus* is spread along the x- and y-axis, meaning PC1 and 2 contain information regarding this genus. But even if data is scattered together like *Trichoderma* and *Verticillium* that doesn't mean that data isn't separable, it just means that the first two principal components don't contain the information needed for separation of classes. To visualize more information a 3D scatter plot also containing the information of the third principal component can help. In the case of ESI negative mode, the samples couldn't be separated well by the first three principal components. More dimensions are needed in cases like this to separate samples. As more than three dimensions can't be visualized other methods like t-SNE are needed, which is explained on

page 77, together with the PCA plot for ESI negative mode in Figure 5.7. PCA plots for ACPI in positive and negative mode are shown in the supporting information Figure 8.1.

Another way to get information by PCA is a loading plot and a biplot. For the theoretical background see page 37. In a biplot the variables are represented as vectors, giving information on which variable gave how much information to which principal component. This can help, for example, to investigate which variables were interesting for the spread of the *Botrytis* samples along with principal component 1. As the feature spaces in the evaluated data set consist of more than 20,000 features, this approach is not feasible. In a loading plot, the features are shown as explained by the principal components. Again, as there are so many features reading a loading plot becomes near impossible. An example of a loading plot for the feature space of ESI positive mode is shown in Figure 5.6.

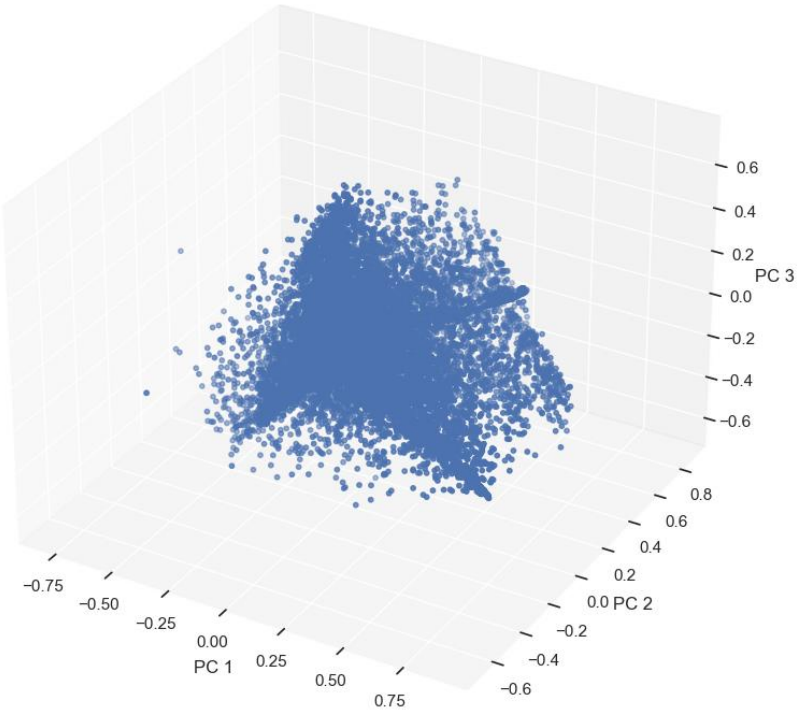


Figure 5.6: Loading plot, all features represented by the first three principal components. ESI positive mode.

The loading plot shows some regions with higher density but overall appears like a ball. This represents the complexity of the samples and the problem, that many of the fungal spore substances might occur in all samples, regardless of genera. Biological samples are highly complex, and features might not always correlate linearly.

In cases when the PCA visualization doesn't seem to show any clusters, another visualization method called t-SNE can be useful. t-SNE is especially helpful if the

relationship between features is complex and non-linear. As visualization with principal components can only show the first three components, it is sometimes not clear if classes are separable, especially if the first components don't explain a lot of the variance. t-SNE can help with that, however as it has many input parameters to control, it can be difficult to get a robust answer. The main parameter is perplexity which should be in the 5 to 50 range (van der Maaten, Hinton, 2008). The perplexity equals the number of neighbors in the learning algorithm and should be bigger with larger datasets. This parameter is critical as a wrong setting can misrepresent the relationship between samples. Choosing the perplexity value with the lowest Kullback-Leibler (KL) divergence can help, as a low KL-divergence indicates that the low dimensionality visualization is similar to the high dimensional distribution (van der Maaten, Hinton, 2008). Other settings like learning rate need to be evaluated as well, as they can influence the results as well. An example where t-SNE is helpful is shown in Figure 5.7. PCA of the data set in ESI negative mode didn't show any clusters whereas t-SNE does. A perplexity of 50 was chosen as it had the lowest KL-divergence.

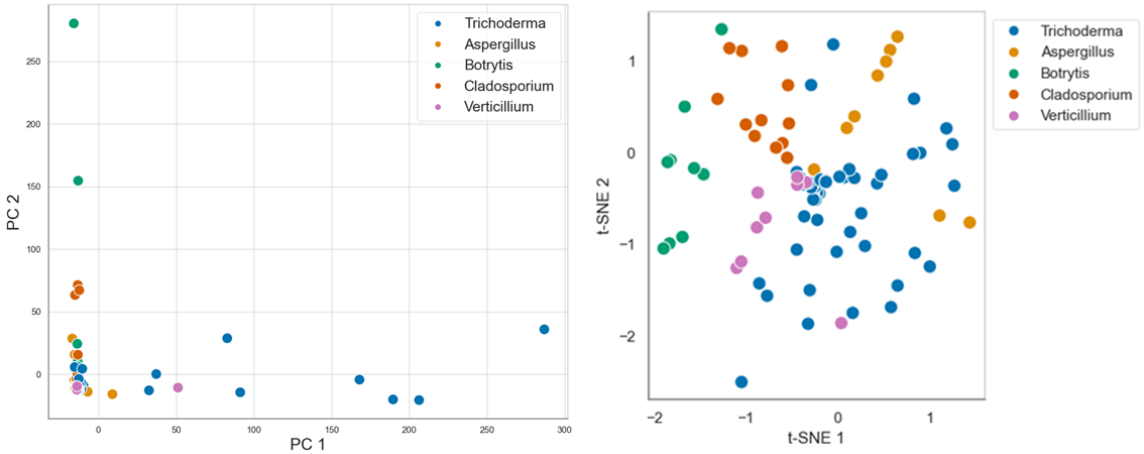


Figure 5.7: PCA (left) and t-SNE (right) for the fungal spore samples measured in negative mode, ESI ionization. Two principal components don't hold enough information needed to separate fungal genera from one another. With t-SNE a two-dimensional visualization is possible.

In the left picture, one can see that no clusters are visualized. T-SNE visualizes the whole dataset in two dimensions and can detect clusters. Therefore t-SNE can help if the data structure is unclear, and the first two or three principal components don't hold enough information.

5.2.3. Unsupervised Machine Learning: Clustering

Unsupervised clustering is helpful if no information about the samples is available, e.g., one doesn't know how many classes are present. Unsupervised clustering doesn't need information from the user nonetheless some parameters need to be optimized. In this work k-means, DBSCAN, and hierarchical clustering analysis (HCA) were used. The theoretical background is described in chapter 2.2.5.

Hierarchical clustering

For distance measurement, the Euclidean distance was used. Other distance measurements like Minkowski or Mahalanobis distance were evaluated but not beneficial for this work. To determine the best linkage method, the cophenetic correlation coefficient was calculated for several linkage methods. The cophenetic correlation factor should be close to one (The MathWorks, 2021a). For theoretical background see Table 2.2 and Formula 7.

Table 5.3 shows results for the clustering of samples and features. Centroid and median linkage produced non-monotonic cluster trees and were therefore not appropriate.

Overall, the average linkage method produced the best results for clustering of the samples with 0.79 being closest to 1. As 0.79 is not a very high factor, the results of the dendrogram representing the samples should be read carefully.

Table 5.3: Cophentic correlation factor for different linkage algorithms.

Linkage	Cophenetic correlation factor for sample space	Cophenetic correlation factor for feature space
average	0.79	0.93
ward	0.47	0.64
single	0.33	0.91
complete	0.60	0.84
weighted	0.70	0.89

The clustering for the features produced higher cophenetic correlation factors, with average linkage being the best as well. As the factor 0.93 is closer to one, it represents the actual relationship between features well. The average linkage methods showed to be the most appropriate for hierarchical clustering of the sample and feature space. This was the case for

all four ionization methods. Results represented by heatmaps and tree diagrams will be shown in chapters 5.3.5 and 5.4.3.

Overall hierarchical clustering showed regions of features that were more intense for certain fungal classes. To further evaluate the features the *van Krevelen* plot was used. As not all features were calculated with a molecular formula and some molecular formulas calculated by MzMine are not reasonable the *van Krevelen* plot won't represent all features.

k-means clustering

Before performing k-means a PCA is used to reduce dimensionality. The k-means results depend on how many PCs are used as an input for the algorithm. To determine the best number of clusters, a so-called elbow plot where the WCSS (Within cluster sum of squares/variance) is plotted against the prospective number of clusters used. The optimum number of clusters is at the point where the steepness of the WCSS plot changes and the WCSS gets minimal. WCSS plots for different example variances are shown in Figure 5.8.

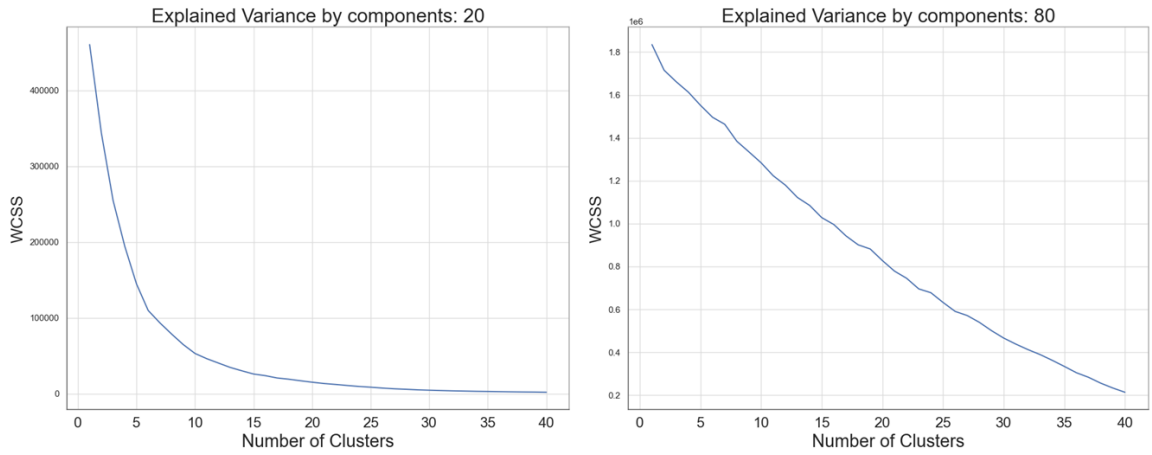


Figure 5.8: WCSS plots for different input of principal components. Left: 5 principal components as input (20 % variance explained). Right: 45 principal components as input (80 % of variance explained). ESI positive mode.

If choosing low dimensional input like in the left picture (5 PCs which explain 20 % of the data) an elbow forms at around 5 clusters. Higher-dimensional inputs don't show a change in steepness; therefore, an optimum number of clusters can't be determined. This is the extreme case on the right side of Figure 5.8, where an input of 45 principal components (80 % variance explained) results in a straight line. This shows that k-means is highly impacted by the curse of dimensionality. To check performance, k-means is evaluated with the different PCs as input. The number of clusters is 5 in all examples (Figure 5.9 and Figure 5.10). Initialization is chosen to be k-means++, which is an improved version of the classic

k-means algorithm (Arthur, Vassilvitskii, 2007). Other parameters like maximum numbers of iterations and the number of times the k-means algorithm will run with different centroids were optimized as well, with relatively high numbers (max_iter = 1000, n_init = 50) to ensure optimum results. For in-depth information on parameters see Pythons sci-kit learn documentation (scikit-learn developers, 2021).

In the following, the k-means results for low-dimensionality input on the one hand and high dimensionality input, on the other hand, are discussed.

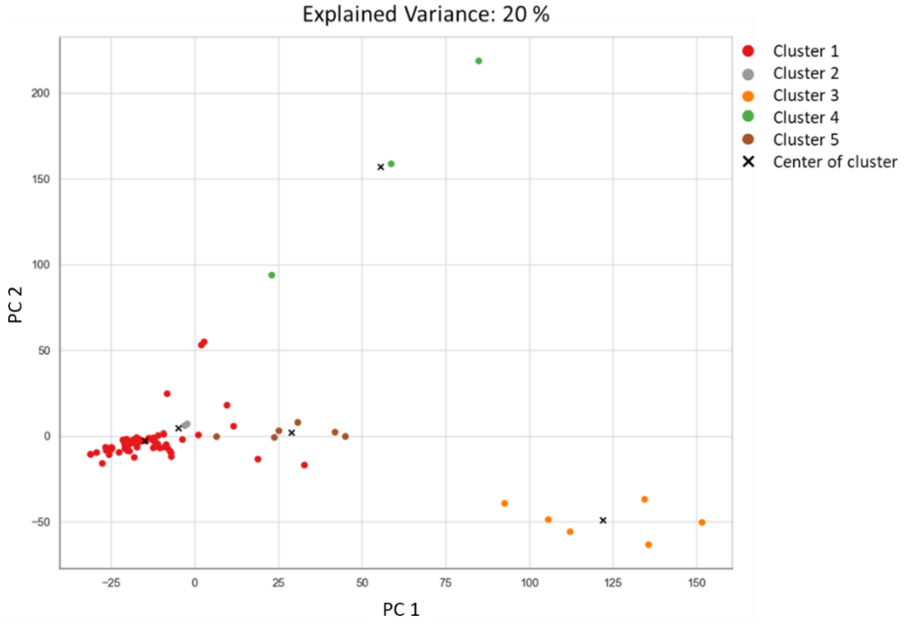


Figure 5.9: k-means clustering for low-dimensional input. 5 PCs, 20 % variance explained. ESI positive mode.

To see the actual class distribution, see Figure 5.5. The plot in Figure 5.9 had 20 % of variance explained as input. All but two *Botrytis* samples form Cluster 3 (orange), most *Cladosporium* samples form cluster 5 (brown), and half of the *Aspergillus* samples form cluster 4 (green). Two samples which are *Trichoderma* were clustered together in the small cluster 2 (grey). The rest of the samples were clustered in the big cluster 1 (red), with no separation between *Trichoderma* and *Verticillium.* The clustering by low-dimensional input resembles the actual distribution but can't differ between *Verticillium* and *Trichoderma.* An input of 9 PCs explaining 30 % of the variance produced the same results. There is possibly not enough information in only 5 or 9 PCs to explain the difference between those two fungal classes.

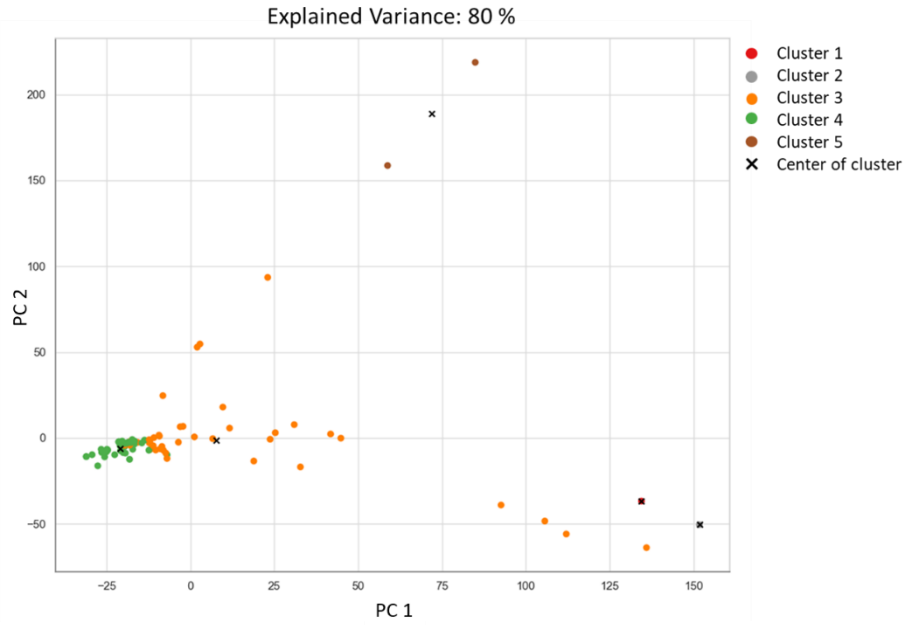


Figure 5.10: k-means clustering for high-dimensional input. 45 PCs, 80 % variance explained. ESI positive mode.

The clustering with 45 PCs explaining 80 % of the variance showed the least sensible results, with two *Botrytis* samples being clustered alone (Cluster 1, red and cluster 2, grey), Two *Aspergillus* samples clustered alone (cluster 5, brown), and the rest of *Cladosporium*; *Verticillium,* and *Trichoderma* being clustered into the orange and green cluster. The green cluster belongs mainly to *Trichoderma* and the orange to *Cladosporium*, *Verticillium,* and the rest of the *Trichoderma* samples. Clustering with 50 % of variance explained also didn't produce sensible results. Also, there two clusters contained only one sample.

Overall k-means clustering apparently can't work well with high dimensional data as present in high-resolution mass spectrometric data. Results are somewhat arbitrary, especially at high dimensions. At lower dimensions results are more meaningful but rely on less information, which might not explain the data set fully, e.g., *Verticillium* and *Trichoderma* weren't separated. The problem is that k-means expects the data to be spherical around the cluster center, which is not true in this work and usually is not true for biological samples in general. A tweak of k-means is k-means with spectral clustering. Spectral clustering consists of two steps including nearest neighbor embedding as the first step and k-means clustering as the last step. It is suitable when the clusters are non-convex. It didn't show better results as *Botrytis*, *Cladosporium,* and *Aspergillus* were clustered together, and *Trichoderma* and *Verticillium* were parted between 4 clusters.

DBSCAN

As k-means results were not very precise, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), another unsupervised clustering method was tested. Core samples with high density are found and expanded from those cores (theoretical background see chapter 2.2.5). The main parameter to be set is epsilon which chooses the maximum distance between two samples to be considered as neighbored. Other parameters like minimum samples are not as crucial but were optimized nonetheless. DBSCAN showed to be very sensitive to the "curse of dimensionality". Input larger than 10 PCs lead to results where more samples were considered noise than an actual sample. Moreover, the samples were put into just one cluster. At high dimensions, distance measurements become artifacts for dimensionality-sensitive algorithms like DBSCAN. For more information on the "curse of dimensionality" see chapter 2.2.4.

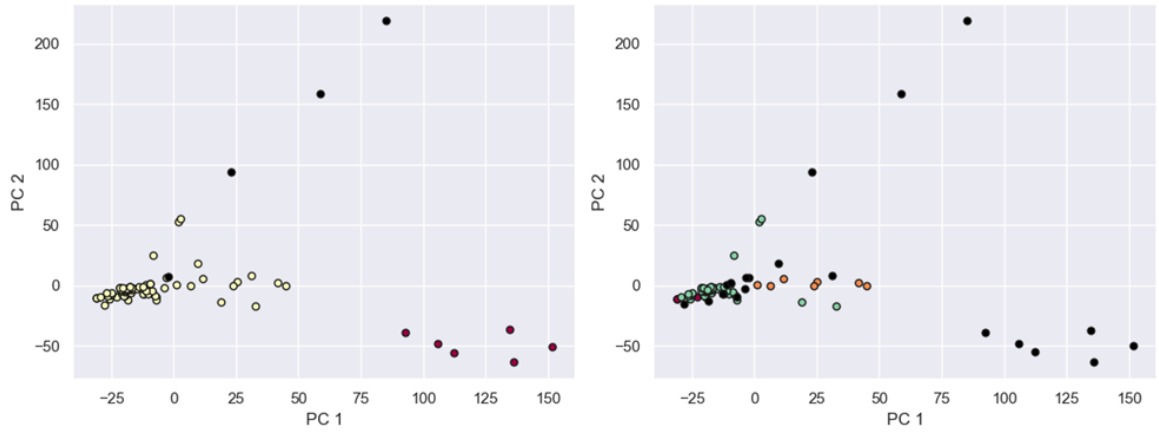


Figure 5.11: DBSCAN for inputs of different dimensionality. Parameters: Epsilon of 40 and minimum sample number of 3. ESI positive mode. Left: PC3 (~ 14 % of variance explained), Right: PC10, (~ 30 % of variance explained).

With a small dimensional input (PC3, 14 % of variance explained) most of the *Botrytis* samples were considered a cluster (red), and *Trichoderma*, *Cladosporium,* and *Verticillium* another (yellow). *Aspergillus* samples were considered noise. This shows that spread-out classes like *Aspergillus* are problematic for DBSCAN as they don't show the density needed to be evaluated as a cluster. At higher dimensionality input (right side of Figure 5.11), *Cladosporium* samples were clustered together (orange), but *Botrytis* and *Aspergillus* were considered noise, as well as some *Trichoderma* and *Verticillium* samples. Optimizing parameters didn't improve the clustering performance. Overall DBSCAN showed worse results than k-means and is not considered a good clustering method for the data sets in this work.

As unsupervised clustering was not able to reliably distinguish between samples correctly, supervised machine learning was used.

5.2.4. Supervised Machine Learning: Classification

For classification two methods, k-nearest neighbor (kNN) and support vector machine (SVM) were tested. The performance of the classification was tested with 10-fold cross-validation. For theoretical background on supervised classification and cross-validation see chapter 2.2.6.

k-nearest neighbor (kNN)

The first algorithm tested was kNN. For kNN the number of principal components needs to be chosen as well as the number of nearest neighbors. Aside from that, kNN doesn't need any input. With kNN, the dimensionality of input should be lower than the number of samples, as kNN is very sensitive to "the curse of dimensionality". Neighbor numbers of 1 and 2 were excluded as regarding only the nearest neighbor of a data point is prone to overfitting. In the example of the full data set measured with ESI in positive mode, the effects of the principal component input and the neighbor number are shown in Figure 5.12.

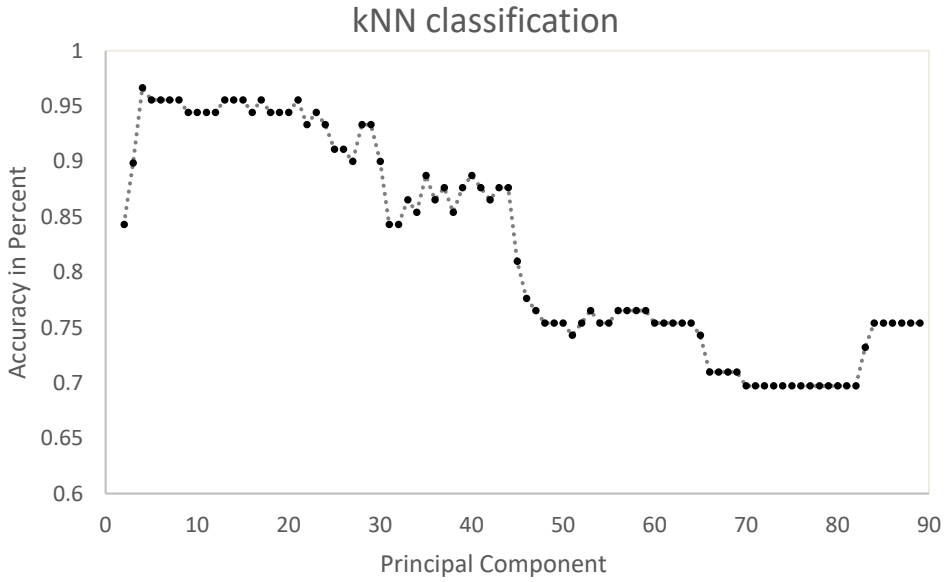


Figure 5.12. kNN accuracy dependence on the dimensionality of input (number of principal components). The graph is shown for PCs 2 to 89. ESI positive mode.

As seen in Figure 5.12 taking the first two or three principal components doesn't explain the data set well enough. Between an input of 4 and 15 principal components, accuracies vary in

the 94 to 96 % range which are very good results. Taking more than 15 principal components results in a loss of accuracy. More dimensions don't increase the classifier's performance but lead to the "curse of dimensionality". The optimal number of principal components varies between ionization methods but is in the range of 10 to 20 principal components for data in this work.

Support Vector Machine (SVM)

For support vector machines the number of dimensions isn't as crucial, as the" kernel trick" takes advantage of high dimensions and isn't as susceptible to the "curse of dimensionality". Especially the linear kernel is very robust and not susceptible to dimensionality or overfitting. Nonetheless, several settings must be chosen with SVMs implementation in sklearn. The kernel, the gamma value, and the error penalty value C see chapter 2.2.6 and sci-kit documentation (scikit-learn developers, 2021). In this work following parameters were tested: Kernel: rbf, linear, polynomial and sigmoid; Gamma value: 1.0, 0.75, 0.5, 0.25, 0.1, 0.01, 0.001, 'auto' and 'scale'; C value: 0.001, 0.01, 0.1, 0.5, 1.0, 10 and 100. The linear kernel is only influenced by the C value, but the other kernels are sensitive to the gamma value. High gamma and low C values are prone to overfitting, as they won't allow misclassification of the training data.

In the following figure, an example for gamma = "auto" and C = 0.01 is shown. Scale gamma value showed similar results, as well as gamma values of 1.

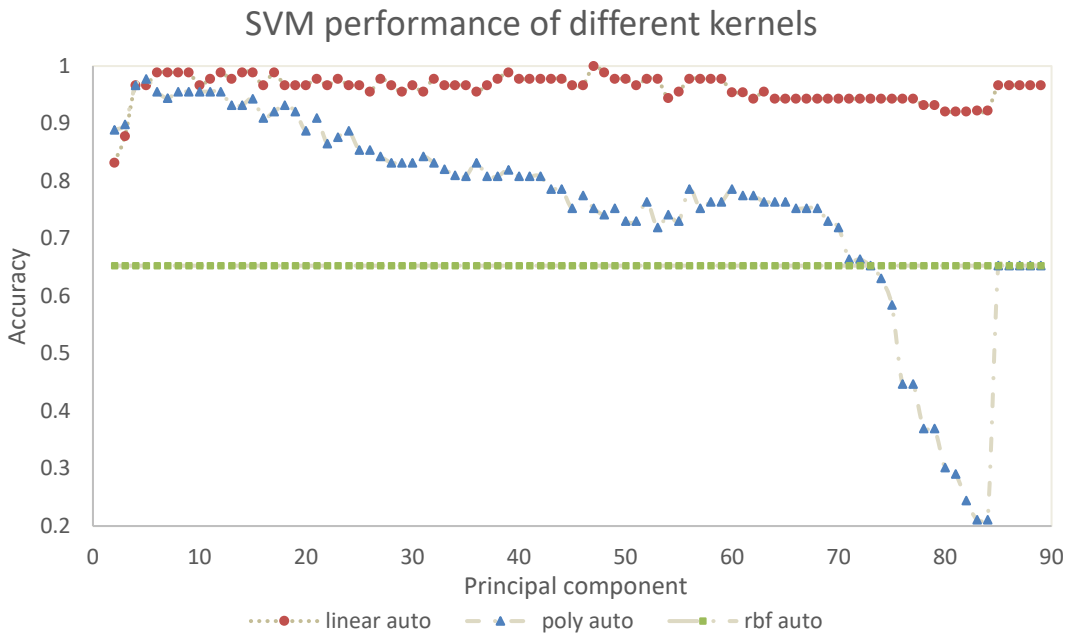


Figure 5.13: SVM accuracy dependence on the dimensionality of input (number of principal components) for three different kernels: linear, polynomial order three, and rbf. Gamma: auto, C = 0.01. ESI positive mode.

The linear kernel shows the best results with accuracies well over 90 %, whereas polynomial works well for lower PCs but worse for high. The rbf kernel doesn't work well, as the C value of 0.01 in this example is set too low. With C values of 1 accuracy values of 93 % are reached for low PC inputs for the rbf kernel. Nonetheless, the linear kernel gives the best results. RBF and polynomial kernel are more easily influenced by noise.

The high accuracy of the linear kernel shows that the classes are linearly separable and that the linear kernel should be chosen for a robust approach. The C value doesn't influence results significantly, indicating that a large margin between classes can be maintained even when misclassification of samples is prevented. The optimization approach was performed for all datasets and all ionization types, but overall, the linear kernel performed best. For all future SVM classifications, the linear kernel was chosen with a C value of 0.01.

Stratified vs. unstratified cross-validation

Not all classes are represented uniformly in the sample set, .g., the class *Trichoderma* is represented by more samples than the other classes. Some classifications have a bias, giving too much weight during training to the most represented class. Stratification of the cross-validation checks that all classes are uniformly distributed between train/test sets in each

fold. The data was shuffled before each train/test split, for both stratified and unstratified cross-validation. Results are shown in Figure 5.14.

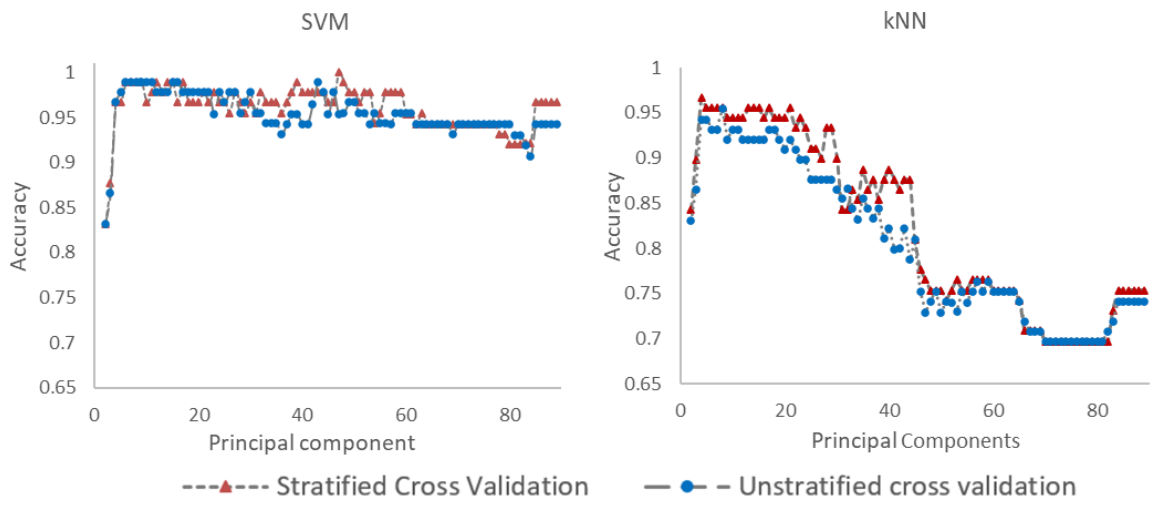


Figure 5.14: Comparison of stratified (red) and unstratified (blue) 10-fold cross-validation for SVM (left) and kNN (right), dataset B, ionization: ESI positive mode.

The results from the stratified k-fold cross validation were only 1 or 2 % points lower than the unstratified, showing that the classifier could robustly separate classes from one another. It also shows that different abundances of the different classes don't influence the classifiers' performance a lot, especially not with SVM classification. Results for all ionization methods can be found in the attachment. Future results will be shown for stratified k-fold cross-validation.

Comparison of datasets

The classification was evaluated with a smaller and a bigger dataset that include samples with lower respective higher possible phenotypical plasticity. Results can indicate if the classifiers react sensibly toward samples from different laboratories, seasons, or measurement periods. This is important as the algorithm's performance should correctly differentiate between different species/genera even if the samples' phenotypes show plasticity. Additionally, some samples were extracted with solvents from a different manufacturer and measured before the orbitrap mass spectrometer was subjected to several maintenance procedures and changes of MS parts. For both datasets, the classifiers were trained and tested by stratified 10-fold-cross validation.

Dataset A is the smaller dataset. It includes only samples of the season Fall/Winter 2021 which have less possible phenotypical plasticity. Dataset B includes more, additional

samples compared to Dataset A, which could also include higher phenotypical plasticity. Dataset B includes *Trichoderma harzianum* strain A and B and *Trichoderma atroviride* samples from 2019/2020 and spring 2021. Some samples (2019/2020) were cultivated partly in another laboratory building, at other seasons, and therefore other humidity. Samples were stored in a chilled culture, therefore dormant between cultivation, which can induce and change in metabolism and therefore can introduce more variation into the data set B. The number of biological replicates for both datasets can be seen in Table 5.4.

Table 5.4: Overview of biological replicates of fungal spore samples.

Genera/Species/ Strain	**Dataset A: Biological replicates from Fall/Winter 2021**	**Dataset B: Additional biological replicates from 2020 and spring 2021**
Aspergillus versicolor	8	Equivalent to dataset A
Botrytis cinerea	8	Equivalent to dataset A
Cladosporium cladosporioides	8	Equivalent to dataset A
Verticillium dahliae	9	Equivalent to dataset A
Trichoderma longibrachiatum	4	Equivalent to dataset A
Trichoderma fasciculatum	4	Equivalent to dataset A
Trichoderma minutisporum	4	Equivalent to dataset A
Trichoderma harzianum strain B	4	Samples from dataset A plus two samples from 2020 and two samples from spring 2021
Trichoderma harzianum strain A	4	Samples from dataset A plus one from 2020 and four from spring 2021
Trichoderma atroviride	4	Samples from dataset A plus five from 2020 and four from spring 2021
Total	57	$57 + 4 + 5 + 9 = 75$

Additionally to biological replicates some samples were measured several times, introducing technical replicates into the data set, where instrument performance and user performance (e.g., dilution of samples) could contribute to the variation of the samples. These samples were treated as technical replicates and included in the datasets.

An overview of biological and technical replicates per ionization method for dataset B is given in Table 5.5. Dataset A has the same numbers of technical replicates except for *Trichoderma harzianum* strain A and B and *Trichoderma atroviride*, where the additional biological replicates were measured.

Table 5.5: Overview of available biological and technical replicates for each ionization method.

Genera/Species/ Strain	Biological + technical replicates for dataset B			
	Ionization: ESI positive	Ionization: APCI positive	Ionization: ESI negative	Ionization: APCI negative
Aspergillus versicolor	12	13	9	6
Botrytis cinerea	8	8	11	8
Cladosporium cladosporioides	8	9	9	9
Vertilillium dahliae	8	9	9	9
Trichoderma longibrachiatum	4	4	3	4
Trichoderma fasciculatum	5	4	6	4
Trichoderma minutisporum	5	4	6	4
Trichoderma harzianum strain B	13	15	14	13
Trichoderma harzianum strain A	12	13	13	16
Trichoderma atroviride	14	17	15	15
Trichoderma total	58	57	57	56
Total	89	96	95	88

Additionally, to exclude bias from the operator and exclude overfitting seven samples were labeled by a colleague and measured. As these samples originate from the same spore harvest as the others and were just aliquoted at the spore stage, they represent technical replicates. Results will be shown in 6.1.2.

Results of stratified 10-fold cross-validation for datasets A and B are shown in Table 5.6. The number of principal components used as input were evaluated during method development and are available in the supporting information see Table 8.6 to Table 8.9.

Table 5.6: Results of stratified 10-fold cross-validation for both datasets. Accuracies are shown with the corresponding standard deviations in brackets.

Classifier	kNN classification accuracy		SVM classification accuracy	
Dataset	Dataset B	Dataset A	Dataset B	Dataset A
ESI positive	0.96 (0.07)	0.94 (0.09)	0.99 (0.03)	0.98 (0.06)
APCI positive	0.94 (0.08)	0.92 (0.08)	0.99 (0.03)	0.94 (0.07)
ESI negative	0.93 (0.07)	0.91 (0.11)	0.94 (0.08)	0.96 (0.07)
APCI negative	0.97 (0.05)	0.97 (0.07)	0.98 (0.04)	0.97 (0.07)

For all ionization methods, accuracies are very high for dataset A but even higher for dataset B. There is only one exception, in ESI negative mode the kNN classifier performs slightly better with dataset A. But in general, these results show that samples with more variability don't decrease the classifiers' performance, but rather increase it. This is probably due to more training instances. That means that the classifiers work robustly and indicates that the introduction of more samples with more variances doesn't lead to problems, but rather gives the classifier more information to work with.

Choosing the classifier

Overall, both classifiers performed very well, regardless of the dataset or if cross-validation was performed stratified or unstratified. Finally, SVM is the better classifier for genus differentiation. SVM reacts more stable with inputs of different dimensionality and showed to be robust with the input of higher variability. This was also the case for species differentiation, as not only genera were differentiated, but also species of the *Trichoderma* genus. Parameters for species differentiation were cross-validated and are available in the supporting information. The results of species differentiation with optimized principal component input are shown in the following table.

Table 5.7: Accuracy results for *Trichoderma* species differentiation with kNN and SVM.

Ionization modes	kNN classification accuracy	SVM classification accuracy
ESI positive	0.95 (0.08)	0.98 (0.05)
APCI positive	0.78 (0.13)	0.9 (0.13)
ESI negative	0.82 (0.12)	0.89 (0.09)
APCI negative	0.87 (0.13)	0.97 (0.07)

For species differentiation, SVM showed clearly better accuracies, especially with APCI ionization. This might be due to the fact, that for species differentiation some species (*T. longibrachiatum, minutisporum, and fasciculatum*) are only presented by a few samples and kNN would need further training data. In total, SVM shows a better performance for species differentiation.

Results of method development

Fungal spore growth was performed on PDA and HMG media at two different temperatures, one at day-night rhythm, and the other one in darkness. Additional *Trichoderma* samples increase the variability of data. Extraction is performed with methanol which showed the overall most extracted compounds, extracted ergosterol very well and was easier to handle than methanol: water. Measurement was performed on a C18 column with methanol as an organic eluent. As it couldn't be determined which ionization method (ESI/APCI positive/negative mode) was the most performant the sample set was measured with all four. Dimensionality reduction will be performed with PCA, and additional visualization with t-SNE can be performed if necessary. Unsupervised clustering allows a first glance at the data. Hierarchical clustering enables visualization of features that are more intense for certain classes, allowing further data evaluation with *van Krevelen* plot, etc. HCA results will be discussed in the following results chapters. Overall supervised machine learning enables the classification of fungal classes and species at very high accuracies, with support vector machine with the linear kernel showing the most promising results. The linear kernel is not prone to overfitting as only the C value can influence the fit. Different C values didn't show an influence on the accuracy results making the linear kernel for this sample set a very robust choice. Parameters and PC input will be chosen as optimized by 10-fold stratified cross-validation. Dataset B will be chosen to present the results for genus differentiation as it includes a higher variance of samples. A validation data set for genus respective class differentiation will be shown in the results. Species differentiation will be performed with all available biological and technical replicates of *Trichoderma* samples.

6. Results and Discussion

6.1. Differentiation between fungal classes and families

Hereafter the results of the differentiation of fungal classes based on the fungal spores are presented. The sample set consists of 5 different taxonomic families from 4 different classes (see Table 4.1). The data set consists of 75 biological replicates and, including technical replicates 90 to 95 samples, depending on the ionization method, see Table 5.5. During method development, methods to achieve distinction by non-target liquid chromatography high-resolution mass spectrometry analysis for the given samples were evaluated. Supervised classification algorithms, especially SVM were found to produce the best results. Additionally, fungal species of the genus *Trichoderma* were differentiated (see chapter 6.2). Furthermore, it was evaluated if certain compounds or fingerprints could be found which were specific for a certain species.

6.1.1. Comparison of SVM results for different ionization methods

Support vector machine was found to be the more robust classification method and showed higher classification accuracies. The mean accuracies over 10-fold stratified cross-validation are shown in Table 6.1. The standard deviation represents the bandwidth of the accuracy for each method during the 10-fold cross-validation.

Table 6.1: Mean Accuracy for classification of fungal genera by support vector machine classification. Accuracy and standard deviation were calculated from 10-fold stratified cross-validation.

Ionization method	Mean accuracy (Standard deviation)
ESI positive mode	0.99 (0.03)
APCI positive mode	0.99 (0.03)
ESI negative mode	0.94 (0.08)
APCI negative mode	0.98 (0.04)

All methods showed a mean classification accuracy of over 90 % after 10-fold cross-validation, giving excellent results. The mean accuracies are accompanied by very low standard deviations, meaning each repeated classification resulted in accuracies in or over the 90 % range. As the train/test split was chosen randomly, the results are independent of which samples were chosen for training or testing of the algorithm.

In positive mode, accuracies were the highest with very low standard deviations of 3 %. In the negative mode, APCI showed higher results than ESI. Also, with ESI in negative mode, a higher number of principal components was needed to achieve the classification. The optimal number of principal components was evaluated during cross-validation. APCI/ESI in positive mode and APCI in negative mode only needed the input of 15 to 18 principal components explaining 50 to 59 % of the data variance. ESI in negative mode needed 30 PCs, explaining 75 % of the variance as input. This indicates that the features detected by ESI in negative mode don't explain the differences between different fungal classes as well.

The confusion matrices allow a closer look at the classification:

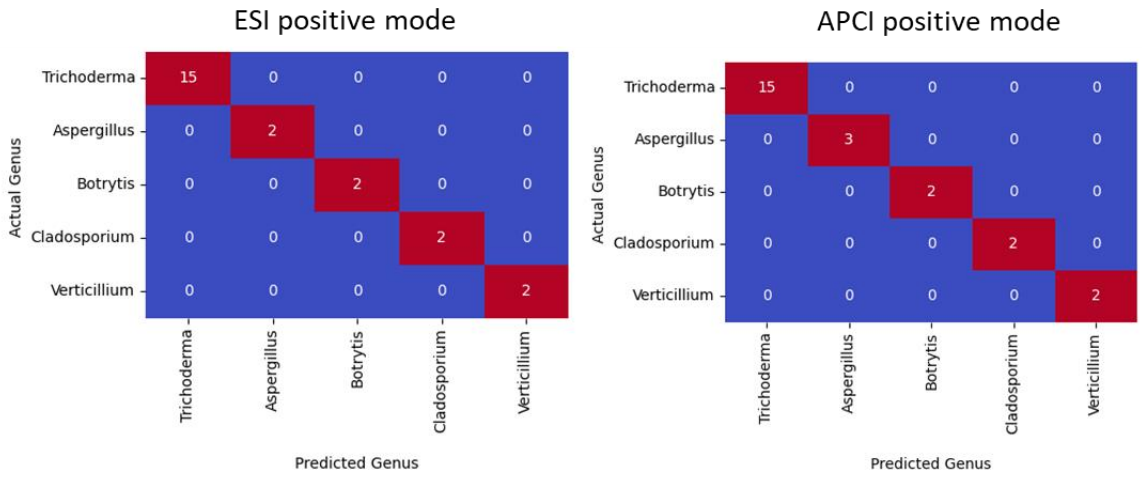


Figure 6.1: Confusion matrices for the differentiation of fungal spore genera by SVM. Left: Ionization by ESI in positive mode. Right: Ionization by APCI in positive mode.

In positive mode, all samples were classified correctly. The classification was performed stratified with 80 % of data per class used for training and 20 % for testing. This indicates that even the small training sample size provided enough information to classify all test data correctly.

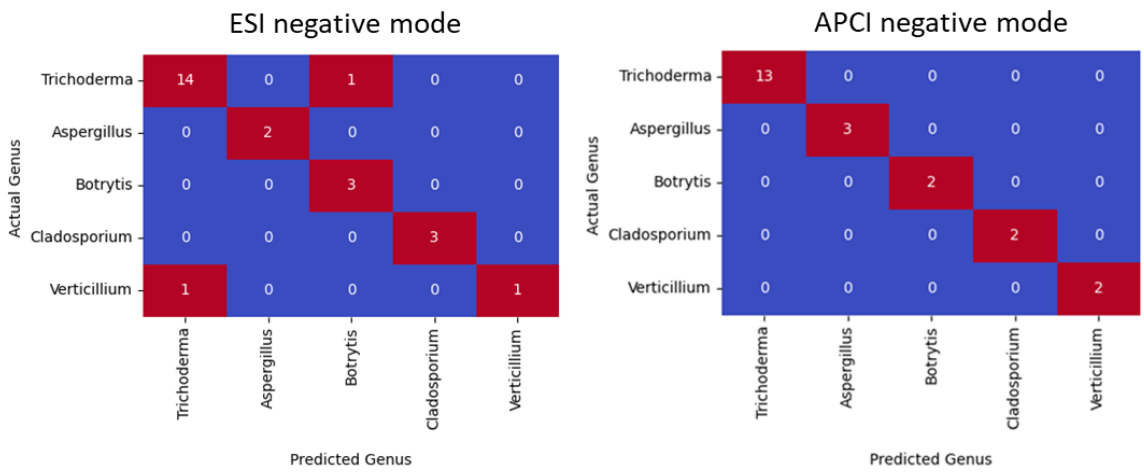


Figure 6.2: Confusion matrices for the differentiation of fungal spore genera by SVM. Left: Ionization by ESI in negative mode. Right: Ionization by APCI in negative mode.

In negative mode APCI showed a correct classification of all samples, whereas with ESI two samples were misclassified, representing an accuracy of 92 %. One *Trichoderma* sample was misclassified as *Botrytis* and one *Verticillium* sample was misclassified as *Trichoderma*. This points to the hypothesis, that ESI negative mode might not contain enough information to separate classes in all cases. Nonetheless, results over 90 % are very promising, especially as it is suggested that more training data might even improve the classifiers' accuracy (see also method development page 89).

6.1.2. Validation of classification

To test the performance of the classifier, six unknown samples for APCI positive/negative and ESI negative mode, respective seven samples for ESI positive mode were tested. The classifier has never been trained with those samples, to that the decision regions formed by the classifier weren't formed with information from the samples themselves. The samples are technical replicates, meaning they are genetically the same and presumably phenotypically very similar to samples the algorithm was trained with. Samples were relabelled by a colleague after harvest, extraction, and dilution so that the user couldn't introduce any bias in the training/test process of the classifier. Parameters for the SVM were chosen according to the parameters evaluated in method development (chapter 5.2.4).

Table 6.2: Results for the classification of validation samples. Classification by SVM with parameters evaluated for the training samples.

Ionization method:	**ESI positive**	**APCI positive**	**ESI negative**	**APCI negative**
Actual Species	**Predicted species by classification**			
Aspergillus versicolor	*Aspergillus*	*Aspergillus*	*Aspergillus*	*Aspergillus*
Trichderma atroviride	*Trichoderma*	*Trichoderma*	*Trichoderma*	*Trichoderma*
Cladosporium cladosporiodes	*Cladosporium*	n/a	n/a	n/a
Trichderma minutisporium	*Trichoderma*	*Trichoderma*	*Trichoderma*	*Trichoderma*
Trichoderma longibrachiatum	*Trichoderma*	*Trichoderma*	*Trichoderma*	*Trichoderma*
Trichoderma harzianum strain A	*Trichoderma*	*Trichoderma*	*Trichoderma*	*Trichoderma*
Verticillium dahliae	*Verticillium*	*Verticillium*	*Verticillium*	*Verticillium*

All ionization methods were able to classify all technical replicates correctly. The *Cladosporium* sample was only available for ESI positive mode measurements. It should be noted that the classification outcome is sensitive to the number of principal components used as input. The results were produced with the principal component number evaluated during the cross-validation but also tested with other principal component inputs. Low PC input that explains 30 % of the data's variance or less didn't produce sensible results. They probably don't hold enough information to differentiate accurately between classes. It indicates, that for unknown samples enough of the data needs to be explained, to enable the classification of unknowns, in this case at least 50 % of the data's variance. It also emphasizes that training with cross-validation is crucial, to prevent under- or overfitting. The number of PCs should be included in the cross-validation. If in question, a higher number of PCs should be chosen, taking advantage of the fact that SVM with a linear kernel is less susceptible to the curse of dimensionality.

Overall classification results are very high, also in the case of unknown samples, the classifier hasn't been trained before. It should be noted that the unknown samples were technical replicates, originating from the same culture as samples that the algorithm has been trained with. That means, that the algorithm was trained with very similar data. Nevertheless, it indicates that fungal spore samples of different classes are linearly

separable. Also, *Trichoderma* samples of different species were classified as *Trichoderma* spp. This indicates that distances within species of the same genus are smaller than distances between different fungal families' classes, despite possible high inter-species diversity. This should be further examined in future projects with more samples.

6.1.3. Genus's differentiation in mixed-species samples

Additionally, it was tested if the SVM could differentiate between mixed samples on a small data set. In actual environmental samples, there will always be a mixture of fungal species present. Mixed samples were created by mixing single species extracts. Samples consist of the following mixtures:

Table 6.3: Composition of mixed samples for classification testing. Mixed from their diluted single-species extracts.

Sample number	Mixture
1	80 % *Aspergillus versicolor* 20 % *Trichoderma harzianum strain A*
2	50 % *Aspergillus versicolor* 50 % *Trichoderma harzianum strain A*
3	20 % *Aspergillus versicolor* 80 % *Trichoderma harzianum strain A*
4	33.3 % *Verticillium* 33,3 % *Trichoderma harzianum strain A* 33.3 % *Trichoderma fasciculatum*

The algorithm is not trained with mixed samples; therefore, the best parameters and the optimum number of principal components couldn't be determined and parameters for single species samples were used. The classification was performed with the SVM and kNN and PC input was chosen that 50 and 80 and 100 % of variance were explained. Low variance input for kNN and high variance input for SVM showed the best results. The best results mean in this case, that samples were classified according to the species with a higher proportion. The classification of evenly mixed samples is not possible with this method that only allows one label for each sample.

Table 6.4 Classification results for mixed samples. The classification was performed with SVM at a PC input that explained the full data set. Correctly predicted samples are marked in green. As sample 2 consists of equal parts of two samples the prediction wasn't coloured.

Sample	ESI positive mode	APCI positive mode	ESI negative mode	APCI negative mode
1	*Aspergillus*	*Aspergillus*	*Aspergillus*	*Aspergillus*
2	*Trichoderma*	*Aspergillus*	*Trichoderma*	*Trichoderma*
3	*Trichoderma*	*Aspergillus*	*Trichoderma*	*Trichoderma*
4	*Trichoderma*	*Trichoderma*	*Trichoderma*	*Trichoderma*

With SVM and principal component input of 100 % variance explained, almost all samples were classified according to the species with the highest proportion. As sample 2 is a 50:50 mixture the classification as either *Aspergillus* or *Trichoderma* won't be rated.

With PC input that explains 80 % variance, input results were worse with only one-third of the samples being classified correctly. With all ionization types sample 4 was classified as *Verticillium*, the lower proportional part. With 50 % variance explained as input results were approximately the same as with 80 % variance input, with one major misclassification: In APCI negative mode sample 3 was classified as *Cladosporium*, which is not present in the sample. This means, that the decision regions formed by the hyperplanes in SVM classification at low information input aren't very suitable to classify mixed samples. It should be examined if the training of the algorithm with mixed-species samples would increase the classifications' performance. Classification by kNN produced the results presented in Figure 6.5.

Table 6.5: Classification results for mixed samples. The classification was performed with kNN at a PC input that explained 50 % of the data set's variance. Correctly predicted samples are marked in green. As sample 2 consists of equal parts of two samples the prediction wasn't colored.

Sample	ESI positive mode	APCI positive mode	ESI negative mode	APCI negative mode
1	*Aspergillus*	*Aspergillus*	*Aspergillus*	*Aspergillus*
2	*Aspergillus*	*Aspergillus*	*Aspergillus*	*Aspergillus*
3	*Trichoderma*	*Aspergillus*	*Trichoderma*	*Trichoderma*
4	*Trichoderma*	*Verticillium*	*Trichoderma*	*Trichoderma*

Almost all samples are classified correctly if using 50 % of the variance. With 80 % variance results are just slightly less accurate, with all of sample 4 being classified as *Verticillium*. With 100 % of variance explained the results were acceptable, except for APCI positive mode. Sample 4 was classified as *Aspergillus* with no *Aspergillus* being present in the sample. Overall kNN showed good results but showed some signs of the "curse of dimensionality" with higher dimensionality data. With mixed samples, the principle of kNN using the *n* nearest neighbors for classification might be beneficial. With SVM, samples need to be within the decision region, but mixed samples might be on the border of decision regions. With kNN samples classification works according to the closest neighbors which might represent the composition of a mixed sample more reliably.

Overall, it is surprising that the algorithm classified samples correctly if it wasn't trained with mixed samples. The sample set was very small but can hint toward the application of machine learning algorithms to find the major fungal class in a sample. This would be preferable with filter samples of air, especially when one fungus dominates the sample. Examples of a fungal spore species dominating the aerosol would be plant pathogens infecting a field. When looking at environmental samples it should be considered that the background/noise might be interfering. Many other substances are present in the air, including other biological or anthropogenic aerosols. Those compounds would probably need to be filtered or excluded beforehand, e.g., by using artificially made anthropogenic aerosol filters as blank.

6.1.4. Evaluation of feature space

Despite results being too complex to analyze by the human eye, the supervised machine learning algorithms find structures in the feature space that enables differentiation between samples of different genera, respective classes. The ionization method doesn't seem to influence the performance of the algorithm extensively, with ESI negative ionization showing slightly lower accuracies than the other ionization methods. To get further information if different ionization methods focus on different biomolecule groups the respective feature spaces were plotted in *van Krevelen* plots.

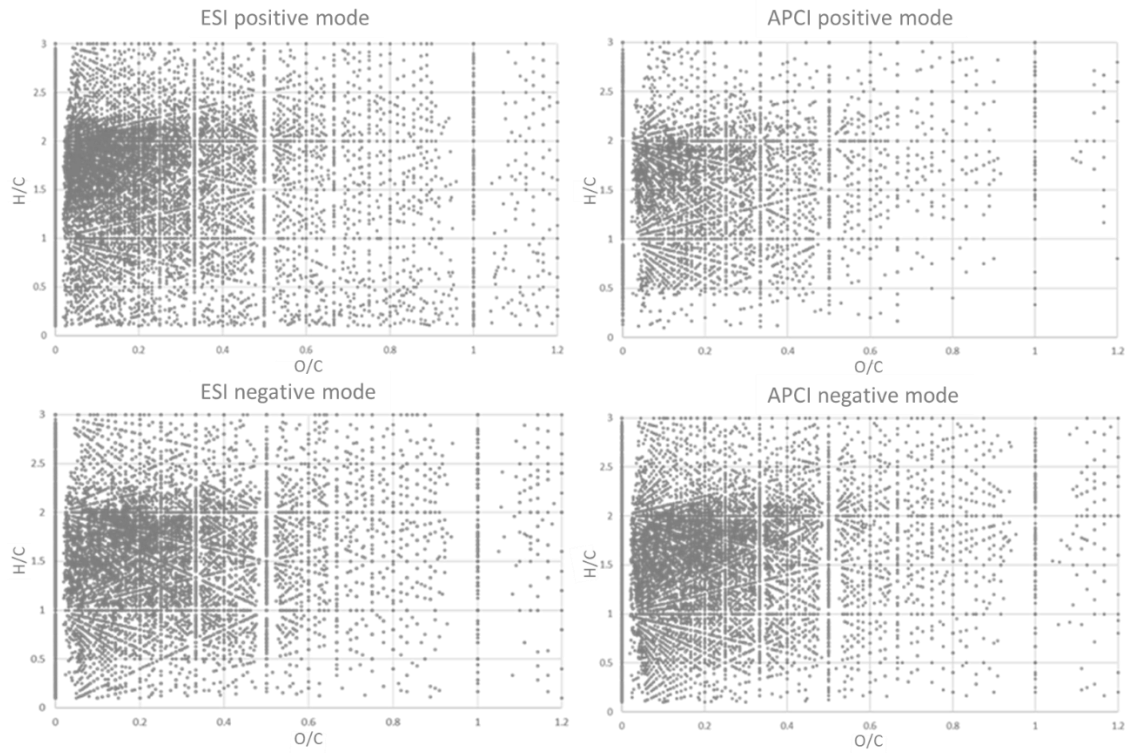


Figure 6.3: *Van Krevelen* plots for all compounds detected in all fungal spore samples by the different ionization methods. Upper left: ESI positive mode. Upper right: APCI positive mode. Lower left: ESI negative mode. Lower right: APCI negative mode.

Figure 5.17 shows the complexity of the feature space. All biomolecule groups are represented according to their O/C and H/C ratio. The largest number of compounds was detected by ESI in positive mode, with 26,047 features spread over ~90 samples, followed by APCI negative mode with 22,383 features. ESI ionization in negative mode produced 18,450 features and APCI in positive mode the least with 11,595 compounds. The *van Krevelen* plots look very similar for all four ionization methods. ESI positive mode and APCI negative mode show higher densities, but this is probably a result of the larger feature space. As the *van Krevelen* plots don't give insights into the feature space as they are too complex, hierarchical clustering analysis is performed.

6.1.5. Hierarchical clustering of features

Hierarchical clustering analysis is used to cluster feature- and sample-wise, to get information if certain features are "responsible" for the distinction between different genera. Heatmaps with large samples and feature cases like in this example get very complex. The samples are represented by the horizontal tree diagram and the features by the vertical.

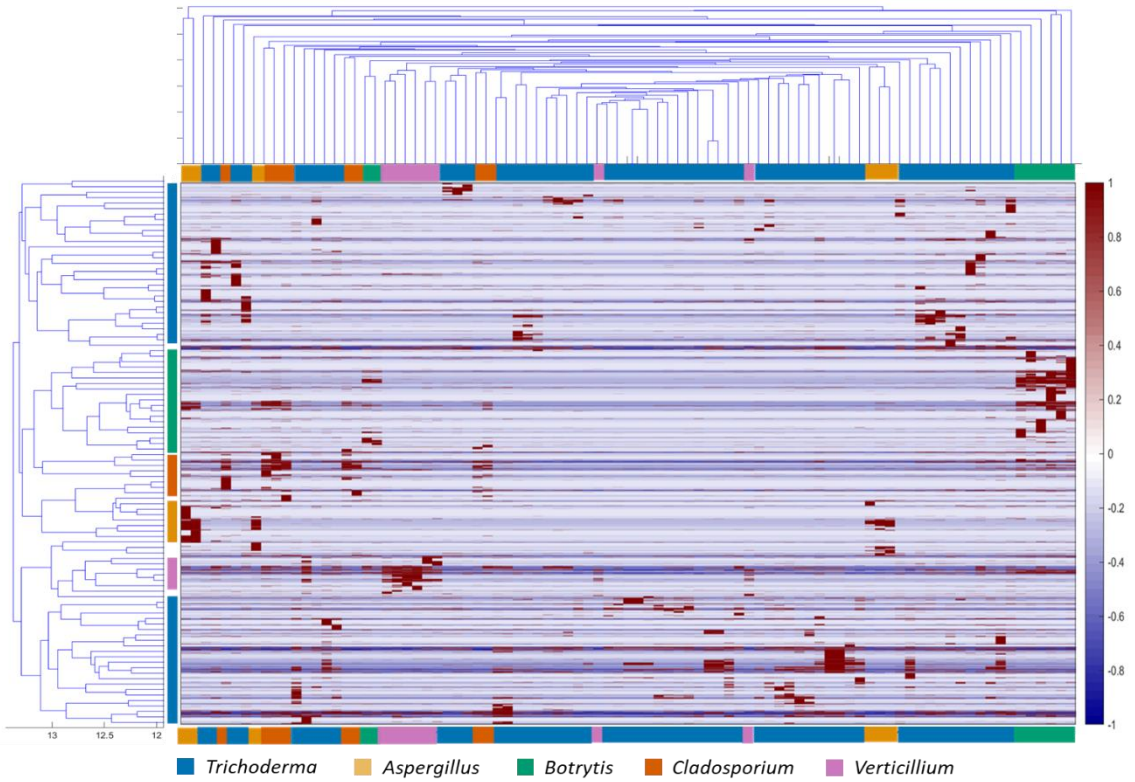


Figure 6.4: HCA of fungal spore samples ionized by ESI positive mode. The legend shows the color coding for the different fungal genera. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

Samples of the same class weren't necessarily clustered together, as shown by the color-coded horizontal bars in the figure. This was expected, as the cophenetic correlation coefficient for sample-wise clustering indicated (see Table 5.3). However, feature-wise classification showed a cophenetic correlation factor close to one, meaning representation by the vertical cluster tree is close to actual distances between features. By looking closely at the heatmap one can see, that certain areas of the heatmap are dark red, indicating a high intensity of features. Some of these features are intense only for samples of a certain fungal class. The vertical color-coded bar shows the areas where features occur more intensely for the respective class. These features could be specific to the fungal class.

In the following also the HCA results for the other ionization methods are shown:

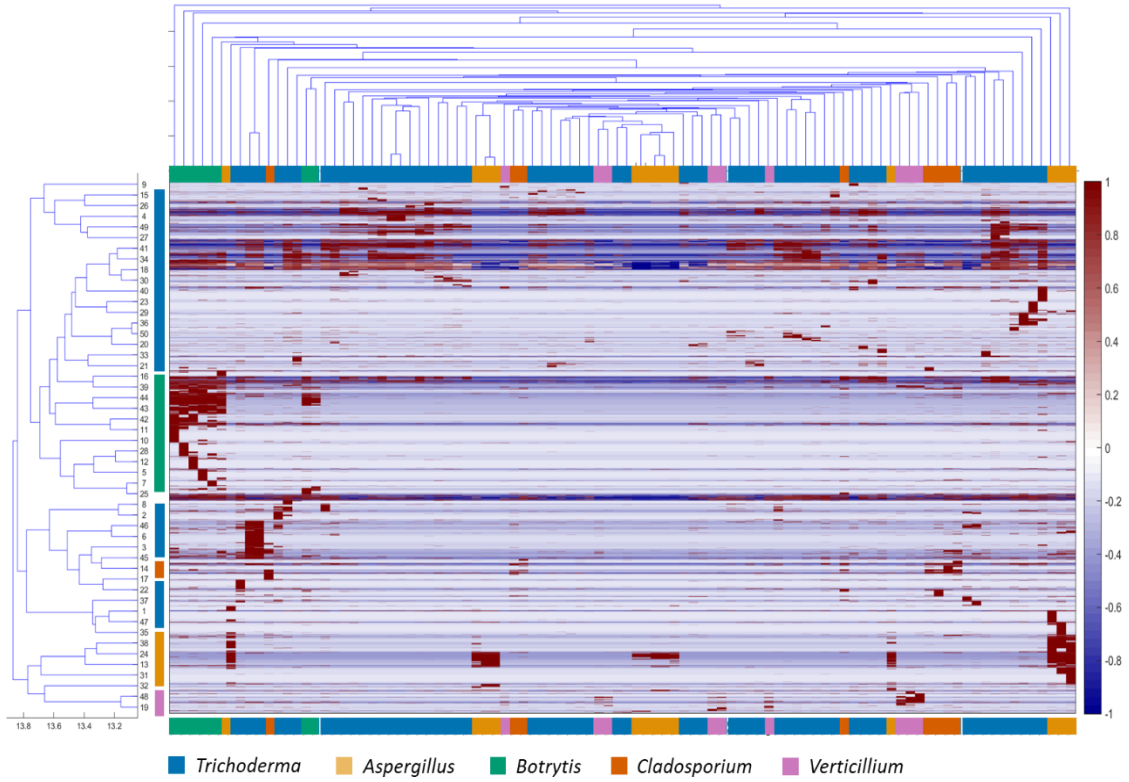


Figure 6.5: HCA of fungal spore samples ionized by APCI positive mode. The legend shows the color coding for the different fungal genera. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

HCA results of APCI positive mode ionization are very similar to ESI positive mode. Samples weren't clustered together, but features show some order according to the fungal genera. Features specific to *Cladosporium* are smaller and less intense when compared to ESI positive mode ionization.

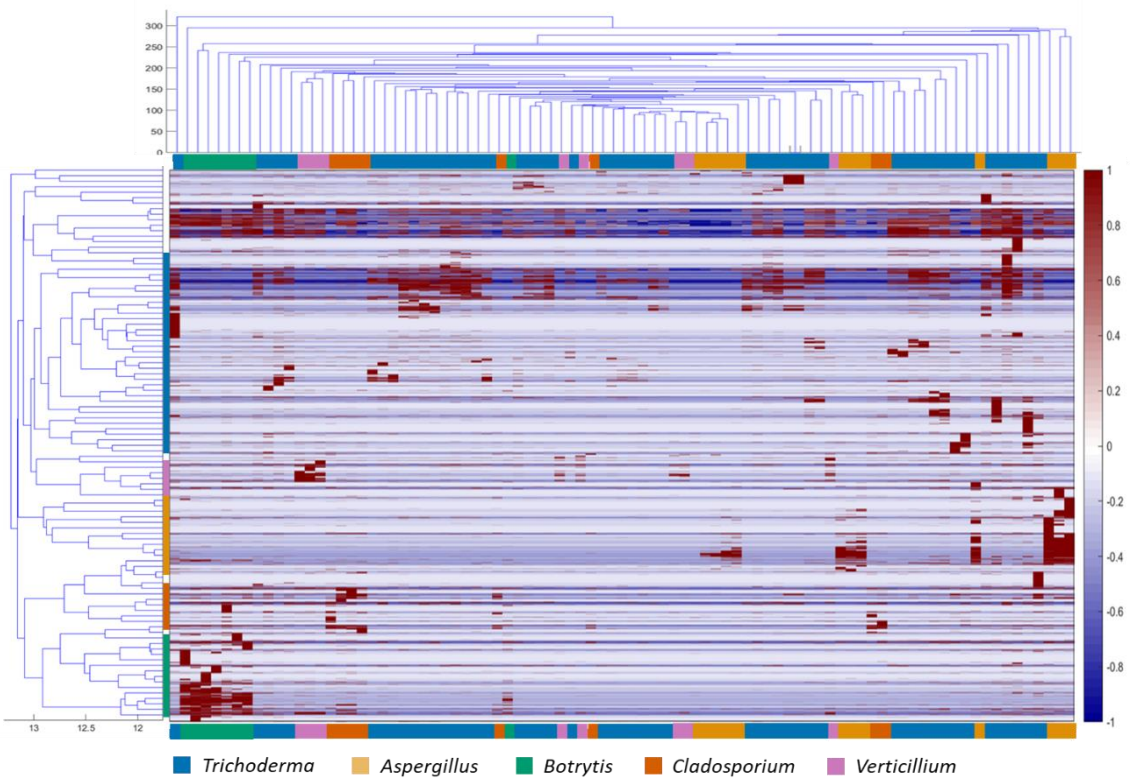


Figure 6.6: HCA of fungal spore samples ionized by APCI negative mode. The legend shows the color coding for the different fungal genera. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

With APCI negative mode ionization results show again, that samples aren't clustered together, but feature regions that are more intense for samples of a specific genus. The upper dark red/dark blue regions represent features that are present in some samples but are absent in others. The presence or absence of those features can't be correlated to environmental conditions like growth media or temperature. As the features are not present in all samples they presumably aren't originating from the primary metabolism. As they are present in samples of different genera it indicates that fungi of different genera produce the same secondary metabolites even if they are not very closely related. This is known from secondary metabolites like melanin which can be present in fungi of different classes (Calvo et al., 2002).

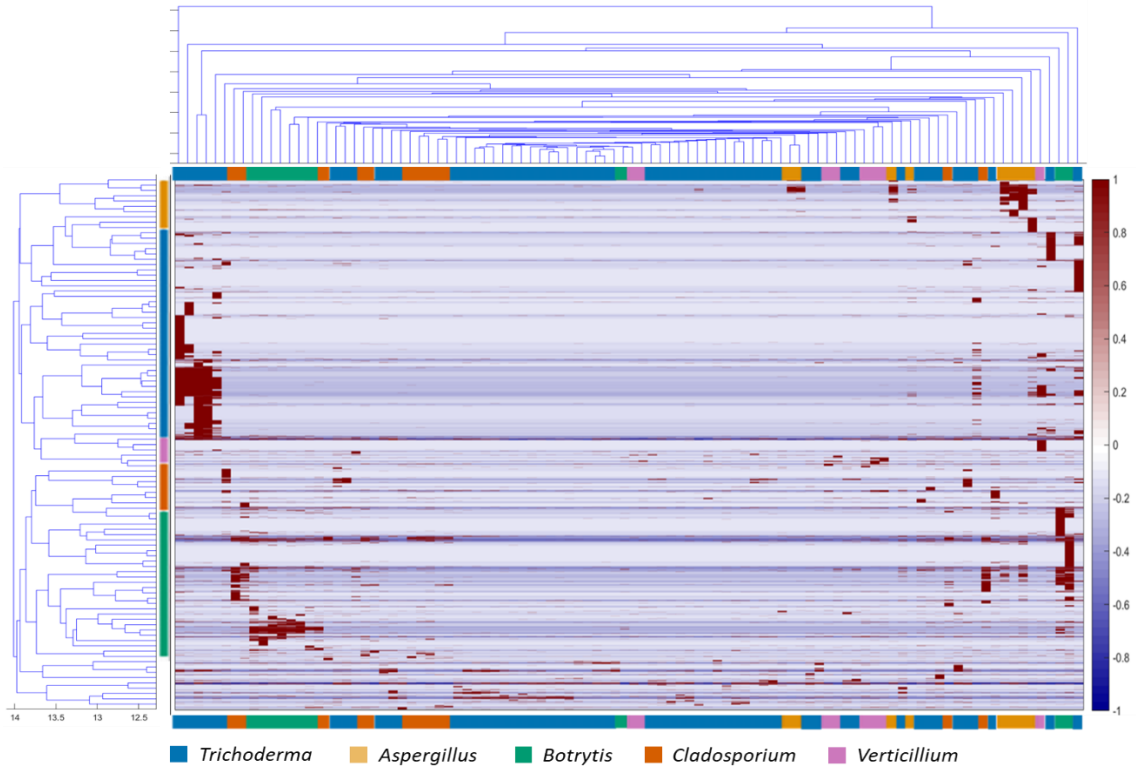


Figure 6.7: HCA of fungal spore samples ionized by ESI negative mode. The legend shows the color coding for the different fungal genera. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

With ESI negative mode the least number of features was detected which is also reflected in the heatmap of the hierarchical clustering analysis. Still, there are feature regions more specific for samples of one genus, but the feature number is quite small. Overall, ESI in negative mode showed lower accuracies and fewer features, indicating that it is not the best method to choose for non-target analysis. This might be because polar molecules ionizable by ESI in negative mode are rather originating from the primary metabolism, like carbohydrates than from the secondary metabolism.

When filtering manually for features that are only present in samples of one fungal species, one can see that those features were all clustered into one cluster. This was the case for the specific feature spaces of all species The problem using these specific feature spaces is discussed in the following chapter and chapter 6.2.3.

6.1.6. Evaluation of possible specific fingerprints

To take a closer look at the more species-specific regions which were clustered by HCA the features were averaged for all respective samples of each species. Shown are the results for ESI positive mode ionization.

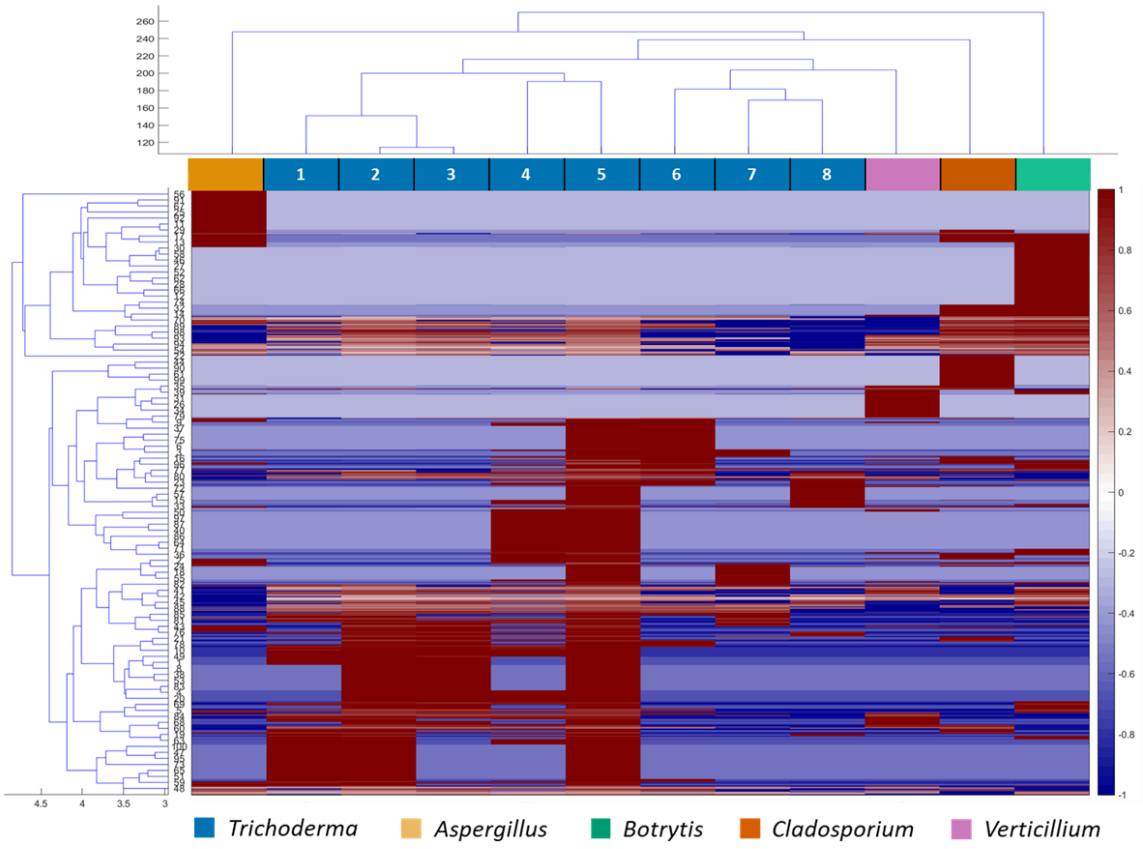


Figure 6.8: Hierarchical clustering analysis for fungal spores of different classes. ESI positive mode. The results of the biological and technical replicates per species were averaged to get a clearer picture of the feature space. The different classes are marked in colour. The different *Trichoderma* species are labeled with a number.

The species-specific feature regions are clearly detectable. What is also more clearly visualized is the differences between samples of different species but the same genus (*Trichoderma*). The different *Trichoderma* species show different high-intensity feature areas, with no feature being present in every single of the *Trichoderma* samples. Even samples of the same species but different strains show differences. This will be further discussed in chapter 6.2.3. It also means that those specific regions for *Aspergillus versicolor* might not be specific for other *Aspergillus* species. Altogether it makes the

determination of a genus-specific fingerprint near impossible as the phenotypical plasticity of each species can be very large.

Despite all the differences in the feature space of different species, the unsupervised clustering shows all *Trichoderma* samples in the clustering diagram next to each other, with height differences between the *Trichoderma* genera being smaller than between different genera/classes. *Verticillium* is more closely to *Trichoderm*a possibly because they belong to the same class. But as the fungis' metabolism is so diverse this might be coincidental. More *Verticillium* species and other samples from the Sordariomycetes class would be needed to support or refute the hypothesis, that hierarchical clustering can represent relationships between fungi of different genera. In some cases, phenotyping has been shown to be supporting taxonomic identification by DNA analysis (Aliferis et al., 2013, Kang, 2011).

6.1.7. Summary of fungal class/family differentiation

The first method to differentiate between fungal classes by non-target LC-HRMS analysis based on fungal spores was developed. Support vector machine with linear kernel worked the best, with cross-validated accuracy results as high as 99 % and very low standard deviations. The sample-set is rather small, but samples showed some phenotypical variety, showing that classification is robust towards inter-species or inter-genera variability, at least in the presented data. Especially that all five *Trichoderma* species, despite showing different intense feature spaces, were still classified into the same genus support the results.

Species or genus differentiation by mass spectrometry has been evaluated before (see chapter 1.2.4). Previous LC-MS or GC-MS studies used the mycelium itself or VOCs emitted by the mycelium to perform genus, order, family, or class differentiation. Some studies used machine learning algorithms to classify species. Müller et al. (Müller et al., 2013) separated 9 different fungal species (Basidiomycetes: *Stropharia, Pholiota, Armillaria Laccaria,* and two *Paxillus* strains. Ascomycetes: *Verticillium, Trichoderma,* and *Cenococcum*) based on the VOC profile with accuracies of 55 to 83 %. Kim et al. (Kim et al., 2016) evaluated *Ascochyta* in a chemotaxonomically approach by LC-MS. Some *Phoma* species were included in the sample set consisting of 45 strains. *Ascochyto* and *Phoma* belong to the same taxonomic order and weren't always separable in the study. The only studies including fungal spores are performed with MALDI-TOF (Becker et al., 2014, Chalupová et al., 2014), but most MALDI studies were performed on filamentous fungal samples. When using fungal spores, concentrations of 2 - 5x10^9 spores/mL were used which is in the same range as in this work (1 x 10^9 spores/mL). Differentiation by MALDI-TOF is based on a peptide

fingerprint (*m/z* 1000 – 20000). The resulting fingerprints are saved in a databank and can be compared to unknown samples. Comparison is performed by the commercially available software Biotyper by Bruker. One study (Lau et al., 2013) used proteins extracted from the fungi's colony to identify clinically relevant mold, mostly *Aspergillus* species. A correct identification on species-level of 88.9 % of samples was reached. With MALDI-TOF different pigmentation of the sample caused trouble (Chalupová et al., 2014), something that was not encountered in this work, probably because ESI/APCI works differently than MALDI. Also, problems when working with variable colony ages were reported, something that wasn't problematic in this work as well.

Fungal spores are presumably not as subjected to phenotypical fluctuations as the mycelium/whole fungus. In general, results of fungal spores are more likely to be transferred to environmental samples than results from pure single-species cultures containing the whole fungus. Fungal spores aren't as susceptible to environmental changes, as fungal spores need to be adapted to and survive many environmental influences. Fungal spores can easily be sampled on filters and extraction with subsequent LC-MS analysis is very fast. This makes differentiation based on fungal spores a very promising approach.

Future investigations should increase the sample set further and get samples grown in different laboratories. Also, including samples that belong to a different order or family of one of the classes would be a promising next step. As the small mixed sample set indicates, that the predominant species in a sample can be classified, more mixed samples at different species ratios should be evaluated. Further environmental influences should be examined and samples from fungi from different origins, e.g., different continents should be included in future data sets. This should induce more phenotypical plasticity, which is important, as fungal samples from different origins can behave very differently and express different metabolites even if the genome is very similar. Additionally, samples should be measured by different orbitrap mass spectrometers (inter-laboratory comparison) to see if the robustness of the method is maintained.

6.2. Differentiation between fungal species

Genus's differentiation by supervised classification methods on the given data set showed promising results. All *Trichoderma* species were classified as *Trichoderma*, even if the samples were grown under different influences, e.g., in different laboratories, two different media and temperatures, and after interim storage in a cooling culture. It was evaluated if SVM or kNN can differentiate on a species level, something that is sometimes difficult to achieve even with DNA analytics (Aliferis et al., 2013, Kang, 2011, Lücking et al., 2020). The results are presented hereafter.

6.2.1. Supervised classification

For species differentiation, the *Trichoderma* samples which were also used in the genus differentiation data set were evaluated. The sample set consists of 42 biological replicates over five species and six strains. Including technical replicates, the sample set consists of 56 to 58 samples depending on the ionization mode. For an exact listing of the species distribution see Table 5.4 and Table 5.5.

For the supervised classification of fungal species in this work, SVM is the better choice as accuracies reached by kNN are always a few percentage points lower than SVM. kNN results are shown in the supplementary information. The lower accuracies of kNN might be due to the small data set. In the case of *T. minutisporum*, *T. fasciculatum,* and *T. longibrachiatum*, only 4 biological replicates, and, including technical replicates, only 5 to 6 instances per species were available. SVM classification accuracies were again calculated from 10-fold stratified cross-validation.

Table 6.6: SVM classification accuracy for the differentiation of *Trichoderma* spores on species level.

Ionization method	Mean accuracy (Standard deviation)
ESI positive mode	0.98 (0.05)
APCI positive mode	0.90 (0.13)
ESI negative mode	0.89 (0.09)
APCI negative mode	0.97 (0.07)

For species differentiation ESI in positive mode and APCI in negative mode show very good results with accuracies of 98 % respective 97 % with low standard deviations. APCI positive

and ESI negative showed lower accuracies, but nonetheless good results with accuracies of 90 % and 89 % with slightly higher standard deviations. Low APCI positive mode accuracies might be connected that the overall detected feature space for the *Trichoderma* samples was only 7934, which is significantly lower than the feature spaces of the other ionization methods. It might be, that *Trichoderma* spp. don't produce that many semi- to nonpolar compounds with functional groups that can be ionized in positive mode. With ESI in positive mode, 17834 features were detected, and with APCI in negative mode 15835 features. ESI in negative mode had a feature space of 12449, which is not significantly lower. Lower accuracies with ESI negative mode coincide with lower accuracies in class differentiation (see chapter 6.1). Again, for ESI negative mode higher number of PCs were needed as input, needing a higher percentage of the explained variance of the feature space to differentiate between samples.

The confusion matrices allow a closer look at the species differentiation. The classifier was trained to look at *Trichoderma harzianum* strain A and B as one class. Strain differentiation is covered in chapter 6.2.4.

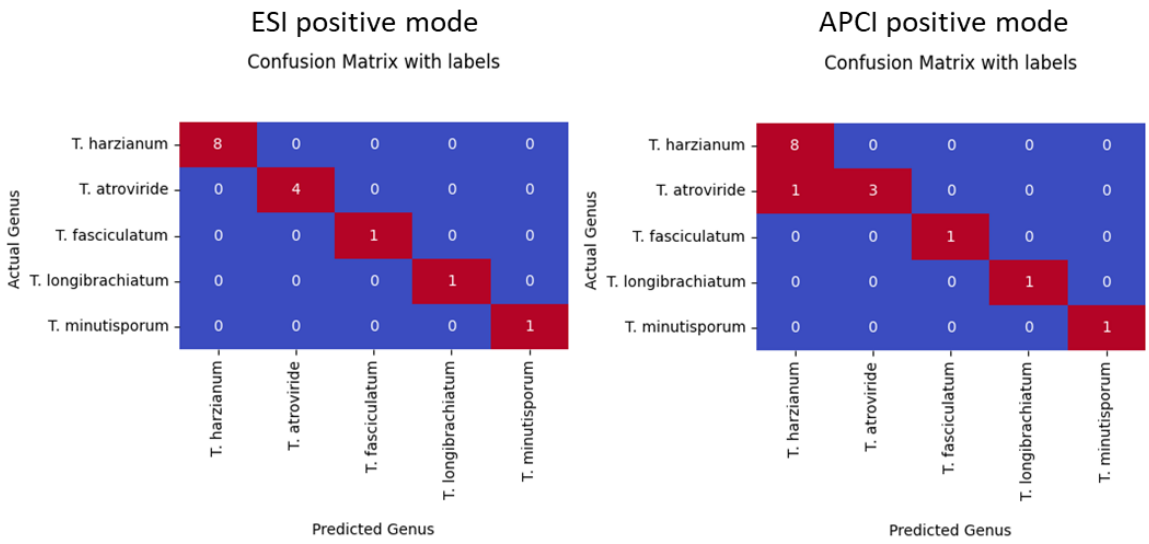


Figure 6.9: Confusion matrices for the differentiation of fungal spore species of the genus Trichoderma by SVM. Left: Ionization by ESI in positive mode. Right: Ionization by APCI in positive mode.

With ESI positive mode the samples were 100 % accurately classified, which is in the standard deviation range of the mean accuracy. With APCI in positive mode, an overall accuracy of 93 % is achieved, meaning one *T. atroviride* samples was misclassified as *T. harzianum*. It should be kept in mind, that an accuracy of 100 % originates from the small

size of the data set. With larger datasets, the classifiers' performance usually increases, but some misclassification cannot be prevented and is normal.

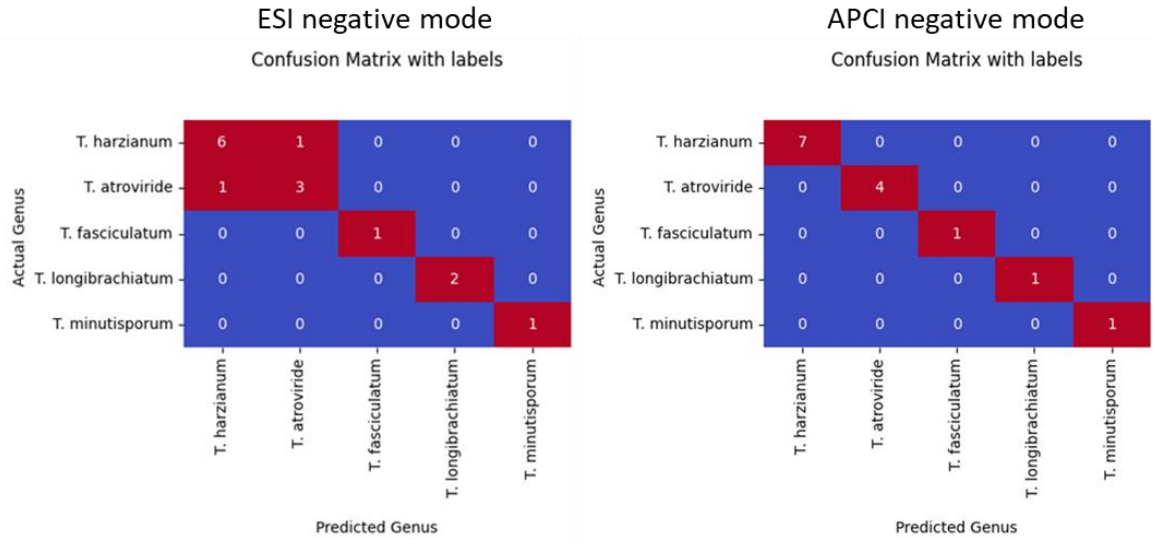


Figure 6.10: Confusion matrices for the differentiation of fungal spore species of the genus Trichoderma by SVM. Left: Ionization by ESI in negative mode. Right: Ionization by APCI in negative mode.

Species differentiation in negative mode shows that with APCI negative mode results are much more accurate than with ESI in negative mode. ESI in negative mode produces overall 87 % accuracy in this example, meaning one *T. atroviride* and one *T. harzianum* sample were misclassified as the respective other class.

Validation of classification

The results are validated with four samples (compare validation of class differentiation, chapter) the classifier has never been trained before. For validation, the classifier is operated with the parameters evaluated during methods development. In the case of SVM with a linear kernel, it only includes the input number of principal components and the C-value.

Table 6.7: Validation of the classification of *Trichoderma* spp. by support vector machine. Linear kernel. Results are shown for a C-value of 0.01. PC input numbers were chosen as evaluated by the 10-fold cross-validation of the train/test sample set.

Ionization method:	**ESI positive**	**APCI positive**	**ESI negative**	**APCI negative**
Actual Species	**Predicted species by classification**			
T. atroviride	*T. atroviride*	*T. atroviride*	*T. atroviride*	*T. atroviride*
T. minutisporium	*T. minutisporum*	*T. minutisporum*	*T. harzianum*	*T. minutisporum*
T. longibrachiatum	*T. longibrachiatum*	*T. longibrachiatum*	*T. longibrachiatum*	*T. longibrachiatum*
T. harzianum strain A	*T. harzianum*	*T. harzianum*	*T. harzianum*	*T. harzianum*

With ESI positive all four validation samples were classified correctly. As an input parameter, the C value was varied between 0.0001 and 10 and the classification remained correct throughout. This indicates that also samples of the same genus, but different species are robustly linearly separable. The C-value doesn't seem to influence classification even at different values meaning ESI positive ionization contains enough information to clearly separate between species. This was also the case for APCI in positive and negative modes, where all samples were classified correctly unregarding the C-parameter. Only with ESI in negative mode, not all validation samples were classified correctly. *T. minutisporum* was misidentified as *T. harzianum*. Additionally, the classification was dependent on the C parameter and only the C value of 0.01 worked correctly. With higher C values also the *T. harzianum* samples were misclassified as *T. atroviride*. Furthermore, is ESI in negative mode the only ionization method which required higher PC numbers as input. Overall seems ESI in negative mode not suitable for the non-target analysis of fungal spores. All other ionization methods produced very good results with the validation set. Future studies should include more samples for training/testing and validation purposes to further improve the robustness of the classification method regarding phenotypical plasticity.

6.2.2. Hierarchical clustering analysis

It was evaluated whether species differentiation can be attributed to certain species-specific features. For that, a hierarchical clustering analysis was performed. In the following, the results for ESI positive mode ionization are shown. The hierarchical clustering results for the other ionization methods are available in the supporting information (see Figure 8.3 to Figure 8.5) as the results are similar.

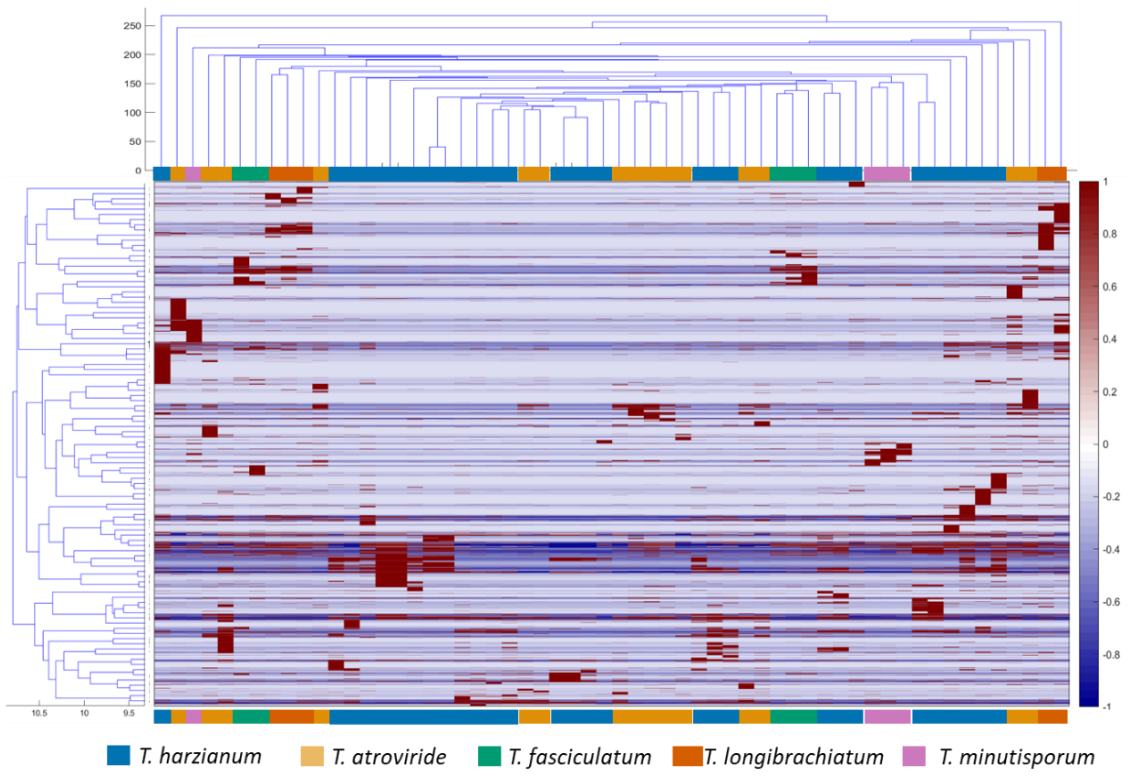


Figure 6.11: Hierarchical clustering analysis of *Trichoderma* species ionized by ESI in positive mode. The legend shows the color coding for the different fungal species. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

Samples were not clustered species-wise and under disregard for the environmental conditions they were grown in. The cophenetic correlation factor is with 0.75 rather low, meaning the horizontal clustering tree doesn't represent the actual distances between species very well.

The cophenetic correlation factor for feature-wise clustering is higher with 0.95. However, there are no joined features present in all samples, which could behave as a general *Trichoderma* specific marker. Diversity within the genus *Trichoderma* is too large. Samples of the same species have some similarities, but these are partly very small. E.g., only 32 compounds are specific for *T. minutisporum*, and present in all 4 biological replicates.

This inter-species variability is also present in e.g., *Aspergillus versicolor*, where close to 100 compounds were present in all *Aspergillus* samples. But when thinking of other *Aspergillus* species one can imagine how few, if any compounds are left which are always available in all samples of all species of a genus. Even two strains of the same species show high variability. Comparing the two strains A and B of *Trichoderma harzianum,* one finds no compound present in all samples of *T. harzianum*, but not present in samples of other species. This diversity is illustrated in the following sub-chapter.

6.2.3. Evaluation of possible species-specific fingerprints

In the following figure, exemplary fingerprints of the inter-species averaged signal intensities for the *Trichoderma* species are shown. The feature space, meaning all compounds expressed by on average by all samples of a species represents the resulting fingerprint. In the following figure, the averaged results for all samples of a species are shown. Fingerprint of the remaining *Trichoderma* species are shown in the supporting information Figure 8.6.

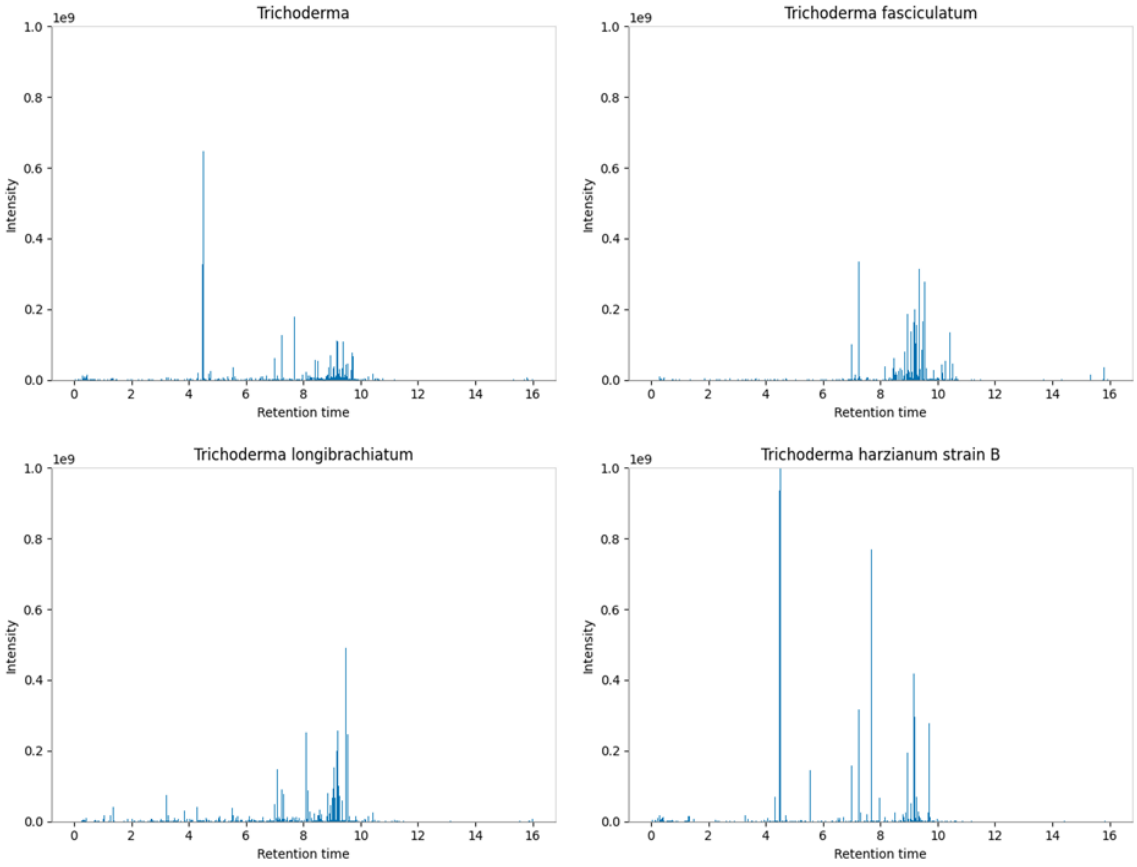


Figure 6.12: Fingerprint of features for *Trichoderma* averaged (upper left), *T. fasciculatum* (upper right), *T. longibrachiatum* (lower left), and *T. harzianum* strain B (lower right). ESI positive mode. The maximum intensity in all figures is 1e9.

The upper left picture shows the averaged signal intensities for all *Trichoderma* species. One could assume that features shown in this fingerprint are typical for *Trichoderma*, however, not all features are present in all species as shown in the heatmap in Figure 6.11. For example, the peak at minute 4.2 is present in *Trichoderma harzianum* strain B in high abundancies but not in the other species, see Figure 6.12 and supporting information Figure 8.6. Nonetheless, some similarities are present, like the denser region around minute 9. These features at this retention time weren't present in e.g., the examined *Aspergillus*

versicolor samples (see Figure 8.2). The presumption is, that all *Trichoderma* samples have this feature region and that it might be specific for *Trichoder*ma spp. This assumption should be validated with a larger sample set, containing more *Trichoderma* species in the future. There are, however, still many features available in the time region from minute 8 to 10, 1546 features were detected on average between all *Trichoderma* species. Plotting these features in a *van Krevelen* plot resulted in a denser region fitting for lipids and peptides, see Figure 6.13.

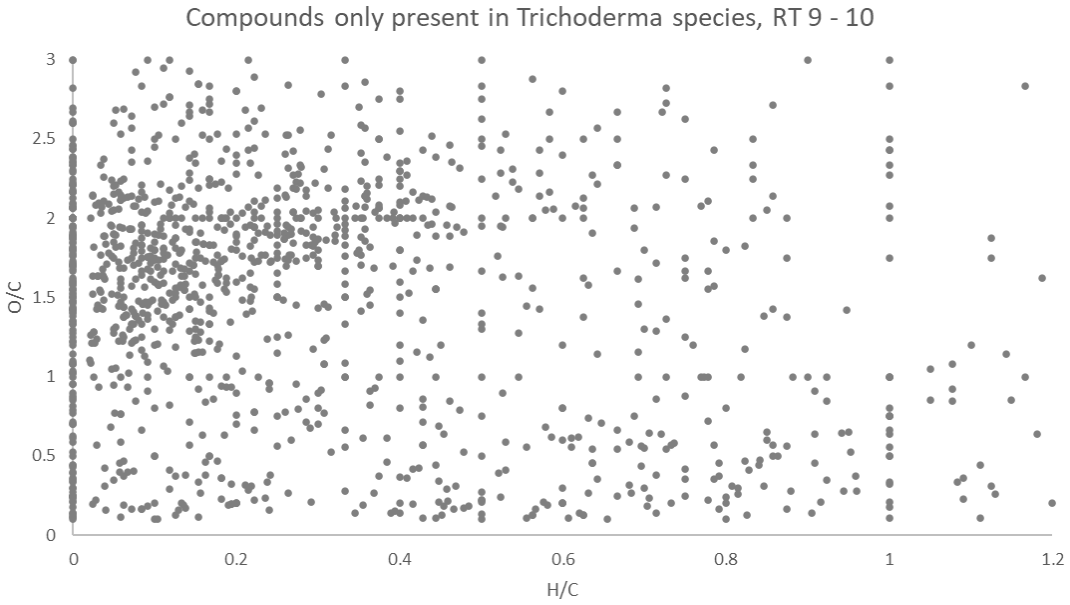


Figure 6.13: *Van Krevelen* plot for features present in *Trichoderma* genus, averaged over all species analyzed in this work. Only features that had a retention time of 9 – 10 minutes are shown.

There are still many features present which do not belong to a specific biomolecule region. Nonetheless, could these be molecules that differ *Trichoderma* from other classes. The abundance of features and the fact that they are not always present in all samples makes it difficult to determine which features are the most important ones. To evaluate reliably which features are more likely to be present in samples of *Trichoderma,* more samples would be needed, possibly also containing several strains to include more inter-species and inter- genera variability. In future investigations, additional machine learning algorithms like Random-Forest could be helpful for research topics like this.

One thing to be considered when evaluating *van Krevelen* plots is, that not all features were calculated with the molecular formula, meaning a *van Krevelen* plot doesn't give the full picture of the metabolome. Different ionization efficiencies of biomolecules can play an additional part when trying to find species-specific regions in a metabolome. Nonetheless,

tendencies of a density of biomolecules of a certain class should be detectable, even if not all biomolecules were ionizable or calculated with a molecular formula.

To examine if these similarities and dissimilarities are also present in samples of the same species but different strains, the averaged intensity values for both strains of *T. harzianum* were plotted.

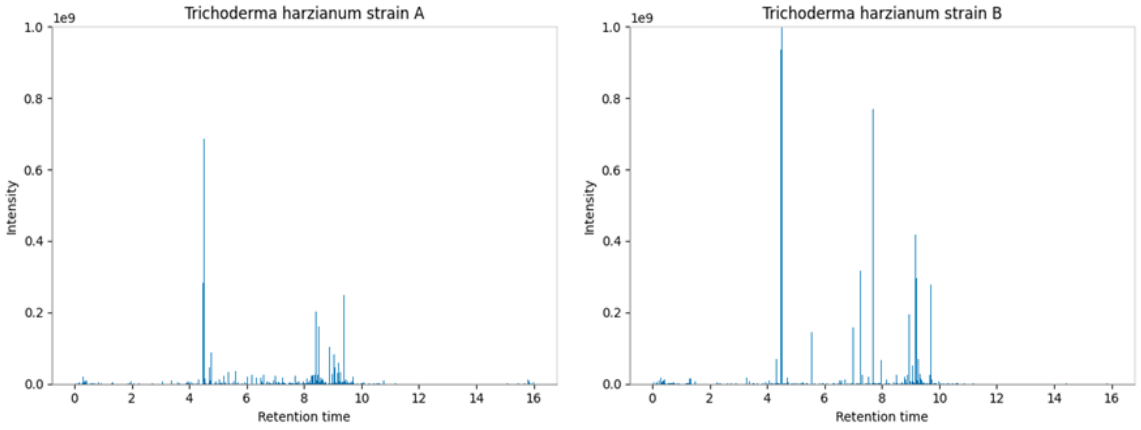


Figure 6.14: Fingerprint of features of *T. harzianum* strain A and B. ESI positive mode. The maximum intensity of 1e9 for all figures.

T. harzianum samples in this work share the intense peak at minute 4.2. Some features present in strain B aren't present in strain A, as the intense peaks at minutes 7 to 8. This emphasizes the inter-species variability. To see if all samples of one strain share the same features the signals for some *T. harzianum* strain B samples were plotted in Figure 6.15. The fingerprint of all T. harzianum strain B samples are available in the appendix Figure 8.7.

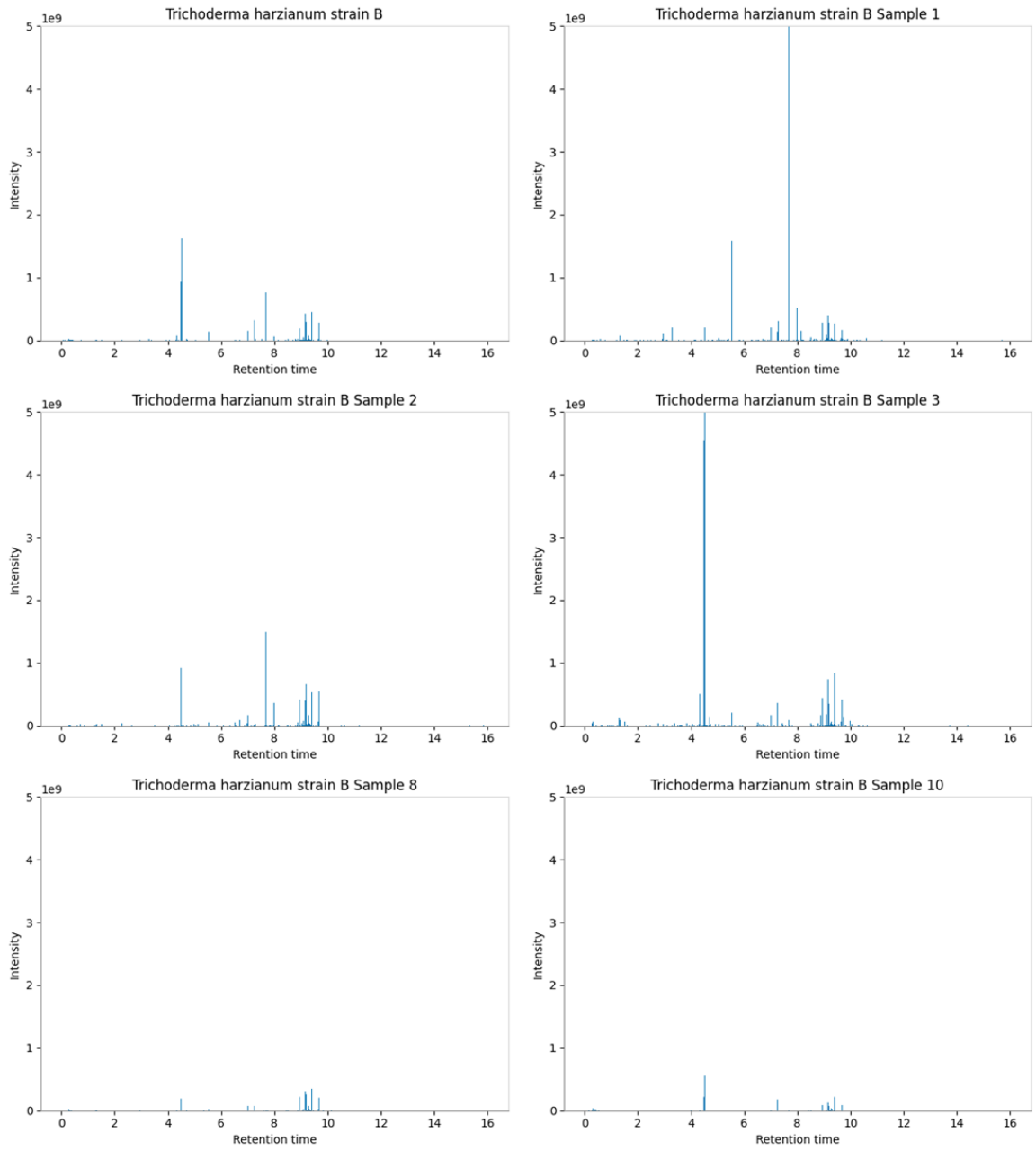


Figure 6.15: Fingerprint of features for different samples of *T. harzianum* B. ESI positive mode. Average for all samples of *T. harzianum* strain B upper left part of the figure. The maximum intensity of 1e9 for all figures.

Some samples present very high signals and others very low signals. This is unrelated to the ergosterol content of the samples and therefore presumably unrelated to the sample's concentration. Additionally, the relative abundancies of signals are different, e.g., Sample 1 shows the peak at min 7.9 as the most intense whereas sample 3 has the peak at min 4.2. as the most intense one. Nonetheless, the spectrums are still similar. It also shows that even the same strain can express different features at different abundancies, and presumably intense features for a specific strain might not be intense or even under a certain threshold for some samples of this strain.

As strains showed a similar pattern but some differences in abundancies of features it was checked if they could be differentiated when both *T. harzianum* strains were treated as their own class. The results are discussed in the following chapter.

6.2.4. Species differentiation including *T. harzianum* strains

The *T. harzianum* class was separated into two for strains A and B. SVM and kNN classification were evaluated by stratified 10-fold cross-validation. kNN classification produced considerably lower accuracy values than SVM. kNN results are available in the supporting information. Best SVM parameters were the same as for the class and species differentiation.

Table 6.8. SVM classification results for strain and species differentiation of *Trichoderma*. Stratified 10-fold cross-validation.

Ionization method	Mean accuracy (Standard deviation)
ESI positive mode	0.94 (0.13)
APCI positive mode	0.82 (0.12)
ESI negative mode	0.84 (0.09)
APCI negative mode	0.95 (0.08)

SVM classification results show similar results as for species differentiation with ESI in positive mode and APCI in negative mode producing the best results. Standard deviations are higher, presumably because fewer training instances were available for *T. harzianum.* In general, the differentiation of more closely related samples can be more difficult, as distances between samples of different species might be small. To see which samples were misclassified the confusion matrices are evaluated.

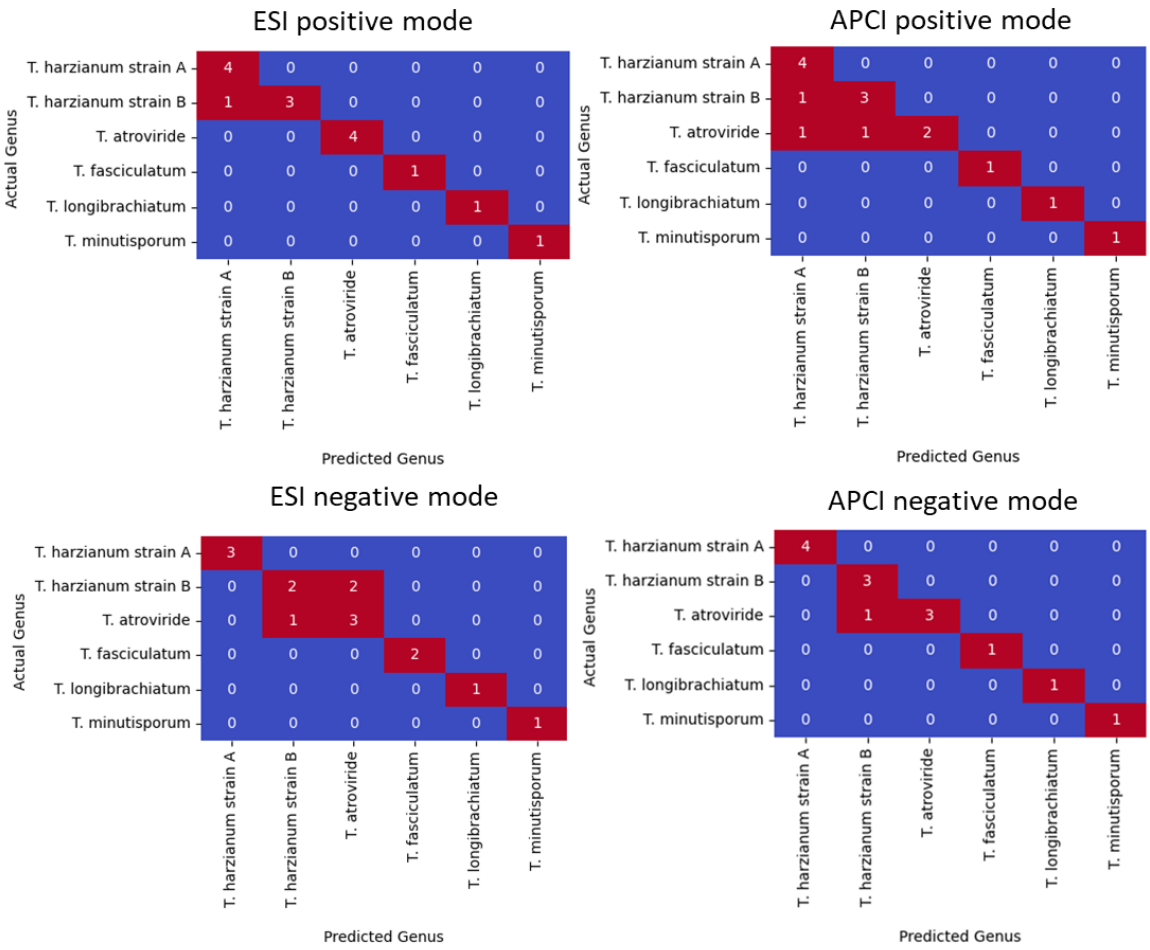


Figure 6.16: Representative correlation matrices for SVM classification of fungal species and strains for *Trichoderma*.

For ESI in positive mode and APCI in negative mode, only one sample was misclassified on average. It was either a mix-up between the two strains of *T.harzianum* or misclassification of *T.atroviride* as mostly *T. harzianum* strain B. It is possible that by splitting *T. harzianum* into two classes the training of the algorithm isn't extensive enough so that *T. harzianum* and *T. atroviride* aren't always distinguishable. For APCI in positive mode and ESI in negative mode accuracies were lower and more samples were misclassified. But also, here the most mix-up happened between *T. harzianum* of both strain and *T. atroviride.* It could also be the case, that for the other species only one sample was used for testing, and with more samples and more inter-species variation results could look different. This should be tested in future studies.

By introducing strain differentiation into the classification, the overall accuracy decreased, probably due to fewer training instances. Nonetheless, the classification differentiated in many cases between the two strains of *T. harzianum*.

6.2.5. Summary of fungal species differentation

The classification was reliable for the sample set even if the inter-species variation was high, as seen in the evaluation of the feature space. Especially ESI in positive mode and APCI in negative mode produced very high accuracies over 95 %. The sample-set is rather small, therefore more samples should be incorporated in the future. Both species and strain differentiation worked well. When misclassification happened it was between *T. harzianum* and *T. atroviride*, which was also the case when *T. harzianum* was treated as one class.

Comparing these results to studies from the literature shows comparable accuracies. Aliferis et al. achieved 83.33 % correct classification of 30 *Rhizoctonia solani* strains by GC-MS (Aliferis et al., 2013). Zwickel et al. evaluated 93 Alternaria strains from 4 species by LC-MS, with a focus on mycotoxin profiles. Species were clustered into high- or low toxin producers and not into species groups. No species-specific mycotoxin profile was detected (Zwickel et al., 2018). Gotthardt et al. analyzed three *Alternaria alternata* isolates and one *Alternaria solani* isolate by a non-target LC-MS metabolomics approach. Also here high variation between biological replicates led to ambiguous unsupervised clustering results (Gotthardt et al., 2020). This "problem" of high inter-species variability was also observed with the *Trichoderma* species used in this work.

Additional applications of non-target metabolome studies are chemotaxonomy by LC- or GC MS. Kang et. al compared the taxonomical classification of *Trichoderma* by chemotaxonomy and by ITS sequencing, based on the mycelium and found concordance between chemotaxonomically and DNA-analysis results. Species or strain differentiation by phenotype is especially interesting as taxonomic classification can be even challenging when using DNA analysis ((Cai, Druzhinina, 2021, Lücking et al., 2020). One recent study showed that the identification of *Trichoderma* species by DNA barcoding is difficult even for experts (Cai, Druzhinina, 2021). Chemotaxonomical approaches can give valuable additional information. It is not clear how well the taxonomy is represented by fungal spores. This could be an additional research question in future work.

Overall, did supervised classification achieve high accuracies for species- and strain differentiation even in the presence of inter-species variability. Future applications could be monitoring of fungal spores which are used as biological plant protectants, e.g., *Trichoderma* spp. In the future, more samples are needed to further investigate strain- and species-level differentiation. Especially *T. longibrachiatum*, *T. fasciculatum,* and *T. minutisporum* were available in only few instances and the sample number of biological replicates should be

increased. A suggested sample size should include 5 strains for at least 5 species each grown under different conditions to get more meaningful results.

6.3. Additional fungal samples

A small sample set of fungal spores on filter samples was available to evaluate with the workflow developed in this work. The fungal spores originate from the Amazonian Rainforest, Brazil, and belong to the taxonomic division Basidiomycetes.

Furthermore, a first investigation of the volatile organic compound profile of *Trichoderma* was conducted.

6.3.1. Basidiomycetes spores from the Amazonian rainforest

Filter fungal spore samples originate from the ATTO (Amazon Tall Tower Observatory) site in the Amazonian rainforest. The research site is in a pristine part of the Amazonian rainforest in the nature reserve "Uatumã Sustainable Development Reserve" 150 km from Manaus. Fungal spores were sampled by a cooperation partner. For further information and experimental see Chapters 4.1 and 4.2. The taxonomic classification of the fungal spore samples is given in the following table.

Table 6.9: Fungal spore samples from the Amazonian rainforest with the taxonomic classification.

Sampling Period	Genus	Family	Order	Class
Wet 2019	*Trametes*	Polyporaceae	Polyporales	Agaricomycetes
Wet 2019	*Picnoporus*			
Wet 2019	-			
Dry 2018	*Ganoderma*	Ganodermataceae		
Dry 2018	Field Mix	mixed	mixed	mixed

All samples belong to the same order and three samples belong to the same family. The order Polyporales usually contains wood-decaying fungi. Taxonomic classification was performed down to genus level except for one sample, which was classified as belonging to the family Polyporaceae. Visual classification at the genus level was not possible. The Polyporaceae samples were sampled in the wet season of 2019, whereas the mixed sample

and the *Ganoderma* sample were sampled in the dry season of 2018. Both positive and negative modes of APCI were applied. Each sample was measured twice (negative mode) or trice (positive mode). As the sample set is quite small, supervised classification is not possible. The unsupervised method of hierarchical clustering analysis was applied to both positive and negative modes. Figure 6.17 shows the results of positive mode ionization by APCI.

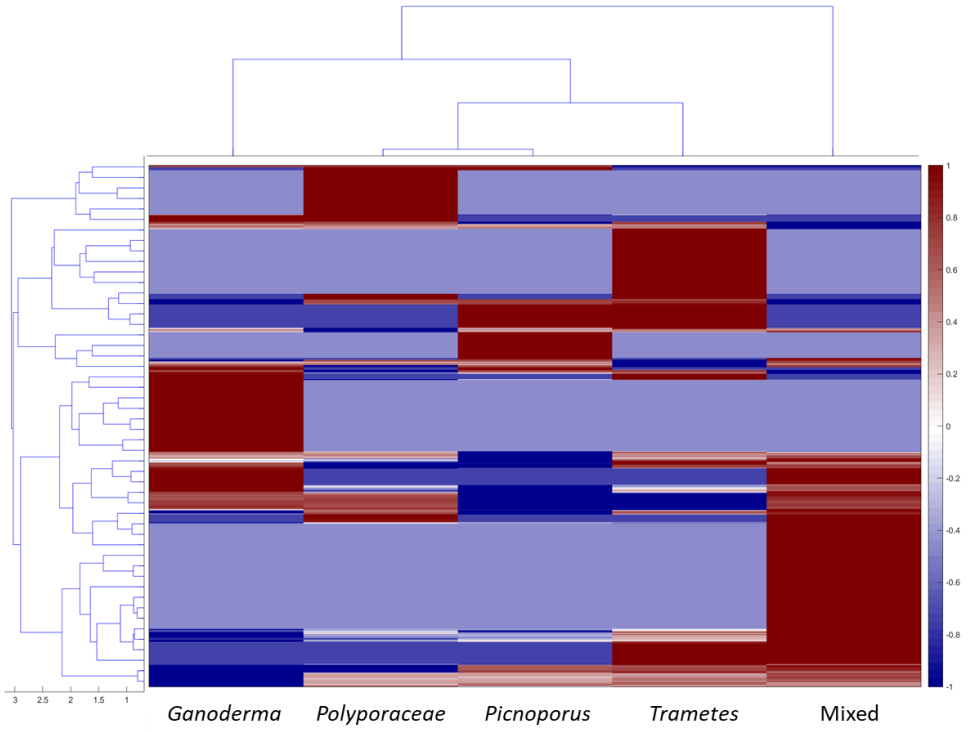


Figure 6.17: Hierarchical clustering analysis of Basidiomycetes spores from ATTO site. APCI positive mode ionization. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

The cophenetic correlation factor of the sample- and feature-wise clustering was 0.98 and 0.95, meaning they represent the actual distances between samples' respective features quite well. One can see that fungi of the family Polyporaceae are clustered very closely together. The sample where the genus wasn't determined is clustered at a very low distance with *Picnoporus*. Only phylogenetic analysis could show if the unknown genus might be closely related to *Picnoporus*. *Ganoderma* was clustered at a larger distance. This could be due to actual less accordance with the other samples, as *Ganoderma* belongs to a different family. Nonetheless is *Ganoderma* still closely related to the other samples as it belongs to the "core polyporoid" clade of the order Polyporales (Hage et al., 2021). The mixed samples show some feature overlap (dark red regions) with *Trametes*, the Polyporcaea sample, and *Ganoderma*. The *Picnoporus* sample has some overlap with the mixed sample in the same

region as the *Ganoderma* sample. Overall, 1179 features were detected, with 92 features detected in all five samples. According to their calculated molecular formula at least 10 of those 92 features were sterol derivates. Both $[M+H]^+$ and $[M+H-H_2O]^+$ ions of ergosterol were present and were identified unambiguously. Of the 92 features, at least 20 had low molecular masses (< *m/z* 100) and molecular formulas fitting to oxygenated unsaturated hydrocarbons. They could originate from isoprene or similar precursors that were emitted from the surrounding vegetation (Bates, Jacob, 2019, Tong et al., 2019). Vegetation markers could be included in these samples as the blank filters were not sampled with air but just exposed shortly to the pump and surrounding air.

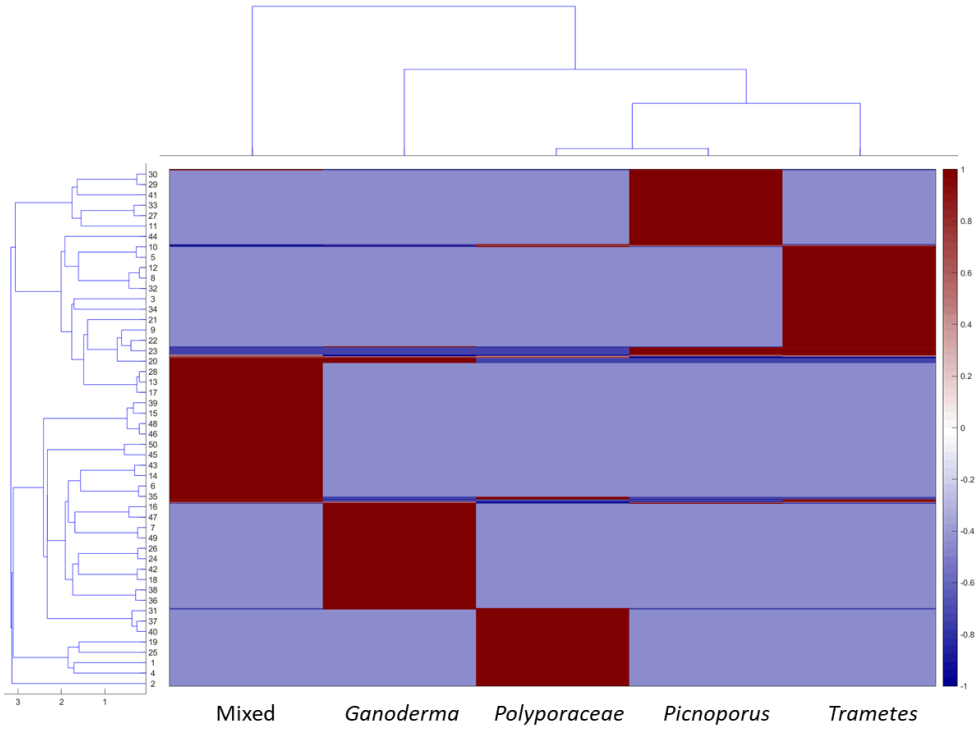


Figure 6.18: Hierarchical clustering analysis of Basidiomycetes spores from ATTO site. APCI negative mode ionization. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

In negative mode, results are similar with Polyporaceae of unknown genus and *Picnoporus* clustered the closest, then *Trametes* and then *Ganoderma*. With negative mode ionization, more features (3578) were detected but there is less overlap between samples. Only 4 features were present in all 5 samples and only 5 features in the 4 single species samples. These features are according to their molecular formula most likely fatty acids which are ubiquitous. Also, here inter-species variability couldn't be considered due to the small sample size.

Summary

Altogether also Basidiomycetes spore showed some distinct features regions for each sample. Spore sampling and extraction by filter sampling worked well and filter samples usually show fewer contaminants when compared to spores from agar plates. With spore harvest from agar plates, small pieces of agar can be incorporated into the sample. This preliminary study indicates that the developed method can be applied to fungal spores sampled from the environment by filters. If features more specific to the fungal spore should be evaluated a blank filter that contains vegetation or other bioaerosol markers should be included in the data processing. Clustering in this small sample set produced results according to the taxonomy, but more samples are needed to include inter-genus variability. LC-MS Analysis of fungal spores can be a non-destructive analysis method that is faster than DNA analysis and could give additional information about the taxonomic classification. Further studies with more samples could also include supervised classification. This supervised classification might also help classify samples of unknown genera like in this case the Polyporaceae further. An interesting approach would be the comparison of fungal spores analysis by DNA data with supervised clustering based on LC-MS or GC-MS data. In the future other supervised classification methods like Random Forest could be included which can give additional information about the relationship between fungal samples.

6.3.2. Fungal volatile organic compounds

Volatile organic compounds (VOC) are commonly associated with microorganisms, including fungi. They derive from the fungi's metabolisms and play an important role in the organism's biology, e.g., having an antibiotic activity or promoting plant growth (Guo et al., 2020b, Müller et al., 2013, Stoppacher et al., 2010). Fungal VOCs consist of different groups of molecules, like terpenoids, alcohols and aldehydes, aromatics, and other heteroaromatic and aliphatic compounds. Odor profiles of fungi might be used as a "fingerprint" or used to find chemical tracers typical for a certain fungal species (Moularat et al., 2008, Müller et al., 2013). The fungi *Trichoderma* is known to produce several VOC, therefore it was tested if the in-lab thermal desorption system was able to detect such fingerprint profiles. The species evaluated was *Trichoderma atroviride.* The experimental information is given in the supporting material (chapter 8.1.4). For theory on thermal desorption see chapter 2.1.2.

Samples were measured by a thermal desorption GC-MS system. Despite the hydrophobic MARKES tubes which are suited for high humidity samples, there was too much water vapor

trapped in the sampling tubes during the thermal desorption process. As the split flow fluctuated during the cryofocussing and injection process, a freezing of the transfer capillary due to water or a similar substance is likely. This resulted in shifted retention times (>2 min) for the samples, especially for the blank sample. The retention time shift is not evenly but is most pronounced in the first minutes of the chromatogram and not detectable at the end, due to the helium flow to the column reaching the setpoint after the transfer capillary is thawed. Due to that, aligning the sample with MZMine was not possible. Several programs were tested to find a suitable alignment solution, but alignment could not be obtained. Therefore, blank subtraction could not be performed by the in-house MATLAB script, also comparison between samples was hindered. A linear temperature-programmed retention index system (LTPRI) according to Van der Dool and Kratz (van den Dool, Dec. Kratz, 1963) was tested to help with data analysis. An alkane standard solution for LTPRI determination was measured. In the first few minutes, retention times fluctuated as well and only for nonane and higher alkanes the retention indices showed a linear response and could be calculated. Unfortunately, the calculated RI for the known compounds in the sample, e.g., ortho-Xylol (890 (literature) to 917 (calculated)) differed substantially. To enable analysis the blank subtraction was performed manually for prominent signals. An overview about the detected mass-to-charge ratios with proposed identity according to the EI-spectra is given in table Table 6.10.

Table 6.10: VOC of *Trichoderma atroviride*. Sources 1) (Guo et al., 2020a), 2) (Stoppacher et al., 2010).

m/z ratio	RT [min]	Sample 1 20 °C, HMG	Sample 2 26 °C, HMG	Sample 3 Kanister, 26 °C, HMG	Substances according to NIST or literature search	Described in literature, source 1 or 2
81	-	NF	NF	8E+04	Methylfuran	-
67.00	3.9	7E+03	6E+02	4E+04	Cyclopropene	-
78	4.2	NF	NF	3E+03	Benzene	Possible contaminant
81, 96	4.9	NF	NF	4E+04	Hexadienal	-
95, 96	5.0	NF	NF	2E+04	Dimethylfuran	-
65, 94	5.4	8E+03	4E+03	4E+04	-	-
91	6.5	NF	4E+03	2E+05	Toluene	Possible contaminant
74, 101	6.8	3E+04	2E+04	3E+04	Butanoic acid	-
106, 91	9.6	4E+03	5E+03	9E+04	Xylol_1	Possible contaminant
106, 91	9.9	1E+04	1E+04	2E+05	Xylol_2	Possible contaminant
106, 91	10.8	5E+03	5E+03	1E+05	Xylol_3	Possible contaminant
121, 91	12.4	3E+03	4E+03	1E+05	Terpen	(1, 2)
105	13.5	1E+04	2E+04	1E+05	Cumen	-
105, 120	13.7	3E+03	5E+03	4E+04	Trimethylbenzol	-
99	14.6	2E+04	7E+04	3E+04	Propylcyclo-hexanol	-
81	14.7	3E+05	7E+03	3E+05	Pentylfuran	(2)
121, 91	15.7	1E+03	8E+03	6E+05	Terpen, possibly ß-phellandrene	(1, 2)
107, 122	15.8	4E+03	6E+03	1E+05	Phenylacetat	-
91, 121	16.2	1E+04	3E+04	3E+06	y-Terpinene	(1, 2)
81, 95	21.5	6E+03	2E+04	1E+05	Cyclopentane or Dodecan	(1, 2)
128	22.2	5E+04	1E+04	2E+05	Napthalen	Possible contaminant
81	22.3	9E+03	1E+03	8E+04	Heptylfuran*	-
147	22.8	3E+04	2E+04	5E+04	-	-
115, 144	23.6	NF	NF	4E+04	Phenylfuran	-
91, 95	25.6	5E+03	9E+03	1E+05	Terpen derivate	(1, 2)
115, 141	26.2	1E+04	2E+03	9E+04	Napthalene derviat*	-
105, 119	31.0	3E+04	9E+03	1E+06	Sesquiterpene, Farnesene	(1, 2)
147, 189	31.3	1E+03	1E+04	2E+04	Sesquiterpene, Zingiberene	(1, 2)
119, 189	31.8	1E+04	1E+04	2E+04	Sesquiterpene	(1, 2)
119, 189	32.1	1E+04	4E+04	5E+04	Sesquiterpene	(1, 2)
119, 161	32.8	3E+04	8E+04	1E+05	Sesquiterpene	(1, 2)
119, 161	33.1	2E+05	3E+05	6E+05	Sesquiterpene	(1, 2)

Not all compounds could be determined clearly, especially as LTPRI shifted and couldn't be used. Suggestions by the NIST library were controlled manually to achieve the best possible identification. Some features were only detected in sample 3, probably due to the larger sample volume from the canister. NIST search together with manual mass spectrum comparison (Linstrom, 1997, Matsuyama, Wasada, 2019) could determine 10 compounds as terpenoids or sesquiterpenoids, which are reported in *Trichoderma* (Guo et al., 2020a) (Stoppacher et al., 2007). Five compounds were determined as furan derivatives, of which pentylfuran was described in the literature (Stoppacher et al., 2010). Some compounds like toluene were also described in the literature but it is not clear if they might occur from the desorption tube, environment, or air and weren't subtracted fully by blank subtraction.

The thermal-desorption GC-MS produced some interesting results and shows that *T. atroviride* emitted volatile organic compounds. To make this approach feasible an improved thermal desorption system, preferably with a more controlled helium flow and moisture trap is needed. A high-resolution mass spectrometer would show better sensitivity and would detect more features, enabling more in-depth profiling. If retention times are stable the data analysis workflow presented in this work could be applied to GC-MS data.

6.4. Non-target LC-MS analysis of electronic cigarettes

6.4.1. Introduction

E-cigarettes are marketed as a "healthy" alternative to traditional cigarettes. Electronic cigarettes consist of an electrical device in which liquids are evaporated, producing smoke which is then inhaled by the user. The liquid contains glycerol and/or propylene glycol, nicotine in variable amounts, and in some cases flavors (El Mubarak et al., 2018). The liquid is combusted by heating to 100 - 250°C degrees by a metal coil (Rowell, Tarran, 2015). During this process, toxic compounds, including aldehydes, nitrosamines, metals, volatile organic compounds, phenolic, and polycyclic aromatic compounds can be formed, especially at high temperatures and if the combustion is "dry" (Cheng, 2014, Farsalinos et al., 2015). Overall e-cigarettes are thought to be less harmful than normal cigarettes, but negative consequences are still unknown (Margham et al., 2016).

The evaluation of possible toxic compounds in e-cigarettes was part of a collaboration with the University Medicine Mainz. In the study (Kuntic et al., 2020) the influence of e-cigarette

consumption on the endothelial function was evaluated. The endothelial function of healthy human subjects was tested while they were consuming smoke from e-cigarettes. Additionally, were e-cigarette liquids and condensates tested on their influence on human endothelial cells. Furthermore, animal studies on mice that lacked phagocytic NADPH oxidase (NOX-2) were performed. E-cigarette vapor exposure caused endothelial dysfunction and induced inflammation and oxidative stress. The NOX-2 pathway was identified as the source of oxidative stress. Negative effects on the cells were significantly more pronounced with condensates than with liquids. This indicates additional toxicity in the condensate, originating from the vaporization process (Kuntic et al., 2020). To identify toxic compounds in e-cigarette liquids and condensates LC-HRMS was used.

6.4.2. Experimental work

Detailed information is available in (Kuntic et al., 2020) and the supporting information (chapter 8.1.4). E-cigarette liquids consisting of 50% propylene glycol and 50 % vegetable glycerol were evaporated with a commercially available e-cigarette and the resulting vapor was condensed. Two liquids were tested, one with 12 mg/~~L~~mL nicotine and one without any nicotine. The resulting liquids and condensates were diluted and measured by LC-HRMS with the ESI source. Columns used were C18- and PFP-(Pentafluorophenyl) phased on the UHPLC system described in this work. Ionization was performed in positive and negative modes with mass spectrometry measurement at high resolution (R=140.000) on the Orbitrap Q Exactive.

HPLC analysis by M. Kuntic presumed the presence of aldehydes, a compound group known to be produced by e-cigarettes and known cell-damaging properties (El Mubarak et al., 2018, Farsalinos et al., 2015, Kim et al., 2014). Many aldehydes in e-cigarettes are very small molecules and show no retention on non-polar columns. Direct detection by the orbitrap Q Exactive with a minimum scan range of m/z 50 is therefore not possible. Aldehyde standards, as well as the e-cigarette condensates, were derivatized by M. Kuntic with DNPH (2,4-Dinitrophenylhydrazine) which increases the molecular weight and enables UV-detection. The orbitrap high-resolution mass spectrometer was then used to identify the aldehyde derivates based on the exact mass together with a retention time comparison of standards.

6.4.3. Results

The DNPH derivatized aldehyde standards, e-cigarette liquid, and condensate were measured on a C18 column in negative mode. Retention times of suspects in liquid and condensate were compared to the retention times of the aldehyde standards as well as the exact mass-to-charge ratios. Five aldehydes were identified unambiguously, one suspect was present in the e-cigarette samples, but no standard was available.

Table 6.11: Analysis of DNPH-aldehyde standards as well as DNPH derivatized e-cigarette liquids and condensates. ESI negative mode, C18 column.

Sample	**Condensate-DNPH**		**Liquid-DNPH**		**Aldehyde-DNPH Standard 1ng/mL**	
Molecule	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Formaldehyde	2.5E+05	7.4E+03	1.2E+04	1.3E+03	1.3E+03	9.1E+02
Acetaldehyde	4.8E+04	3.4E+03	5.4E+03	3.8E+03	0.0E+00	0.0E+00
Acrolein	6.8E+02	9.7E+02	0.0E+00	0.0E+00	2.0E+04	4.5E+03
Propionaldehyde	9.1E+03	4.8E+03	1.1E+03	7.7E+02	5.5E+03	3.2E+03
Butyraldehyde	9.4E+02	6.6E+02	0.0E+00	0.0E+00	3.0E+02	1.5E+03
$[M-H]^- C_{10}H_9N_4O_4$	1.6E+04	1.4E+03	7.6E+02	1.1E+03	n/a	n/a

The LC-MS analysis confirmed the presence of formaldehyde, acetaldehyde, propionaldehyde, and an unknown compound $[M-H]^- C_{10}H_9N_4O_4$ in the e-cigarette liquid. Acrolein and butyraldehyde were either not present in the liquid or below the limit of detection. All aldehydes were present in the condensate with signal intensities at least 8.4-fold higher than in the corresponding liquid. The unknown $[M-H]^- C_{10}H_9N_4O_4$ signal is suspected to be a DNPH adduct of either crotonaldehyde and/or methacrolein. Both aldehydes are known to have harmful effects on the cardiovascular system (Farsalinos et al., 2015, Pei et al., 2014, Samburova et al., 2018). The detected aldehydes especially acrolein showed NOX-2 activation and acrolein protein-adducts were found in lung tissue of mice exposed to e-cigarette vapor (Kuntic et al., 2020).

With aldehydes and especially acrolein potent activators of the NOX-2 pathway were identified. To evaluate if additional compounds are present in the condensates which might be responsible for the toxic effect a semi-target approach was chosen. Data were clustered and visualized by hierarchical clustering. The resulting cluster(s) specific for the

condensates were evaluated if compounds which were reported in the literature for e-cigarettes are present.

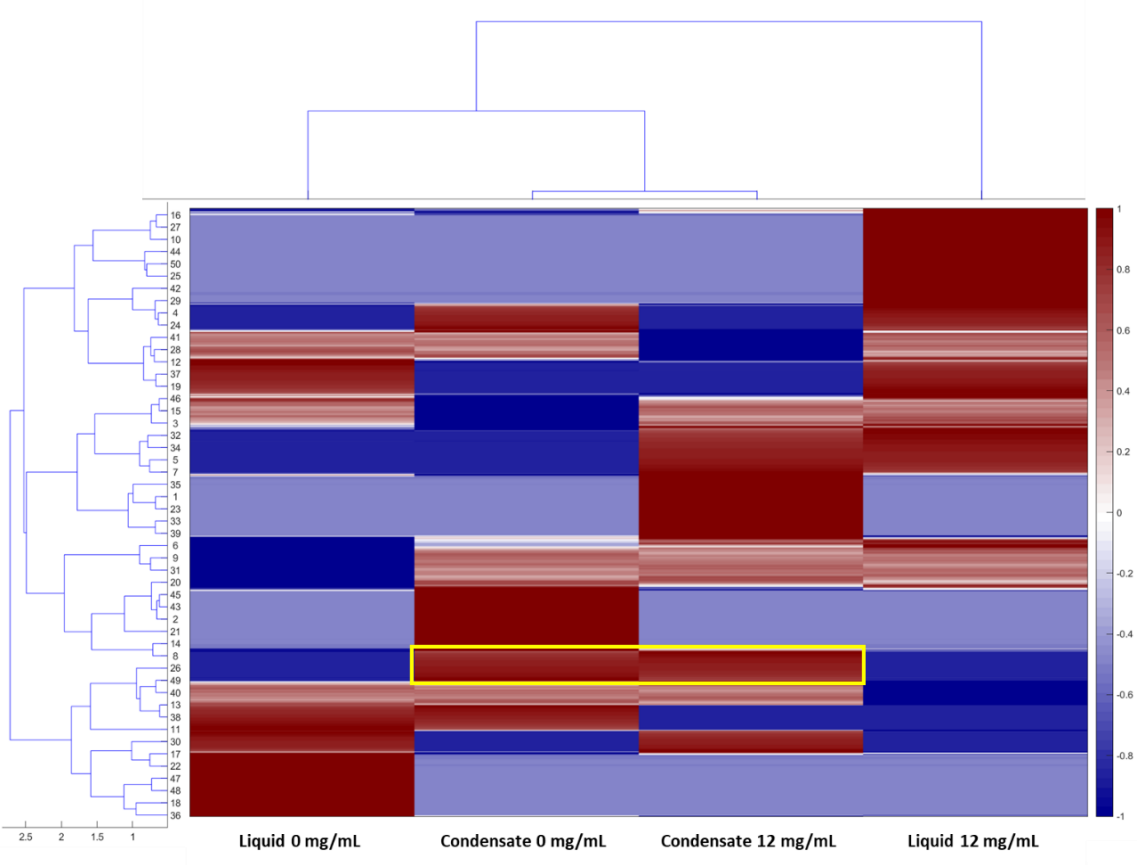


Figure 6.19: Hierarchical clustering analysis of LC-HRMS analysis by ESI positive mode of e-cigarette liquids and condensates. C18 column. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

Both condensates (0 and 12 mg/mL nicotine) are clustered together at a very low distance. They show one significant cluster where compounds are present in both condensates but not in the liquids. This indicates that those compounds were formed during heating and evaporation in the e-cigarette device. A total of 320 compounds were present in both condensates but not the liquid. Around ¼ of the compounds showed no or only minimal retention (< RT 0.6 min). A total of 140 compounds had no calculated molecular formulas. Only 10 % of molecular formulas had reasonable N/C and O/C ratios. Out of these compounds, some were found to match mass-to-charge ratio and calculated molecular formula to compounds described in the literature e.g., 3-methylbutanoate, o-methylbenzaldehyde, or 2-methoxyphenol (Goniewicz et al., 2014, Qasim et al., 2017, Uchiyama et al., 2013).

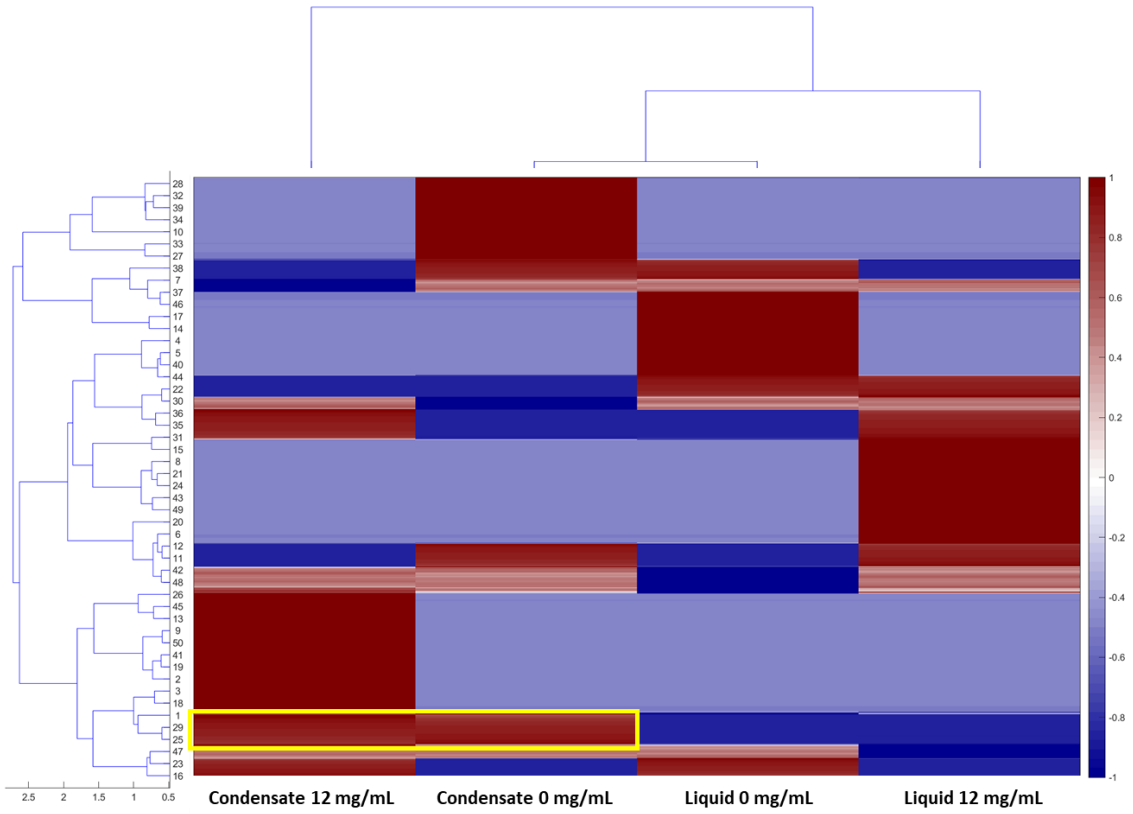


Figure 6.20: Hierarchical clustering analysis of LC-HRMS analysis by ESI negative mode of e-cigarette liquids and condensates. C18 column. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

With negative mode ionization condensate and liquid containing 0 mg/mL nicotine were clustered together. Also, here one cluster represents compounds only present in the condensates. This cluster contains 362 compounds of which 1/3 had very low (< RT 0.6 min) retentions times. Also, here only ~ 10 % had reasonable molecular formula. Compounds were found with mass-to-charge ratio and calculated molecular formula fitting to hydroquinone, methylglyoxal, or also methoxyphenol, which were described in the literature (Goniewicz et al., 2014, Qasim et al., 2017, Uchiyama et al., 2013).

Additional compounds were detected with positive and negative mode ionization in the liquid and condensate containing nicotine. Those suspects are piperidone, imidazole, cresol anatabine, cotinine, and myosmine. These compounds or compounds from the scaffold group have been described in the literature for e-cigarettes or conventional cigarettes (Nicol et al., 2020, Zhu et al., 2015). The three later compounds are nicotine-related alkaloids that originate from the tobacco plant (Flora et al., 2016). As the compounds are presumed according to their exact mass-to-charge ratio they need confirmation by a standard.

The low retention time was a limiting factor in this analysis as some ion suppression was detectable in the low retention time range. The ion suppression is probably caused by well-ionizable compounds present in higher abundances like nicotine and/or the matrix compound glycol and glycerol. A PFP column was tested but didn't show significantly better results. In the future, a more polar column or even a HILIC column should be tested for evaluation of e-cigarette liquids and condensates. Also, the calculation of molecular formulas was not always correct and in ~1/3 of compounds, no molecular formula was calculated. It should be tested if better programs for molecular formula calculation are available or if MZmine 3, published in March 2022 has implementations that improve the molecular formula calculation.

Overall e-cigarettes are less harmful than conventional cigarettes but are not as safe as marketed. Long-term studies are missing but short-term studies showed cytotoxic effects. Also, more severe illnesses like vaping product use-associated lung injury (EVALI) have been reported. More studies are needed on the toxicity of e-cigarettes. Also, liquids containing flavoring or other additives like tetrahydrocannabinol (THC) and vitamin E should be examined (Marques et al., 2021, Rehan et al., 2018). High-resolution mass spectrometry can help identify potentially harmful compounds. Confirmation of a compound with the respective standard is needed but HRMS gives information about the molecular formula of the compounds, minimizing the number of potential targets. Non-target data analysis techniques like hierarchical clustering can facilitate and visualize the detection of compounds that are significant in the condensates or liquids. With more samples comparing different flavorings, etc. this can be especially helpful. Some toxic compounds present in e-cigarette vapor are reported to be below the limit of detection. The development of an enrichment method during samples preparation like e.g., solid phase extraction might be promising.

7. Conclusions and Outlook

In this work, a classification method for the class or species differentiation of fungal spores based on non-target LC-HRMS analysis was successfully developed. To our knowledge, it is the first classification of fungi based on the metabolome of the spores. Samples for class differentiation consisted of four classes from five different families (*Aspergillus versicolor, Botrytis cinerea, Cladosporium cladosporioides, Verticillium dahlia,* and *Trichoderma* spp.). Samples for species differentiation belong to the genus *Trichoderma* and contained five species from six strains. ESI and APCI in positive and negative modes were used to find the most suitable ionization method for the non-target metabolome detection. Several machine learning algorithms, including dimensionality reduction, unsupervised clustering, and supervised classifications were evaluated for data analysis. The final workflow consists of extraction with methanol and a sample-based normalization by spore number and size as the first step. The data analysis includes log-transformation and z-score normalization followed by a dimensionality reduction with principal component analysis. The classification is performed with a support vector machine with a linear kernel. The evaluation of the feature space was conducted with a hierarchical clustering analysis.

Classification resulted in very high accuracies with low standard deviations based on 10-fold cross-validation. Overfitting is unlikely due to the robustness of the classifiers' performance regarding the C-parameter, and the high classification accuracy during cross-validation and for validation samples. The sample-set for the class differentiation included 75 biological replicates and 15 to 20 additional technical replicates, depending on the ionization mode. Validation was performed on additional 6 samples. The sample set for the species differentiation contains 42 biological replicates and depending on the ionization mode ~15 technical replicates. Validation was performed with 4 additional samples. Overall, the ionization with ESI in positive mode provided the best results for both class and species differentiation. For genus differentiation accuracies of 99 % with a standard deviation of 3 % were reached. For species differentiation, the accuracy was 98 % with a standard deviation of 5 %. Also, the other ionization methods, especially APCI in negative mode, provided very high accuracies of over 90 % for genus and strain differentiation. Due to the general availability of the ESI ionization source, ESI is probably the most sensible choice for future applications.

The fungal spore samples were grown under different environmental conditions to induce phenotypical plasticity. Samples grown under different environmental conditions were classified correctly according to the respective species/genus, indicating that the

classification of fungal spores is somewhat robust to environmental influences. This makes the differentiation based on fungal spores a promising approach, especially as the cultivation step can be skipped. This also allows the investigation of fungal spores which usually cannot be cultivated under laboratory conditions (> 80 % of fungal species). Several class distinctive feature regions were visualized and clustered by hierarchical clustering analysis. However, analysis of *Trichoderma* spp. and the two strains of *T. harzianum* showed that presumably characteristic features might not always be present in all samples when species exhibit phenotypical plasticity. With high inter-species and inter-genus variation, the classification of different classes or species cannot be pinpointed to a certain feature space. Machine learning algorithms recognize patterns even if absolute or relative abundances of features vary. This is what makes them so valuable for non-target metabolomic or chemotaxonomic approaches.

Additionally, the performance of the classifier was investigated for mixed species samples on a small sample set. The classifier was able to classify the mixed sample according to the predominant species in a large proportion of the samples. This should be further evaluated on a larger data set but is promising for future field application, where mixed species samples will be common. Future applications could include monitoring fungal allergens or plant pathogens in the air. Also, the dissemination of biological plant protectants could be monitored. Furthermore, filter samples of basidiomycetes spores from the Amazonian rainforest were investigated. Genus distinctive feature regions were available, and the hierarchical clustering was performed according to the taxonomic relationship. Using the fungal spores' metabolome is an interesting approach to support taxonomic classification and hasn't been evaluated yet. Metabolome studies can provide valuable information to improve the classification of closely related species, especially when classic DNA analysis approaches like ITS sequencing are not sufficient. Moreover, a thermal desorption GC-MS system was tested for the investigation of fungal VOCs but showed the need for an improved instrumental setup. Nonetheless are future applications of the developed workflow on GC-HRMS data obvious as it is not only applicable to LC-HRMS data. Also, future application of the developed workflow to other bioaerosol types like bacteria is conceivable. With machine learning algorithms for non-target HRMS data not only biological matrices can be evaluated, but also other fields like medical or environmental matrices. Hierarchical clustering analysis was used in this work to investigate e-cigarette liquids and condensates. Results from e-cigarette analysis were partly published in cooperation with the University Medicine Mainz (Kuntic et al., 2020).

For the application on environmental samples, further investigations should focus on including more samples accounting for additional inter-species and inter-genus variability. Ideally, each genus should be represented by several species and each species by several strains. The samples should originate from different locations, e.g., different plants and different continents. In addition, the workflow could be tested in different laboratories to evaluate the robustness also regarding different Orbitrap instruments. Furthermore, sexual, and asexual spores of the same species could be included, as their metabolism can vary. Overall, the application of machine learning algorithms to non-target LC-MS data has great potential, whether for fungal spore classification, aerosol research, environmental or medical questions.

8. Appendix

8.1. Supporting information

8.1.1. Instruments, chemicals, and programs

Following instruments were used in this work:

Table 8.1: Table of instruments and programs.

Instrument	Model	Manufacturer
Ion source (ESI/APCI)	Ion Max Source, OPTON-20012	Thermo Scientific, Bremen, Germany
High-resolution mass spectrometer	Q Exactive Hybid quadrupole Orbitrap	Thermo Scientific, Bremen, Germany
UHPLC	UltiMate 3000	Thermo Scientific, Bremen, Germany
UHPLC column	Gold Hypersil: C18 50 mm x 2.1 mm x 1.9 µm	Thermo Scientific, Bremen, Germany
Thermal desorption system	In-house build	none
Thermal desorption sampling tubes	Sorbent tubes Odour/sulphur (C6/7 – C30) and Universal (C2/3 – C30)	Markes International GmbH, Offenbach, Germany
Gas Chromatograph	Trace GC 2000	Thermo Fisher, San Jose, CA, USA
Ion Trap Mass spectrometer	Polaris Q ion trap	Thermo Fisher, San Jose, CA, USA
GC column	Rxi 5 MS: 5% diphenyl, 95% dimethyl-polysiloxane, 0.25 µm film thickness, 30 m × 0.25 mm i.d.)	Restek Corp., PA, USA
Centrifuge	Rotanta AP	Andreas Hettich GmbH & Co.KG
Counting Chamber	Neubauer improved	Paul Marienfeld GmbH & Co. KG, Lauda-Königshofen, Germany

Data processing and machine learning were performed with the following tools:

Table 8.2: Programs, programming languages, and libraries used for data processing.

Program	Version	Developer
Preprocessing	MZmine 2.51	Mzmine2 (Pluskal et al., 2010)
Programming Language MATLAB	MATLAB R2020a	The MathWorks, Inc.
Programming Language Python	Python 3.9	Python Software Foundation
PyCharm Integrated Development Environment	2021.1.3 x64	JetBrains s.r.o.
Distribution for Python/ package management	Anaconda 2021.05	Anaconda, Inc.
Scikit-learn Machine Leaning Library	1.0.1	Scikit-learn: Machine Learning in Python (scikit-learn developers, 2021)
Matplotlib Visualization Library	3.5.0	Developer: (Hunter, 2007)
NumPy array processing Library	1.22.0	Developer: (Harris et al., 2020)
pandas data analysis library	1.3.5	(The pandas development team. *pandas-dev/pandas: Pandas 1.4.0rc0*, 2022)
Seaborn statistical data visualization library	0.11.2	Developer: (Waskom, 2021)

For extraction, HPLC and GC measurements following solvents and chemicals were used:

Table 8.3: Table of chemicals and solvents.

Chemical	Manufacturer	Purity
Water (MilliQ)	From a Merck Milli-Q Water purification system	18,2 MΩ·cm, TOC < 1 ppb
Methanol	various	LC-MS Quality
Water (Orbi-grade)	Fisher chemical	LC-MS Quality
Acetonitrile	various	LC-MS Quality
Ethylacetate	various	LC-MS Quality
Formic Acid	various	99% for LC-MS
Miracloth	Millipore, Merck KGgA	-
Potato Dextrose Bouillon	Carl Roth GmbH + Co. KG, Karlsruhe, Germany	-
Pierce LTQ Velos Calibration Solution	Thermo Scientific	-
Helium	various	5.0.

Table 8.4: Parameters for Data Processing with MzMine 2.51.

Step	Setting positive mode	Setting negative mode
Mass Detection	Mass detector: Exact mass Noise level 5.0E4	Mass detector: Exact mass Noise level 5.0E4
FTMS shoulder peaks filter	Filter: Gaussian Mass resolution: 140000	Filter: Gaussian Mass resolution: 140000
ADAP chromatogram builder	Min group size of scans: 3 Group intensity threshold: 5.0e4 Min highest intensity: 5.0e4 m/z tolerance 2.0e-4 or 3.0 ppm	Min group size of scans: 3 Group intensity threshold: 5.0e4 Min highest intensity: 5.0e4 m/z tolerance 2.0e-4 or 3.0 ppm
smoothing	Filter width: 7	Filter width: 7
Chromatogram deconvolution	Wavelets (ADAP) Mz center calculation: Auto S/N threshold: 10 S/N estimator Intensity window SN Min feature height 50,000 Coefficient/area threshold: 80 Peak duration range: 0 – 3 RT wavelet range 0-0.1	Wavelets (ADAP) Mz center calculation: Aut0 S/N threshold: 10 S/N estimator Intensity window SN Min feature height 50,000 Coefficient/area threshold: 80 Peak duration range: 0 – 3 RT wavelet range 0-0.1
Isotopic peaks grouper	m/z tolerance: 3.0e-4, 5.0 ppm Retention time tolerance: 0.2 Maximum charge: 1 Representative isotope: Most intense	m/z tolerance: 3.0e-4, 5.0 ppm Retention time tolerance: 0.2 Maximum charge: 1 Representative isotope: Most intense
Adduct search	Retention time tolerance: 0.2 m/z tolerance 2.0e-4, 3.0 ppm Max relative adduct peak height: 100 % Adducts: M+Na-H, M+MeOH-H, M+H2O-H	Retention time tolerance: 0.2 m/z tolerance 2.0e-4, 3.0 ppm Max relative adduct peak height: 100 % Adducts: M+Na-H, M+MeOH-H, M+H2O-H
Complex search	Mode: M+H+ Retention time tolerance: 0.2 m/z tolerance 2.0e-4 or 3.0 ppm Max complex peak height 100 %	Mode: M+H+ Retention time tolerance: 0.2 m/z tolerance 2.0e-4 or 3.0 ppm Max complex peak height 100 %
Join aligner	m/z tolerance 2.0e-4, 3.0 ppm Weight m/z: 10 Retention time tolerance: 0.2 min Weight RT: 10 Compare isotope pattern: isotope m/z tolerance 5e-4 or 5 ppm Min absolute intensity: 1e3	m/z tolerance 2.0e-4, 3.0 ppm Weight m/z: 10 Retention time tolerance: 0.2 min Weight RT: 10 Compare isotope pattern: isotope m/z tolerance 5e-4 or 5 ppm Min absolute intensity: 1e3 Minimum score: 70 %

	Minimum score: 70 %	
Formula prediction (9 times with decreasing intensity 100 bis 75 % für normal dann nnochmal mit 2 chlor atomen und am ende nochmal mit 9 phosphor und 95 % accuracy	Ionization type: M+H+ m/z tolerance 2.0e-4 or 3 ppm max best formulas per peak 1 max numbers for elements: H:236, C: 156, O:63, S: 14, N:32 Element count heuristics: check H/C ratio, NOPS/C ratios, multiple elements counts RDBE restrictions: 0-40, must be an integer Isotope pattern filter: m/z tolerance 0.001 or 5 ppm, min absolute intensity 1e3, minimum score 100 %	Ionization type: M+H+ m/z tolerance 2.0e-4 or 3 ppm max best formulas per peak 1 max numbers for elements: H:236, C: 156, O:63, S: 14, N:32 Element count heuristics: check H/C ratio, NOPS/C ratios, multiple elements counts RDBE restrictions: 0-40, must be an integer Isotope pattern filter: m/z tolerance 0.001 or 5 ppm, min absolute intensity 1e3, minimum score 100 %
Duplicate peak filter	Filter mode: NEW average, m/z tolerance: 2.0e-4, 3.0 ppm Retention time tolerance: 0.2 min Require same identification	Filter mode: NEW average, m/z tolerance: 2.0e-4, 3.0 ppm Retention time tolerance: 0.2 min Require same identification

Table 8.5: MATLAB and Python Code used in this work.

MATLAB code

Background subtraction

```
%bis subtract_background inklusive, leicht abgewandelt von: Martin %Brüggemann
Leibniz-Institut für Troposphärenforschung e.V. (TROPOS) %brueggemann@tropos.de,
filtering steps von Regina Huesmann, Johannes-%Gutenberg Universität

% check ob csv datei am Ende ein Komma hat (siehe Code Zeile 90-92)
[files, path] = uigetfile('*.csv','Please select a file to convert!',
'C:\Users', 'MultiSelect', 'on');
outdir = uigetdir('C:\Users');
% get blank identifiers from user
disp(' ');disp(' ');
disp('**************************************************');disp(' ');
disp('Please give identifiers of blank samples!');disp(' ');
blank_ID_in = input('Enter now the blank IDs: ','s');
for i = 1:length(files)
    file = string(files(i));
    splittedFile = strsplit(file, ".");
    outputfile = splittedFile(1);
outputpath = strcat(outdir, "\", outputfile, "_Substracted.xlsx");
disp('chosen file:')
disp(file)
disp('output file:')
disp(outputpath)
filename = fullfile(path, file);
for i=1:length(file)
    disp(file(i))
```

```
end
% parameters mass accuracy
tol_mz_abs = 0.0005; % mass tolerance in Da
tol_mz_rel = 3;      % mass tolerance in ppm
RT_tolerance = 0.2;  % retention time tolerance in min
% define parameters for background subtraction
bg_factor = 3;      % multiplier for background signals
% give some user feedback
disp(' ');disp(' ');
disp('**************************************************');disp(' ');
disp(['The following peaklist will be processed:' file]);disp(' ');disp(' ');
disp('The following parameters will be used for all processing:');disp(' ');
disp(['m/z tolerance: ' num2str(tol_mz_abs) ' or ' num2str(tol_mz_rel) ' ppm']);
disp(['retention time tolerance: ' num2str(RT_tolerance) ' min']);
disp(['background subtraction factor: ' num2str(bg_factor)]);disp(' ');
disp('**************************************************');disp(' ');
if strcmp(blank_ID_in,'SK16CW')
    blank_ID = ["ACN" "BW" "A30" "A31" "A43"];
elseif strcmp(blank_ID_in,'SK16F')
    blank_ID = ["ACN" "F15" "F27"];
else
    blank_ID_in = split(blank_ID_in,", ");
    blank_ID_in = string(blank_ID_in)';
    blank_ID = blank_ID_in;
end
disp(' ');disp(' ');
disp('The following blank IDs will be used for blank subtraction:');
disp(blank_ID);disp(' ');
disp('**************************************************');disp(' ');
cutoff_thrshld = 1e5; % peaks that are smaller than this value will be removed
from the dataset
%% read peaklist and get initial data
data = readtable(filename); %the orginal peak data will be saved in this
variable
if ~exist([path, '\Xtract\'], 'dir')
  mkdir([path, '\Xtract\']);
end
% read columns
clmns = data.Properties.VariableNames;
%% remove last column (because of end comma in csv file)
data = data(:,1:end-1);
clmns = clmns(1:end-1);
%% remove duplicates from peaklist and sum up the corresponding peak areas
dummy = data;
clmn_idx_height = find(contains(clmns,'PeakHeight'));
clmn_idx_areas = find(contains(clmns,'PeakArea'));
clearvars data_filtered
i = 0;
while ~isempty(dummy)
    i = i+1;
    %take the maximum of relative and absolute m/z tolerance
    mz_tolerance = max([tol_mz_abs, dummy.rowM_z(1) * tol_mz_rel * 1e-6 ]);

    [~,mz_idx] =
ismembertol(dummy.rowM_z(1),dummy.rowM_z,mz_tolerance,'DataScale',1,'OutputAllIn
dices',1);
    mz_idx = cell2mat(mz_idx);
    [~,RT_idx] =
ismembertol(dummy.rowRetentionTime(1),dummy.rowRetentionTime,RT_tolerance,'DataS
cale',1,'OutputAllIndices',1);
    RT_idx = cell2mat(RT_idx);
    [MF_idx,~] =
find(strcmp(dummy.rowIdentity_mainID_(1),dummy.rowIdentity_mainID_));

    idx = intersect(MF_idx,intersect(mz_idx,RT_idx));
    %idx = intersect(mz_idx,RT_idx);

    data_filtered(i,:) = dummy(idx(1),:);
```

```
    data_filtered(i,clmn_idx_areas) = array2table( sum(
table2array(dummy(idx,clmn_idx_areas)),1 ) );
    data_filtered(i,clmn_idx_height) = array2table(
nanmean(table2array(dummy(idx,clmn_idx_height)),1 ) );

    dummy(idx,:) = [];
end
clear dummy i idx mz_idx RT_idx MF_idx idx clmn_idx_height clmn_idx_areas
%% extract some data from the 'data_filtered' table
% read columns
clmns = data_filtered.Properties.VariableNames;
[~,idx] = find(contains(clmns,'rowIdentity_mainID'),1);
Formula = data_filtered{:,idx};
% get sample names
[~,idx] = find(contains(clmns,'Area'));
samples = data_filtered.Properties.VariableNames(idx);
% get peak areas from table
areas = data_filtered{:,idx};
areas = array2table(areas,'VariableNames',samples);
% get m/z and RT columns
[~,idx] = find(contains(clmns,'rowM_z'),1);
mz = data_filtered{:,idx};
[~,idx] = find(contains(clmns,'rowRetentionTime'),1);
RT = data_filtered{:,idx};
clear idx
%% background subtraction (and removal of blank peaks from peaklist)
[areas_minus_blanks,blanks] = subtract_background(areas,bg_factor); % remove
signals from blank samples
%% prepare data for next step: filtering, calculate average
Ident = string(Formula);
% Return a boolean array named 'empties' (with same dimensions as cell array
% 'A') with true for each empty element and false otherwise
empties = cellfun('isempty',Ident);
% % Now change all the empty cells in A from empty strings '' to double NaN
Ident(empties) = {'"'};
% clearvars -except mz RT clmns areas_minus_blanks Formula Ident samples
data_filtered
%% delete rows with too many zerovalues, hier einstellen wenn auch 2 von 3 areas
größer null ok
%Ident = Formula;
mydata = table2array(areas_minus_blanks);
[rowIdcs, colIdcs]=find(mydata~=0);
[counts, bins] = histcounts(rowIdcs,1:size(mydata,1));
% Wenn alle mit 0 löschen counts ~= 0, ansonsten >= 2 oder 3
binstocount = bins(counts >=2);
%binstocount = bins(counts >=3);
%binstocount gibt die rows an die behalten werden sollen
dataohne0=mydata(binstocount,:);
mzohne0=mz(binstocount);
RTohne0=RT(binstocount);
Identohne0=Ident(binstocount,:);
%clearvars -except mz RT clmns areas_minus_blanks Formula Ident samples newdata
mztokeep RTtokeep Identtokeep
%% calculate average of signal intensities hier ändern wenn man averagen will
Samplemean = mean(dataohne0,2);
% Return a boolean array named 'empties' (with same dimensions as cell array
% 'A') with true for each empty element and false otherwise
datagesamtone0 = [mzohne0 RTohne0 Samplemean];
%% alle Reihen löschen in denen noch ein Complex ist
new = strfind(Identohne0, 'Complex');
TF = cellfun('isempty', new);
sumohnecomplex = Identohne0(TF);
dataohnecomplex = datagesamtone0(TF,:);
clear new TF
%% Alle Reihen löschen in denen noch ein adduct ist
new2 = strfind(sumohnecomplex, 'adduct');
TF = cellfun('isempty', new2);
sumohneadduct = sumohnecomplex(TF);
```

```
dataohneadduct = dataohnecomplex(TF,:);
gesamtohneadduct =[dataohneadduct(:,1:2) sumohneadduct(:,:)
dataohneadduct(:,3:end)];
clear new TF
%clearvars -except ohneadduct sumnew2
%% change Sample Name manually name Table data 1, 2, 3 for later alignment
endtablename = areas_minus_blanks.Properties.VariableNames(1,1);
%%
colNames = [{'mz','RT','Ident'}, endtablename];             %change Sample Name
Table = array2table(gesamtohneadduct,'VariableNames',colNames);
%% write in excel file
writetable(Table,outputpath);
end
```

Background subtract function

```
function [out,blanks] = subtract_background(data,factor,identifier)

if nargin < 3, identifier = "blank"; end %name of blank samples (multiple
arguments possible)

if nargin < 2, factor = 3; end %factor for background subtraction of peak areas
[~,idx] = find(contains(data.Properties.VariableNames,identifier)...
    & contains(data.Properties.VariableNames,'Area')); %get peak area columns of
blanks
blanks = data(:,idx);
avg_blanks = mean(table2array(blanks),2);
AreasToSubtract = avg_blanks * factor;
out = data;      %create new data table from old one
out(:,idx) = []; %remove blanks from data table
for i=1:size(out,2)
    dummy = table2array(out(:,i)) - AreasToSubtract;
    dummy(dummy<0) = 0;
    out(:,i) = array2table(dummy);
end
AreasToSubtract =
array2table(AreasToSubtract,'VariableNames',{'subtracted_area'});
blanks = [blanks, AreasToSubtract]; %make complete table for blank export
%remove rows of blank signals from data table
% idx = [];
% for i=1:size(out,1)
%     dummy = sum(table2array(out(i,20:end)));
%     if dummy <= 1000
%         idx = [idx; i];
%     end
% end
%
% out(idx,:) = [];
end
```

Alignment

```
%% March 2021, Regina Huesmann, Department Chemistry, University of Mainz, AK
Hoffmann
% choose file from which you want to import your data/ which user wants to be
aligned
% get path and filenames from user
[files, path] = uigetfile('*.xlsx','Please files to convert!', 'C:\Users',
'MultiSelect', 'on');
%% give output file path
outdir = uigetdir('C:\Users');
%% give output file name
outputfile = input('Give output filename: ', 's');
%% final name
outputpath = strcat(outdir, "\", outputfile, "_aligned.xlsx");
outputpath2 = strcat(outdir, "\", outputfile, "_rohfassungaligned.xlsx");
```

```
%% read peaklist and get initial data
clearvars joined
for j = 1:length(files)
    file = string(files(j));
    filename = fullfile(path, file);
    opts = detectImportOptions(filename);

    for i = 1:length(opts. VariableTypes)
        if(i == 3)
            opts.VariableTypes{i} = 'string';
        else
            opts.VariableTypes{i} = 'double';
        end
    end

    % all files are joined into one file
    if (j == 1)
        joined = readtable(filename, opts);
    else
        joined =  outerjoin(joined, readtable(filename, opts),
'Mergekeys',true);
    end
end
joined{:,4:end}(isnan(joined{:,4:end})) = 0;
%% parameters for alignment, can be changed to preference of user
tol_mz_abs = 0.0008; % mass tolerance in Da
tol_mz_rel = 7;      % mass tolerance in ppm
RT_tolerance = 2.5;  % retention time tolerance in min
 %% Sort for mz and then RT
C = joined;
RT_sort = sortrows(C,[1 2]);
clear C
%% Alignment via mz and RT and sum up of the corresponding peak areas.
dummy = RT_sort;
clmns = RT_sort.Properties.VariableNames;
clmn_idx_areas = clmns(:,4:end);
clearvars data_filtered
i = 0;
while ~isempty(dummy)
    i = i+1;
    %take the maximum of relative and absolute m/z tolerance
     mz_tolerance = max([dummy.mz(1) * tol_mz_rel * 1e-6 ]);
%    mz_tolerance = max([tol_mz_abs, dummy.mz(1) * tol_mz_rel * 1e-6 ]);
    [~,mz_idx] =
ismembertol(dummy.mz(1),dummy.mz,mz_tolerance,'DataScale',1,'OutputAllIndices',1
);
    mz_idx = cell2mat(mz_idx);
    [~,RT_idx] =
ismembertol(dummy.RT(1),dummy.RT,RT_tolerance,'DataScale',1,'OutputAllIndices',1
);
    RT_idx = cell2mat(RT_idx);
    idx = intersect(mz_idx,RT_idx);
    data_filtered(i,:) = dummy(idx(1),:);
    data_filtered(i,clmn_idx_areas) = array2table( sum(
table2array(dummy(idx,clmn_idx_areas)),1 ) );
    dummy(idx,:) = [];
end
clear dummy i idx mz_idx RT_idx idx clmn_idx_height clmn_idx_areas
%% save two excel files one with the final aligned table and the second one (RT-
sort) to keep the information of sum formulas
writetable(data_filtered,outputpath);
writetable(RT_sort,outputpath2);
```

Hierarchical Clustering

```
% Basic script for clustering by Denis Leppla, Hoffmann Group, Department of
Chemistry, University of Mainz. Modified by Regina Huesmann.
%% import data, Ident/Sum Formula in column 3 will be imported as string, Rest
of data as double
[file, path] = uigetfile('*.xlsx', 'Please select file!', 'C:\Users');
j = 1:length(file);
    file = string(file(j));
    filename = fullfile(path, file);
    opts = detectImportOptions(filename);
    for i = 1:length(opts. VariableTypes)
        if(i == 3)
            opts.VariableTypes{i} = 'string';
        else
            opts.VariableTypes{i} = 'double';
        end
    end
C_complete  = readtable(filename, opts);
%%  Determine output path for excel file
outdir = uigetdir('C:\Users');
%% Determine output path for pictures
outdir2 = uigetdir('C:\Users');
%% Give filename
outputfile = input('Give output filename: ', 's');
outputpath = strcat(outdir, "\", outputfile, "_clustered.xlsx");
%% Clear temporary variables
clear opts
%% find signals per row and delete compounds which are just present in one
sample
ML_matrix = table2array(C_complete(:,4:end));
MassList_final = (C_complete);
Samples = MassList_final.Properties.VariableNames(4:end);
clear ML_matrix l b idx
% log2, vorher shift
ML_matrix =table2array(MassList_final(:,4:end));
datawithshift = ML_matrix + 1;
gelogt = log(datawithshift);
ML_matrix = gelogt;
%%% Calculate z-scores (standardization)
z_scores = normalize(ML_matrix,2, 'zscore');
z_scores_trans = z_scores';
%% create Colormap
% colors = colormap either DIY or from MATLAB of your choice
Compounds = table2cell(MassList_final(:,3));
 Samples = C_complete.Properties.VariableNames(4:end);
%% Cluster analysis for compounds // Plot Dendrogram for compounds
f1 = figure('color',[1 1 1],'units','centimeters','position',[1.5 2 7 25],...
            'paperunits','centimeters','paperposition',[0 0 7 25],...
            'papersize',[7 25]);
dist_1 = pdist(z_scores);
link_1 = linkage(dist_1,'average');
c_oben = cophenet(link_1, dist_1);
leafOrder_1 = optimalleaforder(link_1,dist_1);
[~,D_1] = dendrogram(link_1,50,'reorder',leafOrder_1,'Orientation','left');   %
adjust 2nd input according to number of clusters you want (0 = all)

find_axes = findall(0,'Type','axes');                    % Y axis needs to be
reversed, 'findall' opens the figure editor in the Workspace
set(find_axes,'YDir','reverse')
print('-painters','Dend_comp_60Cluster', '-dsvg', '-r1200')
print('-painters','Dend_comp_60Cluster', '-dmeta', '-r1200')
f1 = gcf
%%
saveas(f1,strcat(outdir2, "\", outputfile, "_dendroseite.png"));
%% rearange rows (compounds) & Find compounds in specific clusters
z_scores_rearange = z_scores(leafOrder_1,:);
Comp_rearange = Compounds(leafOrder_1,:);
```

```
C_complete_reaarange = C_complete(leafOrder_1,:);
%% Number of row transpose leaf_order
leafOrder_1_Liste = transpose(leafOrder_1);
%%
SumFormula = C_complete(:,3);
%% Cluster analysis for samples // Plot Dendrogram for samples %%
%% check Inconsitency
f2 = figure('color',[1 1 1],'units','centimeters','position',[1.5 5 40 7],...
            'paperunits','centimeters','paperposition',[0 0 40 8],...
            'papersize',[40 7]);
dist_2 = pdist(z_scores_trans);
link_2 = linkage(dist_2,'average');
c_seite = cophenet(link_2,dist_2);
leafOrder_2 = optimalleaforder(link_2,dist_2);
[~,D_2] = dendrogram(link_2,0,'reorder',leafOrder_2);
% calculate cophenet
c = cophenet(link_2,dist_2);
clear dist_2 link_2
f2 = gcf
saveas(f2,strcat(outdir2, "\", outputfile, "_dendrooben.png"));
%% rearange rows (compounds)
z_scores_rearange_2 = z_scores_rearange(:,leafOrder_2);
Samples_rearange = Samples(:,leafOrder_2);
%% Plot heatmap WITH arangement!
f3 = figure('color',[1 1 1],'units','centimeters','position',[1.5 2 40 25],...
            'paperunits','centimeters','paperposition',[0 0 40 25],...
            'papersize',[40 25]);
h1=heatmap(z_scores_rearange_2(:,:),'GridVisible','off','CellLabelColor','none',
'Colormap',colors,'XLabel','Sample','YLabel','Compound',...
    'XDisplayLabels',Samples_rearange)
caxis([-1 1])
saveas(h1,strcat(outdir2,  "\", outputfile, "_heatmap2.png"));
close(f3)
clear f3
%% Erstellen einer Tabelle in der alle compounds nach Reihenfolge der cluster
geeordnet sind.
[H,T,outperm]=dendrogram(link_1,100,'reorder',leafOrder_1,'Orientation','left');
T_table = array2table(T);
mat = [T_table C_complete];
mat_reaaranfge = mat(leafOrder_1,:);
%% get number of elements in each formula
sumohneadduct=table2array(mat_reaaranfge(:,4));
A = "C0H0O0";
sumohneadduct = fillmissing(sumohneadduct,'constant', A);
 count_elements = zeros(length(sumohneadduct),9); % we want the order C, H, O,
N, S, P, Cl, Br, I
element_symbol = {'C' 'H' 'O' 'N' 'S' 'P' 'Cl' 'Br' 'I'};
count_others = cell(length(sumohneadduct),1); % to determine if other elements
are present
for j = 1:9

    for i=1:size(sumohneadduct,1)
        dummy = [sumohneadduct{i,1} '##'];    % get formula in row i
        k = strfind(dummy,element_symbol{1,j}); % find position of element
symbol in formula / dummy
        if ~isempty(k) % if element is not present k is empty and we skip this
            k = k(1,1); % take only first occurence of element symbol
            if length(element_symbol{j}) > 1    % for elements like Cl we have
to add one position to k
                k = k+1;
            end
            if ~isnan(str2double(dummy(k+1:k+2)))   % check if element > 9 (two
positions)
                count_elements(i,j) = str2double(dummy(k+1:k+2));
            elseif ~isnan(str2double(dummy(k+1)))   % check if there is a number
after the element
                count_elements(i,j) = str2double(dummy(k+1));
```

```
            else                                        % if there is only the
symbol set the number to 1
                count_elements(i,j) = 1;
            end
        else                                        % if element is not present
set the number to 0
            count_elements(i,j) = 0;
        end
        %look also for other elements (but only in the last round)
        if j==5
            search_other = {'A','a','B','b','D','d','E','e','F','f','G',...
                'g','J','j','K','k','L','M','m','Q','q','R','T','t','U',...
                'u','V','v','W','w','X','x','Y','y','Z','z'};
            k = contains(string(dummy),search_other,'IgnoreCase',false);
            if k == 1
                count_others{i,1} = 'yes';
            else
                count_others{i,1} = 'no';
            end
        end
    end

end
clear dummy i j k
%% alles in eine Tabelle
elements_table = array2table(count_elements);
elements_table.Properties.VariableNames = element_symbol;
%% combine tables
Final_table = [mat_reaaranfge elements_table];
%% save as excel file
writetable(Final_table,outputpath);
%% end
```

Python Code

Principal Component Analysis

```
# import packages
from matplotlib import pyplot as plt
import pandas as pd
import numpy as np
from sklearn.decomposition import PCA
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from mpl_toolkits.mplot3d import Axes3D
from pylab import figure
from mpl_toolkits.mplot3d.proj3d import proj_transform
from matplotlib.text import Annotation
from matplotlib.colors import ListedColormap
import matplotlib.patches as mpatches
from matplotlib.lines import Line2D
# import data, df for data, df1 for labels, insert filepath
df = pd.read_excel(r'C:\Users')
df1 = pd.read_excel(r'C:\Users')
# extrakt data from df, all rows and designated columns
matrix = df.iloc[:, 3:110]
# Transpose data
matrixtrans = matrix.T
# get names of matrix columns
names = matrixtrans.index
newnames = []
for name in names:
```

```
    newnames.append(name.split("_")[0])
# shift +1 then log10 data
matrixshift = matrixtrans +1
matrixlog = np.log2(matrixshift)
# scale with StandardScaler/zscore
scaler = StandardScaler()
x = scaler.fit_transform(matrixlog)
# convert x to data frame
z = pd.DataFrame(x, index=newnames)
# perform PCA
pca_data = PCA(n_components = 10)
principalComponents_data = pca_data.fit_transform(z)
# create Data Frame with all principal component values/ only necessary if
indexing is performed by column header as later in visualization
pcadataDf = pd.DataFrame(data = principalComponents_data, columns = ['PC1',
'PC2', 'PC3','PC4','PC5','PC6','PC7','PC8','PC9','PC10'])
pcadata = pd.DataFrame(data = principalComponents_data, columns = ['PC1', 'PC2',
'PC3','PC4','PC5','PC6','PC7','PC8','PC9','PC10'], index = newnames)
# create labels from df1
y = np.array(df1)
labels = np.ravel(y)
list_of_tuples = list(zip(newnames, labels))
labelnames = pd.DataFrame(list_of_tuples, columns = ['SampleName', 'Index'])
# plot 2D
fig = plt.figure(figsize=(10,8))
fig,ax =plt.subplots()
x_axis = pcadataDf.loc[:,'PC1']
y_axis = pcadataDf.loc[:,'PC2']
ax.set_facecolor('white')
ax.spines['left'].set_color('grey')
ax.spines['bottom'].set_color('grey')
ax.spines['right'].set_color('lightgrey')
ax.spines['top'].set_color('lightgrey')
ax.set_xlabel('PC 1')
ax.set_ylabel('PC 2')
ax.grid(True, color='gainsboro')
colors = ['#0173b2', '#de8f05', '#029e73', '#d55e00', '#cc78bc']
p1 =sns.scatterplot(x_axis, y_axis, hue =
(labelnames.loc[:,'Index']),marker='o', palette = colors,s=40)
lines = [Line2D([0], [0], color=c, linewidth=0, linestyle=None, marker='o') for
c in colors]
labels =['Trichoderma','Aspergillus','Botrytis','Cladosporium','Verticillium']
plt.legend(frameon=False)
leg = plt.legend(lines,labels)
leg.get_frame().set_color('white')
leg.get_frame().set_edgecolor('gainsboro')
plt.title('Cluster by PCA Components')
plt.tight_layout()
plt.show()
# Plot Visualization 3D
fig = plt.figure(figsize=(10,8))
ax = fig.add_subplot(projection='3d')
ax.set_facecolor('white')
xs_axis = pcadataDf.loc[:,'PC1']
ys_axis = pcadataDf.loc[:,'PC2']
zs_axis = pcadataDf.loc[:,'PC3']
cmap_bold = ListedColormap(['#0173b2', '#de8f05', '#029e73', '#d55e00',
'#cc78bc'])
lines = [Line2D([0], [0], color=c, linewidth=0, linestyle=None, marker='o') for
c in colors]
labels =['Trichoderma','Aspergillus','Botrytis','Cladosporium','Verticillium']
plt.legend(frameon=False)
leg = plt.legend(lines,labels,loc='best')
```

```
leg.get_frame().set_color('white')
leg.get_frame().set_edgecolor('gainsboro')
ax.set_facecolor('white')
ax.spines['left'].set_color('grey')
ax.spines['bottom'].set_color('grey')
ax.spines['right'].set_color('lightgrey')
ax.spines['top'].set_color('lightgrey')
ax.set_xlabel('PC 1')
ax.set_ylabel('PC 2')
ax.set_zlabel('PC 3')
ax.grid(True, color='gainsboro')
plt.title('Cluster by PCA Components')
ax.scatter(xs_axis, ys_axis, zs_axis, marker ='o',c =
(labelnames.loc[:,'Index']),cmap=cmap_bold, s=40)
plt.tight_layout()
plt.show()
```

t-SNE

```
# First steps simultaneous to PCA, import data and packages, transpose and
normalize data, make labels

#import additional packages
from sklearn.manifold import TSNE
#perform tsne
n_components = 2
tsne = TSNE(n_components, verbose=2, perplexity=50,
n_iter=5000,learning_rate='auto', random_state=42)
tsne_result = tsne.fit_transform(z)
tsne_result.shape
tsne_result_df = pd.DataFrame({'tsne_1': tsne_result[:,0], 'tsne_2':
tsne_result[:,1], 'label': labels})
#make plot visualize tsne
fig = plt.figure(figsize=(8,8))
fig, ax = plt.subplots(1)
plt.title('Perplexity: 50')
colors = ['#0173b2', '#de8f05', '#029e73', '#d55e00', '#cc78bc']
sns.set_palette((colors))
from matplotlib.lines import Line2D
lines = [Line2D([0], [0], color=c, linewidth=0, linestyle=None, marker='o') for
c in colors]
labelsleg = ['Trichoderma', 'Aspergillus',
'Botrytis','Cladosporium','Verticillium']
p1=sns.scatterplot(x='tsne_1', y='tsne_2', hue='label', data=tsne_result_df,
ax=ax,s=100, palette=colors)
lim = (tsne_result.min()-5, tsne_result.max()+5)
x_axis = tsne_result_df.loc[:,'tsne_1']
y_axis = tsne_result_df.loc[:,'tsne_2']
import matplotlib.ticker as ticker
tick_spacing=1
ax.xaxis.set_major_locator(ticker.MultipleLocator(tick_spacing))
ax.yaxis.set_major_locator(ticker.MultipleLocator(tick_spacing))
ax.tick_params(axis='both',direction='out',color='gainsboro', width=2)
ax.set_xlabel('t-SNE 1')
ax.set_ylabel('t-SNE 2')
ax.set_aspect('equal')
ax.set_facecolor('white')
ax.spines['left'].set_color('grey')
ax.spines['bottom'].set_color('grey')
ax.spines['right'].set_color('lightgrey')
ax.spines['top'].set_color('lightgrey')
ax.grid(True, color='white')
ax.legend(lines, labelsleg, loc='best', facecolor='white',edgecolor='gainsboro',
bbox_to_anchor=(1,1))
```

```
plt.tight_layout()
plt.show()
```

k-Means

```
# First steps simultaneous to PCA, import data and packages, transpose and
normalize data, perform PCA

#import additional packages
from sklearn.cluster import KMeans
# elbow plot to check cluster size for kmeans after pca
wcss = []
for i in range(1,20):
    model=KMeans(n_clusters = i, init = "k-means++", random_state = 42, max_iter
= 1000, n_init = 50)
    model.fit(pcadataDf)
    wcss.append(model.inertia_)
fig = plt.figure(figsize=(10,8))
ax = fig.add_subplot()
ax.set_facecolor('white')
ax.spines['left'].set_color('dimgray')
ax.spines['bottom'].set_color('dimgray')
ax.spines['right'].set_color('darkgrey')
ax.spines['top'].set_color('darkgrey')
ax.grid(True, color='gainsboro')
plt.xlabel('Number of Clusters',fontsize = 15)
plt.ylabel('WCSS',fontsize = 15)
plt.title('Explained Variance by components: 10',fontsize = 15)
plt.plot(range(1,20),wcss)
plt.tight_layout()
plt.show()
# perform kmeans clustering
kmeans = KMeans(n_clusters = 5,                 # Set amount of clusters
                init = 'k-means++',             # Initialization method for
kmeans
                max_iter = 1000,                 # Maximum number of iterations
                n_init = 50,                    # Choose how often algorithm
will run with different centroid
                random_state = 42)               # Choose random state for
reproducibility
pred_y = kmeans.fit_predict(pcadataDf)
y_kmeans =kmeans.predict(pcadataDf)
# Plot the kmeans results
fig = plt.figure(figsize=(10,8))
ax = fig.add_subplot()
ax.set_facecolor('white')
ax.spines['left'].set_color('grey')
ax.spines['bottom'].set_color('grey')
ax.spines['right'].set_color('lightgrey')
ax.spines['top'].set_color('lightgrey')
ax.set_xlabel('PC 1')
ax.set_ylabel('PC 2')
ax.grid(True, color='gainsboro')
plt.tight_layout()
plt.scatter(pcadataDf.loc[:,'PC1'], pcadataDf.loc[:,'PC2'], c=y_kmeans, s=30,
cmap='Set1', label="y_means");
# Plot the clusters
plt.scatter(kmeans.cluster_centers_[:, 0],
            kmeans.cluster_centers_[:, 1],
            s=30, marker='x',                           # Set centroid size
            c='black', label="y_means");
plt.show()
```

DBSCAN

```
# First steps simultaneous to PCA, import data and packages, transpose and
normalize data, perform PCA

# import additional packages
from sklearn.cluster import DBSCAN
#perform DBSCAN
epsilon=40
db=DBSCAN(eps=epsilon, min_samples=3).fit(X)
core_samples_mask = np.zeros_like(db.labels_, dtype=bool)
core_samples_mask[db.core_sample_indices_] = True
labels=db.labels_
# Number of clusters in labels, ignoring noise if present.
n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
n_noise_ = list(labels).count(-1)
no_clusters = len(np.unique(labels) )
no_noise = np.sum(np.array(labels) == -1, axis=0)
print('Estimated no. of clusters: %d' % no_clusters)
print('Estimated no. of noise points: %d' % no_noise)
# Plot result
import matplotlib.pyplot as plt
fig = plt.figure(figsize=(10,8))
ax = fig.add_subplot()
ax.set_facecolor('white')
ax.spines['left'].set_color('grey')
ax.spines['bottom'].set_color('grey')
ax.spines['right'].set_color('lightgrey')
ax.spines['top'].set_color('lightgrey')
ax.grid(True, color='gainsboro')
plt.xlabel('pc1')
plt.ylabel('pc2')
x_axis = pcadataDf.loc[:,'PC1']
y_axis = pcadataDf.loc[:,'PC2']
p1 =sns.scatterplot(x_axis, y_axis, hue = (labelnames.loc[:,'Index']), palette =
"deep")
plt.title('Clusters by PCA Components')
plt.show()
# Black removed and is used for noise instead.
ax = fig.add_subplot()
ax.set_facecolor('white')
ax.spines['left'].set_color('grey')
ax.spines['bottom'].set_color('grey')
ax.spines['right'].set_color('lightgrey')
ax.spines['top'].set_color('lightgrey')
ax.grid(True, color='gainsboro')
ax.set_facecolor('white')
unique_labels = set(labels)
colors = [plt.cm.Spectral(each) for each in np.linspace(0, 1,
len(unique_labels))]
for k, col in zip(unique_labels, colors):
    if k == -1:
        # Black used for noise.
        col = [0, 0, 0, 1]
    ax = fig.add_subplot()
    ax.set_facecolor('white')
    ax.spines['left'].set_color('grey')
    ax.spines['bottom'].set_color('grey')
    ax.spines['right'].set_color('lightgrey')
    ax.spines['top'].set_color('lightgrey')
    ax.grid(True, color='gainsboro')
    class_member_mask = labels == k
    xy = X[class_member_mask & core_samples_mask]
```

```
    ax.set_facecolor('white')
    plt.plot(
        xy[:, 0],
        xy[:, 1],
        "o",
        markerfacecolor=tuple(col),
        markeredgecolor="k",
        markersize=5,
    )
    xy = X[class_member_mask & ~core_samples_mask]
    ax.set_facecolor('white')
    plt.plot(
        xy[:, 0],
        xy[:, 1],
        "o",
        markerfacecolor=tuple(col),
        markeredgecolor="k",
        markersize=5,
    )
ax.set_facecolor('white')
plt.xlabel('PC 1')
plt.ylabel('PC 2')
plt.title("PC5 Eps40 minsamp3", fontsize = 15)
plt.show()
```

Cross Validation, unstratified and stratified, with Support Vector Machine or kNN Classification for a chosen range of principal components

```
# First steps simultaneous to PCA, import data and packages, transpose and
normalize data
#settings for Support Vector Machine
from sklearn.svm import SVC
svm = SVC(kernel='linear', random_state=42, gamma='auto', C=0.01)
# setting for kNN
from sklearn import neighbors
n_neighbors = 3
clf = neighbors.KNeighborsClassifier(n_neighbors)
# make empty list to save later results
cv_scoress = []
AverageAccuracy =[]
# settings for cross validation and stratified kfold validation
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
cv = KFold(n_splits=10, random_state=42, shuffle=True)
skf = StratifiedKFold(n_splits=10, random_state=42, shuffle=True)
#import labels from df1
y = np.array(df1)
y = np.ravel(y)
### perform PCA with number of PC in range, perform cross validation either
unstratified or ### stratified for either kNN or SVM
for i in range(2, 54):
    pca_data = PCA(n_components=i)
    principalComponents_data = pca_data.fit_transform(z)
    # create Data Frame with all principal component values
    pcadataDf = pd.DataFrame(data=principalComponents_data)
    X = np.array(pcadataDf)
    cv_scores = cross_val_score(clf, X, y, scoring='accuracy', cv =skf)
    cv_scoress.append(cv_scores)
AverageAccuracy =[]
StandardDeviation =[]
# calculate mean accuracy and standard deviation, save in results
for scores in cv_scoress:
```

```
    AverageAccuracy.append(np.mean(scores))
    StandardDeviation.append(np.std(scores))

results = pd.DataFrame({'Accuracy':AverageAccuracy,
'StandardDeviation':StandardDeviation})
```

Grid search for best SVM parameters, beginning similar to cross validation,and PCA

```
cv = KFold(n_splits=20, random_state=42, shuffle=True)
from sklearn.model_selection import GridSearchCV
param_grid = {'C': [0.01, 0.1, 0.5, 1, 10, 100],
              'gamma': [1, 0.75, 0.5, 0.25, 0.1, 0.01, 0.001,'auto','scale'],
              'kernel': ['rbf', 'poly', 'linear']}
grid = GridSearchCV(SVC(random_state = 42), param_grid, scoring='accuracy',
cv=cv)
grid.fit(X_train, y_train)
best_params = grid.best_params_
print(f"Best params: {best_params}")
svm_clf = SVC(**best_params)
svm_clf.fit(X_train, y_train)
pred = svm_clf.predict(X_test)
print(f"Accuracy Score: {accuracy_score(y_test, pred) * 100:.2f}%")
```

Visualize classification, here SVM, kNN works the same

```
# First steps simultaneous to PCA, import data and packages, transpose and
normalize data, perform PCA

#perform classification
from sklearn.svm import SVC
svm = SVC(kernel='linear', random_state=0, gamma='auto', C=1)
svm.fit(X_train, y_train)
print('The accuracy of the svm classifier on training data is {:.2f} out of
1'.format(svm.score(X_train, y_train)))
print('The accuracy of the svm classifier on test data is {:.2f} out of
1'.format(svm.score(X_test, y_test)))
from sklearn.metrics import accuracy_score
# instantiate learning model (k = 3)
# predict the response
pred = svm.predict(X_test)
# evaluate accuracy
print("accuracy: {}".format(accuracy_score(y_test, pred)))
# make confusion matrix and visualize
from sklearn.metrics import classification_report, confusion_matrix
cm = (confusion_matrix(y_test, pred))
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred))
import matplotlib.pyplot as pl
cm_df =
pd.DataFrame(cm,index=['Trichoderma','Aspergillus','Botrytis','Cladosporium','Ve
rticillium'], columns
=['Trichoderma','Aspergillus','Botrytis','Cladosporium','Verticillium'])
import seaborn as sns
ax = sns.heatmap(cm_df, annot=True, cmap='coolwarm', vmax=1)
ax.set_title('Confusion Matrix with labels\n\n');
ax.set_xlabel('\nPredicted Genus')
ax.set_ylabel('Actual Genus ');
## Display the visualization of the Confusion Matrix.
plt.tight_layout()
plt.show()

#Visualize SVM results
from matplotlib.colors import ListedColormap
from sklearn import neighbors
```

```
# Create color maps
from mlxtend.plotting import plot_decision_regions
Xvorvis =X[:,0:2]
cmap_light = ListedColormap(['#46B0EA', '#ffb93f', '#34e7b5', '#fe9543',
'#f7b5eb'])
cmap_bold = ListedColormap(['#0173b2', '#de8f05', '#029e73', '#d55e00',
'#cc78bc'])
svm.fit(Xvorvis, y)
x_min, x_max = X[:, 0].min() - 5, X[:, 0].max() + 5
y_min, y_max = X[:, 1].min() - 5, X[:, 1].max() + 5
xx, yy = np.meshgrid(np.arange(x_min, x_max),
                     np.arange(y_min, y_max))
Z = svm.predict(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
fig = plt.figure()
ax = fig.add_subplot()
plt.pcolormesh(xx, yy, Z, cmap=cmap_light, shading='auto')
plt.scatter(X[:, 0], X[:, 1], c=y, cmap=cmap_bold)
ax.scatter(X_test[:, 0], X_test[:, 1], marker='.', color='k')
plt.xlim(xx.min(), xx.max())
plt.ylim(yy.min(), yy.max())
#colors = ['#0173b2','#48beff', '#de8f05', '#029e73', '#d55e00', '#cc78bc']
colors = ['#0173b2', '#de8f05', '#029e73', '#d55e00', '#cc78bc']
lines = [Line2D([0], [0], color=c, linewidth=0, linestyle=None, marker='o') for
c in colors]
#labels =['Trichoderma','Aspergillus','Botrytis','Cladosporium','Verticillium']
labels = ['T. harzianum','T. atroviride', 'T. fasciculatum','T.
longibrachiatum', 'T. minutisporum']
#labels =['T. harzianum 177', 'T. harzianum 178' ,'T. atroviride', 'T.
fasciculatum','T. longibrachiatum', 'T. minutisporum']
#labels = ['Trichoderma', 'Aspergillus',
'Botrytis','Cladosporium','Verticillium']
plt.legend(frameon=False)
leg = plt.legend(lines,labels)
leg.get_frame().set_color('white')
leg.get_frame().set_edgecolor('gainsboro')
plt.title('SVM Classification')
plt.xlabel('PC 1')
plt.ylabel('PC 2')
plt.tight_layout()
plt.show()
```

8.1.2. Supporting information for fungal class differentiation

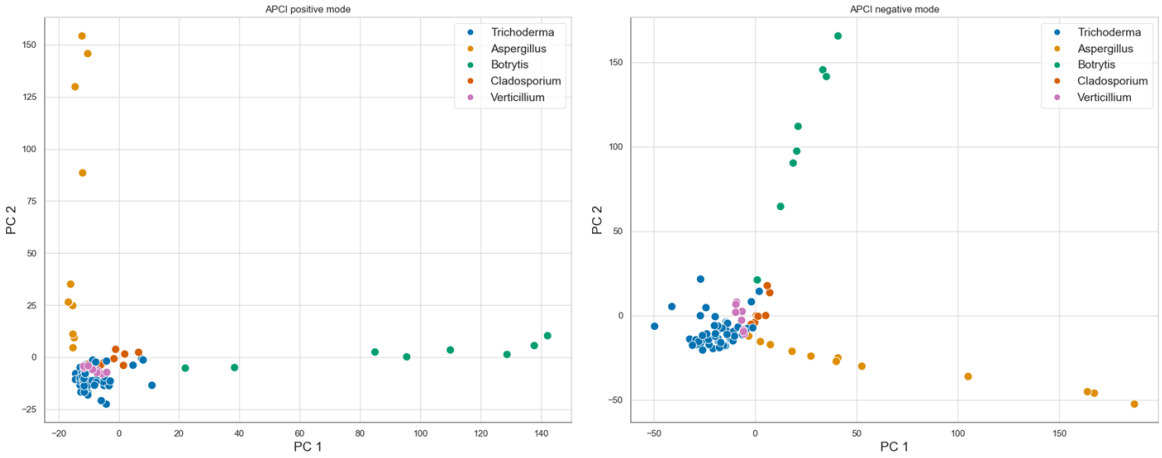


Figure 8.1: PCA for fungal classes. Upper picture: ACPI positive mode, lower picture: APCI negative mode.

Table 8.6: Tables for iterating principal component input to find the best number of PC for SVM and KNN. Stratified 10-fold cross-validation. ESI positive mode.

ESI positive		Dataset B class differentiation	Dataset A class differentation	*Trichoderma* species Differentiation
kNN, n = 3	PC input	15	9	8
	Accuracy	0.96	0.94	0.95
	Std Deviation	0.07	0.09	0.08
SVM linear Kernel C =0.01	PC input	9	10	16
	Accuracy	0.99	0.98	0.98
	Std Deviation	0.03	0.06	0.05

Table 8.7: Tables for iterating principal component input to find the best number of PC for SVM and KNN. Stratified 10-fold cross-validation. APCI positive mode.

APCI positive		Dataset B class differentiation	Dataset A class differentation	*Trichoderma* species Differentiation
kNN, n = 3	PC input	16	15	15
	Accuracy	0.94	0.92	0.78
	Std Deviation	0.08	0.08	0.13
SVM linear Kernel C =0.01	PC input	15	15	16
	Accuracy	0.99	0.94	0.90
	Std Deviation	0.03	0.07	0.13

Table 8.8: Tables for iterating principal component input to find the best number of PC for SVM and KNN. Stratified 10-fold cross-validation. ESI negative mode.

ESI negative		Dataset B class differentiation	Dataset A class differentation	*Trichoderma* species Differentiation
kNN, n = 3	PC input	30	12	12
	Accuracy	0.93	0.91	0.82
	Std Deviation	0.07	0.11	0.12
SVM linear Kernel C =0.01	PC input	33	20	31
	Accuracy	0.94	0.98	0.89
	Std Deviation	0.08	0.07	0.09

Table 8.9: Tables for iterating principal component input to find the best number of PC for SVM and KNN. Stratified 10-fold cross-validation. APCI negative mode.

APCI negative		Dataset B class differentiation	Dataset A class differentation	*Trichoderma* species Differentiation
kNN, n = 3	PC input	18	9	13
	Accuracy	0.97	0.97	0.87
	Std Deviation	0.05	0.07	0.13
SVM linear Kernel C =0.01	PC input	19	9	18
	Accuracy	97	0.95	0.97
	Std Deviation	0.04	0.07	0.07

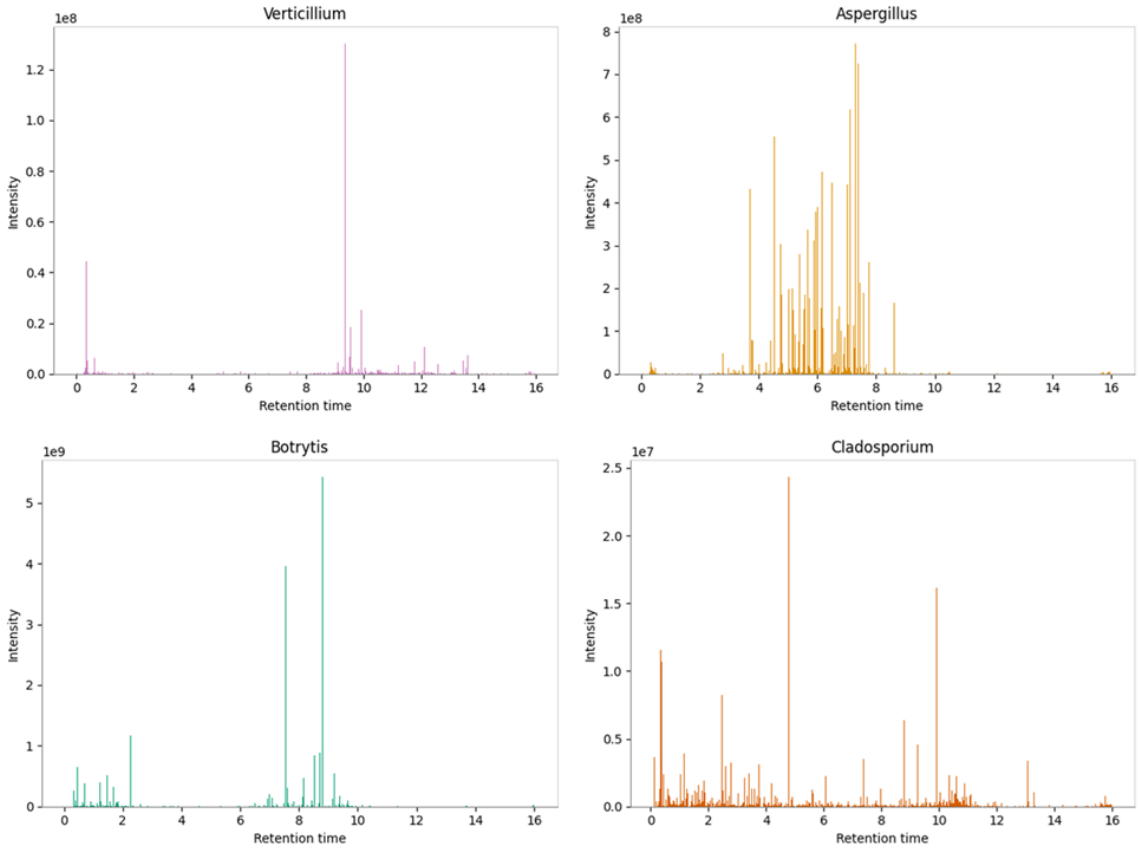


Figure 8.2: Fingerprints for species used for class differentiation. Averaged for all biological and technical replicates of the respective species. Maximum intensity varies for each species.

8.1.3. Supporting information for fungal species differentiation

Table 8.10: SVM classification (linear kernel C = 0.01). Stratified 10-fold cross-validation. Accuracy for the differentiation of *Trichoderma* spores on species level.

Ionization method	PC Input	Mean accuracy (Standard deviation)
ESI positive mode	16	0.98 (0.05)
APCI positive mode	16	0.90 (0.13)
ESI negative mode	31	0.89 (0.09)
APCI negative mode	18	0.97 (0.07)

Table 8.11: SVM, linear Kernel, C = 0.01, stratified 10-fold Cross Validation, *Trichoderma* Species and Strain Differentiation.

Ionization method	PC	% variance explained by PC	Accurary (Standard Deviation)
ESI positive mode	10	45	0.94 (0.13)
APCI positive mode	15	65	0.82 (0.12)
ESI negative mode	34	95	0.84 (0.09)
APCI negative mode	9	50	0.95 (0.08)

Table 8.12: kNN, n=3, stratified 10-fold Cross Validation, *Trichoderma* Species and Strain Differentiation.

Ionization method	PC	% variance explained by PC	Accurary (Std Dev)
ESI positive mode	11	48	0.91 (0.15)
APCI positive mode	26	80	0.75 (0.11)
ESI negative mode	11	70	0.58 (0.20)
APCI negative mode	23	75	0.81 (0.15)

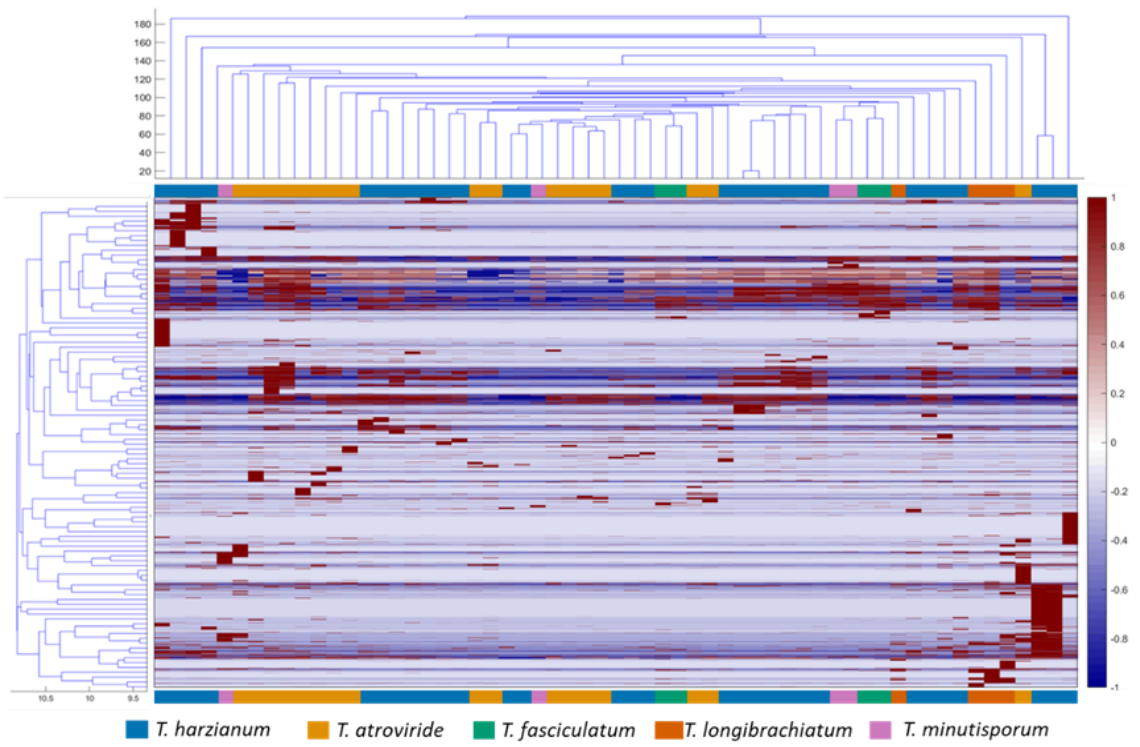


Figure 8.3: Hierarchical clustering analysis, *Trichoderma* species, APCI positive mode. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

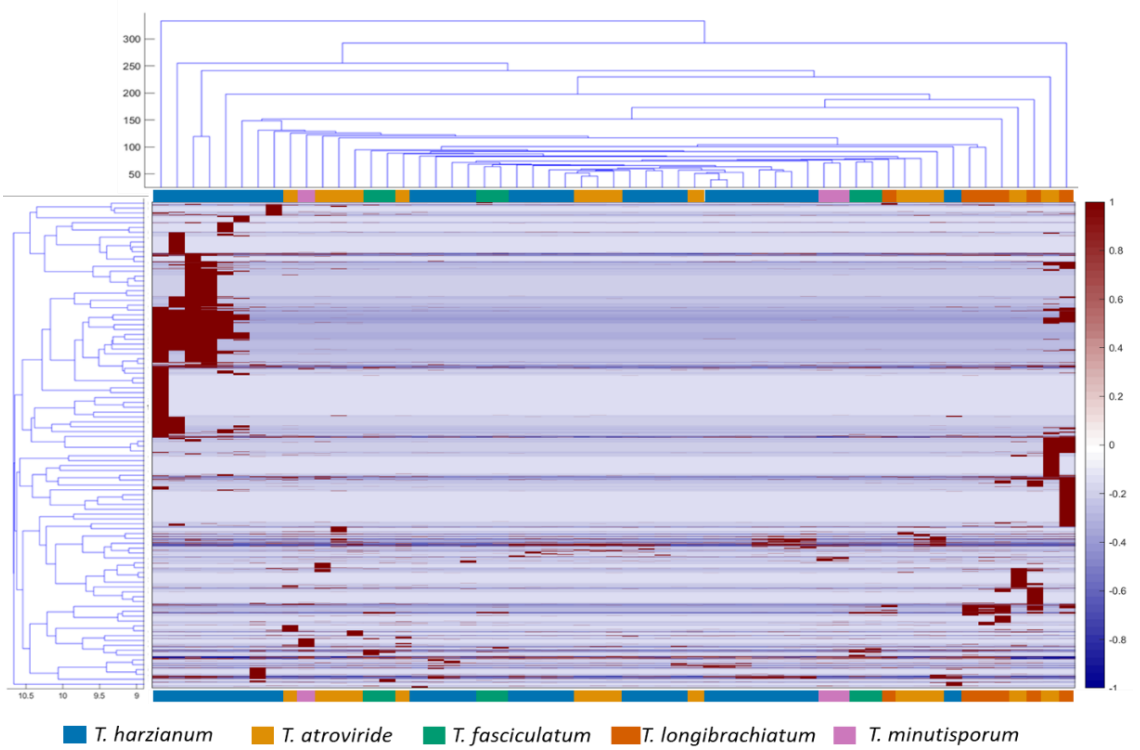


Figure 8.4: Hierarchical clustering analysis, *Trichoderma* species, ESI negative mode. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

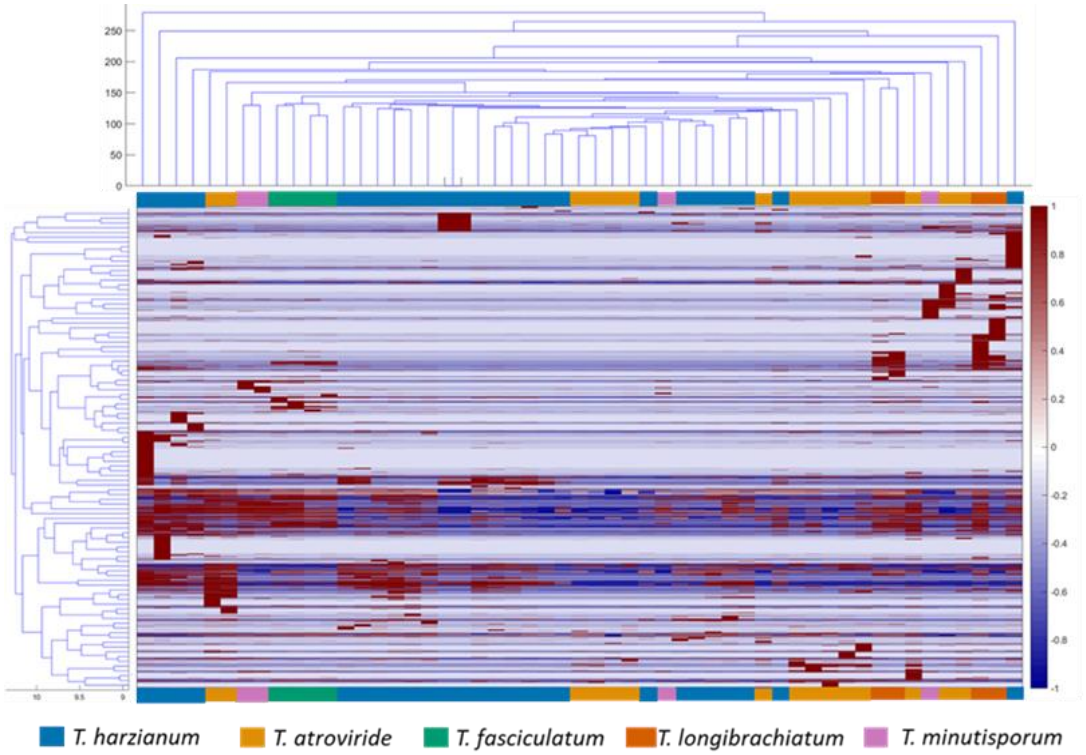


Figure 8.5: Hierarchical clustering analysis, *Trichoderma* species, APCI negative mode. The horizontal tree diagram represents the sample-wise clustering, and the vertical tree diagram the feature-wise clustering.

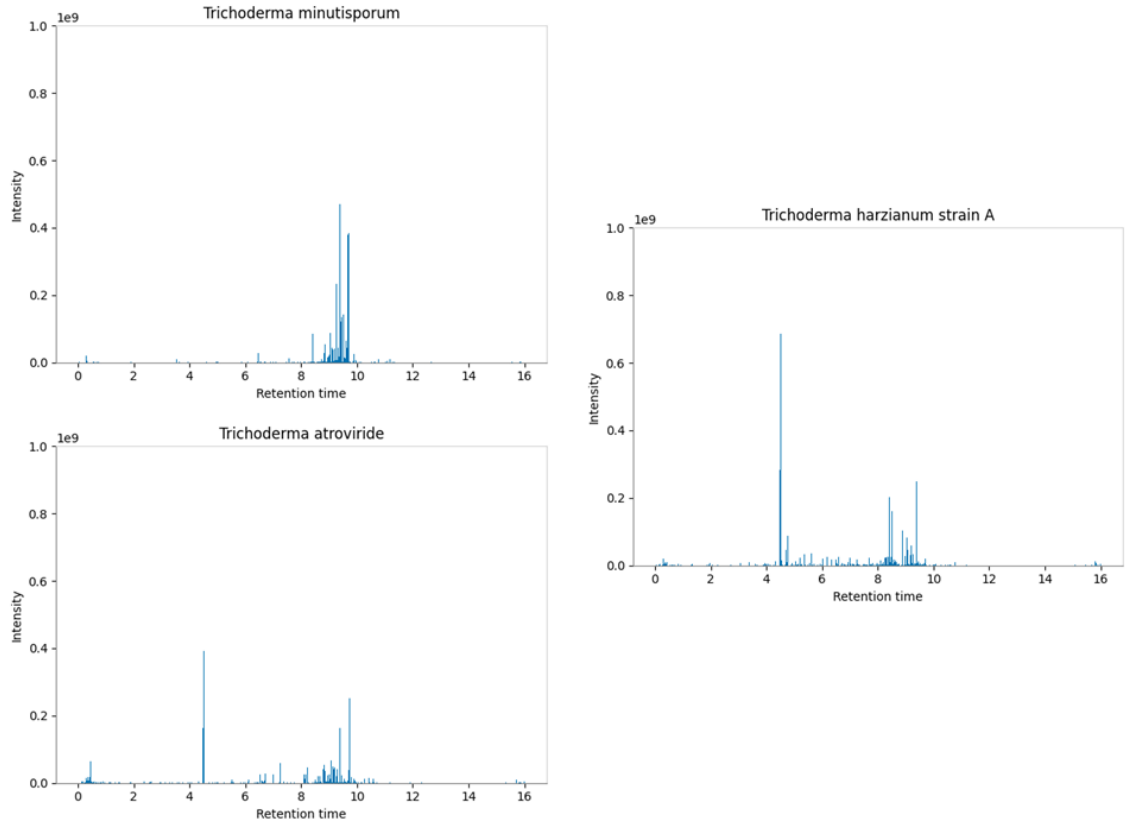


Figure 8.6: «Fingerprint» Spectra of *Trichoderma* spp. ESI positive mode. Fingerprint is obtained from averaging the signal intensity of all samples of the respective species.

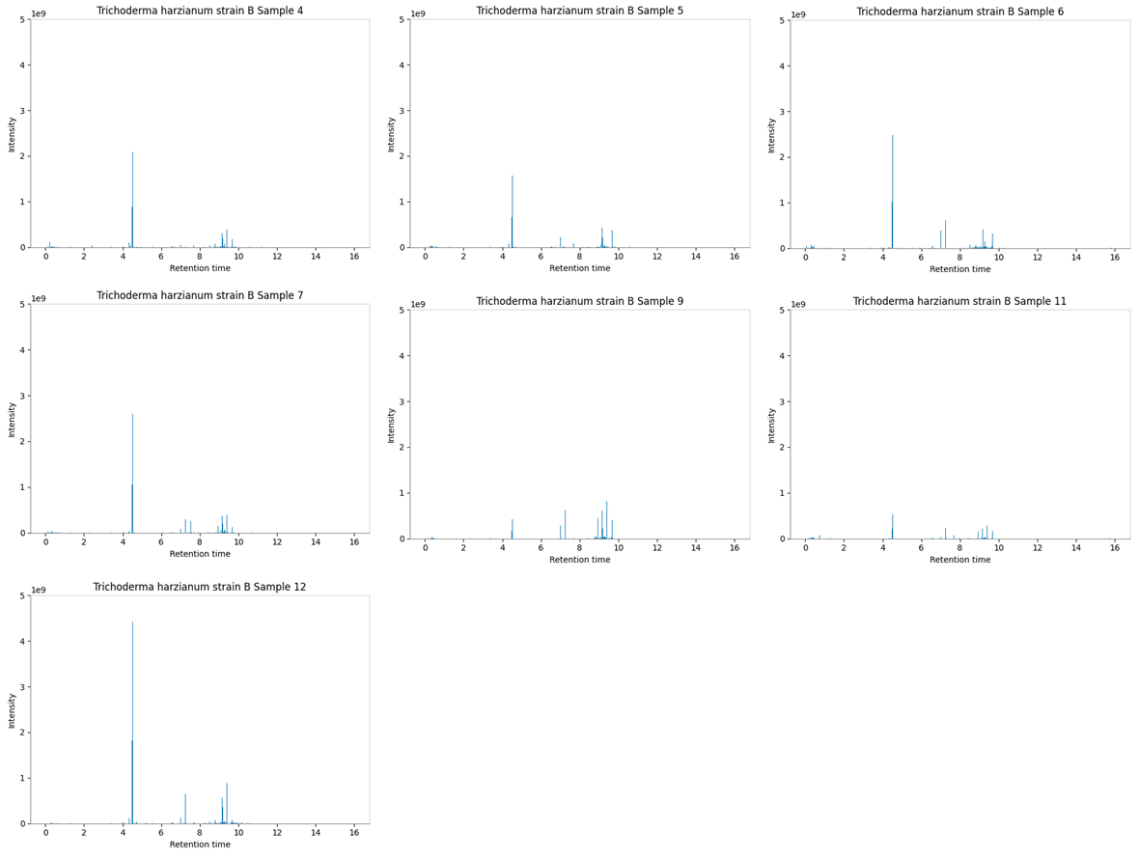


Figure 8.7: «Fingerprints» of *Trichoderma harzianum* strain B samples. Supplement to Figure 6.15. ESI positive mode.

8.1.4. Further supporting information

Experimental thermal desorption GC-MS

For testing of fungal VOC, *Trichoderma atroviride* (Vintec) was grown on four different growth media by the Thines Group (Hendrik Neumann). The growth media CM (Komplett-Medium/ Yeast Glucose Agar), PDA (Potato dextrose agar), OM (Oatmeal) and HMG (Hefe-Malz-Glucose) were tested. Tests were conducted with oatmeal and HMG to determine where the blank signal were lower and which sample showed more intense VOC signals. Standard Agar plates (9 cm) were placed open in a cleaned and sealed exicator and were kept for 30 minutes to allow an equilibration. After 30 minutes sampling was started with clean air (VOC and HEPA filtered) at 100 mL/min. Sampling took place for 30 min (total volume of 3 L) and capped afterwards. HMG showed more VOC signals and less blank signals than OM samples. To obtain better sample/blank ratios bigger samples were grown in Fernbach flask and a canister, all three on HMG medium at 20 °C and 26 °C (Fernbach and Kanister). A blank sample was incubated at 26 °C (Fernbach). The thermal desorption

system consists of an in-house build thermal-desorption device connected to a TRACE GC2000 gas chromatograph. The GC instrument is interfaced to a Polaris Q ion trap mass spectrometer equipped with an external electron ionization (EI) source (Thermo Fisher, San Jose, CA, USA). Previously self-build solid traps containing Carbotrap and Tenax were used, but testing showed that commercially available solid traps (MARKES Universal. C2/3-C30 and Odour/Sulfur. C6/7-C30, Inert-coated) showed better results, e.g., lower blank signals and therefore better limits of detection. The solid traps were desorbed for 10 min at 250 °C, the cooling trap was held at -196 °C, injection took place at 210 °C for 2 min. A Rxi® 5MS fused silica capillary column was installed (Crossbond® 5% diphenyl, 95% dimethyl-polysiloxane, 0.25 µm film thickness, 30 m × 0.25 mm i.d.) (Restek Corp., PA, USA) with Helium as carrier gas and a constant pre-column pressure of 0.5 bar. The ion source was operated at an electron energy of 70 eV and a temperature of 250 °C. The compound identification was performed by comparison with the mass spectral library of the GC/MS data system. Data were acquired and processed using Xcalibur software, version 1.2 (Thermo Fisher). Results were compared to the NIST98 Library. During measurements, a fluctuation in the split flow during desorption was detected, in extreme cases the split flow decreased to zero. Possible explanation is the complete freezing of the transfer capillary, due to excess moisture trapped in the solid traps, or other VOC which are present in excess amounts. When freezing occurred a shift in retention times was detected, most pronounced for the first few minutes of each chromatogram. This is most likely due to shift in the helium flow which could not be controlled externally. The helium flow reached its designated set point only after complete thawing of the transfer capillary, resulting in the extreme retention time shifts.

Experimental e-cigarette liquid

For DNPH derivatization see Kuntic et al 2020. DNPH-adducts of aldehyde standards and DNPH reaction mixtures with e- cigarette liquid or condensate were diluted and subjected to LC- MS analysis. LC-settings were as follows: Thermo Fisher Ultimate 3000 UHPLC; Column: Hypersil Gold C18 column (2.1 x 50 mm 1.9 µm); Eluent A: H2O (98%), acetonitrile (2%), formic acid (400 µL/L); Eluent B: acetonitrile (98%), H2O (2%). The following gradient for the mobile phases was applied: 0 min 2% B; 1 min 35% B; 3.5 min 50% B; 4.5 min 100% B; 6 min 100% B; 6.2 min 2% B. Flow rate was 0.5 mL/min and injection volume was 5 µL. MS-Settings were as follows: Thermo Fisher Orbitrap Exactive; daily calibration ensured a mass accuracy of below 1 ppm; Settings for sample measurements: Full MS; Scan Range 50.0 to 400.0 m/z; Resolution: 140000. Ion Source: Heated Electrospray Ionization ;

Polarity: Negative; Settings: Spray Voltage: 3.3 kV, Sheath Gas Flow Rate: 60 a.u., Auxiliary Gas Flow Rate: 20 a.u., Capillary temperature: 320 °C, Aux Gas heater temperature: 150 °C.

8.2. List of abbreviations

ACN	acetonitrile
APCI	atmospheric pressure chemical ionization
APPI	atmospheric pressure photo ionization
CBS	Centraalbureau voor Schimmelcultures- central bureau of fungal cultures
CCN	cloud condensation nuclei
CPC	condensation particle counter
DBSCAN	density-based spatial clustering of applications with noise
DNA	deoxyribonucleic acid
DSM	Deutsche Sammlung von Mikroorganismen und Zellkulturen – German Collection of microorganisms and cell cultures
EI	electron ionization
ESI	electrospray ionization
EtOAc	ethyl acetate
GC	gas chromatography
H/C	hydrogen to carbon ratio
HCA	hierarchical clustering analysis
HMG	Hefe-Malzextract Agar – yeast malt extract agar
HPLC	high performance liquid chromatography
HRMS	high resolution mass spectrometry
IN	ice nuclei
ITS	internal transcribed spacer
kNN	k- nearest neighbor
LC	liquid chromatography
LOD	limit of detection
LOQ	limit of quantitation
MeOH	methanol
MS	mass spectrometry
m/z	mass-to charge ratio
O/C	oxygen to carbon ratio

PBAP	primary biological aerosol particle
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction
PDA	potato dextrose agar
pg	picogram
ppm	parts per million
PM	particulate matter
PTFE	polytetrafluoroethylene
rbf	radial basis function
RNA	ribonucleic acid
RP	reverse phase
SOA	secondary organic aerosol
spp.	species pluralis
SVM	support vector machine
TEF1α	Translation elongation factor 1α
TIC	total ion count
Tg	teragram
t-SNE	t-distributed stochastic neighbor embedding
UHPLC	ultra high-performance liquid chromatography
VOC	volatile organic compound
WCSS	within cluster sum of square

8.3. List of Figures

Figure 1.1: Size ranges of major biological aerosols. 3
Figure 1.2: Simulated surface annual mean of fungal spore number concentration 6
Figure 1.3: Taxonomy of the kingdom true fungi, 7
Figure 1.4: Examples of secondary metabolites. 9
Figure 1.5: Diversity of airborne fungal spores, as examined by DNA analysis. 12
Figure 1.6: Explanatory workflow for metabolic profiling and fingerprinting. 16
Figure 2.1: Explanatory construction of an HPLC. 18
Figure 2.2: Explanatory scheme of gas chromatography 20
Figure 2.3: Suitable analyte polarity and molecular weight for different ionization sources 22
Figure 2.4: Construction of an electrospray ionization source. 23
Figure 2.5: Schematics of an APCI ionization source. 24
Figure 2.6: Positive and negative ionization formation in the APCI. 24
Figure 2.7: Schematics of an EI ionization source. 25
Figure 2.8: Stability diagram for a 3D ion trap, 26
Figure 2.9: Schematics of the Orbitrap Q Exactive Hybrid Quadrupol Orbitrap mass spectrometer. 27
Figure 2.10: Schematics of the Orbitrap mass analyzer 28
Figure 2.11: Zoom on nominal mass 251.7 in low (left) and high (right) mass resolution 29
Figure 2.12: Normalization strategies divided into sample-based and data-based approaches. 32
Figure 2.13: Explanatory scree plot 36
Figure 2.14: Exemplary score plot on the example of the Fisher Iris Data set. 37
Figure 2.15: Explanatory loading plot with four features with different correlations. 37
Figure 2.16: t-SNE protection of Fisher iris data set. 38
Figure 2.17: Explanatory heat map with tree diagrams to visualize hierarchical clustering results. 41
Figure 2.18: *Van Krevelen* plot with regions for biomolecules marked in color. 42
Figure 2.19: Elbow plot on the example of the Fisher iris data set. 43
Figure 2.20: k-means clustering for the Fisher iris data set. 43
Figure 2.21: Exemplary DBSCAN with core points 44
Figure 2.22: Example of a confusion matrix. 46
Figure 2.23: Hyperplanes separating two classes. 47
Figure 4.1: Tree illustrating relationships between groups of fungi. 53
Figure 4.2: Photograph of fungal cultures before harvest. 55

Figure 4.3: Workflow for sample preparation, extraction, measurement, and data processing of fungal spore samples. 58
Figure 5.1: *van Krevelen* plot for *Trichoderma harzianum* extracts with different solvents. 64
Figure 5.2: *van Krevelen* plot for *Trichoderma harzianum* extracts with different solvents. 65
Figure 5.3: Total ion counts of two samples 73
Figure 5.4: Scree plot showing the variance explained by principal components, 74
Figure 5.5: Visualization of the data as explained by the first two principal components. 75
Figure 5.6: Loading plot, all features represented by the first three principal components. 76
Figure 5.7: PCA (left) and t-SNE (right) for the fungal spore samples 77
Figure 5.8: WCSS plots for different input of principal components. 79
Figure 5.9: k-means clustering for low-dimensional input. 80
Figure 5.10: k-means clustering for high-dimensional input. 81
Figure 5.11: DBSCAN for inputs of different dimensionality. 82
Figure 5.12. kNN accuracy dependence on the dimensionality of input (number of principal components). 83
Figure 5.13: SVM accuracy dependence on the dimensionality of input (number of principal components) for three different kernels: 85
Figure 5.14: Comparison of stratified (red) and unstratified (blue) 10-fold cross-validation for SVM (left) and kNN (right), 86
Figure 6.1: Confusion matrices for the differentiation of fungal spore genera by SVM. 92
Figure 6.2: Confusion matrices for the differentiation of fungal spore genera by SVM. 93
Figure 6.3: *Van Krevelen* plots for all compounds detected in all fungal spore samples 98
Figure 6.4: HCA of fungal spore samples ionized by ESI positive mode. 99
Figure 6.5: HCA of fungal spore samples ionized by APCI positive mode. 100
Figure 6.6: HCA of fungal spore samples ionized by APCI negative mode. 101
Figure 6.7: HCA of fungal spore samples ionized by ESI negative mode. 102
Figure 6.8: Hierarchical clustering analysis for fungal spores of different classes. 103
Figure 6.9: Confusion matrices for the differentiation of fungal spore species of the genus Trichoderma 107
Figure 6.10: Confusion matrices for the differentiation of fungal spore species of the genus Trichoderma by SVM. 108
Figure 6.11: Hierarchical clustering analysis of *Trichoderma* species ionized by ESI in positive mode. 110
Figure 6.12: Fingerprint of features for *Trichoderma* averaged (upper left), *T. fasciculatum* (upper right), *T. longibrachiatum* (lower left), and *T. harzianum* strain B (lower right). 111

Figure 6.13: *Van Krevelen* plot for features present in *Trichoderma* genus, averaged over all species 112
Figure 6.14: Fingerprint of features of *T. harzianum* strain A and B. 113
Figure 6.15: Fingerprint of features for different samples of *T. harzianum* B. 114
Figure 6.16: Representative correlation matrices for SVM classification of fungal species and strains for *Trichoderma*. 116
Figure 6.17: Hierarchical clustering analysis of Basidiomycetes spores from ATTO site. 119
Figure 6.18: Hierarchical clustering analysis of Basidiomycetes spores from ATTO site. 120
Figure 6.19: Hierarchical clustering analysis of LC-HRMS analysis by ESI positive mode of e-cigarette liquids and condensates. 127
Figure 6.20: Hierarchical clustering analysis of LC-HRMS analysis by ESI negative mode of e-cigarette liquids and condensates. 128
Figure 8.1: PCA for fungal classes. 151
Figure 8.2: Fingerprints for species used for class differentiation. 153
Figure 8.3: Hierarchical clustering analysis, *Trichoderma* species, 155
Figure 8.4: Hierarchical clustering analysis, *Trichoderma* species, 155
Figure 8.5: Hierarchical clustering analysis*, Trichoderma* species, 156
Figure 8.6: «Fingerprint» Spectra of *Trichoderma* spp. ESI positive mode. 156
Figure 8.7: «Fingerprints» of *Trichoderma harzianum* strain B samples. 157

8.4. List of Tables

Table 1.1: Estimated aerosol emissions in teragrams per year 2
Table 1.2: Estimated PBAP emissions 4
Table 2.1: Normalization methods by formula. 34
Table 2.2: Typical linkage algorithms with an explanatory picture on the right 40
Table 2.3: Schematics of n-fold cross-validation, adapted from 46
Table 4.1: Table of ascomycetes used in this work. 52
Table 4.2: Table of species used in this work. 54
Table 4.3: Overview of biological replicates of fungal spore samples. 56
Table 4.4: List of basidiomycetes used in this work. 57
Table 4.5: Eluent gradient for UHPLC measurements. Eluent A; Water/acetonitrile, eluent B: Methanol. 59
Table 4.6: Composition of the quality and retention time control sample. 60
Table 4.7: Parameters for Orbitrap and ionization sources. 61

Table 5.1: Spore sizes and calculated approximate volumes for spores of different genera. 66
Table 5.2: Limit of Detection and limit of quantitation for ergosterol with HESI and APCI ionization 69
Table 5.3: Cophentic correlation factor for different linkage algorithms. 78
Table 5.4: Overview of biological replicates of fungal spore samples. 87
Table 5.5: Overview of available biological and technical replicates for each ionization method. 88
Table 5.6: Results of stratified 10-fold cross-validation for both datasets. 89
Table 5.7: Accuracy results for *Trichoderma* species differentiation with kNN and SVM. 89
Table 6.1: Mean Accuracy for classification of fungal genera by support vector machine classification. 91
Table 6.2: Results for the classification of validation samples. 94
Table 6.3: Composition of mixed samples for classification testing. 95
Table 6.4 Classification results for mixed samples. The classification was performed with SVM 96
Table 6.5: Classification results for mixed samples. The classification was performed with kNN 96
Table 6.6: SVM classification accuracy for the differentiation of *Trichoderma* spores on species level. 106
Table 6.7: Validation of the classification of *Trichoderma* spp. by support vector machine. 109
Table 6.8. SVM classification results for strain and species differentiation of *Trichoderma*. 115
Table 6.9: Fungal spore samples from the Amazonian rainforest with the taxonomic classification. 118
Table 6.10: VOC of *Trichoderma atroviride* 123
Table 6.11: Analysis of DNPH-aldehyde standards as well as DNPH derivatized e-cigarette liquids and condensates. 126
Table 8.1: Table of instruments and programs. 133
Table 8.2: Programs, programming languages, and libraries used for data processing. 134
Table 8.3: Table of chemicals and solvents. 134
Table 8.4: Parameters for Data Processing with MzMine 2.51. 135
Table 8.5: MATLAB and Python Code used in this work. 136
Table 8.6: Tables for iterating principal component input to find the best number of PC for SVM and KNN. 151

Table 8.7: Tables for iterating principal component input to find the best number of PC for SVM and KNN. 151
Table 8.8: Tables for iterating principal component input to find the best number of PC for SVM and KNN. 152
Table 8.9: Tables for iterating principal component input to find the best number of PC for SVM and KNN. 152
Table 8.10: SVM classification (linear kernel C = 0.01). 153
Table 8.11: SVM, linear Kernel, C = 0.01, stratified 10-fold Cross Validation, 154
Table 8.12: kNN, n=3, stratified 10-fold Cross Validation, 154

9. References

Abdelfattah, A.; Malacrinò, A.; Wisniewski, M.; Cacciola, S. O.; Schena, L. Metabarcoding: A powerful tool to investigate microbial communities and shape future plant protection strategies. *Biological Control* **2018,** *120*, 1–10.

Abrego, N.; Crosier, B.; Somervuo, P.; Ivanova, N.; Abrahamyan, A.; Abdi, A.; Hämäläinen, K.; Junninen, K.; Maunula, M.; Purhonen, J.; Ovaskainen, O. Fungal communities decline with urbanization-more in air than in soil. *The ISME journal* **2020,** *14* (11), 2806–2815. DOI: 10.1038/s41396-020-0732-1.

Adl, S. M.; Simpson, A. G. B.; Lane, C. E.; Lukeš, J.; Bass, D.; Bowser, S. S.; Brown, M. W.; Burki, F.; Dunthorn, M.; Hampl, V.; Heiss, A.; Hoppenrath, M.; Lara, E.; Le Gall, L.; Lynn, D. H.; McManus, H.; Mitchell, E. A. D.; Mozley-Stanridge, S. E.; Parfrey, L. W.; Pawlowski, J.; Rueckert, S.; Shadwick, L.; Schoch, C. L.; Smirnov, A.; Spiegel, F. W. The Revised Classification of Eukaryotes. *J. Eukaryot. Microbiol.* **2012,** *59* (5), 429–514. DOI: 10.1111/j.1550-7408.2012.00644.x.

Aliferis, K. A.; Cubeta, M. A.; Jabaji, S. Chemotaxonomy of fungi in the Rhizoctonia solani species complex performing GC/MS metabolite profiling. *Metabolomics* **2013,** *9* (S1), 159–169.

Almeida, F.; Rodrigues, M. L.; Coelho, C. The Still Underestimated Problem of Fungal Diseases Worldwide. *Front. Microbiol.* **2019,** *10*, 214. DOI: 10.3389/fmicb.2019.00214.

Andreae, M. O.; Rosenfeld, D. Aerosol–cloud–precipitation interactions. Part 1. The nature and sources of cloud-active aerosols. *Earth-Science Reviews* **2008,** *89* (1-2), 13–41. DOI: 10.1016/j.earscirev.2008.03.001.

Arthur, D.; Vassilvitskii, S. *k-means++: the advantages of careful seeding*; ACM-SIAM symposium on Discrete algorithms., 2007.

Barbara, D. J.; Clewes, E. Plant pathogenic Verticillium species: how many of them are there? *Molecular plant pathology* **2003,** *4* (4), 297–305. DOI: 10.1046/j.1364-3703.2003.00172.x.

Bates, K. H.; Jacob, D. J. A new model mechanism for atmospheric oxidation of isoprene: global effects on oxidants, nitrogen oxides, organic products, and secondary organic aerosol. *Atmos. Chem. Phys.* **2019,** *19* (14), 9613–9640. DOI: 10.5194/acp-19-9613-2019.

Bayram, O.; Braus, G. H. Coordination of secondary metabolism and development in fungi: the velvet family of regulatory proteins. *FEMS microbiology reviews* **2012,** *36* (1), 1–24. DOI: 10.1111/j.1574-6976.2011.00285.x.

Becker, P. T.; Bel, A. de; Martiny, D.; Ranque, S.; Piarroux, R.; Cassagne, C.; Detandt, M.; Hendrickx, M. Identification of filamentous fungi isolates by MALDI-TOF mass

spectrometry: clinical evaluation of an extended reference spectra library. *Medical mycology* **2014,** *52* (8), 826–834. DOI: 10.1093/mmy/myu064.

Begley, T. *Comprehensive Natural Products III*, 3rd ed.; Elsevier: San Diego, 2020.

Beni, A.; Soki, E.; Lajtha, K.; Fekete, I. An optimized HPLC method for soil fungal biomass determination and its application to a detritus manipulation study. *Journal of microbiological methods* **2014,** *103*, 124–130. DOI: 10.1016/j.mimet.2014.05.022.

Biosynthesis and Molecular Genetics of Fungal Secondary Metabolites, Volume 2; Zeilinger, S., Martín, J.-F., García-Estrada, C., Eds.; Fungal Biology; Springer New York: New York, NY, 2015a.

Bissett, J.; Szakacs, G.; Nolan, C. A.; Druzhinina, I.; Gradinger, C.; Kubicek, C. P. New species of Trichoderma from Asia. *Can. J. Bot.* **2003,** *81* (6), 570–586. DOI: 10.1139/b03-051.

Blekherman, G.; Laubenbacher, R.; Cortes, D. F.; Mendes, P.; Torti, F. M.; Akman, S.; Torti, S. V.; Shulaev, V. Bioinformatics tools for cancer metabolomics. *Metabolomics* **2011,** *7* (3), 329–343.

Bonfante, P.; Genre, A. Mechanisms underlying beneficial plant-fungus interactions in mycorrhizal symbiosis. *Nature communications* **2010,** *1*, 48. DOI: 10.1038/ncomms1046.

Boruta, T. Uncovering the repertoire of fungal secondary metabolites: From Fleming's laboratory to the International Space Station. *Bioengineered* **2018,** *9* (1), 12–16. DOI: 10.1080/21655979.2017.1341022.

Bouguettaya, A.; Yu, Q.; Liu, X.; Zhou, X.; Song, A. Efficient agglomerative hierarchical clustering. *Expert Systems with Applications* **2015,** *42* (5), 2785–2797.

Bowers, R. M.; Lauber, C. L.; Wiedinmyer, C.; Hamady, M.; Hallar, A. G.; Fall, R.; Knight, R.; Fierer, N. Characterization of airborne microbial communities at a high-elevation site and their potential to act as atmospheric ice nuclei. *Applied and environmental microbiology* **2009,** *75* (15), 5121–5130.

Brockman, S. A.; Roden, E. V.; Hegeman, A. D. Van Krevelen diagram visualization of high resolution-mass spectrometry metabolomics data with OpenVanKrevelen. *Metabolomics* **2018,** *14* (4), 1052. DOI: 10.1007/s11306-018-1343-y.

Buiarelli, F.; Canepari, S.; Di Filippo, P.; Perrino, C.; Pomata, D.; Riccardi, C.; Speziale, R. Extraction and analysis of fungal spore biomarkers in atmospheric bioaerosol by HPLC-MS-MS and GC-MS. *Talanta* **2013,** *105*, 142–151. DOI: 10.1016/j.talanta.2012.11.006.

Buiarelli, F.; Sonego, E.; Uccelletti, D.; Bruni, E.; Di Filippo, P.; Pomata, D.; Riccardi, C.; Perrino, C.; Marcovecchio, F.; Simonetti, G. Determination of the main bioaerosol components using chemical markers by liquid chromatography–tandem mass spectrometry. *Microchemical Journal* **2019,** *149*, 103974.

C.H. Chen; T.F. Hsieh; 陳俊宏; 謝廷芳. First report of Botrytis cinerea causing gray mold of Jamaica cherry in Taiwan. *1021-9544* **2009**.

Cai, F.; Druzhinina, I. S. In honor of John Bissett: authoritative guidelines on molecular identification of Trichoderma. *Fungal Diversity* **2021,** *107* (1), 1–69. DOI: 10.1007/s13225-020-00464-4.

Calvo, A. M.; Wilson, R. A.; Bok, J. W.; Keller, N. P. Relationship between Secondary Metabolism and Fungal Development. *Microbiology and Molecular Biology Reviews* **2002,** *66* (3), 447–459. DOI: 10.1128/MMBR.66.3.447–459.2002.

Caruana, D. J. Detection and analysis of airborne particles of biological origin: present and future. *The Analyst* **2011,** *136* (22), 4641–4652. DOI: 10.1039/c1an15506g.

Castrillo, J. I.; Oliver, S. G. *Yeast Systems Biology* 759; Humana Press: Totowa, NJ, 2011.

Chalupová, J.; Raus, M.; Sedlářová, M.; Sebela, M. Identification of fungal microorganisms by MALDI-TOF mass spectrometry. *Biotechnology advances* **2014,** *32* (1), 230–241. DOI: 10.1016/j.biotechadv.2013.11.002.

Cheng, T. Chemical evaluation of electronic cigarettes. *Tobacco control* **2014,** *23 Suppl 2*, ii11-7.

Cortesão, M.; Haas, A. de; Unterbusch, R.; Fujimori, A.; Schütze, T.; Meyer, V.; Moeller, R. Aspergillus niger Spores Are Highly Resistant to Space Radiation. *Front. Microbiol.* **2020,** *11*, 560. DOI: 10.3389/fmicb.2020.00560.

D Strycker, B.; Han, Z.; Commer, B.; D Shaw, B.; V Sokolov, A.; O Scully, M. CARS spectroscopy of Aspergillus nidulans spores. *Scientific reports* **2019,** *9* (1), 1789.

Daellenbach, K. R.; Kourtchev, I.; Vogel, A. L.; Bruns, E. A.; Jiang, J.; Petäjä, T.; Jaffrezo, J.-L.; Aksoyoglu, S.; Kalberer, M.; Baltensperger, U.; El Haddad, I.; Prévôt, A. S. H. Impact of anthropogenic and biogenic sources on the seasonal variation in the molecular composition of urban organic aerosols: a field and laboratory study using ultra-high-resolution mass spectrometry. *Atmos. Chem. Phys.* **2019,** *19* (9), 5973–5991.

Dales, R. E.; Munt, P. W. Farmer's Lung Disease. *Canadian Family Physician* **1982,** *28*, 1817–1820.

Deising, H. B.; Gase, I.; Kubo, Y. The unpredictable risk imposed by microbial secondary metabolites: how safe is biological control of plant diseases? *J Plant Dis Prot* **2017,** *124* (5), 413–419.

Després, V.; Huffman, J.; Burrows, S. M.; Hoose, C.; Safatov, A.; Buryak, G.; Fröhlich-Nowoisky, J.; Elbert, W.; Andreae, M.; Pöschl, U.; Jaenicke, R. Primary biological aerosol particles in the atmosphere: a review. *Tellus B: Chemical and Physical Meteorology* **2012,** *64* (1), 15598. DOI: 10.3402/tellusb.v64i0.15598.

Dexter, D. D.; van der Veen, J. M. Conformations of penicillin G: crystal structure of procaine penicillin G monohydrate and a refinement of the structure of potassium penicillin G. *J. Chem. Soc., Perkin Trans. 1* **1978,** *3*, 185–190. DOI: 10.1039/p19780000185.

Di Filippo, P.; Pomata, D.; Riccardi, C.; Buiarelli, F.; Perrino, C. Fungal contribution to size-segregated aerosol measured through biomarkers. *Atmospheric Environment* **2013,** *64*, 132–140. DOI: 10.1016/j.atmosenv.2012.10.010.

Dijksterhuis, J. Fungal spores: Highly variable and stress-resistant vehicles for distribution and spoilage. *Food microbiology* **2019,** *81*, 2–11. DOI: 10.1016/j.fm.2018.11.006.

Domingo-Almenara, X.; Siuzdak, G. Metabolomics Data Processing Using XCMS. *Methods in molecular biology (Clifton, N.J.)* [Online] **2020,** *2104*, 11–24. https://pubmed.ncbi.nlm.nih.gov/31953810/.

Du, X.; Smirnov, A.; Pluskal, T.; Jia, W.; Sumner, S. Metabolomics Data Preprocessing Using ADAP and MZmine 2. *Methods in molecular biology (Clifton, N.J.)* **2020,** *2104*, 25–48. DOI: 10.1007/978-1-0716-0239-3_3.

El Mubarak, M. A.; Danika, C.; Vlachos, N. S.; Farsalinos, K.; Poulas, K.; Sivolapenko, G. Development and validation of analytical methodology for the quantification of aldehydes in e-cigarette aerosols using UHPLC-UV. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association* **2018,** *116* (Pt B), 147–151.

Elbert, W.; Taylor, P. E.; Andreae, M. O.; Pöschl, U. Contribution of fungi to primary biogenic aerosols in the atmosphere: Wet and dry discharged spores, carbohydrates, and inorganic ions. *Atmos. Chem. Phys.* **2007,** *7* (17), 4569–4588. DOI: 10.5194/acp-7-4569-2007.

Enot, D. P.; Lin, W.; Beckmann, M.; Parker, D.; Overy, D. P.; Draper, J. Preprocessing, classification modeling and feature selection using flow injection electrospray mass spectrometry metabolite fingerprint data. *Nature protocols* **2008,** *3* (3), 446–470.

Erler, A.; Riebe, D.; Beitz, T.; Löhmannsröben, H.-G.; Grothusheitkamp, D.; Kunz, T.; Methner, F.-J. Characterization of volatile metabolites formed by molds on barley by mass and ion mobility spectrometry. *Journal of mass spectrometry : JMS* **2020,** *55* (5), e4501. DOI: 10.1002/jms.4501.

Farsalinos, K. E.; Voudris, V.; Poulas, K. E-cigarettes generate high levels of aldehydes only in 'dry puff' conditions. *Addiction (Abingdon, England)* **2015,** *110* (8), 1352–1356.

Feng, K.-C.; Liu, B.-L.; Tzeng, Y.-M. Morphological characterization and germination of aerial and submerged spores of the entomopathogenic fungus Verticillium lecanii. *World Journal of Microbiology and Biotechnology* **2002,** *18* (3), 217–224. DOI: 10.1023/A:1014933229314.

Feofilova, E. P.; Ivashechkin, A. A.; Alekhin, A. I.; Sergeeva, Y. E. Fungal spores: Dormancy, germination, chemical composition, and role in biotechnology (review). *Appl Biochem Microbiol* **2012,** *48* (1), 1–11.

Feussner, K.; Feussner, I. Comprehensive LC-MS-Based Metabolite Fingerprinting Approach for Plant and Fungal-Derived Samples. *Methods in molecular biology (Clifton, N.J.)* **2019,** *1978*, 167–185.

Filzmoser, P.; Walczak, B. What can go wrong at the data normalization step for identification of biomarkers? *Journal of chromatography. A* **2014,** *1362*, 194–205. DOI: 10.1016/j.chroma.2014.08.050.

Fischer, G.; Dott, W. Relevance of airborne fungi and their secondary metabolites for environmental, occupational and indoor hygiene. *Archives of microbiology* **2003,** *179* (2), 75–82. DOI: 10.1007/s00203-002-0495-2.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **1936,** *7* (2), 179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

Flora, J. W.; Wilkinson, C. T.; Sink, K. M.; McKinney, D. L.; Miller, J. H. Nicotine-related impurities in e-cigarette cartridges and refill e-liquids. *Journal of Liquid Chromatography & Related Technologies* **2016,** *39* (17-18), 821–829. DOI: 10.1080/10826076.2016.1266500.

Forsberg, E. M.; Huan, T.; Rinehart, D.; Benton, H. P.; Warth, B.; Hilmers, B.; Siuzdak, G. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. *Nature protocols* **2018,** *13* (4), 633–651.

Frochte, J. *Maschinelles Lernen*. *Grundlagen und Algorithmen in Python*, 2., aktualisierte Auflage; Hanser: München, 2019.

Fröhlich-Nowoisky, J.; Burrows, S. M.; Xie, Z.; Engling, G.; Solomon, P. A.; Fraser, M. P.; Mayol-Bracero, O. L.; Artaxo, P.; Begerow, D.; Conrad, R.; Andreae, M. O.; Després, V. R.; Pöschl, U. Biogeography in the air: fungal diversity over land and oceans. *Biogeosciences* **2012,** *9* (3), 1125–1136. DOI: 10.5194/bg-9-1125-2012.

Fröhlich-Nowoisky, J.; Kampf, C. J.; Weber, B.; Huffman, J. A.; Pöhlker, C.; Andreae, M. O.; Lang-Yona, N.; Burrows, S. M.; Gunthe, S. S.; Elbert, W.; Su, H.; Hoor, P.; Thines, E.; Hoffmann, T.; Després, V. R.; Pöschl, U. Bioaerosols in the Earth system: Climate, health, and ecosystem interactions. *Atmospheric Research* **2016,** *182*, 346–376. DOI: 10.1016/j.atmosres.2016.07.018.

Fröhlich-Nowoisky, J.; Pickersgill, D. A.; Després, V. R.; Pöschl, U. High diversity of fungi in air particulate matter. *Proceedings of the National Academy of Sciences of the United States of America* **2009,** *106* (31), 12814–12819. DOI: 10.1073/pnas.0811003106.

Gerstel. What is thermal desorption? https://www.gerstelus.com/what-is-thermal-desorption-td/.

Gey, M. H. *Instrumentelle Analytik und Bioanalytik;* Springer Berlin Heidelberg: Berlin, Heidelberg, 2015.

Ghosh, P. N.; Fisher, M. C.; Bates, K. A. Diagnosing Emerging Fungal Threats: A One Health Perspective. *Frontiers in genetics* **2018,** *9*, 376. DOI: 10.3389/fgene.2018.00376.

Goniewicz, M. L.; Knysak, J.; Gawron, M.; Kosmider, L.; Sobczak, A.; Kurek, J.; Prokopowicz, A.; Jablonska-Czapla, M.; Rosik-Dulewska, C.; Havel, C.; Jacob, P.; Benowitz, N. Levels of selected carcinogens and toxicants in vapour from electronic cigarettes. *Tobacco control* **2014,** *23* (2), 133–139. DOI: 10.1136/tobaccocontrol-2012-050859.

González-Riano, C.; Dudzik, D.; Garcia, A.; Gil-de-la-Fuente, A.; Gradillas, A.; Godzien, J.; López-Gonzálvez, Á.; Rey-Stolle, F.; Rojo, D.; Ruperez, F. J.; Saiz, J.; Barbas, C. Recent Developments along the Analytical Process for Metabolomics Workflows. *Anal. Chem.* **2020,** *92* (1), 203–226.

Gosselin, M. I.; Rathnayake, C. M.; Crawford, I.; Pöhlker, C.; Fröhlich-Nowoisky, J.; Schmer, B.; Després, V. R.; Engling, G.; Gallagher, M.; Stone, E.; Pöschl, U.; Huffman, J. A. Fluorescent bioaerosol particle, molecular tracer, and fungal spore concentrations during dry and rainy periods in a semi-arid forest. *Atmos. Chem. Phys.* **2016,** *16* (23), 15165–15184. DOI: 10.5194/acp-16-15165-2016.

Gotthardt, M.; Kanawati, B.; Schmidt, F.; Asam, S.; Hammerl, R.; Frank, O.; Hofmann, T.; Schmitt-Kopplin, P.; Rychlik, M. Comprehensive Analysis of the Alternaria Mycobolome Using Mass Spectrometry Based Metabolomics. *Molecular nutrition & food research* **2020,** *64* (3), e1900558. DOI: 10.1002/mnfr.201900558.

Griffin, D. W. Atmospheric movement of microorganisms in clouds of desert dust and implications for human health. *Clinical microbiology reviews* **2007,** *20* (3), 459-77, table of contents. DOI: 10.1128/CMR.00039-06.

Grigoriev, I. V.; Nikitin, R.; Haridas, S.; Kuo, A.; Ohm, R.; Otillar, R.; Riley, R.; Salamov, A.; Zhao, X.; Korzeniewski, F.; Smirnova, T.; Nordberg, H.; Dubchak, I.; Shabalov, I. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic acids research* **2014,** *42* (Database issue), D699-704. DOI: 10.1093/nar/gkt1183.

Grinn-Gofroń, A.; Nowosad, J.; Bosiacka, B.; Camacho, I.; Pashley, C.; Belmonte, J.; Linares, C. de; Ianovici, N.; Manzano, J. M. M.; Sadyś, M.; Skjøth, C.; Rodinkova, V.; Tormo-Molina, R.; Vokou, D.; Fernández-Rodríguez, S.; Damialis, A. Airborne Alternaria and Cladosporium fungal spores in Europe: Forecasting possibilities and relationships with meteorological parameters. *The Science of the total environment* **2019,** *653*, 938–946.

Gross, J. H. *Massenspektrometrie;* Springer Berlin Heidelberg: Berlin, Heidelberg, 2013.

Gummer, J. P. A.; Krill, C.; Du Fall, L.; Waters, O. D. C.; Trengove, R. D.; Oliver, R. P.; Solomon, P. S. Metabolomics protocols for filamentous fungi. *Methods in molecular biology (Clifton, N.J.)* **2012,** *835*, 237–254.

Guo, Y.; Jud, W.; Ghirardo, A.; Antritter, F.; Benz, J. P.; Schnitzler, J.-P.; Rosenkranz, M. Sniffing fungi - phenotyping of volatile chemical diversity in Trichoderma species. *The New phytologist* **2020a,** *227* (1), 244–259. DOI: 10.1111/nph.16530.

Guo, Y.; Jud, W.; Ghirardo, A.; Antritter, F.; Benz, J. P.; Schnitzler, J.-P.; Rosenkranz, M. Sniffing fungi - phenotyping of volatile chemical diversity in Trichoderma species. *The New phytologist* **2020b,** *227* (1), 244–259.

Gutarowska, B.; Skóra, J.; Pielech-Przybylska, K. Evaluation of ergosterol content in the air of various environments. *Aerobiologia* **2015,** *31* (1), 33–44. DOI: 10.1007/s10453-014-9344-4.

Haas, B. J.; Kamoun, S.; Zody, M. C.; Jiang, R. H. Y.; Handsaker, R. E.; Cano, L. M.; Grabherr, M.; Kodira, C. D.; Raffaele, S.; Torto-Alalibo, T.; Bozkurt, T. O.; Ah-Fong, A. M. V.; Alvarado, L.; Anderson, V. L.; Armstrong, M. R.; Avrova, A.; Baxter, L.; Beynon, J.; Boevink, P. C.; Bollmann, S. R.; Bos, J. I. B.; Bulone, V.; Cai, G.; Cakir, C.; Carrington, J. C.; Chawner, M.; Conti, L.; Costanzo, S.; Ewan, R.; Fahlgren, N.; Fischbach, M. A.; Fugelstad, J.; Gilroy, E. M.; Gnerre, S.; Green, P. J.; Grenville-Briggs, L. J.; Griffith, J.; Grünwald, N. J.; Horn, K.; Horner, N. R.; Hu, C.-H.; Huitema, E.; Jeong, D.-H.; Jones, A. M. E.; Jones, J. D. G.; Jones, R. W.; Karlsson, E. K.; Kunjeti, S. G.; Lamour, K.; Liu, Z.; Ma, L.; Maclean, D.; Chibucos, M. C.; McDonald, H.; McWalters, J.; Meijer, H. J. G.; Morgan, W.; Morris, P. F.; Munro, C. A.; O'Neill, K.; Ospina-Giraldo, M.; Pinzón, A.; Pritchard, L.; Ramsahoye, B.; Ren, Q.; Restrepo, S.; Roy, S.; Sadanandom, A.; Savidor, A.; Schornack, S.; Schwartz, D. C.; Schumann, U. D.; Schwessinger, B.; Seyer, L.; Sharpe, T.; Silvar, C.; Song, J.; Studholme, D. J.; Sykes, S.; Thines, M.; van de Vondervoort, P. J. I.; Phuntumart, V.; Wawra, S.; Weide, R.; Win, J.; Young, C.; Zhou, S.; Fry, W.; Meyers, B. C.; van West, P.; Ristaino, J.; Govers, F.; Birch, P. R. J.; Whisson, S. C.; Judelson, H. S.; Nusbaum, C. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. *Nature* **2009,** *461* (7262), 393–398. DOI: 10.1038/nature08358.

Hage, H.; Miyauchi, S.; Virágh, M.; Drula, E.; Min, B.; Chaduli, D.; Navarro, D.; Favel, A.; Norest, M.; Lesage-Meessen, L.; Bálint, B.; Merényi, Z.; Eugenio, L. de; Morin, E.; Martínez, A. T.; Baldrian, P.; Štursová, M.; Martínez, M. J.; Novotny, C.; Magnuson, J. K.; Spatafora, J. W.; Maurice, S.; Pangilinan, J.; Andreopoulos, W.; LaButti, K.; Hundley, H.; Na, H.; Kuo, A.; Barry, K.; Lipzen, A.; Henrissat, B.; Riley, R.; Ahrendt, S.; Nagy, L. G.; Grigoriev, I. V.; Martin, F.; Rosso, M.-N. Gene family expansions and transcriptome signatures uncover fungal

adaptations to wood decay. *Environ Microbiol* **2021,** *23* (10), 5716–5732. DOI: 10.1111/1462-2920.15423.

Harman, G. E. Overview of Mechanisms and Uses of Trichoderma spp. *Phytopathology* **2006,** *96* (2), 190–194. DOI: 10.1094/PHYTO-96-0190.

Harman, G. E.; Howell, C. R.; Viterbo, A.; Chet, I.; Lorito, M. Trichoderma species--opportunistic, avirulent plant symbionts. *Nature reviews. Microbiology* **2004,** *2* (1), 43–56. DOI: 10.1038/nrmicro797.

Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; Del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array programming with NumPy. *Nature* **2020,** *585* (7825), 357–362. DOI: 10.1038/s41586-020-2649-2.

Harris, D. C. *Lehrbuch der quantitativen Analyse*, 8. Auflage; Lehrbuch; Springer Spektrum: Berlin, Heidelberg, 2014.

Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics* **1979,** *28* (1), 100. DOI: 10.2307/2346830.

Hawksworth, D. L.; Lücking, R. Fungal Diversity Revisited: 2.2 to 3.8 Million Species. *Microbiology spectrum* **2017,** *5* (4). DOI: 10.1128/microbiolspec.FUNK-0052-2016.

Headley, J. V.; Peru, K. M.; Verma, B.; Robarts, R. D. Mass spectrometric determination of ergosterol in a prairie natural wetland. *Journal of Chromatography A* **2002,** *958* (1-2), 149–156. DOI: 10.1016/S0021-9673(02)00326-6.

Henderson-Begg, S. K.; Hill, T.; Thyrhaug, R.; Khan, M.; Moffett, B. F. Terrestrial and airborne non-bacterial ice nuclei. *Atmos. Sci. Lett.* **2009,** *4* (7), n/a-n/a. DOI: 10.1002/asl.241.

Hibbett, D. S.; Binder, M.; Bischoff, J. F.; Blackwell, M.; Cannon, P. F.; Eriksson, O. E.; Huhndorf, S.; James, T.; Kirk, P. M.; Lücking, R.; Thorsten Lumbsch, H.; Lutzoni, F.; Matheny, P. B.; McLaughlin, D. J.; Powell, M. J.; Redhead, S.; Schoch, C. L.; Spatafora, J. W.; Stalpers, J. A.; Vilgalys, R.; Aime, M. C.; Aptroot, A.; Bauer, R.; Begerow, D.; Benny, G. L.; Castlebury, L. A.; Crous, P. W.; Dai, Y.-C.; Gams, W.; Geiser, D. M.; Griffith, G. W.; Gueidan, C.; Hawksworth, D. L.; Hestmark, G.; Hosaka, K.; Humber, R. A.; Hyde, K. D.; Ironside, J. E.; Kõljalg, U.; Kurtzman, C. P.; Larsson, K.-H.; Lichtwardt, R.; Longcore, J.; Miadlikowska, J.; Miller, A.; Moncalvo, J.-M.; Mozley-Standridge, S.; Oberwinkler, F.; Parmasto, E.; Reeb, V.; Rogers, J. D.; Roux, C.; Ryvarden, L.; Sampaio, J. P.; Schüssler, A.; Sugiyama, J.; Thorn, R. G.; Tibell, L.; Untereiner, W. A.; Walker, C.; Wang, Z.; Weir, A.; Weiss, M.; White, M. M.; Winka, K.; Yao, Y.-J.; Zhang, N. A higher-level phylogenetic classification of the Fungi. *Mycological*

Research [Online] **2007,** *111* (Pt 5), 509–547. https://www.sciencedirect.com/science/article/pii/S0953756207000615.

Ho, C. S.; Lam, C. W. K.; Chan, M. H. M.; Cheung, R. C. K.; Law, L. K.; Lit, L. C. W.; Ng, K. F.; Suen, M. W. M.; Tai, H. L. Electrospray ionisation mass spectrometry: principles and clinical applications. *The Clinical biochemist. Reviews* **2003,** *24* (1), 3–12.

Hoffmann, E. de. *Mass Spectrometry. Principles and Applications*, 3rd ed.; John Wiley & Sons Incorporated: Newy York, 2013.

Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R. The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry : JMS* **2005,** *40* (4), 430–443. DOI: 10.1002/jms.856.

Huffman, J. A.; Perring, A. E.; Savage, N. J.; Clot, B.; Crouzy, B.; Tummon, F.; Shoshanim, O.; Damit, B.; Schneider, J.; Sivaprakasam, V.; Zawadowicz, M. A.; Crawford, I.; Gallagher, M.; Topping, D.; Doughty, D. C.; Hill, S. C.; Pan, Y. Real-time sensing of bioaerosols: Review and current perspectives. *Aerosol Science and Technology* **2019,** *78* (24), 1–31.

Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007,** *9* (3), 90–95. DOI: 10.1109/MCSE.2007.55.

Index Fungorum. *Species Fungorum for CoL+*, 2020. DOI: 10.15468/TS7WSB.

Jacobson, M. Z.; Streets, D. G. Influence of future anthropogenic emissions on climate, natural emissions, and air quality. *J. Geophys. Res.* **2009,** *114* (D8), 955. DOI: 10.1029/2008JD011476.

James, T. Y.; Kauff, F.; Schoch, C. L.; Matheny, P. B.; Hofstetter, V.; Cox, C. J.; Celio, G.; Gueidan, C.; Fraker, E.; Miadlikowska, J.; Lumbsch, H. T.; Rauhut, A.; Reeb, V.; Arnold, A. E.; Amtoft, A.; Stajich, J. E.; Hosaka, K.; Sung, G.-H.; Johnson, D.; O'Rourke, B.; Crockett, M.; Binder, M.; Curtis, J. M.; Slot, J. C.; Wang, Z.; Wilson, A. W.; Schüßler, A.; Longcore, J. E.; O'Donnell, K.; Mozley-Standridge, S.; Porter, D.; Letcher, P. M.; Powell, M. J.; Taylor, J. W.; White, M. M.; Griffith, G. W.; Davies, D. R.; Humber, R. A.; Morton, J. B.; Sugiyama, J.; Rossman, A. Y.; Rogers, J. D.; Pfister, D. H.; Hewitt, D.; Hansen, K.; Hambleton, S.; Shoemaker, R. A.; Kohlmeyer, J.; Volkmann-Kohlmeyer, B.; Spotts, R. A.; Serdani, M.; Crous, P. W.; Hughes, K. W.; Matsuura, K.; Langer, E.; Langer, G.; Untereiner, W. A.; Lücking, R.; Büdel, B.; Geiser, D. M.; Aptroot, A.; Diederich, P.; Schmitt, I.; Schultz, M.; Yahr, R.; Hibbett, D. S.; Lutzoni, F.; McLaughlin, D. J.; Spatafora, J. W.; Vilgalys, R. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **2006,** *443* (7113), 818–822. DOI: 10.1038/nature05110.

Janssen, R. H. H.; Heald, C. L.; Steiner, A. L.; Perring, A. E.; Huffman, J. A.; Robinson, E. S.; Twohy, C. H.; Ziemba, L. D. Drivers of the fungal spore bioaerosol budget: observational

analysis and global modeling. *Atmos. Chem. Phys.* [Online] **2021a,** *21* (6), 4381–4401. https://acp.copernicus.org/articles/21/4381/2021/.

Janssen, R. H. H.; Heald, C. L.; Steiner, A. L.; Perring, A. E.; Huffman, J. A.; Robinson, E. S.; Twohy, C. H.; Ziemba, L. D. Drivers of the fungal spore bioaerosol budget: observational analysis and global modeling. *Atmos. Chem. Phys.* **2021b,** *21* (6), 4381–4401. DOI: 10.5194/acp-21-4381-2021.

Kang, D. Chemotaxonomy of Trichoderma spp. Using Mass Spectrometry-Based. *J. Microbiol. Biotechnol.* **2011,** *21* (1), 5–13. DOI: 10.4014/jmb.1008.08018.

Keller, N. P. Fungal secondary metabolism: regulation, function and drug discovery. *Nature reviews. Microbiology* **2019,** *17* (3), 167–180. DOI: 10.1038/s41579-018-0121-1.

Keller, N. P.; Turner, G. *Fungal Secondary Metabolism* 944; Humana Press: Totowa, NJ, 2012.

Kew, W.; Blackburn, J. W. T.; Clarke, D. J.; Uhrín, D. Interactive van Krevelen diagrams - Advanced visualisation of mass spectrometry data of complex mixtures. *Rapid communications in mass spectrometry : RCM* **2017,** *31* (7), 658–662. DOI: 10.1002/rcm.7823.

Kim, K.-H.; Kabir, E.; Jahan, S. A. Airborne bioaerosols and their impact on human health. *Journal of environmental sciences (China)* **2018,** *67*, 23–35. DOI: 10.1016/j.jes.2017.08.027.

Kim, M.; Han, C.-H.; Lee, M.-Y. NADPH oxidase and the cardiovascular toxicity associated with smoking. *Toxicological research* **2014,** *30* (3), 149–157. DOI: 10.5487/TR.2014.30.3.149.

Kim, S.; Kramer, R. W.; Hatcher, P. G. Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram. *Anal. Chem.* **2003,** *75* (20), 5336–5344. DOI: 10.1021/ac034415p.

Kim, W.; Peever, T. L.; Park, J.-J.; Park, C.-M.; Gang, D. R.; Xian, M.; Davidson, J. A.; Infantino, A.; Kaiser, W. J.; Chen, W. Use of metabolomics for the chemotaxonomy of legume-associated Ascochyta and allied genera. *Scientific reports* **2016,** *6*, 20192. DOI: 10.1038/srep20192.

Kind, T.; Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics* **2007,** *8*, 105. DOI: 10.1186/1471-2105-8-105.

Kluger, B.; Lehner, S.; Schuhmacher, R. Metabolomics and Secondary Metabolite Profiling of Filamentous Fungi. In *Biosynthesis and Molecular Genetics of Fungal Secondary Metabolites, Volume 2*; Zeilinger, S., Martín, J.-F., García-Estrada, C., Eds.; Fungal Biology; Springer New York: New York, NY, 2015; pp 81–101. DOI: 10.1007/978-1-4939-2531-5_6.

Koutroumbas, K.; Theodoridis, S. *Pattern recognition*, 4th ed.; Academic Press: Burlington, MA, London, 2010.

Krug, H. F.; Wick, P. Nanotoxicology: An interdisciplinary challenge. *Angewandte Chemie (International ed. in English)* **2011,** *50* (6), 1260–1278. DOI: 10.1002/anie.201001037.

Kruve, A. Semi-quantitative non-target analysis of water with liquid chromatography/high-resolution mass spectrometry: How far are we? *Rapid communications in mass spectrometry : RCM* **2019,** *33 Suppl 3*, 54–63. DOI: 10.1002/rcm.8208.

Kumar, A.; Attri, A. K. Characterization of fungal spores in ambient particulate matter: A study from the Himalayan region. *Atmospheric Environment* **2016,** *142*, 182–193. DOI: 10.1016/j.atmosenv.2016.07.049.

Kuntic, M.; Oelze, M.; Steven, S.; Kröller-Schön, S.; Stamm, P.; Kalinovic, S.; Frenis, K.; Vujacic-Mirski, K.; Bayo Jimenez, M. T.; Kvandova, M.; Filippou, K.; Al Zuabi, A.; Brückl, V.; Hahad, O.; Daub, S.; Varveri, F.; Gori, T.; Huesmann, R.; Hoffmann, T.; Schmidt, F. P.; Keaney, J. F.; Daiber, A.; Münzel, T. Short-term e-cigarette vapour exposure causes vascular oxidative stress and dysfunction: evidence for a close connection to brain damage and a key role of the phagocytic NADPH oxidase (NOX-2). *European Heart Journal* **2020,** *41* (26), 2472–2483. DOI: 10.1093/eurheartj/ehz772.

Kvakkestad, V.; Sundbye, A.; Gwynn, R.; Klingen, I. Authorization of microbial plant protection products in the Scandinavian countries: A comparative analysis. *Environmental Science & Policy* **2020,** *106*, 115–124.

Lau, A. F.; Drake, S. K.; Calhoun, L. B.; Henderson, C. M.; Zelazny, A. M. Development of a clinically comprehensive database and a simple procedure for identification of molds from solid media by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Journal of Clinical Microbiology* **2013,** *51* (3), 828–834. DOI: 10.1128/JCM.02852-12.

Lau, A. P.; Lee, A. K.; Chan, C. K.; Fang, M. Ergosterol as a biomarker for the quantification of the fungal biomass in atmospheric aerosols. *Atmospheric Environment* **2006,** *40* (2), 249–259. DOI: 10.1016/j.atmosenv.2005.09.048.

Lee, A. K.; Lau, A. P.; Cheng, J. Y.; Fang, M.; Chan, C. K. Source identification analysis for the airborne bacteria and fungi using a biomarker approach. *Atmospheric Environment* **2007,** *41* (13), 2831–2843. DOI: 10.1016/j.atmosenv.2006.11.047.

Li, T.-Y.; Liu, B.-H.; Chen, Y.-C. Characterization ofAspergillus spores by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2000,** *14* (24), 2393–2400. DOI: 10.1002/1097-0231(20001230)14:24<2393:AID-RCM178>3.0.CO;2-9.

Li, Y.; Steenwyk, J. L.; Chang, Y.; Wang, Y.; James, T. Y.; Stajich, J. E.; Spatafora, J. W.; Groenewald, M.; Dunn, C. W.; Hittinger, C. T.; Shen, X.-X.; Rokas, A. A genome-scale phylogeny of the kingdom Fungi. *Current biology : CB* **2021,** *31* (8), 1653-1665.e5. DOI: 10.1016/j.cub.2021.01.074.

Liebal, U. W.; Phan, A. N. T.; Sudhakar, M.; Raman, K.; Blank, L. M. Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites* **2020,** *10* (6). DOI: 10.3390/metabo10060243.

Linstrom, Peter. *NIST Chemistry WebBook, NIST Standard Reference Database 69*, 1997. DOI: 10.18434/T4D303.

Livera, A. M. de; Dias, D. A.; Souza, D. de; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T. P. Normalizing and integrating metabolomics data. *Anal. Chem.* **2012,** *84* (24), 10768–10776. DOI: 10.1021/ac302748b.

Lohmann, U.; Feichter, J. Global indirect aerosol effects: a review. *Atmos. Chem. Phys.* **2005,** *5* (3), 715–737. DOI: 10.5194/acp-5-715-2005.

Lücking, R.; Aime, M. C.; Robbertse, B.; Miller, A. N.; Ariyawansa, H. A.; Aoki, T.; Cardinali, G.; Crous, P. W.; Druzhinina, I. S.; Geiser, D. M.; Hawksworth, D. L.; Hyde, K. D.; Irinyi, L.; Jeewon, R.; Johnston, P. R.; Kirk, P. M.; Malosso, E.; May, T. W.; Meyer, W.; Öpik, M.; Robert, V.; Stadler, M.; Thines, M.; Vu, D.; Yurkov, A. M.; Zhang, N.; Schoch, C. L. Unambiguous identification of fungi: where do we stand and how accurate and precise is fungal DNA barcoding? *IMA fungus* **2020,** *11*, 14. DOI: 10.1186/s43008-020-00033-z.

Maciá-Vicente, J. G.; Shi, Y.-N.; Cheikh-Ali, Z.; Grün, P.; Glynou, K.; Kia, S. H.; Piepenbring, M.; Bode, H. B. Metabolomics-based chemotaxonomy of root endophytic fungi for natural products discovery. *Environ Microbiol* **2018,** *20* (3), 1253–1270.

Madelin, T. M. Fungal aerosols: A review. *Journal of Aerosol Science* **1994,** *25* (8), 1405–1412.

Madla, S.; Miura, D.; Wariishi, H. Optimization of Extraction Method for GC-MS based Metabolomics for Filamentous Fungi. *J Microbial Biochem Technol* **2012,** *04* (01).

Margham, J.; McAdam, K.; Forster, M.; Liu, C.; Wright, C.; Mariner, D.; Proctor, C. Chemical Composition of Aerosol from an E-Cigarette: A Quantitative Comparison with Cigarette Smoke. *Chemical research in toxicology* **2016,** *29* (10), 1662–1678. DOI: 10.1021/acs.chemrestox.6b00188.

Marques, P.; Piqueras, L.; Sanz, M.-J. An updated overview of e-cigarette impact on human health. *Respiratory research* **2021,** *22* (1), 151. DOI: 10.1186/s12931-021-01737-5.

Martin Müller. *Bestimmung von Ergosterol als Markersubstanz für Pilzsporen*: Mainz, 2016.

Matsuyama, S., Wasada, N. Spectral Database for Organic Compounds, SDBS. https:// sdbs.db.aist.go.jp/ (accessed April 6, 2022).

McLaughlin, D. J.; Hibbett, D. S.; Lutzoni, F.; Spatafora, J. W.; Vilgalys, R. The search for the fungal tree of life. *Trends in Microbiology* **2009,** *17* (11), 488–497. DOI: 10.1016/j.tim.2009.08.001.

Meinicke, P.; Lingner, T.; Kaever, A.; Feussner, K.; Göbel, C.; Feussner, I.; Karlovsky, P.; Morgenstern, B. Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms for Molecular Biology : AMB* **2008,** *3*, 9. DOI: 10.1186/1748-7188-3-9.

Merkl, R. *Bioinformatik*. *Grundlagen, Algorithmen, Anwendungen*, 3. vollst. überarb. u. erw. Auflage; EBL-Schweitzer; Wiley-VCH: Weinheim, 2015.

Merluzzi, V. J. *The Search for Anti-Inflammatory Drugs*. *Case Histories from Concept to Clinic;* Birkhäuser Boston: Boston, MA, 1995.

Meyer, W.; Irinyi, L.; Hoang, M. T. V.; Robert, V.; Garcia-Hermoso, D.; Desnos-Ollivier, M.; Yurayart, C.; Tsang, C.-C.; Lee, C.-Y.; Woo, P. C. Y.; Pchelin, I. M.; Uhrlaß, S.; Nenoff, P.; Chindamporn, A.; Chen, S.; Hebert, P. D. N.; Sorrell, T. C. Database establishment for the secondary fungal DNA barcode translational elongation factor 1α (TEF1α) 1. *Genome* **2019,** *62* (3), 160–169. DOI: 10.1139/gen-2018-0083.

Miller, J. D.; Young, J. C. The use of ergosterol to measure exposure to fungal propagules in indoor air. *American Industrial Hygiene Association journal* **1997,** *58* (1), 39–43. DOI: 10.1080/15428119791013062.

Misra, B. B. Data normalization strategies in metabolomics: Current challenges, approaches, and tools. *European journal of mass spectrometry (Chichester, England)* **2020,** *26* (3), 165–174. DOI: 10.1177/1469066720918446.

Moore, D.; Robson, G. D.; Trinci, A. P. J. *21st century guidebook to fungi*, Second edition; Cambridge University Press: Cambridge, New York, Melbourne, New Delhi, Singapore, 2020.

Moularat, S.; Robine, E.; Ramalho, O.; Oturan, M. A. Detection of fungal development in a closed environment through the identification of specific VOC: demonstration of a specific VOC fingerprint for fungal development. *The Science of the total environment* **2008,** *407* (1), 139–146.

Müller, A.; Faubert, P.; Hagen, M.; Castell, W. zu; Polle, A.; Schnitzler, J.-P.; Rosenkranz, M. Volatile profiles of fungi--chemotyping of species and ecological functions. *Fungal genetics and biology : FG & B* **2013,** *54*, 25–33.

Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: an overview. *WIREs Data Mining Knowl Discov* **2012,** *2* (1), 86–97.

Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak

Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Anal. Chem.* **2017,** *89* (17), 8689–8695. DOI: 10.1021/acs.analchem.7b01069.

Naranjo-Ortiz, M. A.; Gabaldón, T. Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biological reviews of the Cambridge Philosophical Society* **2019,** *94* (6), 2101–2137. DOI: 10.1111/brv.12550.

Nicol, J.; Fraser, R.; Walker, L.; Liu, C.; Murphy, J.; Proctor, C. J. Comprehensive Chemical Characterization of the Aerosol Emissions of a Vaping Product Based on a New Technology. *Chemical research in toxicology* **2020,** *33* (3), 789–799. DOI: 10.1021/acs.chemrestox.9b00442.

Nizkorodov, S. A.; Laskin, J.; Laskin, A. Molecular chemistry of organic aerosols through the application of high resolution mass spectrometry. *Physical chemistry chemical physics : PCCP* [Online] **2011,** *13* (9), 3612–3629. https://pubs.rsc.org/en/content/articlehtml/2011/cp/c0cp02032j.

Oliveira, M.; Ribeiro, H.; Delgado, L.; Fonseca, J.; Castel-Branco, M. G.; Abreu, I. Outdoor allergenic fungal spores: comparison between an urban and a rural area in northern Portugal. *Journal of investigational allergology & clinical immunology* **2010,** *20* (2), 117–128.

Ovaskainen, O.; Abrego, N.; Somervuo, P.; Palorinne, I.; Hardwick, B.; Pitkänen, J.-M.; Andrew, N. R.; Niklaus, P. A.; Schmidt, N. M.; Seibold, S.; Vogt, J.; Zakharov, E. V.; Hebert, P. D. N.; Roslin, T.; Ivanova, N. V. Monitoring Fungal Communities With the Global Spore Sampling Project. *Front. Ecol. Evol.* **2020,** *7*.

Overy, D. P.; Bayman, P.; Kerr, R. G.; Bills, G. F. An assessment of natural product discovery from marine (sensu strictu) and marine-derived fungi. *Mycology* **2014,** *5* (3), 145–167. DOI: 10.1080/21501203.2014.931308.

Pace, L.; Boccacci, L.; Casilli, M.; Fattorini, S. Temporal variations in the diversity of airborne fungal spores in a Mediterranean high altitude site. *Atmospheric Environment* **2019,** *210*, 166–170.

Pandey, R.; Usui, K.; Livingstone, R. A.; Fischer, S. A.; Pfaendtner, J.; Backus, E. H. G.; Nagata, Y.; Fröhlich-Nowoisky, J.; Schmüser, L.; Mauri, S.; Scheel, J. F.; Knopf, D. A.; Pöschl, U.; Bonn, M.; Weidner, T. Ice-nucleating bacteria control the order and dynamics of interfacial water. *Science advances* **2016,** *2* (4), e1501630. DOI: 10.1126/sciadv.1501630.

Pasanen; Yli-Pietila; Kalliokoski; Tarhanen. Ergosterol content in various fungal species and biocontaminated building materials. *Applied and environmental microbiology* **1999,** *65* (1), 138–142.

Payne, J. W.; Jakes, R.; Hartley, B. S. The primary structure of alamethicin. *The Biochemical journal* **1970,** *117* (4), 757–766. DOI: 10.1042/bj1170757.

Pei, Z.; Zhuang, Z.; Sang, H.; Wu, Z.; Meng, R.; He, E. Y.; Scott, G. I.; Maris, J. R.; Li, R.; Ren, J. α,β-Unsaturated aldehyde crotonaldehyde triggers cardiomyocyte contractile dysfunction: role of TRPV1 and mitochondrial function. *Pharmacological research* **2014,** *82*, 40–50. DOI: 10.1016/j.phrs.2014.03.010.

Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics* **2010,** *11*, 395. DOI: 10.1186/1471-2105-11-395.

Pöschl, U. Atmospheric aerosols: Composition, transformation, climate and health effects. *Angewandte Chemie (International ed. in English)* **2005,** *44* (46), 7520–7540. DOI: 10.1002/anie.200501122.

Pöschl, U.; Shiraiwa, M. Multiphase chemistry at the atmosphere-biosphere interface influencing climate and public health in the anthropocene. *Chemical reviews* **2015,** *115* (10), 4440–4475. DOI: 10.1021/cr500487s.

Principal component analysis; Jolliffe, I. T., Ed., 2. ed.; Springer Series in Statistics; Springer: New York, NY, 2002.

Processing Metabolomics and Proteomics Data with Open Software; Winkler, R., Ed.; New Developments in Mass Spectrometry; Royal Society of Chemistry: Cambridge, 2020.

Qasim, H.; Karim, Z. A.; Rivera, J. O.; Khasawneh, F. T.; Alshbool, F. Z. Impact of Electronic Cigarettes on the Cardiovascular System. *Journal of the American Heart Association* **2017,** *6* (9). DOI: 10.1161/JAHA.117.006353.

Qin, Q.-M.; Vallad, G. E.; Subbarao, K. V. Characterization of Verticillium dahliae and V. tricorpus Isolates from Lettuce and Artichoke. *Plant disease* **2008,** *92* (1), 69–77. DOI: 10.1094/PDIS-92-1-0069.

Rämä, T.; Quandt, C. A. Improving Fungal Cultivability for Natural Products Discovery. *Front. Microbiol.* **2021,** *12*, 706044. DOI: 10.3389/fmicb.2021.706044.

Rehan, H. S.; Maini, J.; Hungin, A. P. S. Vaping versus Smoking: A Quest for Efficacy and Safety of E-cigarette. *Current drug safety* **2018,** *13* (2), 92–101. DOI: 10.2174/1574886313666180227110556.

Reino, J. L.; Guerrero, R. F.; Hernández-Galán, R.; Collado, I. G. Secondary metabolites from species of the biocontrol agent Trichoderma. *Phytochem Rev* **2007,** *7* (1), 89–123. DOI: 10.1007/s11101-006-9032-2.

Riches, Eleanor, Eleanor Riches, Steve Bajic, Efstathios Elia, John Langley, Julie Herniman. MS Atmospheric Pressure Ionisation Sources: Their Use and Applicability. https://www.chromatographytoday.com/article/ion-chromatography-ic/58/waters-corporation/pms-atmospheric-pressure-ionisation-sources-their-use-and-applicabilityp/2240 (accessed November 25, 2021).

Rivas-Ubach, A.; Liu, Y.; Bianchi, T. S.; Tolić, N.; Jansson, C.; Paša-Tolić, L. Moving beyond the van Krevelen Diagram: A New Stoichiometric Approach for Compound Classification in Organisms. *Anal. Chem.* **2018,** *90* (10), 6152–6160. DOI: 10.1021/acs.analchem.8b00529.

Rivera-Mariani, F. E.; Bolaños-Rosero, B. Allergenicity of airborne basidiospores and ascospores: need for further studies. *Aerobiologia* **2012,** *28* (2), 83–97.

Robert Koch-Institut. Epidemiologischer Steckbrief zu SARS-CoV-2 und COVID-19. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Steckbrief.html?nn=13490888 (accessed January 3, 2021).

Rosenberg, E. The potential of organic (electrospray- and atmospheric pressure chemical ionisation) mass spectrometric techniques coupled to liquid-phase separation for speciation analysis. *Journal of Chromatography A* **2003,** *1000* (1-2), 841–889. DOI: 10.1016/s0021-9673(03)00603-4.

Rowell, T. R.; Tarran, R. Will chronic e-cigarette use cause lung disease? *American journal of physiology. Lung cellular and molecular physiology* **2015,** *309* (12), L1398-409. DOI: 10.1152/ajplung.00272.2015.

Rupp, S.; Weber, R. W. S.; Rieger, D.; Detzel, P.; Hahn, M. Spread of Botrytis cinerea Strains with Multiple Fungicide Resistance in German Horticulture. *Front. Microbiol.* **2016,** *7*, 2075. DOI: 10.3389/fmicb.2016.02075.

Samanipour, S.; Kaserzon, S.; Vijayasarathy, S.; Jiang, H.; Choi, P.; Reid, M. J.; Mueller, J. F.; Thomas, K. V. Machine learning combined with non-targeted LC-HRMS analysis for a risk warning system of chemical hazards in drinking water: A proof of concept. *Talanta* **2019,** *195*, 426–432. DOI: 10.1016/j.talanta.2018.11.039.

Samburova, V.; Bhattarai, C.; Strickland, M.; Darrow, L.; Angermann, J.; Son, Y.; Khlystov, A. Aldehydes in Exhaled Breath during E-Cigarette Vaping: Pilot Study Results. *Toxics* **2018,** *6* (3).

Schnelle-Kreis, J.; Sklorz, M.; Herrmann, H.; Zimmermann, R. Atmosphärische Aerosole: Quellen, Vorkommen, Zusammensetzung. *Chem. Unserer Zeit* **2007,** *41* (3), 220–230. DOI: 10.1002/ciuz.200700414.

scikit-learn developers. scikit-learn: Machine Learning in Python. https://scikit-learn.org/stable/index.html (accessed April 7, 2022).

Seinfeld, J. H.; Pandis, S. N. *Atmospheric Chemistry and Physics*. *From Air Pollution to Climate Change*, 2. Aufl.; Wiley-Interscience: s.l., 2012.

Singh, L. P.; Gill, S. S.; Tuteja, N. Unraveling the role of fungal symbionts in plant abiotic stress tolerance. *Plant signaling & behavior* **2011,** *6* (2), 175–191. DOI: 10.4161/psb.6.2.14146.

Sleighter, R. L.; Hatcher, P. G. The application of electrospray ionization coupled to ultrahigh resolution mass spectrometry for the molecular characterization of natural organic matter. *Journal of mass spectrometry : JMS* **2007,** *42* (5), 559–574. DOI: 10.1002/jms.1221.

Smedsgaard, J.; Nielsen, J. Metabolite profiling of fungi and yeast: from phenotype to metabolome by MS and informatics. *Journal of experimental botany* **2005,** *56* (410), 273–286. DOI: 10.1093/jxb/eri068.

Snyder, L. R. Classification of the solvent properties of common liquids. *Journal of Chromatography A* **1974,** *92* (2), 223–230. DOI: 10.1016/S0021-9673(00)85732-5.

Sood, M.; Kapoor, D.; Kumar, V.; Sheteiwy, M. S.; Ramakrishnan, M.; Landi, M.; Araniti, F.; Sharma, A. Trichoderma: The "Secrets" of a Multitalented Biocontrol Agent. *Plants (Basel, Switzerland)* **2020,** *9* (6).

Spatafora, J. W.; Aime, M. C.; Grigoriev, I. V.; Martin, F.; Stajich, J. E.; Blackwell, M. The Fungal Tree of Life: from Molecular Systematics to Genome-Scale Phylogenies. *Microbiol Spectr* **2017,** *5* (5). DOI: 10.1128/microbiolspec.FUNK-0053-2016.

Spracklen, D. V.; Heald, C. L. The contribution of fungal spores and bacteria to regional and global aerosol number and ice nucleation immersion freezing rates. *Atmos. Chem. Phys.* **2014,** *14* (17), 9051–9059. DOI: 10.5194/acp-14-9051-2014.

Srzednicki, G.; Craske, J.; Nimmuntavin, C.; Mantais, L. G.; Wattananon, S. Determination of ergosterol in paddy rice using solid phase extraction. *J. Sci. Food Agric.* **2004,** *84* (15), 2041–2046. DOI: 10.1002/jsfa.1909.

Stoppacher, N.; Kluger, B.; Zeilinger, S.; Krska, R.; Schuhmacher, R. Identification and profiling of volatile metabolites of the biocontrol fungus Trichoderma atroviride by HS-SPME-GC-MS. *Journal of microbiological methods* **2010,** *81* (2), 187–193.

Stoppacher, N.; Reithner, B.; Omann, M.; Zeilinger, S.; Krska, R.; Schuhmacher, R. Profiling of trichorzianines in culture samples of Trichoderma atroviride by liquid chromatography/tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2007,** *21* (24), 3963–3970. DOI: 10.1002/rcm.3301.

Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Oresic, M. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics* **2007,** *8*, 93.

Tedersoo, L.; Sánchez-Ramírez, S.; Kõljalg, U.; Bahram, M.; Döring, M.; Schigel, D.; May, T.; Ryberg, M.; Abarenkov, K. High-level classification of the Fungi and a tool for evolutionary ecological analyses. *Fungal Diversity* **2018,** *90* (1), 135–159. DOI: 10.1007/s13225-018-0401-0.

The Genus Aspergillus. From taxonomy and genetics to industrial application : [proceedings of a symposium held under the auspices of the Federation of European Microbiological Societies, April 5 - 8, 1993, in Canterbury, Kent, United Kingdom; Powell, K. A., Ed.; FEMS symposium 69; Plenum Pr: New York, NY, 1994.

The MathWorks, Inc. Cophenetic correlation coefficien. https://de.mathworks.com/help/stats/cophenet.html (accessed March 3, 2022).

The MathWorks, Inc. linkage: Agglomerative hierarchical cluster tree. https://de.mathworks.com/help/stats/linkage.html (accessed March 6, 2022).

The Merck index. An encyclopedia of chemicals, drugs, and biologicals; O'Neil, M. J., Ed., 14. ed.; Merck handbooks; Merck: Whitehouse Station, NJ, 2006.

The pandas development team. *pandas-dev/pandas: Pandas 1.4.0rc0;* Zenodo, 2022.

Thermo Fisher Scientific Inc. *Exactive Series Operating Manual (P/N BRE0012255, Revision A)*, 2017. https://assets.thermofisher.com/TFS-Assets/CMD/manuals/man-bre0012255-exactive-series-manbre0012255-en.pdf.

Thines, E.; Anke, H.; Weber, R. W. Fungal secondary metabolites as inhibitors of infection-related morphogenesis in phytopathogenic fungi. *Mycological Research* **2004,** *108* (1), 14–25. DOI: 10.1017/S0953756203008943.

Tong, H.; Zhang, Y.; Filippi, A.; Wang, T.; Li, C.; Liu, F.; Leppla, D.; Kourtchev, I.; Wang, K.; Keskinen, H.-M.; Levula, J. T.; Arangio, A. M.; Shen, F.; Ditas, F.; Martin, S. T.; Artaxo, P.; Godoi, R. H. M.; Yamamoto, C. I.; Souza, R. A. F. de; Huang, R.-J.; Berkemeier, T.; Wang, Y.; Su, H.; Cheng, Y.; Pope, F. D.; Fu, P.; Yao, M.; Pöhlker, C.; Petäjä, T.; Kulmala, M.; Andreae, M. O.; Shiraiwa, M.; Pöschl, U.; Hoffmann, T.; Kalberer, M. Radical Formation by Fine Particulate Matter Associated with Highly Oxygenated Molecules. *Environ. Sci. Technol.* **2019,** *53* (21), 12506–12518. DOI: 10.1021/acs.est.9b05149.

Uchiyama, S.; Ohta, K.; Inaba, Y.; Kunugita, N. Determination of carbonyl compounds generated from the E-cigarette using coupled silica cartridges impregnated with hydroquinone and 2,4-dinitrophenylhydrazine, followed by high-performance liquid chromatography. *Analytical sciences : the international journal of the Japan Society for Analytical Chemistry* **2013,** *29* (12), 1219–1222. DOI: 10.2116/analsci.29.1219.

Ulrich, S.; Biermaier, B.; Bader, O.; Wolf, G.; Straubinger, R. K.; Didier, A.; Sperner, B.; Schwaiger, K.; Gareis, M.; Gottschalk, C. Identification of Stachybotrys spp. by MALDI-TOF mass spectrometry. *Fresenius J Anal Chem* **2016,** *408* (27), 7565–7581. DOI: 10.1007/s00216-016-9800-9.

van den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **2006,** *7* (1).

van den Dool, H.; Dec. Kratz, P. A generalization of the retention index system including linear temperature programmed gas—liquid partition chromatography. *Journal of Chromatography A* **1963,** *11*, 463–471. DOI: 10.1016/S0021-9673(01)80947-X.

van der Helm, D.; Baker, J. R.; Eng-Wilmot, D. L.; Hossain, M. B.; Loghry, R. A. Crystal structure of ferrichrome and a comparison with the structure of ferrichrome A. *J. Am. Chem. Soc.* **1980,** *102* (12), 4224–4231. DOI: 10.1021/ja00532a039.

van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008,** *2008* (9), 2579–2605.

Veselkov, K. A.; Vingara, L. K.; Masson, P.; Robinette, S. L.; Want, E.; Li, J. V.; Barton, R. H.; Boursier-Neyret, C.; Walther, B.; Ebbels, T. M.; Pelczer, I.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal. Chem.* **2011,** *83* (15), 5864–5872.

Wang, D. Y.-C.; Kumar, S.; Hedges, S. B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B* **1999,** *266* (1415), 163–171. DOI: 10.1098/rspb.1999.0617.

Waskom, M. seaborn: statistical data visualization. *JOSS* **2021,** *6* (60), 3021. DOI: 10.21105/joss.03021.

Webster, J.; Weber, R. *Introduction to Fungi*, 3rd ed.; Cambridge University Press: Leiden, 2007.

Weete, J. D.; Abril, M.; Blackwell, M. Phylogenetic distribution of fungal sterols. *PloS one* **2010,** *5* (5), e10899. DOI: 10.1371/journal.pone.0010899.

Weinberg, Zack. SVM separating hyperplanes. https://en.wikipedia.org/wiki/File:Svm_separating_hyperplanes_(SVG).svg (accessed March 6, 2022).

Wentura, D.; Pospeschill, M. *Multivariate Datenanalyse*. *Eine kompakte Einführung*; Lehrbuch; Springer: Wiesbaden, 2015.

Winder, C. L.; Dunn, W. B. Fit-for-purpose quenching and extraction protocols for metabolic profiling of yeast using chromatography-mass spectrometry platforms. *Methods in molecular biology (Clifton, N.J.)* **2011,** *759*, 225–238.

Womack, A. M.; Artaxo, P. E.; Ishida, F. Y.; Mueller, R. C.; Saleska, S. R.; Wiedemann, K. T.; Bohannan, B. J. M.; Green, J. L. Characterization of active and total fungal communities in the atmosphere over the Amazon rainforest. *Biogeosciences* **2015,** *12* (21), 6337–6349. DOI: 10.5194/bg-12-6337-2015.

Wu, B.; Hussain, M.; Zhang, W.; Stadler, M.; Liu, X.; Xiang, M. Current insights into fungal species diversity and perspective on naming the environmental DNA sequences of fungi. *Mycology* **2019,** *10* (3), 127–140. DOI: 10.1080/21501203.2019.1614106.

Wu, Y.; Li, L. Sample normalization methods in quantitative metabolomics. *Journal of chromatography. A* **2016,** *1430*, 80–95. DOI: 10.1016/j.chroma.2015.12.007.

Wulff, J. E.; Mitchell, M. W. A Comparison of Various Normalization Methods for LC/MS Metabolomics Data. *ABB* **2018,** *09* (08), 339–351.

Xie, H.; Wang, X.; van der Hooft, J. J.; Medema, M. H.; Chen, Z.-Y.; Yue, X.; Zhang, Q.; Li, P. Fungi population metabolomics and molecular network study reveal novel biomarkers for early detection of aflatoxigenic Aspergillus species. *Journal of hazardous materials* **2022,** *424* (Pt A), 127173. DOI: 10.1016/j.jhazmat.2021.127173.

Yamamoto, N.; Bibby, K.; Qian, J.; Hospodsky, D.; Rismani-Yazdi, H.; Nazaroff, W. W.; Peccia, J. Particle-size distributions and seasonal diversity of allergenic and pathogenic fungi in outdoor air. *The ISME journal* **2012,** *6* (10), 1801–1811. DOI: 10.1038/ismej.2012.30.

Yi, L.; Dong, N.; Yun, Y.; Deng, B.; Ren, D.; Liu, S.; Liang, Y. Chemometric methods in data processing of mass spectrometry-based metabolomics: A review. *Analytica chimica acta* **2016,** *914*, 17–34.

Zeilinger, S.; Gruber, S.; Bansal, R.; Mukherjee, P. K. Secondary metabolism in Trichoderma – Chemistry meets genomics. *Fungal Biology Reviews* **2016,** *30* (2), 74–90. DOI: 10.1016/j.fbr.2016.05.001.

Zeilinger, S.; Martín, J.-F.; García-Estrada, C. *Biosynthesis and Molecular Genetics of Fungal Secondary Metabolites, Volume 2;* Springer New York: New York, NY, 2015b.

Zhu, Y.; Ren, H.; Wei, Y.; Bie, Z.; Ji, L. Determination of Imidazole, 4-Methylimidazole, and 2-Methylimidazole in Cigarette Additives by Ultra-High Performance Liquid Chromatography. *Analytical Letters* **2015,** *48* (17), 2708–2714. DOI: 10.1080/00032719.2015.1045593.

Zimek, A.; Schubert, E.; Kriegel, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analy Data Mining* **2012,** *5* (5), 363–387. DOI: 10.1002/sam.11161.

Zin, N. A.; Badaluddin, N. A. Biological functions of Trichoderma spp. for agriculture applications. *Annals of Agricultural Sciences* **2020,** *65* (2), 168–178.

Zubarev, R. A.; Makarov, A. Orbitrap mass spectrometry. *Anal. Chem.* **2013,** *85* (11), 5288–5296. DOI: 10.1021/ac4001223.

Zwickel, T.; Kahl, S. M.; Rychlik, M.; Müller, M. E. H. Chemotaxonomy of Mycotoxigenic Small-Spored Alternaria Fungi – Do Multitoxin Mixtures Act as an Indicator for Species Differentiation? *Front. Microbiol.* **2018,** *9*.

10. List of related poster presentations and publications

[deleted in the public version]

11. Acknowledgements

[deleted in the public version]

12. Curriculum Vitae

[deleted in the public version]