DISSERTATION

Structural aspects of hyperspectral imaging data: a case study on microplastics analysis from the viewpoint of chemometrics

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Doktors der technischen Wissenschaften
unter der Leitung von

Ao.Univ.Prof. Mag. Dr. Johann Lohninger
E164
Institut für Chemische Technologien und Analytik

eingereicht an der Technischen Universität Wien,
Fakultät für Technische Chemie
von

Dipl.-Ing. Benedikt Hufnagl, BSc
e1025978

Wien, am 30/03/2022

Benedikt Hufnagl

DISSERTATION

Structural aspects of hyperspectral imaging data: a case study on microplastics analysis from the viewpoint of chemometrics

carried out for the purpose of obtaining the degree of

Doctor technicae

submitted at TU Wien,
Faculty of Technical Chemistry
by

Benedikt Hufnagl

under the supervision of
Prof. Dr. Hans Lohninger
Institute of Chemical Technologies and Analytics
Vienna University of Technology
Austria

Acknowledgements

The basis of this research dates back to the day I met my thesis supervisor Hans Lohninger in a lab course where he held a presentation about statistics. Back then it made me quite curious that a chemistry professor talks so much about mathematics and statistics. This raised my interest in the field of chemometrics which has been my main focus area throughout my studies. Hans mentored me throughout my bachelor and master thesis and I can say that I learned most of my chemometric 'hard skills' while working with him, such as programming, software development, fundamentals of statistics and machine learning. However, Hans' biggest contribution to this thesis is his commitment as a mentor. He was always there, when I needed his help in any matter that arose. I am extremely grateful for his help and dedication during all the years we have known each other.

When I originally envisioned this thesis I expected it to become some rather complex discourse on chemometrics. Well, it turned out quite differently and this is mostly because of the support from the Purency team. Purency gave me the chance to meet and talk to many researchers across the world that do microplastics analysis, which would never have been possible if I had tried it on my own. This allowed me to better understand the user perspective of chemometric tools and also inspired me to write the thesis from a more holistic viewpoint on the general matter of microplastics analysis. Many thanks to the team members at Purency: Michael Stibi, Aurelia Liechtenstein, Valerie Hengl, Lukas Wander, Michaela Radl and Susanna Di Vita for their support. There are many things one cannot learn or achieve without a strong team in the back.

I also thank Lukas Wander for our fruitful discussions. When I sat at the computer late at night to write some more lines I often saw via MS Teams that Lukas was also working on his PhD thesis. That kept me going, even though I was very tired most of the time.

Special thanks to the research group at the University of Bayreuth. To Sven Ritschar I say thanks for inviting me to collaborate on a joint research project about Eisenia fetida. Thanks to Heghnar Martirosyan for the valuable feedback on the machine learning models. I thank Martin Löder for his help in setting up research collaborations and his continuing support. Without his help many things would not be as they are today. Special thanks to Christian Laforsch for assessing this thesis, for providing guidance regarding the peer review of the papers and the help with the introduction of the last publication.

I thank Gabriele Eder and Lidija Lukacic for the valuable insights into FTIR spectroscopy. Thanks also to Michael Washüttl for setting up the collaboration and the continuing support. Many thanks to Elisavet Kanaki, Franziska Fischer and Dieter Fischer for the collaboration on Raman spectroscopy.

I would also like to extend my gratitude to Jana Weißer for the joint work on detecting microplastics in bottled water. I greatly enjoyed our fruitful discussions about machine learning and how to apply it to microplastics analysis. I also thank Karl Glas for his support with legal matters.

To Dieter Steiner I say thanks for his help in understanding random forests. I also thank Lisa Weijler and Alexandra Walter for the insightful discussions.

I have greatly benefited from the participation at ISO/TC 61/SC 14/WG 4. Special thanks to Claus-Gerhard Bannick and Ulrike Braun for supporting me. I also thank Natalia Ivleva, Jes Vollertsen and Sebastian Primpke for their moral support and the collaboration on data processing.

Thanks to Gernot Friedbacher for participating as a member of the examination committee.

Last but not least I thank my family. I am deeply grateful to Tina for her everlasting moral support and her patience. It took quite a long time to complete these studies. I thank my dog Graham for sitting right next to me during the long hours of work. To my parents Andrea and Manfred I say thanks for inspiring me to pursue studies at TU Wien. I also thank my parents-in-law Birgit, Bettina, Roland and Wolfgang and my sister-in-law Alisa for supporting me and their help with looking after my children.

Abstract

Microplastics is an ubiquitous contaminant that has been detected in almost any environmental habitat on earth. However, due to the lack of data the potentially harmful effects on the environment and human health can still not be determined. In the past five years the interest in the topic has risen dramatically, where drivers include the research community, policy makers but also the aware customer and thus also certain industries. The demand for better and comparable data drives the need for harmonization and standardization of existing analytical methods. Currently, the most widely applied instrumental approach is microspectroscopy based on FTIR, Raman or QCL. These methods have in common that the analysis of the spectroscopic data is based on spectral library search.

While spectral reference databases for spectral library search can be easily bought on the market and are usually included in the instrument software the identification of microplastics based on this approach has been criticised for poor data quality and incomparability. For this reason, the research community has started to compile publicly available reference databases and is currently discussing parameter settings for spectral library search to improve quality as well as comparability.

Looking at the state of the art of data science and the broad use of machine learning in our day-to-day life it is somewhat surprising that only with a few exceptions this methodology has not been applied for microplastics detection and quantification as an alternative to spectral library search. This leads to the question whether there are certain structural aspects of microplastics data which make the application of machine learning difficult and also how these obstacles can be overcome.

Within this thesis the data originating from μFTIR imaging measurements has been studied from the viewpoint of unsupervised and supervised learning in order to provide answers to these questions. This was done by developing a novel graph-based clustering approach as well as different random forest based classifiers. The insights gained from the development and the results from both methods show that the creation of annotated training data for supervised learning is far from trivial. Representative samples which are required for sampling training examples are hard to come by. Further, the task of annotating the sampled training data requires expert knowledge in the fields of spectroscopy as well as chemometrics and is therefore prone to subjectivity and labeling errors. Among the achievements of this thesis is the creation of a random forest classifier that shows superior performance, both with respect to data quality and throughput rate. While the effort for creating these classifiers should not be underestimated the results show that machine learning brings significant advantages with respect to the analysis of microplastics, such as increased speed and scalability, which is key to allow for large scale monitoring of this environmental contaminant.

Kurzfassung

Mikroplastik ist ein allgegenwärtiger Umweltschadstoff, der in fast jedem Lebensraum auf der Erde nachgewiesen wurde. Aufgrund des Mangels an representativen Daten lassen sich die potenziell schädlichen Auswirkungen auf die Umwelt und die menschliche Gesundheit jedoch noch immer nicht bestimmen. In den letzten fünf Jahren ist das Interesse an diesem Thema dramatisch gestiegen, wobei die Forschungsgemeinschaft, die politischen Entscheidungsträger, aber auch der Konsument und damit bestimmte Industriebranchen die Nachfrage nach besseren und vergleichbaren Daten treiben. Der derzeit am häufigsten angewandte instrumentelle Ansatz ist die Mikrospektroskopie auf der Basis von FTIR, Raman und QCL. Diese Methoden haben gemeinsam, dass die Analyse der spektroskopischen Daten auf automatischen Abgleichen mit Spektrenbibliotheken beruht.

Während Spektrenbibliotheken leicht auf dem Markt erhältlich sind und in der Regel auch in der Gerätesoftware enthalten sind, ist die Identifizierung von Mikroplastik auf der Grundlage dieses Ansatzes wegen der schlechten Datenqualität und schlechten Vergleichbarkeit immer häufiger kritisiert worden. Aus diesem Grund hat die Forschungsgemeinschaft begonnen, öffentlich zugängliche Spektrenbibliotheken zusammenzustellen und diskutiert derzeit die Parametereinstellungen für den automatischen Abgleich in Spektralbibliotheken, um die Qualität und die Vergleichbarkeit zu verbessern.

Betrachtet man den aktuellen Stand der Technik im bereich Data Science und die breite Anwendung des maschinellen Lernens in unserem täglichen Leben, so ist es etwas verwunderlich, dass diese Methodik nur mit wenigen Ausnahmen bisher keine Anwendung für die Analyse von Mikroplastik als Alternative zur Spektralbibliotheken gefunden hat. Dies wirft letzlich die Frage auf, ob es bestimmte strukturelle Aspekte in Mikroplastikdaten gibt, die die Anwendung von maschinellem Lernen erschweren und weiters, wie diese Hindernisse überwunden werden können.

Im Rahmen dieser Doktorarbeit wurden die aus μFTIR Imaging stammenden Daten unter dem Gesichtspunkt des unüberwachten und überwachten Lernens untersucht, um Antworten auf diese Fragen zu finden. Dies geschah durch die Entwicklung eines neuartigen graphenbasierten Clustering-Ansatzes sowie verschiedener Random-Forest-basierter Klassifikatoren. Die aus der Entwicklung gewonnenen Erkenntnisse und die Ergebnisse beider Methoden zeigen, dass die Erstellung von annotierten Trainingsdaten für das überwachte Lernen alles andere als trivial ist. Repräsentative Proben, die für die Auswahl von Trainingsdaten benötigt werden, sind schwer zu beschaffen. Außerdem erfordert die Annotation der Trainingsdaten Expertenwissen in den Bereichen Spektroskopie und Chemometrie und ist anfällig für Subjektivität und Flüchtigkeitsfehler. Zu den Errungenschaften dieser Arbeit gehört die Entwicklung eines Random-Forest-Klassifikators, der sowohl in Bezug auf die Datenqualität als auch auf die Durchsatzrate eine hervorragende Leistung zeigt. Während der Aufwand für die die Erstellung dieser Klassifikatoren nicht zu unterschätzen ist, zeigen die Ergebnisse, dass maschinelles Lernen erhebliche Vorteile für die Analyse von Mikroplastik mit sich bringt, wie z.B. eine höhere Geschwindigkeit und Skalierbarkeit, die für eine groß angelegte Überwachung dieses Kontaminanten notwendig ist.

For my family;
Tina, Olivia, Aaron & Graham.

Contents

1 About this thesis

This doctoral thesis follows the cumulative form. Contrary to a monograph peer-reviewed papers are an essential part of the thesis. In the printed version of this document the publications are supplied as annexes. The public web version only contains the text body of this document. Table 1 summarizes the publications which make up the central matter of this thesis. Summaries of the listed publications can be found in section 7.

Table 1: Peer-reviewed publications that are cumulated to form this doctoral thesis.

Citation	DOI	Title	Journal
Hufnagl and Lohninger (2020)	10.1016/j.aca.2019.10.071	A graph-based clustering method with special focus on hyperspectral imaging	Analytica Chimica Acta
Hufnagl et al. (2019)	10.1039/C9AY00252A	A methodology for the fast identification and monitoring of microplastics in environmental samples using random decision forest classifiers	Analytical Methods
Hufnagl et al. (2022)	10.1021/acs.estlett.1c00851	Computer-Assisted Analysis of Microplastics in Environmental Samples Based on μFTIR Imaging in Combination with Machine Learning	Environmental Science & Technology Letters
Weisser et al. (2021)	10.3390/w13060841	From the Well to the Bottle: Identifying Sources of Microplastics in Mineral Water	Water
Ritschar et al. (2021)	10.1007/s00418-021-02037-1	Classification of target tissues of Eisenia fetida using sequential multimodal chemical analysis and machine learning	Histochemistry and Cell Biology

2 Introduction

Microplastics (Thompson et al., 2004) is a global pollutant which has been detected in almost any environmental habitat on earth, as high up as the peak of Mount Everest (Napper et al., 2020) and as deep as the Mariana trench (Peng et al., 2018). While microplastics was initially associated with ocean pollution, studies addressing terrestrial ecosystems soon revealed that fresh water (Eerkes-Medrano et al., 2015), soil (Möller et al., 2020) and glaciers (Ambrosini et al., 2019) are also contaminated. As microplastics are ingested by aquatic organisms they also traverse the food chain. In the human body microplastics have been detected in placentas of unborn babies (Ragusa et al., 2021) as well as in stool (Schwabl et al., 2019).

Given the ubiquity of microplastics on earth and in organisms there are concerns regarding the effect on the environment and ecosystem as well as potential health effects to humans. Even though the topic of microplastics has attracted a lot of attention from researchers over the past fifteen years, well-founded statements regarding their toxicity cannot be made so far. The reasons for this are manifold. One aspect is the fact that the detection and quantification of microplastics is a complex and not fully resolved analytical problem. Because of this there is currently no broadly accepted measurement methodology. Provencher et al. (2020) further criticized the lack of quality and lack of inter-comparability of a large portion of the published microplastics literature and emphasized the need for harmonization as well as standardization.

Even though standards are currently in development at the International Organization for Standardization (ISO) as well as other national standardization bodies and first interlaboratory proficiency tests have been conducted (Van Mourik et al., 2021; Belz et al., 2021; DeFrond et al., 2022) there is still a long way to go to reach global consensus on definitions and methodological aspects. The following sections will focus on the methodological issues of microplastics analysis to put the associated research papers of this doctoral thesis into a broader context within the microplastics research domain.

2.1 How to define microplastics?

The definition of microplastics is a disputed matter, even when it comes to defining a size range (Hartmann et al., 2019). Perhaps the most well-known size definition was given by the National Oceanic and Atmospheric Administration (NOAA) (Arthur et al., 2009) as all plastic particles which are less than 5 mm in diameter. Considering the SI system this definition might cause confusion as the 'micro' in microplastics can also be associated with a range of 1 μm to 1000 μm. The motivation for the definition by NOAA lies in the assumption that plastic particles below 5 mm are more likely to be ingested by organisms.

In recent years plastic particles which are smaller than 1 μm have also come into the focus of attention. Plastic particles below that size have been coined ‘nanoplastics’ which is another confusing term, as nanoparticles are commonly defined to be between 1 nm and 100 nm. ISO/TR 21960:2020 (2020) distinguishes between nanoplastics ($1nm - 1000nm$), microplastics ($1\mu m - 1000\mu m$) and *large* microplastics ($1mm - 5mm$) to provide a categorization framework that follows the SI system but takes the NOAA definition into account.

Regarding chemical composition the term ‘plastics’ introduces additional issues as the definition of plastics does not include elastomers such as rubber or silicone. Within microplastics research, however, man-made polymers are often included even though they do not fall under the definition of plastic. Similar issues arise when considering plastics with a high content of additives, co-polymers, composites, surface coatings and tire wear. From a scientific perspective a too narrow definition might pose a hindrance for future research while on the other hand a globally accepted definition is an essential prerequisite for regulations and policy making. (Hartmann et al., 2019)

2.2 Detection and quantification approaches

From the perspective of instrumental analytical chemistry the detection of microplastics poses difficulties, especially, if the target size is less than 500 μm (Hidalgo-Ruz et al., 2012). Below that particle size manipulation of individual particles becomes more and more difficult. Picking up individual particles, which would allow a spectroscopic analysis via ATR-FTIR, is thus no longer feasible. Further particle numbers increase with decreasing size which makes manual approaches unsuitable for gaining representative data. Different instrumental approaches have thus been developed over the course of the recent years, which are becoming ever more performant in terms of detection limit as well as throughput rate (Käppler et al., 2016; Renner et al., 2017b; Silva et al., 2018; Xu et al., 2019; Primpke et al., 2020a; Möller et al., 2020; Ivleva, 2021). The more prominent approaches can be roughly summarized as spectroscopy, thermoanalytical methods and microscopy.

Light microscopy has been criticized to be very prone to bias with error rates estimated to be as high as 70% (Hidalgo-Ruz et al., 2012). In a recent interlaboratory proficiency test which was conducted by the Southern California Coastal Water Research Project (SCCWRP) (DeFrond et al., 2022) light microscopy was ruled out as an applicable method in the current Californian monitoring program. Irrespective of that light microscopy remains an important tool for estimating the total number of particles (Belz et al., 2021) and is often included in microspectroscopy devices applied for microplastics analysis. Besides light microscopy fluorescense microscopy has gained increased interest as a surrogate method for detecting microplastics. Here microplastics are selectively dyed using Nile Red as a fluorescence marker (Maes et al., 2017; Süssmann et al., 2021).

The most prominent thermoanalytical methods (La Nasa et al., 2020; Becker et al., 2020) include pyrolysis gas chromatography - mass spectroscopy (py-GC-MS) (Picó and Barceló, 2020) and thermal extraction-desorption gas chromatography - mass spectroscopy (TED-GC-MS) (Dümichen et al., 2017; Eisentraut et al., 2018; Bannick et al., 2019). The main advantages of these methods include the possibility to skip extensive sample preparation and the high degree of automation. Thereby a very high sample throughput can be achieved. Microplastics contamination is thus measured as a mass value which is a typical measurand in environmental monitoring. Particle size classification is possible by using sieve cascades. However, both methods sometimes cannot selectively differentiate between certain types of polymers and environmental matrix components.

Spectroscopy is currently the largest and most frequently applied instrument group. Microplastics in the size range of *large microplastics* down to 200 μm can be detected and quantified using NIR hyperspectral imaging (Serranti et al., 2018; Shan et al., 2019). As a manual approach ATR-FTIR can be applied with ease for particles larger than 500 μm (Löder et al., 2015). Below 500 μm microspectroscopy approaches such as μFTIR and μRaman become predominant for microplastics detection. These approaches have in common that the sample is usually purified using a preparation scheme (Renner et al., 2017b; Möller et al., 2020) and then subsequently concentrated on a membrane filter. See figure 1 and 3 for a light microscopy image of an environmental sample. The spectroscopic measurement is then conducted by measuring the particulate remnants on the filter surface either by imaging or point-wise measurement. Microspectroscopy brings the advantage, that next to polymer type information also particle size and shape can be measured.

2.2.1 An overview of microspectroscopy devices

While the analysis of microplastics larger than 500 μm is a rather straightforward matter (speaking from an instrumental perspective), the range below 200 μm, especially below 50 μm, is a complex analytical problem. By the end of 2019 spectrometer manufacturers began with positioning certain microspectroscopy devices as solutions for microplastics analysis. However, from the scientific literature we can gather that so far not a single approach can address all issues raised when

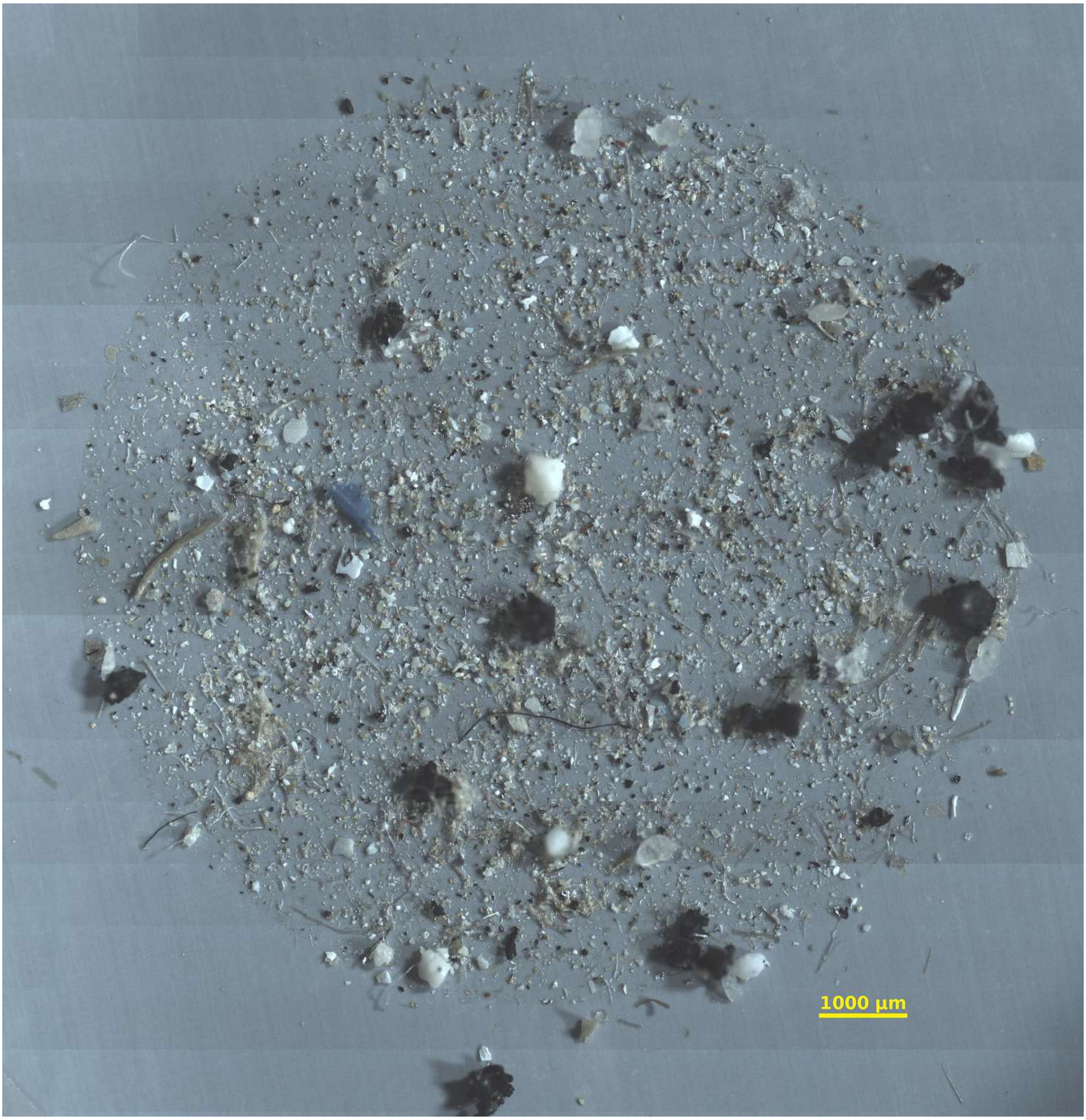


Figure 1: Waste water treatment plant outlet sample prepared with the enzymatic digestion protocol developed by Löder et al. (2017). The largest particles visible in the image are below 500 μm and can still be characterized using μFTIR imaging in transmission mode. Above that the particles have to be measured manually using ATR-FTIR. Note, that only a very small fraction of the present particles are microplatics. See figure 3 for a zoomed-in view of the sample. *Courtesy M. G. J. Löder and C. Laforsch, University of Bayreuth.*

analyzing microplastics. Figure 2 gives an overview of devices which are currently in broader use or in the focus of more recent publications.

A key distinguishing aspect between the methodologies is the way in which the instrument measures the filter sample spatially. We can here differentiate between imaging and single point approaches. Devices capable of imaging usually rely on an array detector which can either be one-dimensional (aka. line array (LA)), or two-dimensional (aka. focal plane array (FPA)). Usually, large areas of interest or, ideally, the entire filter are mapped and the resulting tiles are assembled to form a

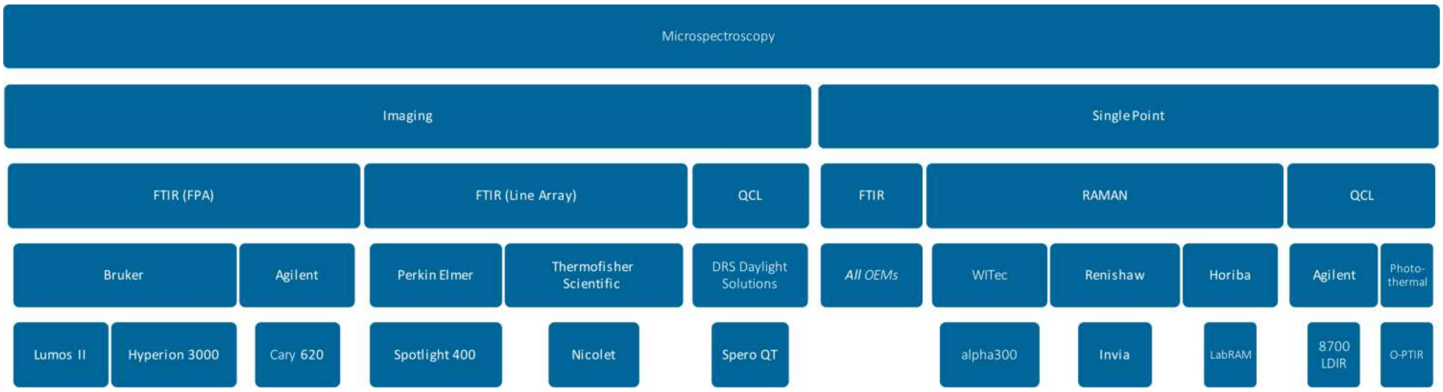


Figure 2: Technical segmentation of the current market showing devices in broader use and upcoming technologies which are prominently promoted for microplastics analysis. The domain of single point FTIR devices is summarized as ‘All OEMs’ as these are usually distributed as product variations of their imaging-capable counterparts.

hyperspectral image.

Devices which lack array detectors can, of course, also be used to measure hyperspectral images. However, in the case of microplastics analysis the time requirement for the measurement will be in the range of days to weeks. Because of this constraint single point devices use a different approach to measure particles on the filter. Here, the particles are detected before the spectroscopic measurement is performed. This can be done by capturing an image of the filter in visible light or by measuring an intensity map at a certain wavelength. The particle contours can then be detected by means of binarization or a watershed algorithm (Anger et al., 2019; Brandt et al., 2020). Subsequently, spectra are measured at the center of the particles. Depending on the number of particles present on the filter this approach allows an analysis within a reasonable amount of time, though limiting the spectroscopic measurement to a subsample of particles is often necessary (Anger et al., 2018).

In both the imaging and the single point domain there are different spectroscopic techniques in use. As discussed above Raman can potentially be used to capture hyperspectral images, though practically it has gained little relevance (Käppler et al., 2016). This niche is thus dominated by focal plane array (FPA) and line array (LA) based systems in combination with Fourier-transform infrared (FTIR). FPA-based μFTIR devices include the Bruker Hyperion 3000 (Löder et al., 2015), the Bruker Lumos II and the Agilent Cary 620 (da Silva et al., 2020). LA-based μFTIR are the Perkin Elmer Spotlight 400 and the Thermofisher Nicolet iM10 MX (Xu and Gowen, 2019). A special case within this niche is the DRS Daylight Spero QT which is a quantum cascade laser (QCL) based system making use of an FPA bolometer (Primpke et al., 2020c).

Within the single point domain both FTIR and Raman play an important role. In figure 2 the μFTIR devices are not further specified as they are usually distributed as product variations like the imaging μFTIRs. For example, the Perkin Elmer Spotlight 200 is a single point device which can be upgraded to a Spotlight 400. In the same way Lumos II can be bought with a FPA detector or a single detector installed.

The application of Raman microspectroscopy, on the other hand, has gained increased interest as it forms a complementary technique to FTIR and comes with certain advantages (Xu et al., 2019). First, Raman allows the detection of particles down to 1 μm. It therefore covers the full size range of microplastics which fall under the ISO definition. Secondly, samples do not have to be dried carefully before the measurement as Raman is less affected by water. Thirdly, Raman is more robust regarding colored or black microplastics and further allows to identify additives such as fillers and pigments (Käppler et al., 2016). Examples of devices include the WITec alpha300, the Renishaw Invia and the Horiba LabRAM (Brandt et al., 2020; Primpke et al., 2020a).

QCL-based systems are also represented by the Agilent 8700 LDIR and the Photothermal O-PTIR. The 8700 LDIR is a rather new device but has already been applied in microplastics analysis (Scircle et al., 2020; Mughini-Gras et al., 2021; Li et al., 2021; Belz et al., 2021). It uses a QCL as an illumination source for reflection measurements. The O-PTIR, on the other hand, uses the QCL to induce the photothermal effect and allows to measure both an infrared as well as a Raman signal concurrently with superior quality (Barrett et al., 2020). So far there are only few cases where it has been applied for microplastics analysis though the technology seems to be promising.

2.2.2 The challenges of measuring microplastics using microspectroscopy

Depending on the used spectroscopic technique there are different issues that arise when measuring microplastics, especially when dealing with particle sizes below 50 μm. Regarding μFTIR the signal-to-noise ratio of reflectance spectra is usually

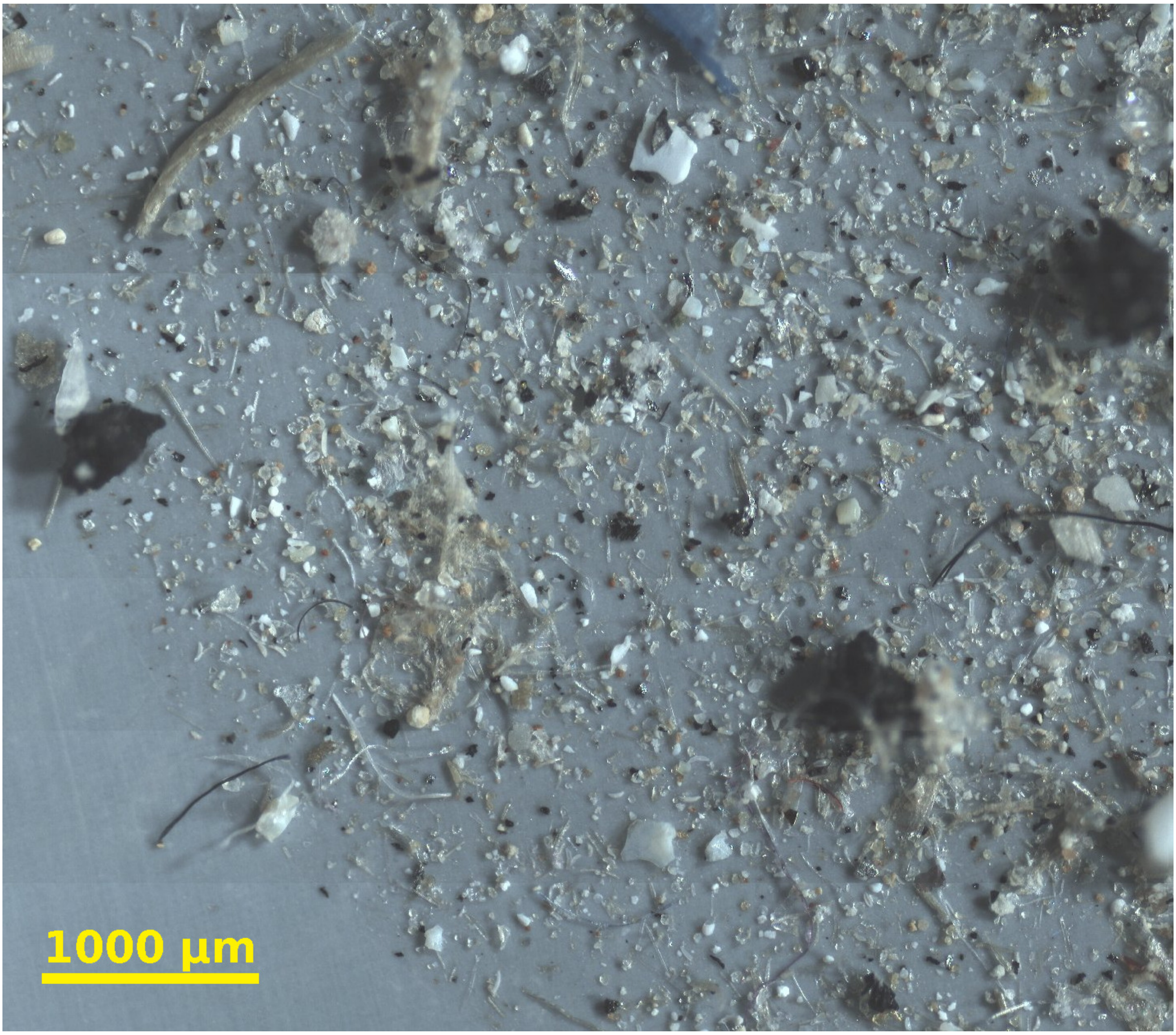


Figure 3: Waste water treatment plant outlet sample prepared with the enzymatic digestion protocol developed by Löder et al. (2017). Even though particles smaller than 10 μm are visible in the light microscopy image a characterization using μFTIR in transmission mode is often impossible due to strong scattering effects. See figure 1 for a full view of the sample. *Courtesy M. G. J. Löder and C. Laforsch, University of Bayreuth.*

too low to allow a reasonably accurate identification of microplastics. For this reason it is common to measure the filter membranes in transmission mode. This, however, comes with the disadvantage that spectra of thick particles are sometimes no longer recognizable due to total absorption. It can also be the case that total absorption arises due to additives which are strongly absorbing, such as black carbon and pigments. The effect also plays a role in reflectance measurements if the light path is reflected underneath the particle. In this case the optical path length through the particle is twice as long. The resulting transflection spectra may thus also be affected by total absorption. The QCL-based 8700 LDIR is a case where this effect sometimes hampers the identification of particles.

Scattering effects arise if the wavelength of the radiation source comes close to the particle size. In μFTIR this is very prominent with particles in a size range of 5 μm to 30 μm. The spectra of such particles often exhibit strong baseline distortions and sometimes peak deformations known as Mie scattering (Bassan et al., 2009). Scattering also occurs with certain filter types such as silicon wafers and steel meshes at the pores.

Raman, on the other hand, suffers from similar problems due to fluorescence and low signal-to-noise ratios. Further, different excitation lasers create different Raman spectra which adds to the complexity of building suitable spectral reference libraries.

Apart from the spectroscopic technique there are other problems related to microplastics and the embedding matrix. Polymer ageing due to mechanical stress and oxidative weathering often occurs in environmental microplastics. As a result changes in the characteristic vibrational bands can be observed, which in more severe cases, might hinder the identification. Additives such as pigments and fillers overlap the signals created by the polymer backbone. Especially in μRaman spectroscopy these substances might be stronger Raman scatterers than the polymer.

The used membrane filters also have a strong effect on the measurement. Usually, the filter material is transparent only within a certain wavelength range. Because of this the spectral range for transmission FTIR has to be limited (e.g. if aluminum oxide filters are used). (Xu et al., 2019; Ivleva, 2021)

2.2.3 The state of the art of data processing

Xu et al. (2019) observed that chemometric techniques are often neglected within the microplastics literature even though a lot of research is based on spectroscopy. According to Renner et al. (2018) only 25% of papers reported the data analysis approach and only 2% gave a more detailed description. The predominant data analysis approach for microplastics identification in microspectroscopy is *spectral library search*. On the one hand this is not surprising as search engines as well as commercial databases are available in most OEM software packages. Further, spectral library search is, after all, one of the workhorses in day-to-day lab work for identifying unknown substances. On the other hand library search, or more precisely, the literature which used this technique for microplastics analysis, has been criticized for two main reasons.

First, there is no consensus on what kind of databases should be used. Even though there are many commercial polymer databases available, most of them are not tailored towards microplastics detection. As already discussed earlier, many effects (e.g. ageing, total absorption, ...) are not represented in these databases, though more recently dedicated commercial microplastics databases have been made available in the marked. The research community reacted to this problem by publishing self-made databases (Primpke et al., 2018; Munno et al., 2020; Cowger et al., 2021).

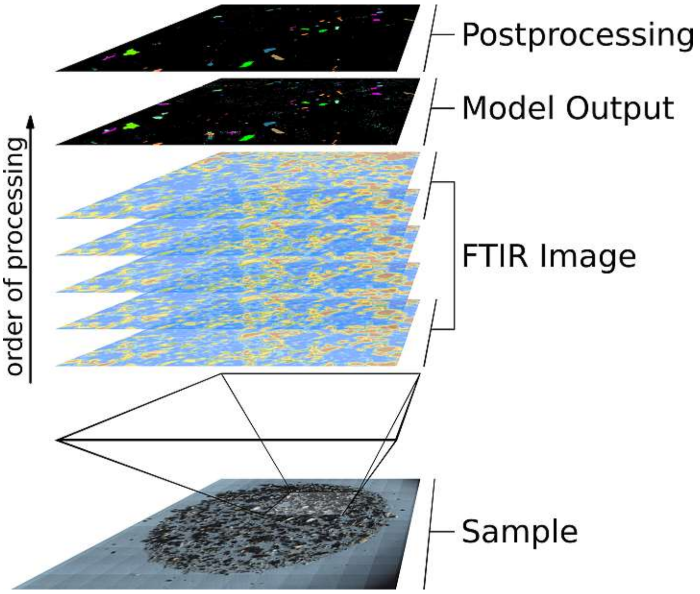


Figure 4: Schematic drawing of the measurement process and data processing in a μFTIR imaging setup. Using focal plane array or line array detectors the filter surface is measured to create a FTIR hyperspectral image. The data is then processed by means of a machine learning model to detect the microplastics based on their characteristic vibrational signatures. Published by *Hufnagl et al. (2022) under CC BY 4.0.*

Secondly, library search algorithms have been found to lack robustness regarding said effects independently of the used databases (Renner et al., 2017b). For this reason attempts have been made to improve the results on the algorithmic level (Renner et al., 2017a; Primpke et al., 2017; Renner et al., 2019; Kedzierski et al., 2019; Primpke et al., 2020b). Additionally, there is no consensus both within research and standardization on the search criterion for computing the *hit quality index* (HQI) as well as the threshold that should be applied to discard low-quality spectra. In the view of the author this is perhaps the most critical point for the following reason. Some literature recommends a HQI threshold of 0.7 without giving a verifiable explanation why this value was chosen (Renner et al., 2017b). By tuning this value both the false-negative (FN) and the false-positive (FP) rate is affected. Without a proper evaluation of this value the microplastics content may either be overestimated or underestimated. However, if the HQI is fixed to a certain value, the FN and FP rate ultimately also depends on the database and search criterion used. However, the research community and standardization still continues the discussion of setting a suitable threshold, failing to see that such a definition makes sense only if the database and the search criterion are fixed as well.

Xu et al. (2019) further criticizes that databases are made up of reference spectra which have been measured with a certain spectroscopic technique, however, when the user applies the database for microplastics detection the used spectroscopic technique might be a different one. A typical example is the database published by Primpke et al. (2018) which contains lots of ATR-FTIR spectra even though the intended application is for transmission μFTIR. The database is also available in the device software of the Agilent 8700 LDIR even though this instrument is QCL-based. In the view of the author this critique is well justified, however, there is another aspect to this. Suppose there is a library search routine validated for spectroscopic method A (including a fixed database, search criterion and HQI threshold) with known FN and FP rates. In that case the library search routine might still be applied for identifying microplastics using spectroscopic method B. However, the argument, that one still uses a validated method would be wrong and misleading. In that case the FN and FP rates will have to be redetermined.

It can be summarized that there are various issues regarding spectral library search that are still unresolved. On the other hand said problems are well-understood within the chemometrics and data science community. Taking the viewpoint of machine learning, spectral library search can be categorized as a special case of the k nearest neighbors classifier (kNN), namely $k = 1$. As a representative of instance-based learning kNN is also a distance-based learner, where different measures such as e.g. Pearson correlation or Euclidean distance can be applied to determine the similarity between the training data and the unknown object. While kNN is a well-established technique and is often used as a benchmark, it is in many cases preferable to use different learners.

As spectra are high-dimensional data, distance metrics can be severely affected by the *Curse of Dimensionality*, a term coined by Bellman (1961) to describe the strange behavior of distance metrics within high-dimensional feature spaces. It can be observed, that with increasing dimensionality the distances between objects of a dataset become more and more similar (Jimenez and Landgrebe, 1998). As a result the performance of distance-based classifiers decreases with increasing dimensionality. Related to this effect is the *popular nearest neighbors* phenomenon (Radovanović et al., 2010), where a few objects of the training data become overly popular as nearest neighbors and thus dominate the classification result. Even though these effects where not specifically studied as part of this doctoral thesis, Primpke et al. (2018) described different issues when assembling their database. Firstly, the polymers HDPE and LDPE could not be distinguished due to an artifact caused by the ATR crystal. Only after the artifact was removed the two polymers could be distinguished. Considering, that the artifact was only a diminutive feature of the spectra and did not overlap with any of the vibrational bands it is quite astounding that it had such a severe effect. Secondly, the authors describe a long process of iteratively removing and adding spectra to the database because the presence or absence of entries created cross-dependencies in the results. Both issues point towards the discussed issues with high-dimensional feature spaces.

Even though machine learning has only played a minor role in microplastics analysis so far there are some examples of its application. In the domain of NIR hyperspectral imaging Serranti et al. (2018) and Shan et al. (2019) trained classification models based on partical least squares discriminant analysis (PLS-DA) (Wold et al., 2001; Lee et al., 2018) and Support Vector Machine (SVM) (Cortes and Vapnik, 1995) respectively. Paul et al. (2018) and Hahn et al. (2019) developed pre-screening methods based on bulk spectra of the samples using SVM and PLS-R respectively. The latter two works are one of the few examples of regression models.

Within the domain of μFTIR imaging, however, only spectral library search approaches (Primpke et al., 2017, 2018; Liu et al., 2019) had been published at that time. *Hufnagl et al. (2019)* described a random forest based approach (Breiman, 2001) for detecting five common polymer types using Bruker Hyperion 3000. Da Silva et al. (2020) compared PLS-DA and soft independent modeling of class analogies (SIMCA) for nine polymer types using Agilent Cary 620. *Weisser et al. (2021)* used the methodology of *Hufnagl et al. (2019)* to derive random forests for identifying sources of microplastics in drinking water. This model was also able to detect nine common polymers using Agilent Cary 620. The to date most advanced random forest which can detect more than twenty-one polymers was published by *Hufnagl et al. (2022)* for the devices Bruker Hyperion 3000 and Bruker Lumos II (see figure 4 for a schematic drawing of the analysis process). Other spectral library search based methods for μFTIR imaging which have been published after *Hufnagl et al. (2019)* include Primpke et al. (2020b) and Corradini et al. (2021).

This doctoral thesis also deals with unsupervised learning. *Hufnagl and Lohninger (2020)* studied the structural aspects of microplastics data using a newly developed clustering technique. Wander et al. (2020) applied uniform manifold approximation and projection (UMAP) as well as k-means and hierarchical density based clustering (HDBSCAN) in a workflow to analyze datasets from Agilent Cary 620 and Bruker Hyperion 3000.

Other examples of machine learning include Kedzierski et al. (2019) who developed a classifier using kNN (in this case $k > 1$) for ATR-FTIR. The training data was reused by de Medeiros Back et al. (2022) to benchmark random forests, SVM, logistic regression, kNN, decision trees and Gaussian naive bayes. Kumar et al. (2021) developed a random forest based approach for Raman spectroscopy. Recently, Brandt et al. (2021) demonstrated the potential of autoencoding neural networks (Hinton and Salakhutdinov, 2006; Van Der Maaten et al., 2009) for noise reduction and removal of spectral distortions in spectra originating from different techniques.

2.2.4 μFTIR Imaging: Pros and Cons

As depicted in figure 2 μFTIR imaging can be divided into focal plane array (FPA) based systems and line array (LA) based systems. In order to apply μFTIR imaging the sample needs to be purified using a sample preparation scheme. The literature describes various approaches to selectively remove and digest the environmental matrix which in turn increases the relative abundance of microplastics. One common step is density separation which will remove sediments from the sample. After that follows the chemical removal of the matrix by either alkaline, acidic, oxidative or enzymatic treatment (Renner et al., 2017b; Möller et al., 2020). In this thesis the samples have been prepared using an enzymatic purification scheme developed by Löder et al. (2017). While the treatment with enzymes is rather time consuming and costly it comes with the advantage that microplastics are not damaged or lost because of the use of strong chemicals (Hurley et al., 2018). On the other hand there are certain matrix components which cannot be digested using enzymes. As a result there are usually more remnants after sample preparation in comparison to other preparation schemes. In the context of this doctoral thesis this circumstance is in so far beneficial for developing machine learning based approaches as the data thus contains various disruptive factors which help to make the models more robust and generalizable.

After preparation the sample is concentrated on an IR transparent filter. Common examples include aluminum oxide (aka. Anodisc) and monocrystalline silicon wafers. Less common are polytetrafluoroethylene (PTFE) and metal meshes. Anodiscs become IR intransparent below 1250 cm^{-1}. Above 3595 cm^{-1} interferences with water induce a lot of noise. Because of this the measured spectral range is limited accordingly. The measurement setup for Bruker Hyperion 3000 is described in more detail in Löder et al. (2015).

The result of the measurement is a hyperspectral image with a spectral resolution of 4 cm^{-1} and a lateral resolution of *ca.* $11\mu\mathrm{m} \times 11\mu\mathrm{m}$. Typical dimensions are 1024×1024 pixels though they can be considerably larger. The given characteristics are achievable for Bruker Hyperion 3000, Bruker Lumos II and Agilent Cary 620, which are all FPA-based systems. For LA-based systems both the spectral resolution as well as the imaged area are usually considerably reduced in order to conserve measurement time.

Imaging, especially FPA-based imaging, comes with the advantage that the entire sample can be measured. As discussed earlier it is common in single point approaches to measure only a subsample of particles. Further, the whole contour of the particle is measured while single point approaches usually rely on a single spectrum at the center of the particle. This comes with the advantage that more information is available and also allows to deal with agglomerates of particles, which pose a problem for the image segmentation algorithms that locate the individual particles in single point approaches. Additionally, there is no measurement time dependence on the number of particles on the filter. Because of this also very complex samples such as sewage sludge or sediments require the same amount of measurement time as say drinking water. On the other hand this is also one of the arguments why μRaman is a preferred candidate for the monitoring of drinking water as it makes little sense to measure samples with μFTIR imaging if only a few particles are present on the filter.

However, μFTIR comes at a cost. The data which is produced during the measurement requires often more than 5 GB of memory (there are examples of up to 70 GB) and the number of pixels (or spectra respectively) is usually in the millions. Because of this the data analysis is non-trivial and requires a lot of time using the current state of the art. The perhaps most widely used software for this purpose is siMPle (Primpke et al., 2020b), a spectral library search based approach. According to Primpke et al. (2020a) the analysis time needed to process a single sample is about 4 h to 48 h. Here the analysis time that is needed for the expert audit of the data is not included. While FPA-based μFTIR imaging comes with many advantages from an instrumental perspective, the large amounts of data produced by a single measurement create a bottleneck for data analysis. This has prevented the success of this method so far.

3 Research questions and goals of this thesis

The previous sections gave a brief introduction to the problem of microplastics analysis with special focus on analytical instruments, typical challenges that arise with microspectroscopy and the current state of data processing approaches. A common thread running through all discussed aspects is that there is no single approach, be it instrumental or data analytical, that provides a solution for the problem of microplastics detection and quantification in all its aspects. This is in part rooted in the fact that there are conflicting definitions of microplastics. Moreover, there are concerns voiced that it is too early to standardize and harmonize methods and that further research and development is needed.[1]

[1] The author observed that at the SETAC Europe 31st Annual Meeting Virtual Conference the question was raised for whom the methods should be standardized and harmonized. Interestingly, this question could not be answered. This is also in accordance with the polls conducted at the JRC online symposium (09.09.21) *Challenges of microplastics analysis - bridging state of the art and policy needs* which revealed that most attendees felt that most of the focus should be laid on method development.

Comparing the state of the art of data processing in microplastics analysis with other fields which make use of spectroscopy (e.g. remote sensing, process analytical technologies, ...) it is indeed interesting to find that only a few researchers made use of machine learning approaches for microplastics analysis and not a single one of the proposed methods has been used in a broader sense (e.g. like monitoring studies). On the other hand Renner et al. (2017b) stressed the need to develop more robust spectral library search approaches to address the already discussed issues. In the view of the author this raises the question why machine learning methods have been used so scarcely in this field and what potential impact the broader use of it could have on microplastics research and monitoring, especially if one considers successful applications from other fields (Plaza et al., 2009; Zhai et al., 2021).

From these considerations the author deduces the following research questions:

1. What are the structural aspects of microplastics measurement data which govern the performance of machine learning approaches?

2. What are key obstacles that need to be overcome to apply machine learning more broadly for microplastics detection and quantification?

Considering the different microscopy devices and spectroscopic techniques listed in figure 2 μFTIR imaging is ideal to provide answers for these research questions as the datasets contain a wealth of data examples for testing different hypotheses and the lateral context (the fact that we are dealing with images) adds another possibility to validate results based on human expert knowledge.

The goals of this thesis are thus to

1. develop a clustering algorithm for the study of the structural aspects of hyperspectral imaging data and

2. build a classifier for microplastics detection and quantification to gain knowledge about the methodological challenges.

4 Methodology

In the more classical sense machine learning algorithms (Hastie et al., 2009) can be divided into *supervised* and *unsupervised* learning approaches. Supervised learning infers a function from labeled training data, which is a dataset of objects and associated labels. The perhaps most matured example is classification (Domingos, 2012). Unsupervised learning, on the other hand, has to infer such a function by self-discovering patterns within the training dataset, as there are no labels provided with the data. These differences come with the consequence that unsupervised learning is often applied in exploratory settings or situations, where labeled training data is difficult to come by. Expert annotation, the process by which training data is labeled, can be costly and time consuming (Sheng et al., 2008). Further, experts may overlook certain aspects of the data which might become visible by using unsupervised learning techniques such as cluster analysis (Jain et al., 1999; Xu and Wunsch, 2005; Jain, 2010). Supervised learning is thus preferable in situations where a task with a definable objective has to be repeated multiple times.

Before pursuing the aim of building a classifier for detecting microplastics a profound understanding of structural aspects of μFTIR imaging data should be obtained. Combining the knowledge obtained from applying both unsupervised and supervised learning further allows to compare and validate two very antithetic approaches with each other. For this reason the first research which was done for this doctoral thesis focused on cluster analysis.

4.1 Cluster analysis using a novel graph-based approach

Even though cluster analysis is a representative of unsupervised learning there are usually some assumptions that have to be made by an expert from which decisions can be derived, such as the kind of data preprocessing, the applied similarity measure and the selection of the clustering algorithm. In the case of μFTIR measurement data of membrane filters there is already a lot of information available which can be used to formulate assumptions about structural aspects of the data and the objective of the clustering:

- The objective of the cluster analysis is the following: given a μFTIR image the task is to detect microplastics in the sample and assign them to different clusters according to their polymer type.

- The number of clusters is unknown in advance.

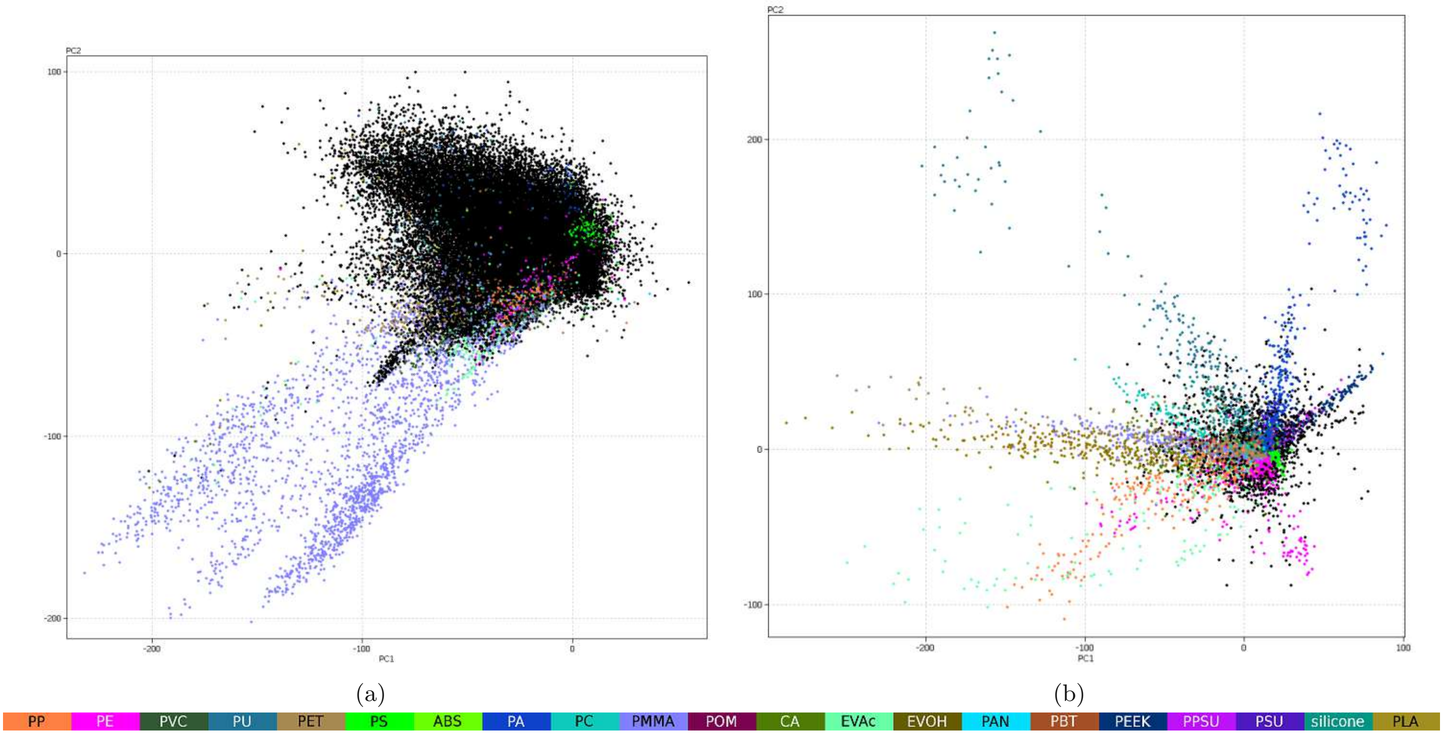
- Other particles which are present on the membrane filter originate from the environment and can therefore be very heterogeneous regarding their chemical composition.

- The spectra (or pixels respectively) of the membrane filter have been background corrected.
- Due to the limited lateral resolution of FPA-based μFTIR imaging, or because particles may overlap, mixed spectra are to be expected.

From this information we can now derive certain assumptions about the datasets. A basic model for a cluster that corresponds to a chemical compound (e.g. microplastics of the same polymer type) is a hyperellipsoidal data distribution. One end of the hyperellipsoid structure is closely located at the coordinate origin while the other end protrudes into the d-dimensional feature space. Moving from the origin along the main axis of the hyperellipsoid the spectral absorption increases. The extent of the other axes correspond to variations in the spectra and noise. The spectra of the membrane filter form a hyperspherical data distribution at the origin. Because of the expected mixture effects the clusters will not be clearly separated from each other, but will be connected by spectra which can be seen as linear combinations of the different chemical compounds and the filter.

Another assumption is about the spatial extent of the clusters. Consider the case that we have a very homogeneous constituent on the filter, then its spectra will show only minor variations which may result in a very compact cluster. On the other hand, inhomogeneous material, such as remnants from the environment, may cause diffuse clusters of large extent. Further, there might be marginal differences in the spectra which may be indicative of different species.

Figure 5 depicts two PCA plots where structural features of a μFTIR imaging dataset and a training dataset for classification are depicted. Even though these PCA plots are two-dimensional projections of a very high dimensional feature space the visible structures seem to support the stated assumptions.



(a)

(b)

(c) class colors

Figure 5: PCA plots of (a) a μFTIR imaging dataset *(Hufnagl et al., 2019)* which is available through 10.5281/zenodo.2555732 and (b) the training dataset which was used to derive the random forest model in *Hufnagl et al. (2022)*. The colors in (a) correspond to the classification result obtained by the random forest and in (b) they correspond to the expert labels. For both datasets the first derivative of the spectra was used to compute the PCA plot.

A common approach in the development and comparison of clustering algorithms is to apply them to 2D datasets where point clouds depict different forms of separability problems (Zahn, 1971; Veenman et al., 2002; Gionis et al., 2007; Fu and Medico, 2007; Jain and Law, 2005; Chang and Yeung, 2008). Even though separating clusters in the well-established benchmarks is very challenging, the stated assumptions about the structural aspects of μFTIR imaging data is hardly reflected in them. For this reason a benchmarking of common clustering algorithms based on datasets which incorporate such structures was already conducted in the author's diploma thesis (Steindl, 2018, p. 67-77).

Another topic of that diploma thesis was the development of a new graph-based clustering algorithm dubbed *graph-based competitive clustering* (GBCC) which was also tested on these datasets. GBCC follows the idea that clusters can be separated by using the gradient of distances between consecutive pairs of observations along a path to determine the longest distance as a cluster border. The concept is independent of the scales of the clusters, meaning that a border can be determined independently of the relative extent of the clusters. It was found, that only GBCC and KNSC (Shi and Malik, 2000), which is a representative of spectral clustering (Von Luxburg, 2007), could solve certain separability problems. However, the application of GBCC to hyperspectral imaging datasets yielded rather unsatisfying results as the original methodology of detecting dense centers for starting the algorithm was, back then, not very sophisticated. The method proved to be not sensitive enough to detect minor variations of density in the data, which is key for detecting microplastics in μFTIR datasets.

In *Hufnagl and Lohninger (2020)* GBCC was revised to address the stated issues. First the concept behind the center detection was changed. Secondly, the source code was rewritten from scratch and optimized in multiple iterations to increase its computational speed. A specialized sparse matrix structure was developed, as the standard implementation in MATLAB (www.mathworks.com) proved to be too slow for the clustering step. Further, the graph computation was parallelized to speed up the computation of the k nearest neighbor graph. GBCC was then compared with k-means, DBSCAN (Ester et al., 1996) and KNSC on a multi-challenge 2D dataset (Hufnagl and Lohninger, 2019). In order to assess GBCC's performance with respect to hyperspectral imaging problems it was applied to a multimodal hyperspectral image of precipitated atmospheric particulate matter combining Raman spectroscopy with energy dispersive electron probe X-ray (EDX) (Ofner et al., 2015) as well as a μFTIR imaging dataset of microplastics *(Hufnagl et al., 2019)*. The former dataset has been thoroughly studied using PCA, hierarchical clustering (HCA), k-means and vertex component analysis (VCA) (Nascimento and Dias, 2005) which allowed the comparison of GBCC's results with those from an independently working expert. As research and development for the random forest classifiers was done in parallel to this work the results published in *Hufnagl et al. (2019)* could be compared to the one obtained with GBCC.

4.2 Classification using random forests

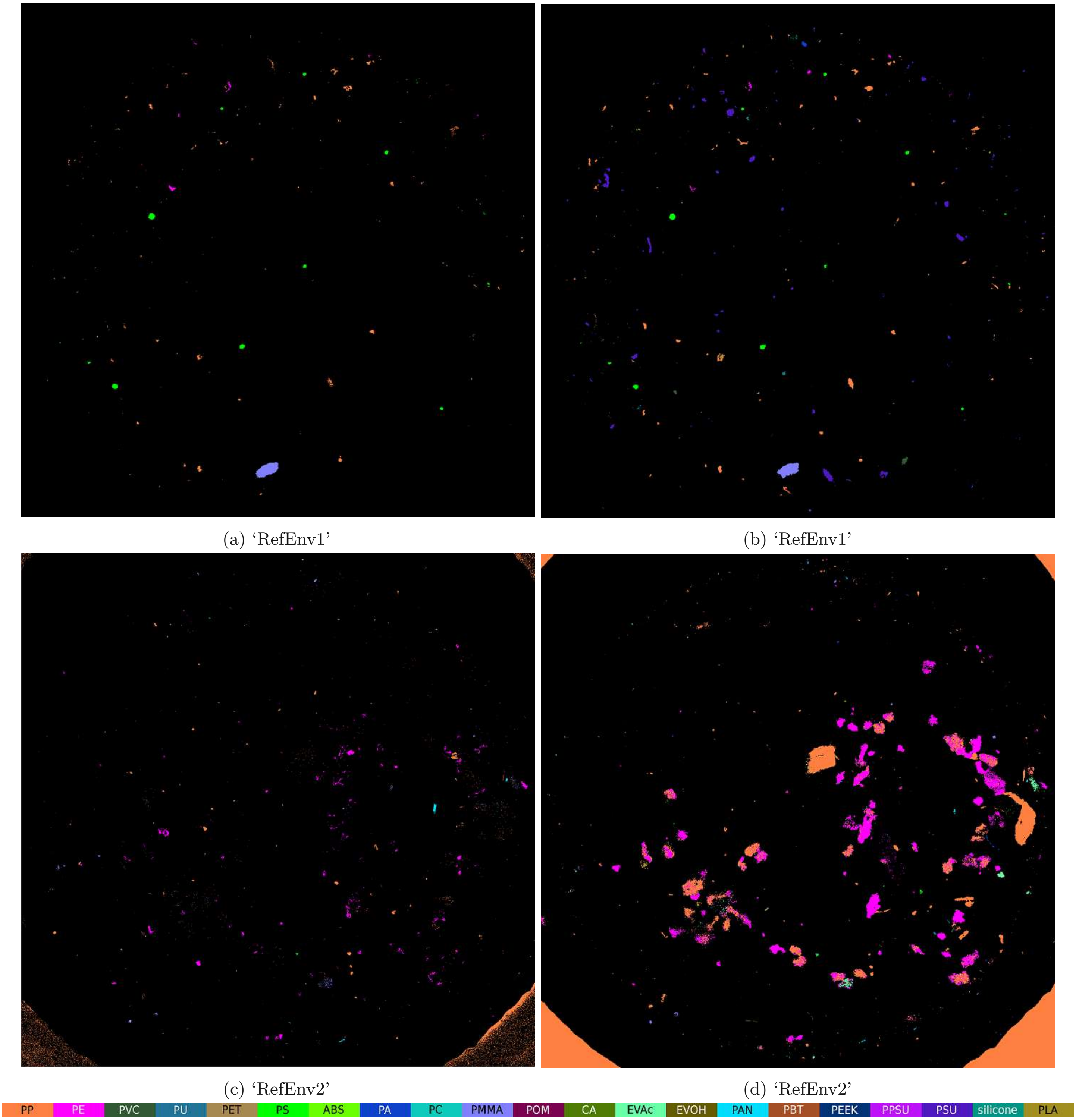
Leveraging the knowledge gained from the previously described clustering experiments a preliminary study *(Hufnagl et al., 2019)* for the development of classifiers was conducted to understand the behavior of random forests (Breiman, 2001) in the context of microplastics classification. The decision to use random forests was based on their performance regarding variable selection in high-dimensional classification problems and their ability to solve non-linear problems.

For the creation of a labeled training dataset the author formed a group of three experts, including himself, to establish a *ground truth* for the polymer types polyethylene (PE), polypropylene (PP), poly(methyl methacrylate) (PMMA), polyacrylonitrile (PAN) and polystyrene (PS). It was assumed from the beginning that in order to create a broadly applicable classifier enough variability has to be ensured in the training data to also reflect differing expert opinions. As a source for labeled microplastics spectra a spiked sample measured with a Bruker Hyperion 3000 was used. Spectra from the matrix and the filter surface were sampled as well.

A common approach in the data processing of spectra is to subject them to a form of preprocessing, such as baseline correction or noise filtering (Renner et al., 2018). In this study a different approach was taken by applying spectral descriptors (SPDCs) (Lohninger and Ofner, 2014). In short, SPDCs are simple mathematical functions which can be defined by a spectroscopy expert to map characteristic vibrational patterns in the spectra, such as a certain peak, to a single descriptive number. By doing so for multiple classes of microplastics the expert creates a set of functions which can be used to map the original raw data feature space into a descriptor space. This usually also means a significant dimensionality reduction from hundreds of spectral sampling points to usually less than 50 SPDCs. Further, this method also has a de-correlating effect on the data, which is always beneficial for the performance of classifiers.

In an initial experiment it could be shown that when the inferred random forest was applied to other datasets from the literature (Primpke et al., 2018) the model performed poorly and created lots of false positives, especially when considering the matrix and thick particles due to total absorption. By sampling additional matrix spectra the performance of the random forest model improved significantly. However, datasets with microplastics that exhibit strong total absorption, but where, on the other hand, the ground truth can be assured, where not available at that time, which is why this matter was addressed later on. Classification results which have been obtained with this model are depicted in figures 6a and 6c. Performance metrics of the random forest model where computed by means of out-of-bag estimates (Breiman, 2001; Biau and Scornet, 2016) as well as by binary confusion matrices, which were obtained from a separate test dataset.

In a follow-up study *(Weisser et al., 2021)* the method described in *Hufnagl et al. (2019)* was adapted for the context of microplastics detection in mineral water. Here a training dataset for the polymers polyethylene (PE), polypropylene (PP), polyethylene terephthalate (PET), polyvinylchloride (PVC), polystyrene (PS), polytetrafluoroethylene (PTFE), polyamide (PA), polylactic acid (PLA), ethylene vinyl alcohol (EvOH) was assembled. Further, cellulose fibers from lab coats and wipes as well as skin particles were sampled to account for possible contamination during sample preparation. The data was



(a) ‘RefEnv1’

(b) ‘RefEnv1’

(c) ‘RefEnv2’

(d) ‘RefEnv2’

(e) class colors

Figure 6: A comparison of classification results on two μFTIR datasets using the random forest models. (a) and (c) have been analyzed using the model of *Hufnagl et al. (2019)* whereas (b) and (d) have been analyzed using the extended model of *Hufnagl et al. (2022)*. ‘RefEnv1’ and ‘RefEnv2’ were published by Primpke et al. (2018) under CC BY 4.0. (a) and (c) are available in the supporting information of *Hufnagl et al. (2019)* under CC BY 3.0.

measured using an Agilent Cary 620. In *Hufnagl et al. (2019)* the validation was based on binary confusion matrices. While these give a more detailed insight into the behavior of each binary classifier the cross-dependencies between the classifiers can not be assessed. For this reason a script for computing multi-class confusion matrices was implemented to validate the new model. Global multi-class performance measures (Ballabio et al., 2018) as well as sensitivity, specificity and precision of the ensemble of classifiers could thus be computed.

Using the gathered experience from the previous studies the original training dataset used in *Hufnagl et al. (2019)*, which consisted of 3270 spectra, was iteratively enhanced to address different issues such as total absorption, Mie scattering and polymer ageing. In the latter case it was found, that the model sometimes had difficulties identifying weathered microplastics in real environmental samples. Even though in the encountered cases the spectral signatures still allowed an easy identification of the polymer type, parts of the particles were not identified or the result was very noisy. Considering, that the original training data was based on spiked samples where microplastics had been created artificially from reference materials the original model was thus biased. By using a larger collection of real world samples the robustness of the model could be improved for the more common polymers by also sampling weathered microplastics spectra. Further, the model was also enhanced with training data sampled from Bruker Lumos II datasets. In this way the model became applicable for both types of devices. In *Hufnagl et al. (2022)* the number of spectra in the training data already exceeded 12000 spectra of which 50% were matrix spectra. This made the model very robust with respect to different kinds of matrices, including sediment, soil, compost and sewage sludge. Further, the number of polymer types which can be identified by the model was increased to 21, making it the model with the broadest applicability so far. In figure 6 the random forest models from *Hufnagl et al. (2019)* and *Hufnagl et al. (2022)* are compared based on two well-studied datasets from the literature.

In order to reduce the effect of label noise in the microplastics training spectra a special auditing scheme was developed. Label noise (aka. class noise) (Nettleton et al., 2010; Frénay and Verleysen, 2014) arises if the annotating expert makes a mistake when defining the label of an observation of the training data. There are different reasons why label noise may occur. It may be that the annotator is distracted at a certain moment and simply sets the wrong label. Another reason may be that the data that shall be annotated is difficult to interpret which makes the process subjective. Reidsma and op den Akker (2008) showed the effect subjectivity has on the inferred classifiers by using datasets which have been annotated by multiple experts for a natural language processing task. Not surprisingly, the classifiers performed better when they were applied to test data annotated by the same expert, while they performed inferior on other test data as a consequence of the subjectivity of the other experts. This experiment inspired a data auditing process, which was used in *Hufnagl et al. (2022)* to reduce possible label noise from the training data. From each annotating expert a random forest model was inferred based on his or her training data. This model was then applied to the training datasets provided by the other annotating experts. If e is the number of experts, then each spectrum thus received an expert label plus $e - 1$ votes from the random forest models. In this way labeling errors could be identified easily because of the low agreement of votes cast by the models and the label given by the expert.

In the context of this thesis *Ritschar et al. (2021)* forms the basis for future research regarding the detection of microplastics in tissues. In this work the method described in *Hufnagl et al. (2019)* was adapted for segmenting tissue sections of the model organism *Eisenia fetida* using μFTIR imaging in order to identify target tissue for ecotoxicological analysis. In a second step MALDI-MS imaging is applied for detecting species which are indicative of metabolic changes. The tissue was segmented into three classes, which were 'muscle', 'digestive system' and 'other tissue'. A fourth class accounted for the background. In order to ensure that the random forest model generalizes well on new datasets different sections and different worms were used as a source for training data. The inferred model was then validated by means of Monte Carlo cross validation (Xu and Liang, 2001; Westad and Marini, 2015). Figure 7 shows the application of the random forest model to different thin-sections.

5 Results and Discussion

The following discussion of the results will lay the focus on aspects of the cumulated papers (see table 1) which are relevant in the context of the research questions and goals of this thesis and the related microplastics and machine learning research literature. Additional findings which are not central to this thesis are mentioned in section 7. The reader is advised to read the cumulated papers before proceeding.

5.1 Structural aspects of microplastics data

5.1.1 Subjective annotation and label noise

If we follow the aim of building a supervised machine learning approach for performing a specific task we have to provide information to the learner which describes the objective in terms of labeled training data. In the context of this doctoral thesis this means that we have to sample spectra from μFTIR datasets and annotate them using predefined class labels. Even though this seems trivial at first glance there are various effects that make the interpretation of the spectra very difficult. As discussed in sections 2.2.2, 2.2.4 and 4.1 we have to deal with e.g. polymer ageing, total absorption, Mie scattering and overlapping particles.

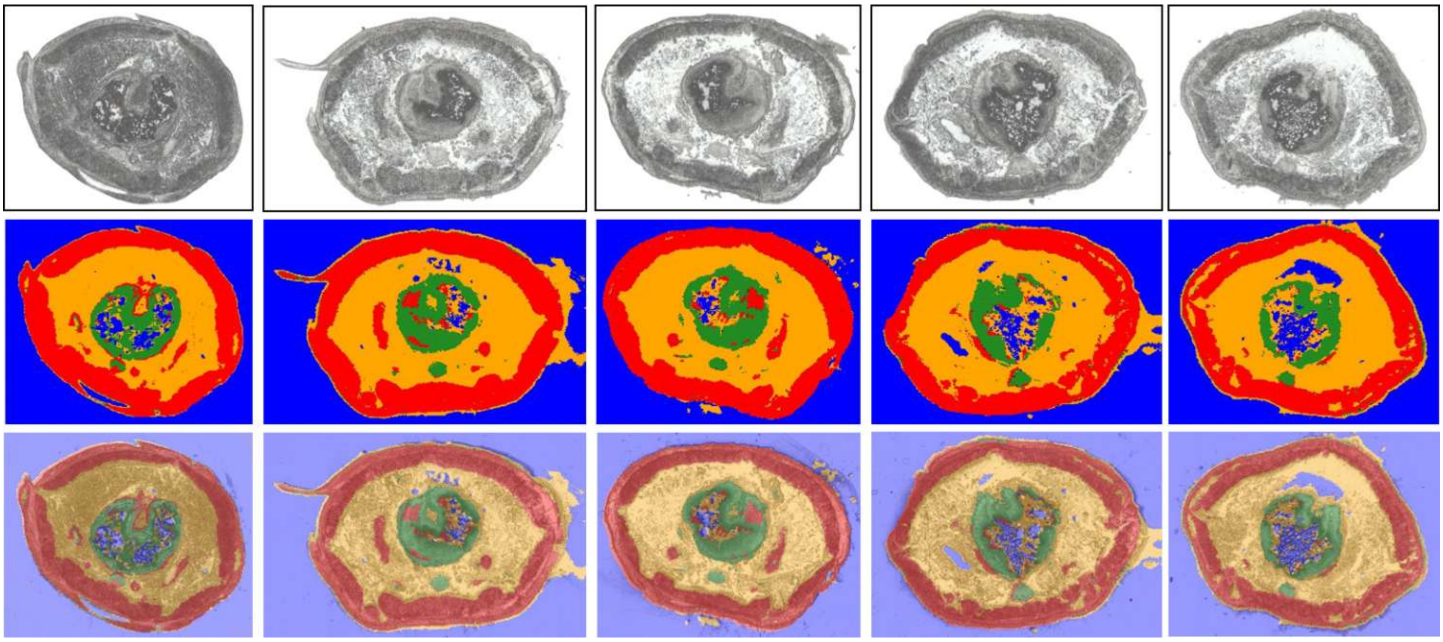


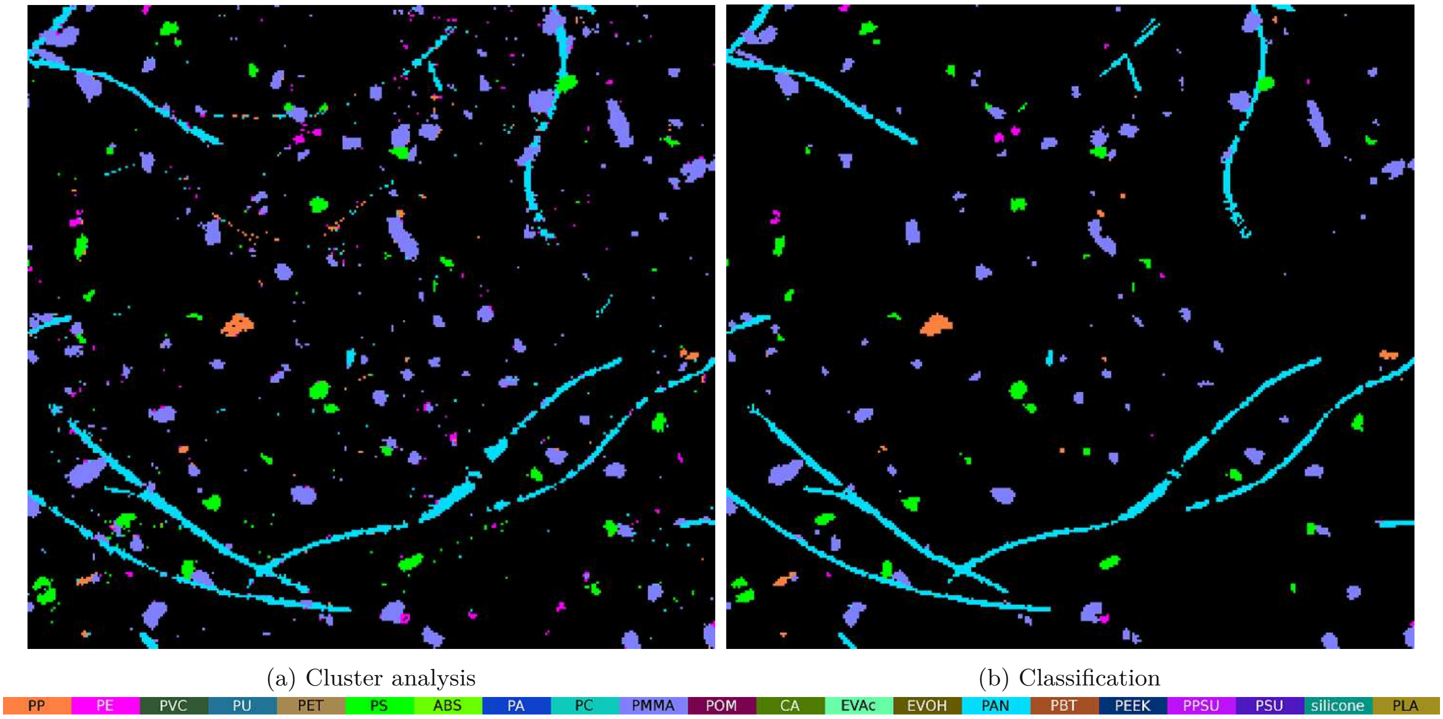
Figure 7: Tissue sections of *Eisenia fetida* measured using μFTIR and classified with random forests. blue=‘background’, red=‘muscle’, green=‘digestive system’, orange=‘other tissue’. *Adapted from* Ritschar et al. (2021) *under CC BY 4.0.*

As a simple example consider the case where the filter just contains artificial microplastics with no matrix present. Here, the decision what spectra should be chosen for representing the particle edges is still problematic, because weak remnants of the spectral signatures of a certain microplastics particle might still occur in pixels which are already relatively distant to the observed edge in the light microscopy image. Thus, it is not obvious for an expert which pixels should be attributed to the particle edge or the filter. If we now consider the effects that we are confronted with in a real world sample the task of sampling and annotating spectra becomes even more difficult. Because of this experts might also disagree regarding the way spectra are sampled and regarding the chosen labels.

With the advent of crowdsourcing solutions (Welinder et al., 2010) that facilitate the acquisition of large quantities of non-expert labels for applications such as natural language tasks (Snow et al., 2008) and computer vision (Tommasi et al., 2017) there has been an increased awareness of the problem that both the data as well as the labels may be erroneous and biased (Frénay and Verleysen, 2014). One might argue that subjective tasks, such as describing the emotional content of newspaper headlines, is far removed from annotating polymer spectra yet there is literature to be found also in the natural sciences which address similar issues (Millard and Richardson, 2015; Malossini et al., 2006; Hughes et al., 2004; Albert and Dodd, 2004; Smyth et al., 1995).

Subjectivity in the context of mathematical modeling is also discussed by Hennig (2010), who takes a constructivist perspective and distinguishes between our personal reality and the observer-independent reality to which we have no access to. In the context of machine learning the labeled training data is sometimes referred to as the *ground truth*. Yet beyond a purely mathematical context this term is misleading as we cannot define an observer-independent truth. This ultimately also applies to the annotation of spectra as could be observed when assembling the training data annotated by different experts *(Hufnagl et al., 2019)*. If our goal is to create a classifier which makes predictions that are accepted by many different users it therefore makes sense to ensure that different personal truths regarding the sampled and annotated spectra are reflected within the training datasets.

Besides that the process of annotating spectra is subjective, it is also tedious, considering the large number of spectra which have been sampled for building the random forest classifiers. Inevitably, the annotating expert will sooner or later make mistakes due to distraction or fatigue. It is therefore important to consider that these labeling errors can have an effect on the classification. In that context the type of machine learning approach has to be considered because algorithms exhibit different levels of robustness regarding the label noise effect (Folleco et al., 2008). Frénay and Verleysen (2014) give a review about different ways to deal with label noise. The already discussed approach which was used in *(Hufnagl et al., 2022)* provided a simple and yet effective approach to spot labeling errors in the training data.



(a) Cluster analysis

(b) Classification

(c) class colors

Figure 8: A comparison of GBCC's clustering result (a) *(Hufnagl and Lohninger, 2020)* with the classification result (b) as published by *Hufnagl et al. (2019)*.

5.1.2 Clustering vs. classification

The impossibility of defining an observer-independent truth also concerns cluster analysis. In a later paper Hennig (2015) discussed the question 'What are the true clusters?' as comparing and validating clustering algorithms ultimately depends on what we define as a cluster or the 'true' clusters. It is thus not surprising that there is no general definition of a cluster. It follows that what we perceive as a 'good' clustering in the end depends on the goal which we want to achieve and does not constitute some fundamental observer-independent truth.

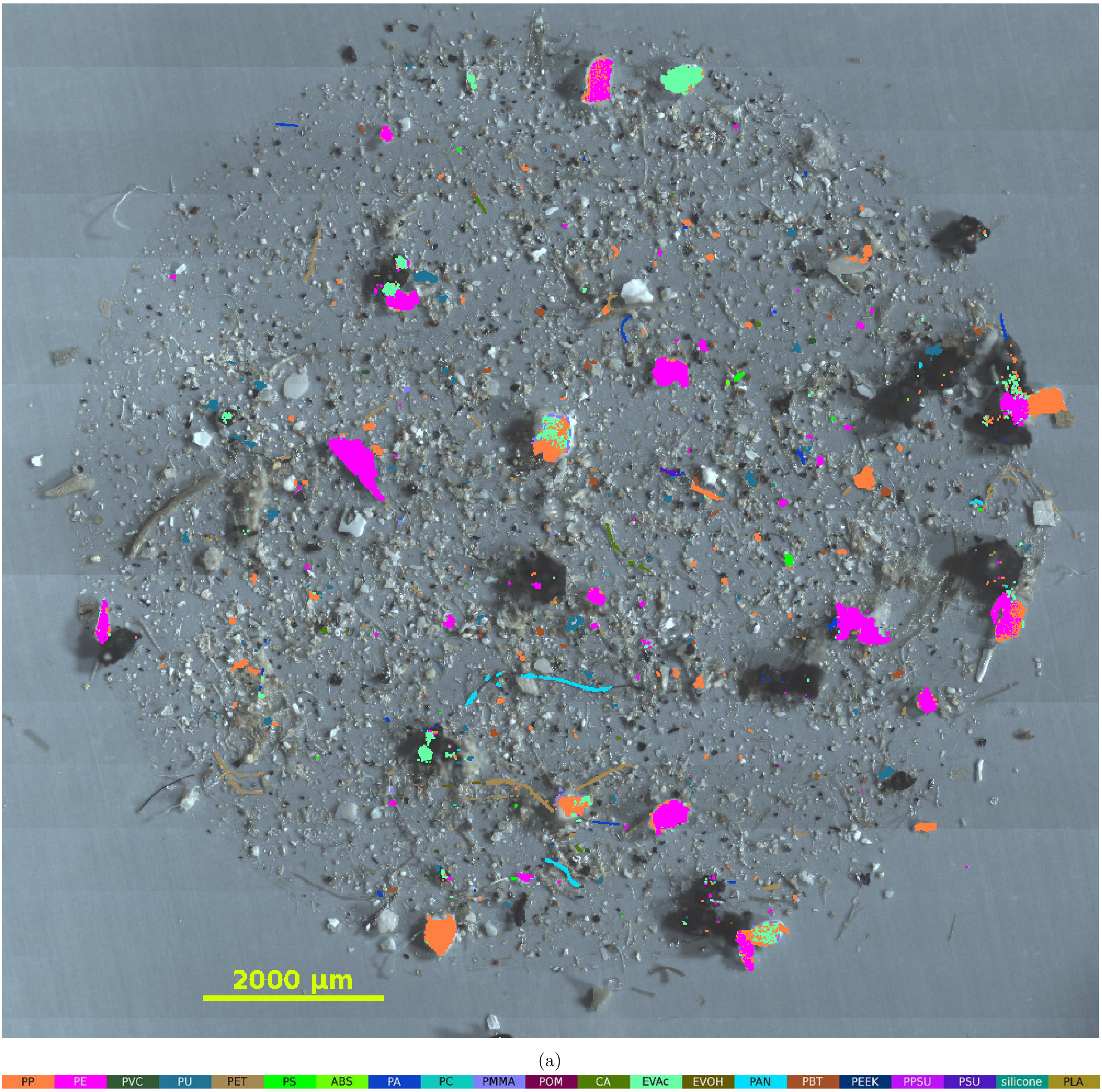
However, within the literature that deals with the development of clustering algorithms it is still a common validation approach to apply the algorithm to labeled training datasets that originally have been designed for a specific classification task. The result of the clustering algorithm is then compared to the labels of the training data, which is used as a ground truth to compute error rates. The problem with this approach is, that it is a mere assumption that class labels and the data structure coincide in some way (Von Luxburg et al., 2012). Further, as with classification algorithms, the data that is the target of the analysis might have a very different structure.

This is evident if one compares figures 5a and 5b where PCA plots of a μFTIR dataset and the training dataset used in *Hufnagl et al. (2022)* are depicted. Even though both plots show the hyperellipsoid structures which have been described in section 4.1 the labeled training data is a very incomplete picture of the μFTIR data. The reason is simply that the experts chose not only the labels but also the spectra. Therefore, the labeled training data is only a subsample that depends on the ability of the expert to define a label. The experts didn't sample spectra where they were unsure what label should be assigned. This is typically the case if one encounters a spectrum which resembles a polymer but also looks very similar to some matrix component.

In *Hufnagl and Lohninger (2020)* the clustering result obtained with GBCC was not compared to labeled training data, but instead to the classification result obtained by the random forest model of *Hufnagl et al. (2019)* as can be seen in figure 8. This validation approach is not better than using labeled training data as a benchmark, because the criticism voiced by Von Luxburg et al. (2012) also translates to this case, but the comparison allowed an interesting insight. By comparing the found microplastics it can be seen, that the particles which have been detected by means of cluster analysis are often slightly larger than the ones detected by means of classification. An explanation for these differences might be, that the variability of the spectra at the particle edges is considerably higher than at the center. Therefore, it is very difficult to sample enough spectra that the edges are well reflected in the training data. Further, the choice of adding a certain spectrum from the edge to a microplastics class is not always obvious, as the characteristic spectral signature might already be very weak and distorted because of scattering effects. Once the random forest is trained the decision boundaries are fixed. GBCC, however, assigned these spectra due to their proximity in the feature space. It will therefore be more flexible regarding mixture effects.

In conclusion it proved to be a useful comparison as GBCC highlighted weaknesses in the classification results that can not be detected by means of classification performance measures (Ballabio et al., 2018). It thus makes sense to validate the performance of a classifier with unsupervised learning approaches such as cluster analysis. In a similar way the results published by Wander et al. (2020), who used PCA, k-means and UMAP to cluster the dataset 'RefEnv2' (Primpke et al., 2018), can be compared to the random forest results shown in figures 6c and 6d.

5.1.3 Spectral variability of microplastics and matrix components



(a)

(b) class colors

Figure 9: Waste water treatment plant outlet sample analyzed using random forests. Published by *Hufnagl et al. (2022) under CC BY 4.0.*

The perhaps most striking example why the variability of the microplastics spectra needs to be modeled can be seen by comparing figures 6c and 6d. Here the model developed in *Hufnagl et al. (2019)* could not detect any of the larger (and thus thicker) particles while the improved model described in *Hufnagl et al. (2022)* could detect them and even distinguish between PP and PE. However, there are some particles where the class affiliation is noisy, meaning that different polymers have been assigned at neighboring pixels. The reason for this is that with increasing total absorption there is less and less information which can be used to distinguish between the polymers. For example PP, PE and EVAc look very similar if one compares spectra which exhibit strong total absorption. Because of this thick particles sometimes show a noisy mixture of these three classes as can be seen in figure 9. Figures 6a and 6b also show that the increase in training data also had a positive effect on smaller microplastics if one closely compares the detected contours in both images.

Overall the models performance with respect to false positive rates has improved considerably by ensuring that matrices from different environmental origins have been incorporated in the training data. The broad applicability regarding different matrices is also shown in *Hufnagl et al. (2022, fig. 1)*, where the random forest has been applied to plankton, sediment, soil, compost and other matrices. However, cases of false positives remain which are often caused by noisy total absorption spectra of matrix components. These misclassified cases arise because a very noisy spectrum will traverse the decision trees of the random forest and arrive at a leave node of a polymer with a certain non-zero probability. While this seems unlikely one should keep in mind that we are dealing with millions of spectra that are classified during the analysis.

5.1.4 Device bias

An important aspect regarding the problem of building a machine learning model that generalizes well across different μFTIR devices could be identified during the research conducted for the publication by *Weisser et al. (2021)*. The model developed in *Hufnagl et al. (2019)* could detect microplastics in the data that was measured using the Agilent Cary 620, however, with a decreased performance. While Bruker Hyperion 3000 and Agilent Cary 620 are very similar devices, there are certain differences which create slightly different data. These include, for example, the optical pathway, the detector and the apodization function. In *Hufnagl et al. (2022)* the training data was already extended to also include spectra from other devices such as Bruker Lumos II. It was found that the performance of the model improved for these devices, however, further research is needed to better understand the influence certain differences between the μFTIR devices have on the classification result. A comparative study is insofar difficult as this would require a lab where different μFTIR imaging devices are located next to each other. A transport of filter samples between labs would not make sense in that case because some particles would move or worse be lost, making the data incomparable.

5.2 Performance of the learning approaches

5.2.1 Throughput rate

A major factor that governs the throughput rate regarding the number of samples which can be processed within a certain time frame is the computational speed of the algorithms. In this context clustering is very slow compared to classification. The computation times required to cluster the dataset shown in figure 8 using GBCC *(Hufnagl and Lohninger, 2020)* was more than a day. The specific hardware used for the computation was the same as the one used in *Hufnagl et al. (2019)*. The main bottleneck was the computation of the kNN graph even though the process is parallelized. In direct communications with the first author of Wander et al. (2020) it was found that the long computation times are probably related to the used code base, which was MATLAB R2016b. According to the author the python scripts for UMAP finished computations on much larger datasets within a few hours, even though UMAP also relies on kNN graphs. While there might be room for further improvement regarding the computational speed GBCC remains an exploratory analysis approach and is unsuitable for the processing of multiple samples on a routine basis.

Regarding classification the situation is quite different. The performance assessment conducted in *Hufnagl et al. (2019)* based on the random forest for six binary classifiers allowed an estimation of computation times for 20 polymers and 10^6 pixels, which yielded 15 min. In *Hufnagl et al. (2022)* this estimation proved to be correct as the random forest for 22 binary classifiers requires about 20 min. By using a statistical detection method for masking the pixels that cover the filter surface the computational time can often be reduced by more than 50%. The computation time is thus no longer a bottleneck for the analysis. The software siMPle (Primpke et al., 2020b), which is based on spectral library search, requires more than 4 hours in comparison (Primpke et al., 2020a).

Another important aspect is the number of parameters that needs to be chosen by the user for the application of the approach. In siMPle the user has to optimize more than 3 parameters for samples of different matrix types. Considering that the computation time is rather long this can become a very tedious iterative process. The design philosophy behind the random forest model is that all parameters are fixed and that no optimization has to be done by the user. This can save the user a lot of time if the samples have a high variability.

5.2.2 Statistical performance

The performance assessments based on confusion matrices for both models show that the most difficult training problem is the differentiation between microplastics and other spectra while wrong assignments among polymer classes are rare *(Hufnagl et al., 2019, fig. 3)(Hufnagl et al., 2022, tab. S2)*. This highlights the importance of the 'Non-Polymer' or 'Other' class regarding the overall result. False positives, meaning spectra which are wrongly identified as microplastics, and false negatives, true microplastics which are not detected, are in the end the most important performance indicator. However, even though the accuracy of the random forest for 21 polymer classes is convincing, one should keep in mind that the conducted Monte Carlo cross validation (Xu and Liang, 2001) has its limitations. As discussed more in-depth by Westad and Marini (2015) a different approach to cross validation would be to split the samples according to the measurement devices, the matrices or any kind of other factor that may have a strong influence on the data. This would give more information about the stability of the model and show sources of variation that need special attention for future development. Further, cross validation is commonly known to cause overfitting, if it is used iteratively during the development process, which in turn causes an underestimation of the classification error. Therefore a truly independent test dataset, ideally assembled by an independent research group, would be more suitable for error rate estimation.

6 Scientific Contribution

6.1 Conference Contributions

- Peeken et al. (2021)
- Hufnagl et al. (2021)
- Wander et al. (2021)
- Speaker at the *Agilent - Microplastics in the Environment Virtual Symposium 2021* on the topic 'Machine Learning vs. Databases: A question of speed, accuracy and scalability'; (30th of September 2021)
- Speaker at the *Final online workshop MISSOURI project* on the topic 'Robust Ultra-Fast Analysis of Microplastics in Large μFTIR Imaging Datasets using Machine Learning'; (15th of October 2021)

6.2 Application studies

Earlier as well as later versions of the random forest model have been used in environmental application studies (Frei et al., 2019; Kumar et al., 2021; Möller et al., 2021; Teichert et al., 2021). Further, Dong et al. (2022) used the model of *(Hufnagl et al., 2022)* in a study to compare analysis results obtained with the devices Bruker Lumos II, Agilent 8700 LDIR and WITec alpha300 R.

7 Summaries of the scientific publications

7.1 Hufnagl and Lohninger (2020)

Title A graph-based clustering method with special focus on hyperspectral imaging

Authors Benedikt Hufnagl and Hans Lohninger

Abstract A common trait of the more established clustering algorithms such as k-means and HCA is their tendency to focus mainly on the bulk features of the data which causes minor features to be attributed to larger clusters. For hyperspectral imaging this has the consequence that substances which are covered by only a few pixels tend to be overlooked and thus cannot be separated. If small lateral features such as particles are the research objective this might be the reason why cluster analysis fails. Therefore we propose a novel graph-based clustering algorithm dubbed GBCC which is sensitive to small variations in data density and scales its clusters according to the underlying structures. The analysis of the proposed method covers a comparison to k-means, DBSCAN and KNSC using a 2D artificial dataset. Further the method is evaluated on a multisensor image of atmospheric particulate matter composed of Raman and EDX data as well as an FTIR image of microplastics.

Findings and impact Within that work 'gradient-separable' was discussed as a new separation problem for clustering and a new 2D clustering benchmark dataset was introduced. The structural aspects of hyperspectral imaging data in the context of the applied similarity measures and clusters of large differences in scale was discussed, which highlighted the importance of understanding the relation between the aim of the cluster analysis and the used methodology. The comparison of established clustering algorithms with the new graph-based algorithm 'GBCC' using the 2D benchmark dataset showed that there are vast differences in the results regarding gradient-separability and showed that only graph-based algorithms like GBCC and KNSC scaled their clusters according to the underlying data structures. The comparison of GBCC's results with Ofner et al. (2015) demonstrated the usability of the center detection step as a means to detect chemical compounds and thus provides an alternative to spectral endmember extraction techniques such as VCA. GBCC also proved to work in high-dimensional feature spaces as could be demonstrated on the Raman dataset, which contains 1024 spectral variables. Seen in the broader context of graph-based clustering algorithms GBCC contributes to the clustering of directed graphs, meaning that edges are not bi-directional.

Contribution B.H. wrote the manuscript and performed the associated research, including the methodological development, implementation of algorithms and analysis and interpretation of the results. H.L. supervised the research and critically reviewed the manuscript. B.H.'s estimated contribution is 95%.

Citations CrossRef: 3; Google Scholar: 7; (retrieved on the 3rd of January 2022)

7.2 Hufnagl et al. (2019)

Title A methodology for the fast identification and monitoring of microplastics in environmental samples using random decision forest classifiers

Authors Benedikt Hufnagl, Dieter Steiner, Elisabeth Renner, Martin G. J. Löder, Christian Laforsch and Hans Lohninger

Abstract A new yet little understood threat to our ecosystems is microplastics. These microscopic particles accumulate in our oceans and in the end may find their way into the food chain. Even though their origin and the laws governing their formation have become ever more clear fast and reliable methodologies for their analysis and identification are still lacking or at an early stage of development. The first automatic approaches to analyze µFTIR images of microplastics which have been enriched on membrane filters are promising and provide the impetus to put further effort into their development. In this paper we present a methodology which allows discrimination between different polymer types and measurement of their abundance and their size distributions with high accuracy. In particular we apply random decision forest classifiers and compute a multiclass model for the polymers polyethylene, polypropylene, poly(methyl methacrylate), polyacrylonitrile and polystyrene. Further classification results of the analyzed µFTIR images are given for comparability. The study also briefly discusses common issues that can arise in classification such as the curse of dimensionality and label noise.

Findings and impact The herein conducted research revealed that the most challenging classification problem is the separation of the matrix and the microplastics classes. Ambiguous spectra make the task of annotating the training data subjective which may be a cause for label noise. Strong total absorption hampers the detection of thick microplastics but not all types of polymers are affected to a similar degree. This work provided preliminary proof that random forests are well-suited for detecting microplastics and microfibers. The throughput rate highlighted the advantages of model-based machine learning over library search based approaches.

Contribution B.H. wrote the manuscript and performed a major part of the associated research, including the methodological development, implementation of algorithms and analysis and interpretation of the results. D.S and H.L also contributed to the design of the training data. H.L. further implemented algorithms and supervised the research. D.S., E.R., M.L., C.L. and H.L. critically reviewed the manuscript. B.H.'s estimated contribution is 55%.

Citations CrossRef: 36; Google Scholar: 45; (retrieved on the 3rd of January 2022)

7.3 Hufnagl et al. (2022)

Title Computer-Assisted Analysis of Microplastics in Environmental Samples Based on μFTIR Imaging in Combination with Machine Learning

Authors Benedikt Hufnagl, Michael Stibi, Heghnar Martirosyan, Ursula Wilczek, Julia N. Möller, Martin G. J. Löder, Christian Laforsch and Hans Lohninger

Abstract The problem of automating the data analysis of microplastics following a spectroscopic measurement such as focal plane array (FPA)-based micro-Fourier transform infrared (FTIR), Raman, or QCL is gaining ever more attention. Ease of use of the analysis software, reduction of expert time, analysis speed, and accuracy of the result are key for making the overall process scalable and thus allowing nonresearch laboratories to offer microplastics analysis as a service. Over the recent years, the prevailing approach has been to use spectral library search to automatically identify spectra of the sample. Recent studies, however, showed that this approach is rather limited in certain contexts, which led to developments for making library searches more robust but on the other hand also paved the way for introducing more advanced machine learning approaches. This study describes a model-based machine learning approach based on random decision forests for the analysis of large FPA-µFTIR data sets of environmental samples. The model can distinguish between more than 20 different polymer types and is applicable to complex matrices. The performance of the model under these demanding circumstances is shown based on eight different data sets. Further, a Monte Carlo cross validation has been performed to compute error rates such as sensitivity, specificity, and precision.

Findings and impact In this study the original random forest model was extended from 5 to 21 polymer types. Further a rich variety of matrix spectra was added to the training data which in the end contained more than 12000 annotated spectra. The resulting random forest model performed very well considering the sensitivity, specificity and precision of the respective classes. The robustness regarding total absorption could be increased significantly by creating samples with very thick particles. The auditing scheme allowed to reduce label noise in the training data and thus proved to be a useful tool. The work also demonstrated the advantage of superimposing the light microscope image of the sample with the μFTIR image for improved dual control of the results. By including both spectra from Bruker Hyperion 3000 and Bruker Lumos II a cross-device model could be created successfully.

Contribution B.H. drafted the manuscript and performed a major part of the associated research, including the methodological development, implementation of algorithms and analysis and interpretation of the results. J.M. and C.L. also contributed text to the introduction. J.M. and M.L. created spiked samples. H.L. further implemented algorithms and supervised the research. M.S., H.M., U.W., J.M., M.L., C.L., H.L. critically reviewed the manuscript. B.H.'s estimated contribution is 50%.

Citations CrossRef: 0; Google Scholar: 2; ResearchGate: 3 (retrieved on the 19th of March 2022)

7.4 Weisser et al. (2021)

Title From the Well to the Bottle: Identifying Sources of Microplastics in Mineral Water

Authors Jana Weisser, Irina Beer, Benedikt Hufnagl, Thomas Hofmann, Hans Lohninger, Natalia P. Ivleva and Karl Glas

Abstract Microplastics (MP) have been detected in bottled mineral water across the world. Because only few MP particles have been reported in ground water-sourced drinking water, it is suspected that MP enter the water during bottle cleaning and filling. However, until today, MP entry paths were not revealed. For the first time, this study provides findings of MP from the well to the bottle including the bottle washing process. At four mineral water bottlers, five sample types were taken along the process: raw and deferrized water samples were filtered in situ; clean bottles were sampled right after they left the bottle washer and after filling and capping. Caustic cleaning solutions were sampled from bottle washers and MP particles isolated through enzymatic and chemical treatments. The samples were analyzed for eleven synthetic and natural polymer particles ≥ 11 μm with Fourier-transform infrared imaging and random decision forests. MP were present in all steps of mineral water bottling, with a sharp increase from <1 MP L^{-1} to 317 ± 257 MP L^{-1} attributed to bottle capping. As 81% of MP resembled the PE-based cap sealing material, abrasion from the sealings was identified as the main entry path for MP into bottled mineral water.

Findings and impact In the context of this doctoral thesis this study contributed to a better understanding of how the original methodology for the development of classifiers can be transferred to other devices. Indeed, the application of the original random forest model to data measured using the Agilent Cary 620 showed that the model could only generalize well within bounds. These experiments laid the groundwork for later developments to extend the model to Bruker Lumos II. Further, the development of additional scripts allowed a more thorough validation based on Monte Carlo cross validation. This approach was then also applied to the later studies cited in this thesis.

Contribution B.H. contributed by advising on the methodological level to assist in the development of a new random forest model, however, he was not involved in the sampling or annotation of training data. This was done by the colleagues at TU Munich independently. B.H. further contributed new algorithms for data processing as well as for validation and critically reviewed the manuscript. His estimated contribution is 10%.

Citations CrossRef: 9; Google Scholar: 10; (retrieved on the 3rd of January 2022)

7.5 Ritschar et al. (2021)

Title Classification of target tissues of Eisenia fetida using sequential multimodal chemical analysis and machine learning

Authors Sven Ritschar, Elisabeth Schirmer, Benedikt Hufnagl, Martin G. J. Löder, Andreas Römpp and Christian Laforsch

Abstract Acquiring comprehensive knowledge about the uptake of pollutants, impact on tissue integrity and the effects at the molecular level in organisms is of increasing interest due to the environmental exposure to numerous contaminants. The analysis of tissues can be performed by histological examination, which is still time-consuming and restricted to target-specific staining methods. The histological approaches can be complemented with chemical imaging analysis. Chemical imaging of tissue sections is typically performed using a single imaging approach. However, for toxicological testing of environmental pollutants, a multimodal approach combined with improved data acquisition and evaluation is desirable, since it may allow for more rapid tissue characterization and give further information on ecotoxicological effects at the tissue level. Therefore, using the soil model organism Eisenia fetida as a model, we developed a sequential workflow combining Fourier transform infrared spectroscopy (FTIR) and matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI-MSI) for chemical analysis of the same tissue sections. Data analysis of the FTIR spectra via random decision forest (RDF) classification enabled the rapid identification of target tissues (e.g., digestive tissue), which are relevant from an ecotoxicological point of view. MALDI imaging analysis provided specific lipid species which are sensitive to metabolic changes and environmental stressors. Taken together, our approach provides a fast and reproducible workflow for label-free histochemical tissue analyses in E. fetida, which can be applied to other model organisms as well.

Findings and impact This work demonstrated how random forests can be used for segmenting tissues in *Eisenia fetida*. In order to ensure that the model generalizes well across different samples the variability of the spectra was modeled by sampling different thin-sections. As the detection of microplastics in tissues is an upcoming research topic this work represent a first step towards solving this problem.

Contribution B.H. contributed by advising on the methodological level to assist in the development of a new random forest model. The sampling and annotation of training data was done independently at the University of Bayreuth by S.R.. B.H. and S.R. worked together to iteratively improve the training data so that the model works across different thin-sections. B.H. also wrote the section ‘Random forest statistical performance assessment’ and conducted the validation of the final random forest model. His estimated contribution is 15%.

Citations CrossRef: 0; Google Scholar: 0; (retrieved on the 3rd of January 2022)

8 Conclusion

To date the analysis of microplastics remains a challenging analytical problem where many different instrumental approaches are applied to address certain aspects such as quantifying particle numbers, analyzing shape or polymer type, degradation state, size distributions and mass balances. Within the domain of microspectroscopy there already exists a rich variety of analytical instruments for FTIR, Raman and QCL spectroscopy. What all of these approaches have in common is the way in which data analysis of the measured spectra is handled. Spectral library search is the predominant approach with only a few exceptions where advanced chemometric techniques such as unsupervised or supervised learning have been used. Within this thesis both clustering and classification algorithms have been studied with respect to μFTIR imaging data in order to provide answers to the following research questions:

1. What are the structural aspects of microplastics measurement data which govern the performance of machine learning approaches?
2. What are key obstacles that need to be overcome to apply machine learning more broadly for microplastics detection and quantification?

The performance of clustering approaches mostly depends on their ability to detect clusters which contain only a few objects, as microplastics are rather scarce within the data. At the same time the algorithms also need to be able to deal with the clusters originating from the matrix components which make up the bulk of the data. If this strong imbalance cannot be handled the microplastics will most likely be masked by other compounds on the filter. In a similar way the performance of classification ultimately depends on the quality of the training data with respect to the variability of the matrix and physical measurement effects such as Mie scattering and total absorption. False positives are a typical sign that certain structural aspects have not been modeled well.

Considering the effort which was put into the development of the random forest classifiers the biggest obstacles for building broadly applicable classification models is the collection of representative filter samples as well as the sampling, annotating and auditing of training data. Depending on the intended applicability domain the effort increases considerably if one wants to address both different matrices as well as different devices. This obstacle is not limited to μFTIR imaging but translates to Raman and QCL as well.

While building broadly applicable classifiers seems overly complex and resource intensive it is nonetheless an endeavor worth considering. The performance of spectral library search is governed by a trade-off between analytical quality and analysis speed. Further, the research community still hasn't reached a consensus on how to set appropriate thresholds for HQI indices and what kind of databases should be used. The reason for this is rooted in the problem that different reference databases yield different results and also in the Curse of Dimensionality, which decreases the performance of distance metrics in high-dimensional feature spaces.

Supervised learning offers an alternative way as it imposes no limit on the number of reference spectra which can be used and further uses such methods as variable selection and dimensionality reduction to yield significantly better performance on high-dimensional data. In order to fully exploit the potential of supervised learning for microplastics detection the goal should be to build a broadly applicable model which can perform the task independently of the device and the matrix. Considering that the demand for microplastics research and analysis is also driven by certain industrial sectors and a first monitoring campaign has started in the USA, high-throughput methodologies will become more and more important in the future. In this context supervised machine learning provides both the necessary analysis speed as well as the technical scalability to meet these demands.

References

Albert, P. S. and L. E. Dodd
2004. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60(2):427–435.

Ambrosini, R., R. S. Azzoni, F. Pittino, G. Diolaiuti, A. Franzetti, and M. Parolini
2019. First evidence of microplastic contamination in the supraglacial debris of an alpine glacier. *Environmental pollution*, 253:297–301.

Anger, P. M., L. Prechtl, M. Elsner, R. Niessner, and N. P. Ivleva
2019. Implementation of an open source algorithm for particle recognition and morphological characterisation for microplastic analysis by means of Raman microspectroscopy. *Analytical Methods*, 11(27):3483–3489.

Anger, P. M., E. von der Esch, T. Baumann, M. Elsner, R. Niessner, and N. P. Ivleva
2018. Raman microspectroscopy as a tool for microplastic particle analysis. *TrAC Trends in Analytical Chemistry*, 109:214–226.

Arthur, C., J. Baker, and H. Bamford
2009. Proceedings of the international research workshop on the occurrence, effects, and fate of microplastic marine debris, september 9-11, 2008. Technical report.

Ballabio, D., F. Grisoni, and R. Todeschini
2018. Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems*, 174:33–44.

Bannick, C. G., R. Szewzyk, M. Ricking, S. Schniegler, N. Obermaier, A. K. Barthel, K. Altmann, P. Eisentraut, and U. Braun
2019. Development and testing of a fractionated filtration for sampling of microplastics in water. *Water research*, 149:650–658.

Barrett, J., Z. Chase, J. Zhang, M. M. B. Holl, K. Willis, A. Williams, B. D. Hardesty, and C. Wilcox
2020. Microplastic pollution in deep-sea sediments from the great australian bight. *Frontiers in Marine Science*, 7:808.

Bassan, P., H. J. Byrne, F. Bonnier, J. Lee, P. Dumas, and P. Gardner
2009. Resonant mie scattering in infrared spectroscopy of biological materials–understanding the 'dispersion artefact'. *Analyst*, 134(8):1586–1593.

Becker, R., K. Altmann, T. Sommerfeld, and U. Braun
2020. Quantification of microplastics in a freshwater suspended organic matter using different thermoanalytical methods–outcome of an interlaboratory comparison. *Journal of Analytical and Applied Pyrolysis*, 148:104829.

Bellman, R.
1961. Adaptative control processes.

Belz, S., I. Bianchi, C. C., H. Emteborg, F. Fumagalli, O. Geiss, D. Gilliland, A. Held, U. Jakobsson, R. La Spina, D. Mehn, Y. Ramaye, P. Robouch, J. Seghers, B. Sokull-Kluettgen, E. Stefaniak, and J. Stroka
2021. *Current status of the quantification of microplastics in water - Results of a JRC-BAM inter-laboratory comparison study on PET in water*. Luxembourg: Publications Office of the European Union.

Biau, G. and E. Scornet
2016. A random forest guided tour. *Test*, 25(2):197–227.

Brandt, J., L. Bittrich, F. Fischer, E. Kanaki, A. Tagg, R. Lenz, M. Labrenz, E. Brandes, D. Fischer, and K.-J. Eichhorn
2020. EXPRESS: High-Throughput Analyses of Microplastic Samples Using Fourier Transform Infrared and Raman Spectrometry. *Applied Spectroscopy*, P. 0003702820932926.

Brandt, J., K. Mattsson, and M. Hassellöv
2021. Deep Learning for Reconstructing Low-Quality FTIR and Raman Spectra - A Case Study in Microplastic Analyses. *Analytical Chemistry*.

Breiman, L.
2001. Random forests. *Machine learning*, 45(1):5–32.

Chang, H. and D.-Y. Yeung
2008. Robust path-based spectral clustering. *Pattern Recognition*, 41(1):191–203.

Corradini, F., N. Beriot, E. Huerta-Lwanga, and V. Geissen
2021. uFTIR: An R package to process hyperspectral images of environmental samples captured with μFTIR microscopes. *SoftwareX*, 16:100857.

Cortes, C. and V. Vapnik
1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Cowger, W., Z. Steinmetz, A. Gray, K. Munno, J. Lynch, H. Hapich, S. Primpke, H. De Frond, C. Rochman, and O. Herodotou
2021. Microplastic spectral classification needs an open source community: Open specy to the rescue! *Analytical Chemistry*.

da Silva, V. H., F. Murphy, J. M. Amigo, C. Stedmon, and J. Strand
2020. Classification and quantification of microplastics (< 100 μm) using a focal plane array–fourier transform infrared imaging system and machine learning. *Analytical Chemistry*, 92(20):13724–13733.

de Medeiros Back, H., E. C. V. Junior, O. E. Alarcon, and D. Pottmaier
2022. Training and evaluating machine learning algorithms for ocean microplastics classification through vibrational spectroscopy. *Chemosphere*, 287:131903.

DeFrond, H., L. Thornton-Hampton, S. Kotar, K. Gesulga, C. Matuch, W. Lao, C. Rochman, and C. Wong
2022. Microplastics interlaboratory methods comparison study to provide recommendations for monitoring microplastics in drinking water in the state of california. *Chemosphere*. manuscript in preparation.

Domingos, P.
2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Dong, M., Z. She, X. Xiong, and Z. Luo
2022. Automated analysis of microplastics based on vibrational spectroscopy: Are we measuring the same metrics? *Analytical and Bioanalytical Chemistry*.

Dümichen, E., P. Eisentraut, C. G. Bannick, A.-K. Barthel, R. Senz, and U. Braun
2017. Fast identification of microplastics in complex environmental samples by a thermal degradation method. *Chemosphere*, 174:572–584.

Eerkes-Medrano, D., R. C. Thompson, and D. C. Aldridge
2015. Microplastics in freshwater systems: a review of the emerging threats, identification of knowledge gaps and prioritisation of research needs. *Water research*, 75:63–82.

Eisentraut, P., E. Dümichen, A. S. Ruhl, M. Jekel, M. Albrecht, M. Gehde, and U. Braun
2018. Two birds with one stone—fast and simultaneous analysis of microplastics: microparticles derived from thermoplastics and tire wear. *Environmental Science & Technology Letters*, 5(10):608–613.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al.
1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, Pp. 226–231.

Folleco, A., T. M. Khoshgoftaar, J. Van Hulse, and L. Bullard
2008. Identifying learners robust to low quality data. In *IEEE International Conference on Information Reuse and Integration*, Pp. 190–195. IEEE.

Frei, S., S. Piehl, B. Gilfedder, M. Löder, J. Krutzke, L. Wilhelm, and C. Laforsch
2019. Occurence of microplastics in the hyporheic zone of rivers. *Scientific reports*, 9(1):1–11.

Frénay, B. and M. Verleysen
2014. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

Fu, L. and E. Medico
2007. Flame, a novel fuzzy clustering method for the analysis of dna microarray data. *BMC bioinformatics*, 8(1):3.

Gionis, A., H. Mannila, and P. Tsaparas
2007. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):4.

Hahn, A., G. Gerdts, C. Völker, and V. Niebühr
2019. Using FTIRS as pre-screening method for detection of microplastic in bulk sediment samples. *Science of the total environment*, 689:341–346.

Hartmann, N. B., T. Huffer, R. C. Thompson, M. Hassellöv, A. Verschoor, A. E. Daugaard, S. Rist, T. Karlsson, N. Brennholt, M. Cole, M. P. Herrling, M. C. Hess, N. P. Ivleva, A. L. Lusher, and M. Wagner
2019. Are we speaking the same language? Recommendations for a definition and categorization framework for plastic debris. *Environmental science & technology*, 53(3):1039–1047.

Hastie, T., R. Tibshirani, and J. H. Friedman
2009. *The elements of statistical learning*, Springer series in statistics, 2. ed. edition. New York: Springer.

Hennig, C.
2010. Mathematical models and reality: A constructivist perspective. *Foundations of Science*, 15(1):29–48.

Hennig, C.
2015. What are the true clusters? *Pattern Recognition Letters*, 64:53–62.

Hidalgo-Ruz, V., L. Gutow, R. C. Thompson, and M. Thiel
2012. Microplastics in the marine environment: a review of the methods used for identification and quantification. *Environmental science & technology*, 46(6):3060–3075.

Hinton, G. E. and R. R. Salakhutdinov
2006. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Hufnagl, B. and H. Lohninger
2019. A multi-challenge clustering benchmark dataset embedding large differences in spatial extent. Zenodo. https://doi.org/10.5281/zenodo.2583762. [dataset].

Hufnagl, B. and H. Lohninger
2020. A graph-based clustering method with special focus on hyperspectral imaging. *Analytica chimica acta*, 1097:37–48.

Hufnagl, B., H. Lohninger, M. G. J. Löder, and C. Laforsch
2021. Automated Ultra-Fast Analysis of Microplastics in Large μFTIR Imaging Datasets From Environmental Samples Via RDF Classifiers. In *Abstract Book SETAC Europe 31st annual meeting*, P. 54. SETAC Europe.

Hufnagl, B., D. Steiner, E. Renner, M. G. J. Löder, C. Laforsch, and H. Lohninger
2019. A methodology for the fast identification and monitoring of microplastics in environmental samples using random decision forest classifiers. *Analytical Methods*, 11:2277–2285.

Hufnagl, B., M. Stibi, H. Martirosyan, U. Wilczek, J. N. Möller, M. G. Löder, C. Laforsch, and H. Lohninger
2022. Computer-Assisted Analysis of Microplastics in Environmental Samples Based on μFTIR Imaging in Combination with Machine Learning. *Environmental Science & Technology Letters*, 9(1):90–95.

Hughes, N. P., S. J. Roberts, and L. Tarassenko
2004. Semi-supervised learning of probabilistic models for ecg segmentation. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, Pp. 434–437. IEEE.

Hurley, R. R., A. L. Lusher, M. Olsen, and L. Nizzetto
2018. Validation of a method for extracting microplastics from complex, organic-rich, environmental matrices. *Environmental science & technology*, 52(13):7409–7417.

ISO/TR 21960:2020
2020. Plastics — Environmental aspects — State of knowledge and methodologies. Standard, International Organization for Standardization, Geneva, CH.

Ivleva, N. P.
2021. Chemical analysis of microplastics and nanoplastics: Challenges, advanced methods, and perspectives. *Chemical Reviews*, 121(19):11886–11936.

Jain, A. K.
2010. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Jain, A. K. and M. H. Law
2005. Data clustering: A user's dilemma. In *International conference on pattern recognition and machine intelligence*, Pp. 1–10. Springer.

Jain, A. K., M. N. Murty, and P. J. Flynn
1999. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Jimenez, L. O. and D. A. Landgrebe
1998. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 28(1):39–54.

Käppler, A., D. Fischer, S. Oberbeckmann, G. Schernewski, M. Labrenz, K.-J. Eichhorn, and B. Voit
2016. Analysis of environmental microplastics by vibrational microspectroscopy: FTIR, Raman or both? *Analytical and bioanalytical chemistry*, 408(29):8377–8391.

Kedzierski, M., M. Falcou-Préfol, M. E. Kerros, M. Henry, M. L. Pedrotti, and S. Bruzaud
2019. A machine learning algorithm for high throughput identification of FTIR spectra: Application on microplastics collected in the Mediterranean Sea. *Chemosphere*, 234:242–251.

Kumar, B. V., L. A. Löschel, H. K. Imhof, M. G. Löder, and C. Laforsch
2021. Analysis of microplastics of a broad size range in commercially important mussels by combining FTIR and Raman spectroscopy approaches. *Environmental Pollution*, 269:116147.

La Nasa, J., G. Biale, D. Fabbri, and F. Modugno
2020. A review on challenges and developments of analytical pyrolysis and other thermoanalytical techniques for the quali-quantitative determination of microplastics. *Journal of Analytical and Applied Pyrolysis*, 149:104841.

Lee, L. C., C.-Y. Liong, and A. A. Jemain
2018. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst*, 143(15):3526–3539.

Li, L., X. Zhao, Z. Li, and K. Song
2021. COVID-19: Performance study of microplastic inhalation risk posed by wearing masks. *Journal of hazardous materials*, 411:124955.

Liu, F., K. B. Olesen, A. R. Borregaard, and J. Vollertsen
2019. Microplastics in urban and highway stormwater retention ponds. *Science of The Total Environment*, 671:992–1000.

Löder, M. G. J., H. K. Imhof, M. Ladehoff, L. A. Löschel, C. Lorenz, S. Mintenig, S. Piehl, S. Primpke, I. Schrank, C. Laforsch, and G. Gerdts
2017. Enzymatic purification of microplastics in environmental samples. *Environmental science & technology*, 51(24):14283–14292.

Löder, M. G. J., M. Kuczera, S. Mintenig, C. Lorenz, and G. Gerdts
2015. Focal plane array detector-based micro-fourier-transform infrared imaging for the analysis of microplastics in environmental samples. *Environmental Chemistry*, 12(5):563–581.

Lohninger, H. and J. Ofner
2014. Multisensor hyperspectral imaging as a versatile tool for image-based chemical structure determination. *Spectroscopy Europe*, 26(5):6–10.

Maes, T., R. Jessop, N. Wellner, K. Haupt, and A. G. Mayes
2017. A rapid-screening approach to detect and quantify microplastics based on fluorescent tagging with nile red. *Scientific reports*, 7(1):1–10.

Malossini, A., E. Blanzieri, and R. T. Ng
2006. Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics*, 22(17):2114–2121.

Millard, K. and M. Richardson
2015. On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. *Remote sensing*, 7(7):8489–8515.

Möller, J. N., I. Heisel, A. Satzger, E. C. Vizsolyi, S. J. Oster, S. Agarwal, C. Laforsch, and M. G. Löder
2021. Tackling the challenge of extracting microplastics from soils: A protocol to purify soil samples for spectroscopic analysis. *Environmental Toxicology and Chemistry*.

Möller, J. N., M. G. Löder, and C. Laforsch
2020. Finding microplastics in soils: A review of analytical methods. *Environmental Science & Technology*, 54(4):2078–2090.

Mughini-Gras, L., R. Q. van der Plaats, P. W. van der Wielen, P. S. Bauerlein, and A. M. de Roda Husman
2021. Riverine microplastic and microbial community compositions: A field study in the netherlands. *Water Research*, 192:116852.

Munno, K., H. De Frond, B. O'Donnell, and C. M. Rochman
2020. Increasing the accessibility for characterizing microplastics: Introducing new application-based and spectral libraries of plastic particles (slopp and slopp-e). *Analytical chemistry*, 92(3):2443–2451.

Napper, I. E., B. F. Davies, H. Clifford, S. Elvin, H. J. Koldewey, P. A. Mayewski, K. R. Miner, M. Potocki, A. C. Elmore, A. P. Gajurel, and R. C. Thompson
2020. Reaching new heights in plastic pollution—preliminary findings of microplastics on mount everest. *One Earth*, 3(5):621–630.

Nascimento, J. M. and J. M. Dias
2005. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE transactions on Geoscience and Remote Sensing*, 43(4):898–910.

Nettleton, D. F., A. Orriols-Puig, and A. Fornells
2010. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review*, 33(4):275–306.

Ofner, J., K. A. Kamilli, E. Eitenberger, G. Friedbacher, B. Lendl, A. Held, and H. Lohninger
2015. Chemometric analysis of multisensor hyperspectral images of precipitated atmospheric particulate matter. *Analytical chemistry*, 87(18):9413–9420.

Paul, A., L. Wander, R. Becker, C. Goedecke, and U. Braun
2018. High-throughput NIR spectroscopic (NIRS) detection of microplastics in soil. *Environmental Science and Pollution Research*, Pp. 1–11.

Peeken, I., E. Bergami, I. Corsi, B. Hufnagl, C. Katlein, T. Krumpen, M. Löder, Q. Wang, and C. Wekerle
2021. The role of sea ice for plastic pollution in the arctic. In *EGU General Assembly Conference Abstracts*, Pp. EGU21–13366.

Peng, X., M. Chen, S. Chen, S. Dasgupta, H. Xu, K. Ta, M. Du, J. Li, Z. Guo, and S. Bai
2018. Microplastics contaminate the deepest part of the world's ocean. *Geochem. Perspect. Lett*, 9:1–5.

Picó, Y. and D. Barceló
2020. Pyrolysis gas chromatography-mass spectrometry in environmental analysis: Focus on organic matter and microplastics. *TrAC Trends in Analytical Chemistry*, P. 115964.

Plaza, A., J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni
2009. Recent advances in techniques for hyperspectral image processing. *Remote sensing of environment*, 113:S110–S122.

Primpke, S., S. H. Christiansen, W. Cowger, H. De Frond, A. Deshpande, M. Fischer, E. Holland, M. Meyns, B. A. O'Donnell, B. Ossmann, M. Pittroff, G. Sarau, B. M. Scholz-Böttcher, and K. Wiggin
2020a. Express: Critical assessment of analytical methods for the harmonized and cost efficient analysis of microplastics. *Applied Spectroscopy*, P. 0003702820921465.

Primpke, S., R. K. Cross, S. M. Mintenig, M. Simon, A. Vianello, G. Gerdts, and J. Vollertsen
2020b. EXPRESS: Toward the Systematic Identification of Microplastics in the Environment: Evaluation of a New Independent Software Tool (siMPle) for Spectroscopic Analysis. *Applied Spectroscopy*, P. 0003702820917760.

Primpke, S., M. Godejohann, and G. Gerdts
2020c. Rapid identification and quantification of microplastics in the environment by quantum cascade laser-based hyperspectral infrared chemical imaging. *Environmental Science & Technology*, 54(24):15893–15903.

Primpke, S., C. Lorenz, R. Rascher-Friesenhausen, and G. Gerdts
2017. An automated approach for microplastics analysis using focal plane array (FPA) FTIR microscopy and image analysis. *Analytical Methods*, 9(9):1499–1511.

Primpke, S., M. Wirth, C. Lorenz, and G. Gerdts
2018. Reference database design for the automated analysis of microplastic samples based on Fourier transform infrared (FTIR) spectroscopy. *Analytical and bioanalytical chemistry*, 410(21):5131–5141.

Provencher, J. F., G. A. Covernton, R. C. Moore, D. A. Horn, J. L. Conkle, and A. L. Lusher
2020. Proceed with caution: The need to raise the publication bar for microplastics research. *Science of the Total Environment*, 748:141426.

Radovanović, M., A. Nanopoulos, and M. Ivanović
2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Ragusa, A., A. Svelato, C. Santacroce, P. Catalano, V. Notarstefano, O. Carnevali, F. Papa, M. C. A. Rongioletti, F. Baiocco, S. Draghi, E. D'Amore, D. Rinaldo, M. Matta, and E. Giorgini
2021. Plasticenta: First evidence of microplastics in human placenta. *Environment international*, 146:106274.

Reidsma, D. and R. op den Akker
2008. Exploiting 'subjective'annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, Pp. 8–16.

Renner, G., A. Nellessen, A. Schwiers, M. Wenzel, T. C. Schmidt, and J. Schram
2018. Data preprocessing & evaluation used in the microplastics identification process: A critical review & practical guide. *TrAC Trends in Analytical Chemistry*.

Renner, G., P. Sauerbier, T. C. Schmidt, and J. Schram
2019. Robust automatic identification of microplastics in environmental samples using FTIR microscopy. *Analytical chemistry*, 91(15):9656–9664.

Renner, G., T. C. Schmidt, and J. Schram
2017a. A new chemometric approach for automatic identification of microplastics from environmental compartments based on FT-IR spectroscopy. *Analytical chemistry*, 89(22):12045–12053.

Renner, G., T. C. Schmidt, and J. Schram
2017b. Analytical methodologies for monitoring micro (nano) plastics: which are fit for purpose? *Current Opinion in Environmental Science & Health*, Pp. 55–61.

Ritschar, S., E. Schirmer, B. Hufnagl, M. G. Löder, A. Römpp, and C. Laforsch
2021. Classification of target tissues of eisenia fetida using sequential multimodal chemical analysis and machine learning. *Histochemistry and Cell Biology*, Pp. 1–11.

Schwabl, P., S. Köppel, P. Königshofer, T. Bucsics, M. Trauner, T. Reiberger, and B. Liebmann
2019. Detection of various microplastics in human stool: a prospective case series. *Annals of internal medicine*, 171(7):453–457.

Scircle, A., J. V. Cizdziel, L. Tisinger, T. Anumol, and D. Robey
2020. Occurrence of microplastic pollution at oyster reefs and other coastal sites in the Mississippi sound, USA: impacts of freshwater inflows from flooding. *Toxics*, 8(2):35.

Serranti, S., R. Palmieri, G. Bonifazi, and A. Cózar
2018. Characterization of microplastic litter from oceans by an innovative approach based on hyperspectral imaging. *Waste Management*, 76:117–125.

Shan, J., J. Zhao, Y. Zhang, L. Liu, F. Wu, and X. Wang
2019. Simple and rapid detection of microplastics in seawater using hyperspectral imaging technology. *Analytica Chimica Acta*, 1050:161–168.

Sheng, V. S., F. Provost, and P. G. Ipeirotis
2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, Pp. 614–622.

Shi, J. and J. Malik
2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905.

Silva, A. B., A. S. Bastos, C. I. Justino, J. P. da Costa, A. C. Duarte, and T. A. Rocha-Santos
2018. Microplastics in the environment: Challenges in analytical chemistry - a review. *Analytica chimica acta*, 1017:1–19.

Smyth, P., U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi
1995. Inferring ground truth from subjective labelling of venus images. In *Advances in neural information processing systems*, Pp. 1085–1092.

Snow, R., B. O'Connor, D. Jurafsky, and A. Y. Ng
2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, Pp. 254–263. Association for Computational Linguistics.

Steindl, B.
2018. Graph-based competitive clustering: a clustering algorithm for hyperspectral images in process analytical technologies. diploma thesis, TU Wien, Wien.

Süssmann, J., T. Krause, D. Martin, E. Walz, R. Greiner, S. Rohn, E. K. Fischer, and J. Fritsche
2021. Evaluation and optimisation of sample preparation protocols suitable for the analysis of plastic particles present in seafood. *Food Control*, 125:107969.

Teichert, S., M. G. Löder, I. Pyko, M. Mordek, C. Schulbert, M. Wisshak, and C. Laforsch
2021. Microplastic contamination of the drilling bivalve hiatella arctica in arctic rhodolith beds. *Scientific Reports*, 11(1):1–12.

Thompson, R. C., Y. Olsen, R. P. Mitchell, A. Davis, S. J. Rowland, A. W. John, D. McGonigle, and A. E. Russell
2004. Lost at sea: where is all the plastic? *Science*, 304(5672):838–838.

Tommasi, T., N. Patricia, B. Caputo, and T. Tuytelaars
2017. A deeper look at dataset bias. In *Domain Adaptation in Computer Vision Applications*, Pp. 37–55. Springer.

Van Der Maaten, L., E. Postma, and J. Van den Herik
2009. Dimensionality reduction: a comparative review. *J Mach Learn Res*, 10:66–71.

Van Mourik, L., S. Crum, E. Martinez-Frances, B. van Bavel, H. Leslie, J. de Boer, and W. Cofino
2021. Results of WEPAL-QUASIMEME/NORMANs first global interlaboratory study on microplastics reveal urgent need for harmonization. *Science of the Total Environment*, 772:145071.

Veenman, C. J., M. J. T. Reinders, and E. Backer
2002. A maximum variance cluster algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, 24(9):1273–1280.

Von Luxburg, U.
2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Von Luxburg, U., R. C. Williamson, and I. Guyon
2012. Clustering: Science or art? In *Proceedings of ICML workshop on unsupervised and transfer learning*, Pp. 65–79. JMLR Workshop and Conference Proceedings.

Wander, L., B. Hufnagl, H. Lohninger, E. Kanaki, F. Fischer, D. Fischer, H. Martirosyan, M. G. J. Löder, and C. Laforsch
2021. Fast and Reproducible Analysis of Infrared and Raman Spectra of Microplastics Using Machine Learning. In *Abstract Book SETAC North America 42nd annual meeting*, P. 343. SETAC North America.

Wander, L., A. Vianello, J. Vollertsen, F. Westad, U. Braun, and A. Paul
2020. Exploratory analysis of hyperspectral FTIR data obtained from environmental microplastics samples. *Analytical Methods*, 12(6):781–791.

Weisser, J., I. Beer, B. Hufnagl, T. Hofmann, H. Lohninger, N. P. Ivleva, and K. Glas
2021. From the well to the bottle: Identifying sources of microplastics in mineral water. *Water*, 13(6):841.

Welinder, P., S. Branson, P. Perona, and S. J. Belongie
2010. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, Pp. 2424–2432.

Westad, F. and F. Marini
2015. Validation of chemometric models–a tutorial. *Analytica chimica acta*, 893:14–24.

Wold, S., M. Sjöström, and L. Eriksson
2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130.

Xu, J.-L. and A. A. Gowen
2019. Investigation of plasticizer aggregation problem in casein based biopolymer using chemical imaging. *Talanta*, 193:128–138.

Xu, J.-L., K. V. Thomas, Z. Luo, and A. A. Gowen
2019. FTIR and Raman imaging for microplastics analysis: state of the art, challenges and prospects. *TrAC Trends in Analytical Chemistry*, P. 115629.

Xu, Q.-S. and Y.-Z. Liang
2001. Monte carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11.

Xu, R. and D. Wunsch
2005. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.

Zahn, C. T.
1971. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on computers*, 100(1):68–86.

Zhai, H., H. Zhang, L. Pingxiang, and L. Zhang
2021. Hyperspectral image clustering: Current achievements and future lines. *IEEE Geoscience and Remote Sensing Magazine*.