Machine learning classification of microbial community compositions to predict anthropogenic pollutants in the Baltic Sea

Kumulative Dissertation
zur
Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Rostock

vorgelegt von
René Janßen, geb. am 12.9.1986 in Goch
aus Rostock
Rostock, 18.09.2020

https://doi.org/10.18453/rosdok_id00002897

Gutachter:

Prof. Dr. Matthias Labrenz, Leibniz-Institut für Ostseeforschung Warnemünde, Sektion Biologische Meereskunde

Prof. Dr. Rudolf Amann, Bremen, Max-Planck-Institut für Marine Mikrobiologie, Abteilung Molekulare Ökologie

Prof. Dr. Alexander Probst, Universität Duisburg-Essen,
Fachbereich Umweltmikrobiologie und Biotechnologie

Ass.-Prof. Stephen Techtmann, PhD, Houghton, Michigan Technological University, Department of Biological Sciences

Jahr der Einreichung: 2020

Jahr der Verteidigung: 2020

Table of contents

Summary

Since their emergence about 4 billion years ago, bacteria have developed an unrivaled variety that allows them to conquer every conceivable habitat. Microbial metabolic processes are the basis of all life on earth. The physiological diversity of microbial communities enables specific and distinguishable reactions to environmental stimuli. Conversely, this means that by exploring the communities, conclusions can be drawn about their environment. Modern high-throughput sequencing methods such as next generation sequencing allow for the determination of community composition at the taxonomic level via phylogenetic marker genes. Similarly, the functional potential of a community can be determined by the totality of existing genes as well as the corresponding activity profile by sequencing the gene transcripts. These methods produce a large amount of data, which has to be interpreted by bioinformatics and multivariate analysis. Machine learning (ML) methods can be used to identify from this amount of information the relevant part for the recognition of specific environmental stimuli. These methods try to train a model that links the input (microbial community information) with the output (environmental stimulus). In the context of this PhD thesis it was investigated whether the analysis of microbial community data by ML can predict contamination. As specific environmental stimuli, contamination events by glyphosate and 2,4,6-trinitrotoluene (TNT) in the Baltic Sea were studied. The potentials and limitations of this approach were explored. The relevant parts of the community were examined in detail to determine whether there are actual interactions between bacteria and pollutants. Only in this case the observation of the relevant bacteria can provide indications of contamination, a so-called indicative microbial fingerprint. Furthermore, it was investigated whether microbial communities react with a delay to the decrease of a contaminant, because in this case the community would still indicate the pollutant, although the latter could not be detected analytically.

The microbial communities were described by 16S rRNA (gene) amplicon sequencing and additionally DNA shotgun sequencing. Further metabolites and environmental factors as well as cell counts and, if applicable, geochemical and sedimentological parameters were determined. The two methods Random Forest and Artificial Neural Network were used to predict the presence of the contaminants from the community data. Statistical and bioinformatic analyses were conducted to evaluate the ecological and biological significance of the ML results. The R package phyloseq2ML was developed to facilitate the use and analysis of microbiological data sets for machine learning. The described approach was first tested in the laboratory and then in field experiments. In the 140-day laboratory trial, the herbicide glyphosate was added to a continuously operated Baltic Sea-imitating microcosm. The presence of glyphosate could be predicted by the microbial community with

up to 99.9 % accuracy, with Random Forest consistently providing more accurate predictions. Glyphosate affected microbial succession and was degraded to the metabolite AMPA; an increase in cell count, the metabolite AMPA and the required *gox* gene were detected. The potentially responsible organisms were identified by ML and statistical models. A selection of a few bacterial taxa achieved on average better predictions than by using the entire community composition, for glyphosate the genus *Parvibaculum* alone was sufficient due to the simple experimental design. It was also shown that free-living bacteria were more often, but for shorter durations, affected by glyphosate than those existing in the biofilm. Most of the measured responses to glyphosate ended while the herbicide was still detectable at 1 µM and it was concluded that the concentrations detected in the Baltic Sea are not sufficiently valuable as a food source to be degraded. The environmental samples came from the munitions dumpsite Kolberger Heide near Kiel and were contaminated with various explosives. Prediction of TNT with up to 84 % (balanced) accuracy was more challenging due to the multiple influences a natural habitat is exposed to, complicated by the complexity of sedimentary communities, sample composition and low concentrations of TNT in the $pmol \cdot g^{-1}$ range. Nevertheless, 25 decision-relevant genera could be identified, which allowed more accurate predictions than the use of sediment information such as grain size distribution, element contents or sum parameters such as total nitrogen. Based on the misclassifications it was possible to determine from which regions of the Kolberger Heide further samples are needed and which samples were potentially formerly contaminated with TNT. The results of my PhD thesis demonstrate the potential to predict environmental influences, more precisely, contamination events in the Baltic Sea by ML-analyzed microbial communities. Taxa contributing to indicative fingerprints could be identified, but a higher number of samples is necessary for final confirmation. It was recognized that dependencies (e.g. spatial) between ecological samples allow overoptimistic prediction accuracies. However, their occurrence is pervasive in experiments investigating ecological hypotheses. In order to identify potential dependencies and to estimate their influence as well as to draw conclusions for ecology from ML-relevant information, interpretable ML methods should be prioritized. It was shown that the implementation of the presented approach into regular monitoring operations would improve assessment of the environmental state, is possible both in terms of methodology and resources and in return offers the required extension of the sample size for ML.

Zusammenfassung

Bakterien haben seit ihrer Entstehung vor ca. 4 Milliarden Jahren eine unerreichte Vielfalt entwickelt, die es ihnen gestattet, jeden denkbaren Lebensraum zu erobern. Mikrobielle Stoffwechselprozesse sind Grundlage jeglichen Lebens auf der Erde. Die physiologische Diversität mikrobieller Gemeinschaften ermöglicht spezifische Reaktionen auf Umweltreize. Im Umkehrschluss bedeutet dies, dass über die Erkundung der Gemeinschaften Rückschlüsse auf deren Umwelt gezogen werden können. Moderne Hochdurchsatz-Sequenziermethoden wie das Next generation sequencing erlauben die Ermittlung der Gemeinschaftszusammensetzung auf taxonomischer Ebene über phylogenetische Markergene. Ebenso ist das funktionelle Potential einer Gemeinschaft über die Gesamtheit der vorhandenen Gene als auch das entsprechende Aktivitätsprofil durch die Sequenzierung der Gentranskripte zugänglich. Diese Methoden produzieren eine unübersichtliche Menge an Daten, die mithilfe von bioinformatischer Aufbereitung und multivariater Analyse interpretiert werden muss. Verfahren des maschinellen Lernens (ML) können eingesetzt werden, um aus dieser Menge an Informationen den relevanten Anteil zur Erkennung spezifischer Umweltreize zu identifizieren. Diese Verfahren versuchen selbstständig ein Modell zu trainieren, das den Input (mikrobielle Gemeinschaftsinformationen) mit dem Output (Umweltreiz) verknüpft. Im Rahmen dieser Doktorarbeit wurde untersucht, ob die Analyse von mikrobiellen Gemeinschaftsdaten durch ML eine Vorhersage von Kontaminationsereignissen ermöglicht. Als spezifische Umweltreize wurden Kontaminationsereignisse durch Glyphosat und 2,4,6-Trinitrotoluol (TNT) in der Ostsee studiert, die für die dicht besiedelte Region von besonderer Relevanz sind. Dabei wurden die Potentiale und Limitationen dieses Ansatzes ausgelotet. Die relevanten Anteile der Gemeinschaft wurden im Detail untersucht, um festzustellen, ob es sich um tatsächliche Wechselwirkungen zwischen Bakterien und Schadstoffen handelt. Nur in diesem Falle kann die Beobachtung der entsprechenden Bakterien Indikationen für eine Kontamination liefern, einen sogenannten indikativen, mikrobiellen Fingerabdruck. Weiterhin war es Gegenstand der Untersuchungen, ob mikrobielle Gemeinschaften verzögert auf das Verschwinden eines Kontaminanten reagieren, da in diesem Falle die Gemeinschaft immer noch den Schadstoff indiziert, obwohl der selbige nicht mehr analytisch feststellbar wäre. Die mikrobiellen Gemeinschaften wurden per 16S rRNA (Gen) Amplikonsequenzierung und zusätzlich DNA-Shotgunsequenzierung beschrieben. Außerdem wurden weitere Metabolite und Umweltfaktoren sowie Zellzahlen und gegebenenfalls geochemische und sedimentologische Parameter ermittelt. Die beiden Methoden Random Forest und Artificial Neural Network wurden eingesetzt, um aus den Gemeinschaftsdaten eine Präsenz der Kontaminanten vorherzusagen. Statistische und

bioinformatische Analysen wurden angewandt, um die ökologische bzw. biologische Sinnhaftigkeit der ML-Ergebnisse zu evaluieren. Das R Package phyloseq2ML wurde entwickelt, um den Einsatz und die Analyse mikrobiologischer Datensätze für Maschinelles Lernen zu vereinfachen. Der beschriebene Ansatz wurde zuerst im Labor und dann im Feldversuch erprobt. Im 140 Tage währenden Laborversuch wurde das Herbizid Glyphosat zu einem kontinuierlich betriebenen Ostsee-nachempfundenen Mikrokosmos gegeben. Die Anwesenheit von Glyphosat konnte durch die mikrobielle Gemeinschaft mit bis zu 99,9 % Genauigkeit vorhergesagt werden, wobei Random Forest durchgängig präzisere Vorhersagen erstellte. Glyphosat beeinflusste die mikrobielle Sukzession und wurde zum Metaboliten AMPA abgebaut; ein Anstieg der Zellzahl, der Metabolit AMPA sowie das benötigte *gox* Gen wurden nachgewiesen. Die potentiell verantwortlichen Organismen konnten durch ML und statistische Modelle übereinstimmend identifiziert werden. Eine Auswahl weniger bakterieller Taxa als Input für die Modelle erreichte im Schnitt bessere Vorhersagen als unter Einsatz der gesamten Gemeinschaftszusammensetzung, für Glyphosat reichte aufgrund des simplen Versuchsdesigns allein die Gattung *Parvibaculum*. Ebenfalls konnte gezeigt werden, dass freilebende Bakterien häufiger, aber kürzer von Glyphosat beeinflusst wurden als im Biofilm existierende. Die gemessenen Reaktionen auf Glyphosat endeten bereits größtenteils während das Herbizid noch mit 1 µM nachweisbar war und es wurde geschlussfolgert, dass die in der Ostsee nachgewiesenen Konzentrationen nicht ausreichend wertvoll als Nahrungsquelle sind, um abgebaut zu werden. Die Umweltproben stammten aus dem Munitionsversenkungsgebiet Kolberger Heide nahe Kiel und waren kontaminiert mit verschiedenen Sprengstoffe. Die Vorhersage von TNT mit bis zu 84 % (balancierter) Genauigkeit gestaltete sich anspruchsvoller aufgrund der vielfältigen Einflüsse, denen ein natürliches Habitat ausgesetzt ist, weiterhin erschwert durch die Komplexität der im Sediment angesiedelten Gemeinschaften, die Probenzusammenstellung und die niedrige Konzentration von TNT im $pmol \cdot g^{-1}$ Bereich. Nichtsdestotrotz konnten 25 entscheidungsrelevante Gattungen ermittelt werden, die genauere Vorhersagen erlaubten als die Nutzung der Sedimentinformationen wie z.B. Korngrößenverteilung, Elementgehalte oder Summenparameter wie Gesamtstickstoff. Anhand der Fehlklassifikationen konnte ermittelt werden, aus welchen Regionen der Kolberger Heide weitere Proben benötigt werden und welche Proben potentiell ehemals mit TNT kontaminiert *waren*. Die Ergebnisse meiner Doktorarbeit demonstrieren das Potential, Umwelteinflüsse, genauer, Kontaminationsereignisse in der Ostsee durch ML-analysierte mikrobielle Gemeinschaften vorhersagen zu lassen. Taxa, die zu indikativen Fingerprints beitragen, konnten ermittelt werden, zur endgültigen Bestätigung ist jedoch eine höhere Probenzahl notwendig. Es wurde erkannt, dass Abhängigkeiten (z.B. räumliche) zwischen ökologischen Proben überoptimistische Vorhersagegenauigkeiten ermöglichen. Ihr

Vorkommen ist aber für Experimente, die ökologische Zusammenhänge untersuchen, weit verbreitet. Sowohl um potentielle Abhängigkeiten zu erkennen und ihren Einfluss abzuschätzen, als auch um ökologische Zusammenhänge aus den ML-relevanten Informationen ermitteln zu können, sollten interpretierbare ML-Methoden priorisiert werden. Es wurde gezeigt, dass die Implementierung des dargestellten Ansatzes in den regulären Monitoringbetrieb sowohl in Bezug auf die Methodik als auch ressourcentechnisch möglich ist und im Gegenzug die benötigte Ausweitung der Probenmenge anbietet.

General introduction

Formation of Earth and life

The earth was formed out of solar nebula about 4.54 billion years ago (Dalrymple, 2001). The earliest specimen of the genus *Homo* has been estimated to be about 2.8 million years old (Villmoare et al., 2015), *Homo sapiens* about 300,000 years. Such geological time spans are difficult to imagine; it appears that humankind existed and experienced evolutionary forces over a long time. To put this impression into perspective, it has been confirmed that the microfossils of the earliest life forms are 3.5 billion years old (Bernard and Papineau, 2014), with more finds gathered near submarine-hydrothermal vents indicating microfossils between 3.77 and potentially up to 4.3 billion years ago (Dodd et al., 2017). These microfossils, potentially originated “only” 200 million years after formation of Earth, are remnants of the first microorganisms. Today’s bacteria are descendants of those ancient microorganisms (Di Giulio, 2003), which have become the most abundant life forms on earth. They were subjected to the mechanisms of evolution for billions of years. The selective processes took rapid effect due to average bacterial generation times of hours to days (Vieira-Silva and Rocha, 2010). An unmatched variety of physiologies evolved, allowing prokaryotes to conquer by now virtually all environments on Earth, including habitats no other life forms are equipped for. To give some examples of the extraordinary capabilities of bacteria, alive cells, millions of years old, have been reported from samples taken 2.5 km below the sea sediment surface, where they were estimated to reproduce every 10,000 years (Inagaki et al., 2015). Microbial life has been retrieved from deep marine sediment, where almost no energy source was available, and subsequently stimulated and promoted to grow after 100 million years (Morono et al., 2020). Furthermore, bacterial cells survive freezing, or more specific, vitrification of the cell interior. This quality is exploited for long term storage of microbial cultures, but also enables them to survive in permafrost for millions of years (Christner et al., 2003). On the opposite side of the thermal scale, Archaea can grow at up to 122 °C and Eubacteria at up to 100 °C (Clarke, 2014). Bacteria of the family *Deinococcaceae* possess efficient DNA repair mechanisms sufficient to live within the cooling system of nuclear reactors as well as to survive clinical instrument sterilization via irradiation (Makarova et al., 2001). Growth of the acidophile archaeum *Ferroplasma acidarmanus* has been reported at pH 0 (Dopson et al., 2004). The currently known limits of microbial life demonstrate their range of ecological niches and life styles. However, bacteria living in ordinary conditions are still remarkable. To understand the importance of microorganisms to a given habitat (including the whole Earth, ultimately), their size, usually in the range of micrometers, must be considered an advantage. Therefore, 1 mL of sea water may contain more than 1,000,000 cells (Heinänen, 1991) of 1000s of species, and

1 g of sediment holds over 10,000,000,000 cells (Braun et al., 2016). Their cell size of, on average, 0.1 µm³, allows them to efficiently interact with their surroundings by diffusion, depending on the surface to volume ratio (Schulz and Jørgensen, 2001). These features are the reasons why microorganisms are the driving force of the biogeochemical cycles, or, as Falkowski et al. (2008) state "[…] Earth's redox state is an emergent property of microbial life on a planetary scale".

Bacterial lifestyles and strategies

Bacteria virtually always co-exist in microbial communities, displaying a variety of lifestyles. For example, the principles of oligotrophic and copiotrophic growth strategies, referring to bacteria adapted to environments with less and more nutrients available, respectively, have been discussed by Koch (2001). In terms of co-existence, one can distinguish free-living cells and surface-colonization via biofilms (Rieck et al., 2015). The majority of bacterial life occurs in biofilms, displaying high cell abundances and activities (Costerton et al., 1995). Nonetheless, free-living and biofilm-involved (also named planktonic and sessile) lifestyles can be expressed by the same organisms (Marshall, 2013). Biofilms can be formed at any kind of surface or interface and are as ubiquitous as bacteria themselves (Flemming and Wuertz, 2019).

Bacteria living in biofilms on sediments, are commonly referred to as particle-associated (Meyer-Reil, 1994; Rieck et al., 2015). Sediments provide characteristics which affect the physical conditions of a habitat and ultimately the microbial community composition. Most notably the grain size distribution affects the penetration depth of oxygen into the sediment and the begin of reducing conditions, a major selection criterion for microbial communities (Broman et al., 2017). The composition and shape of the sediment grains is important, as it, together with the grain size, determines which material may adsorb to the particles and thereby defining what nutrients are bioavailable (Zinke et al., 2018). Furthermore, sediments and even single sediment grains comprise microhabitats, allowing for the coexistence of e.g. aerobic and anaerobic species in mm range (Edlund, 2007).

Reactions of microbial communities towards changing environments

Summarizing the previously described findings, bacteria can be found everywhere. They co-exist in communities, they are very old and have therefore developed a broad range of physiological traits. Due to their contribution to the biogeochemical cycles as well as being the foundation of the food web, they are of indispensable value for the environment and actively shaping it. Classical community ecology prioritized determining the (environmental) factors that shape community composition (Paliy and Shankar, 2016). Having thus assembled an extensive knowledge about these factors, one can now in return explore the

potential to derive information about a specific environment solely from the microbial community composition. Interactions between the environment and microbial communities are the foundation for the studies of this thesis. The physiological diversity of bacteria leads to specific, distinguishable assemblages for ecological niches, from the gut of insects (Douglas, 2015) to the anoxic regions of the Baltic Sea (Thureborn et al., 2016). Such assemblages are not static, they inherit levels of intrinsic variability; the “normal operating range” (Orwin and Wardle, 2004). Deviations from this range indicate a condition of stress, caused e.g. by a disturbance. Disturbance is defined as either a) indirectly affecting the environment of a community, e.g. a saltwater inflow into a brackish system and thereby changing the osmotic conditions for the bacterial cell (Bergen et al., 2018), or b) directly affecting the microbial community itself (Rykiel Jr., 1985; Glasby and Underwood, 1996), for example, due to the availability of hydrocarbons during an oil spills (Smith et al., 2015). Furthermore, disturbances can be distinguished as pulse and press. A pulse is a short-term stressor whereas a press effects the system over a longer period of time or even continuously (Shade et al., 2012). A common case of disturbance is an anthropogenic contamination event, where foreign substances are introduced to an ecosystem. The pollutants may be of synthetic or natural origin, but the concentration is significantly above the natural background (e.g. radioactivity, hydrocarbons, heavy metals), sometimes also described as “degree of contamination” (Shirani et al., 2020). Depending on the type and strength of disturbance, specific members of the community can take advantage of e.g. nutrient availability, and outgrow their competitors (Lindh and Pinhassi, 2018) or are capable of degrading substances useless or even harmful for other members (Fahy et al., 2005). Bacteria can express different genes to adjust their metabolism or to exchange genetic information via horizontal gene transfer to overcome the effects of the disturbance (Thureborn et al., 2016). Ultimately, all of these scenarios potentially result in the same outcome: the abundances of taxa change, hence, the microbial community composition is altered. Resilience in that context is defined as the rate at which a community returns to its original composition after being disturbed, also known as community recovery. Resistance describes the degree to which a microbial composition remains the same amidst a disturbance. However, disturbances may also have initiated microbial succession towards another stable state of the ecosystem, as microbial communities can have multiple stable states (Shade et al., 2012).

A resilient microbial community in recovery reflects an environmental condition which no longer prevails (the disturbance) and which therefore could not be detected by direct analyses. In consequence, the microbial community maintains information about a disturbance for the duration of community recovery. Resilient community compositions have

been exploited in that manner to detect former oil spills, after the hydrocarbon levels had returned to background levels (Smith et al., 2015).

Accessing microbial communities via next generation sequencing

Being able to analyze microbial community compositions is the requirement to access the information they contain. For a long time, bacterial strains had to be isolated to enable further investigations. However, the majority of bacteria are yet not cultivatable (reviewed by Bodor et al., 2020). Clone libraries (Green and Sambrook, 2012) and fingerprinting techniques such as denaturing gradient gel electrophoresis (Fischer and Lerman, 1980) and single-strand conformation polymorphism (Orita et al., 1989) enabled enquiry into the dominant taxa independent of cultivating. In recent years, the sequencing coverage of microbial communities has improved significantly by the advent of next generation sequencing (NGS) technology, namely the 454 Pyrosequencing (reviewed in Clarke, 2005; Leamon and Rothberg, 2009) and then the Illumina sequencing by synthesis platforms (Caporaso et al., 2011). Depending on the sequencing depth and habitat complexity, NGS provides utilizable information on taxa of relative abundances < 0.1 % (Janßen et al., 2019b). Cultivation techniques are still and will be essential to investigate the physiology and other characteristics of bacteria. Yet the phylogenetic identity and functional potential of organisms are encoded by the genome and the expressed genes constitute the transcriptome. Therefore, sequencing of DNA or RNA allows us to identify which bacteria are present, what they are capable of (both DNA) and what genetic functions they are actually utilizing (RNA). The methods used for this thesis include i) the sequencing of a specific region determined by a primer set (“amplicon” sequencing) and ii) shotgun sequencing, a primer-less method.

In amplicon sequencing, the target region is flanked by a set of oligonucleotide primers and becomes specifically amplified by PCR before sequencing. This approach is commonly applied to sequence several functional genes, but most important is probably the 16S rRNA gene (Caporaso et al., 2011). Due to its essential, and therefore conserved sequence regions, it was possible to design primers targeting parts of the 16S rRNA gene within a large variety of prokaryotes (Takahashi et al., 2014). Amplicons ensure that the capacity of the sequencing device can be fully utilized by sequencing only the targeted regions, avoiding spending capacity on irrelevant sequences. Therefore, amplicon sequencing is the method of choice to retrieve microbial community compositions or the abundance of specific functional genes. The genetic data of a sample prepared for sequencing is called a library; the output consists of the sequences and their abundances per library. Bioinformatic pipelines (e.g. mothur by Schloss et al., 2009, DADA2 by Callahan et al., 2016) query 16S

rRNA data bases (e.g. SILVA by Yilmaz et al., 2014) with sequence similarity-comparing algorithms to provide a phylogenetic classification and a taxonomic annotation of such sequences.

Shotgun sequencing refers to the method of breaking up the total DNA or reverse transcribed RNA into fractions of appropriate size to sequence it in totality (Leamon and Rothberg, 2009). The DNA of an environmental sample is represented by its metagenome, the RNA as metatranscriptome, respectively. Compared to amplicon sequencing, shotgun sequencing results in a larger amount of sequencing data, followed by more complex bioinformatic processing (e.g. MetaSPAdes by Nurk et al., 2017; MetaWRAP by Uritskiy et al., 2018). In return, both metagenomes and transcriptomes provide information not only about the taxonomic composition of a sample, independent of primer sequences, but of all encoded or transcribed information. Further processing steps involve the functional annotation of genes (e.g. Prokka by Seemann, 2014) and the recreation of metabolic pathways (e.g. MinPath by Ye and Doak, 2011) using appropriate data bases (e.g. MetaCyc by Caspi et al., 2020). Binning tools collect sequences of similar nucleotide composition and sequencing coverage in an attempt to recompose distinct genomes from metagenomes (e.g. CONCOCT by Alneberg et al., 2014; MaxBin by Wu et al., 2014). A table comprised of the absolute counts of either genes, transcripts or taxa per library is the primary outcome of amplicon and shotgun sequencing, respectively.

Microbiological data sets have specific characteristics

For reliable and sound interpretation of data generated by sequencing, it has to be kept in mind that such tables are not exact representations of the community composition of the sampled habitat due to a variety of reasons, here ordered along the sample processing: biological systems may react to the sampling itself, therefore the state while sampled has to be preserved (Charvet et al., 2019). The sampling of a habitat usually cannot and should not be exhaustive, but representative. However, it is not trivial to provide evidence for a representative sampling. Laboratory processing introduces further bias, such as the protocol used for extracting the nucleic acids, the chosen primer set in terms of amplicon sequencing, the sequencing depth and the sequencing itself. Bioinformatic programs designed for large data amounts involve heuristics and probability (e.g. the read mapper kallisto by Bray et al., 2016), which possibly leads to varying results for the same analysis step. The resulting library therefore contains a subset of the microbial community in form of compositional data (Gloor et al., 2017). The underlying probability distribution remains largely unknown. An additional issue is that a large number of bacteria are not described yet.

Machine learning to classify microbial community compositions

In order to analyze microbial community compositions to learn about their environment, NGS allows the retrieval of a wealth of information, although the described limits and constraints apply. Machine learning (ML) is a field of statistics and data analyses with proven capability to find relationships and patterns in complex data sets of unknown distribution. ML is also called statistical learning or algorithmic modelling. Bzdok (Bzdok, 2017; Bzdok et al., 2018) has described an increased usage of ML in general biology-related sciences. More specifically, microbial community data has been increasingly analyzed and used for predictions by ML algorithms (reviewed by Qu et al., 2019). For a short introduction, statistical analysis involves on the one hand descriptive statistics to understand the gathered data, using measures such as range, mean or median. These results are specific for the collected samples. On the other hand, statistical inference aims to draw further conclusions beyond the exact collected samples, such as how likely is a measured effect to occur by chance.

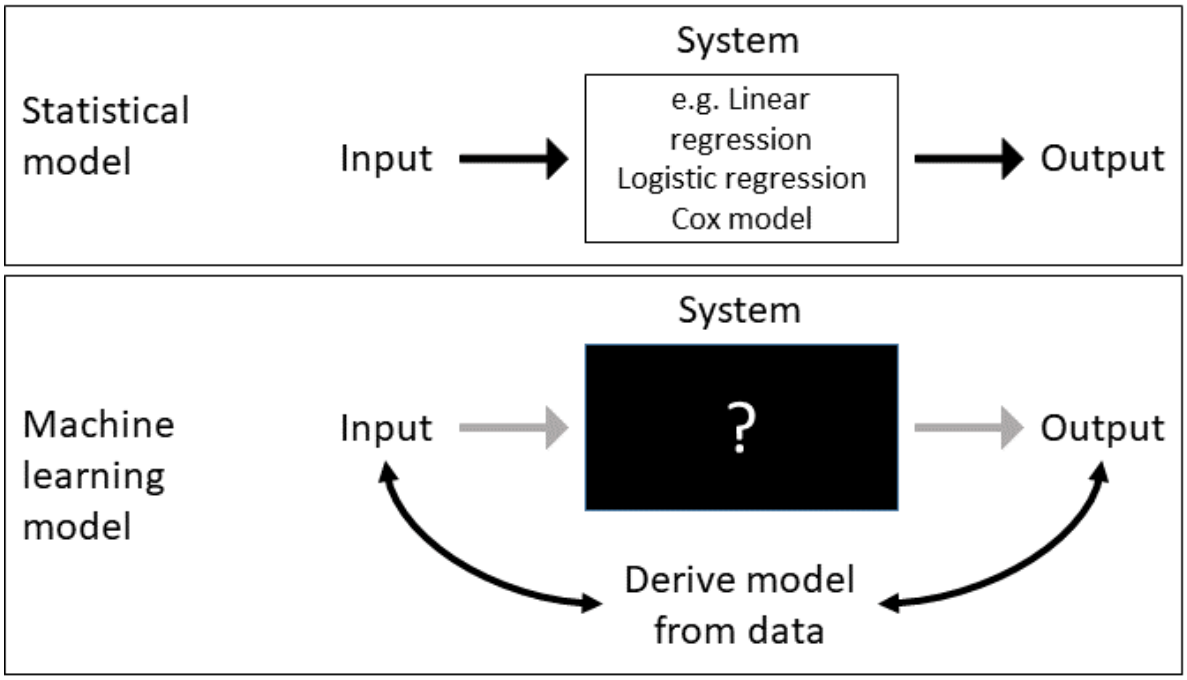


Figure A: Comparison on the description of an unknown system by statistical and ML models. Statistical models are chosen *a priori* and therefore require certain knowledge or assumptions about the system they describe. ML derives a model from the data, the resulting model does not claim to be a true representation of the system. Modified after *Breiman (2001b) and Bzdok (2017)*.

Statistical inference requires a statistical model (Figure A), which can essentially be described as a set of assumptions about the probability distribution of the population the sample was drawn from. To illustrate, after a coin toss the chance for each side facing upwards is ½, which is already a statistical model. It allows the calculation of further data points without additional sampling (or coin tossing). It is important to notice that any statistically inferred conclusion from the measured data is based on the chosen model

(Steel et al., 2013). Parametric and non-parametric tests can be applied (Oksanen, 2015) to test e.g. for significance, confidence intervals or hypotheses rejection.

In most data sets and especially in environmental biological systems, there are more variables and outcomes than “flipped coin lands on either side” (Økland, 2007). Data collection for microbiological and ecological studies is not trivial, as described earlier, and both time and money expensive, with the data itself being compositional (Gloor et al., 2017). The resulting limited sample size aggravates the estimation of potentially complex underlying probability distributions and thus, the selection of the appropriate statistical model (Økland, 2007).

ML prioritizes the detection of generalized patterns for predictions on new data sets over inference and interpretability: “Statistics draws population inferences from a sample, and ML finds generalizable predictive patterns.” (Bzdok et al., 2018). Statistical inferences are therefore drawn from the whole data, which however complicates the detection of irrelevant information (“noise”); including noise results in a phenomenon called overfitting (Dietterich, 1995). As a simple example drawn from image analysis, imagine different pictures of a chair. Let us define that all chairs have four legs, a seat and a back. These features represent a generalized pattern (an abstraction of a chair), allowing a model to classify unknown images as showing chairs. Yet, an overfitted model would additionally include irrelevant or misleading data such as the background of the image, the color of the legs or the material of the seat. As consequence, the overfitted model cannot identify chairs that do not exactly look like the one earlier presented to it. Data sets subjected to ML analysis are split into training and test data to identify overfitting. Cross validation performs multiple splits and serves the same purpose, it is applied by ML and also by statistical models (Stone, 1974). ML performance is evaluated based on its predictive power about the holdout test data (or otherwise new and unseen data), assuming that “Higher predictive accuracy indicates capturing of underlying mechanisms.” (Breiman, 2001b). In comparison, statistical models use *p* values, goodness of fit and analysis of residuals for their validation. Statistical inference leads to the single, best solution for the whole data set, based on *a priori* chosen model assumptions, to perform model-driven hypothesis testing (Figure A). A potential problem is the existence of similarly good alternative solutions, which however would lead to different inferred conclusions (Breiman, 2001b). ML in contrast uses data-driven learning algorithms, where “hypothesis” has a different meaning. A hypothesis is a single possible state of a ML model, much like an allele is a single possible state of a gene. The parameters of the model, defining its state, are adjusted during learning from data: while training, the model explores the sum of hypotheses, the hypothesis space, for the optimal solution (also called function approximation). To conclude, ML models are rather “derived” from linking

input to output data (Bzdok, 2017). Due to its model-deriving concept, ML is usually deemed non-parametric. The only statistical assumption for various ML algorithms is that the variables are independent and identically distributed (i.i.d.), which is often violated for real world cases (Dundar et al., 2007). Furthermore, Økland (2007) states that "samples with statistically desirable properties will be ecologically irrelevant".

I find it important to mention that the distinction between statistic models and statistical/ machine learning was and still is the topic of heated discussions. Breiman referenced this controversy in his 2001 article "Statistical modeling: The Two Cultures", where he gave detailed examples. He discussed in favor of embracing the use of ML and answering comments from several critics. A comprehensive comparison including both terminology and examples is provided in this non-peer-reviewed blog post by Matthew Stewart, who addresses the claim that machine learning and statistics are identical: https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3. He uses the example of linear regression, which can be applied by machine learning as well as by statistical models, to explain the potential of conflating both terms.

Shallow and deep machine learning algorithms

Machine learning algorithms belong to the larger field of artificial intelligence and are important tools for data mining, data analysis and prediction (Mitchell, 1997). They can be categorized in shallow- and deep-learning algorithms. This distinction stems originally from artificial neural networks (ANN), which could comprise several processing layers to manipulate the data. An undecided number of required processing layers made the ANN "deep" (Schmidhuber, 2015). The definition is maybe more clearly expressed by naming everything "shallow" which is not an ANN with multiple processing layers. An extensive review, foremost on deep learning, but including a detailed ML timeline can be found in Schmidhuber (2015). The beginnings of ML include the first symbolic ANN described by Minsky in 1951. Those models did not learn until backpropagation with gradient descent was invented. This enabled the use of the difference ("loss") between the predicted and true values to improve the model. Today, vast amounts of training data and ML software implementations are publicly and freely available. The hardware is sufficiently powerful and in parts specifically designed (e.g. chips like the tensor processing units) for the calculation of deep learning models.

Throughout the thesis, the focus was to train ML models with community composition data. Additionally, in Chapter III, environmental parameters were included as independent variables. The data used for machine learning is ambiguously described as features; the

meaning differs between shallow and deep learning. In shallow learning, the independent variables are equivalent to features. However, such variables include manually manipulated ones. For example, total organic carbon (TOC) is commonly not measured directly, but calculated from total carbon (TC) and total inorganic carbon (TIC). This process is called feature engineering and is deemed essential for the predictive success of ML models. This is intuitive, as TOC provides different and potentially more relevant information than the measured variables TC and TIC. In contrast, deep learning methods automatically generate abstract representations, sometimes also called features, from raw input data (Zhong et al., 2019). Shallow learning algorithms do not engineer new features from the provided independent variables.

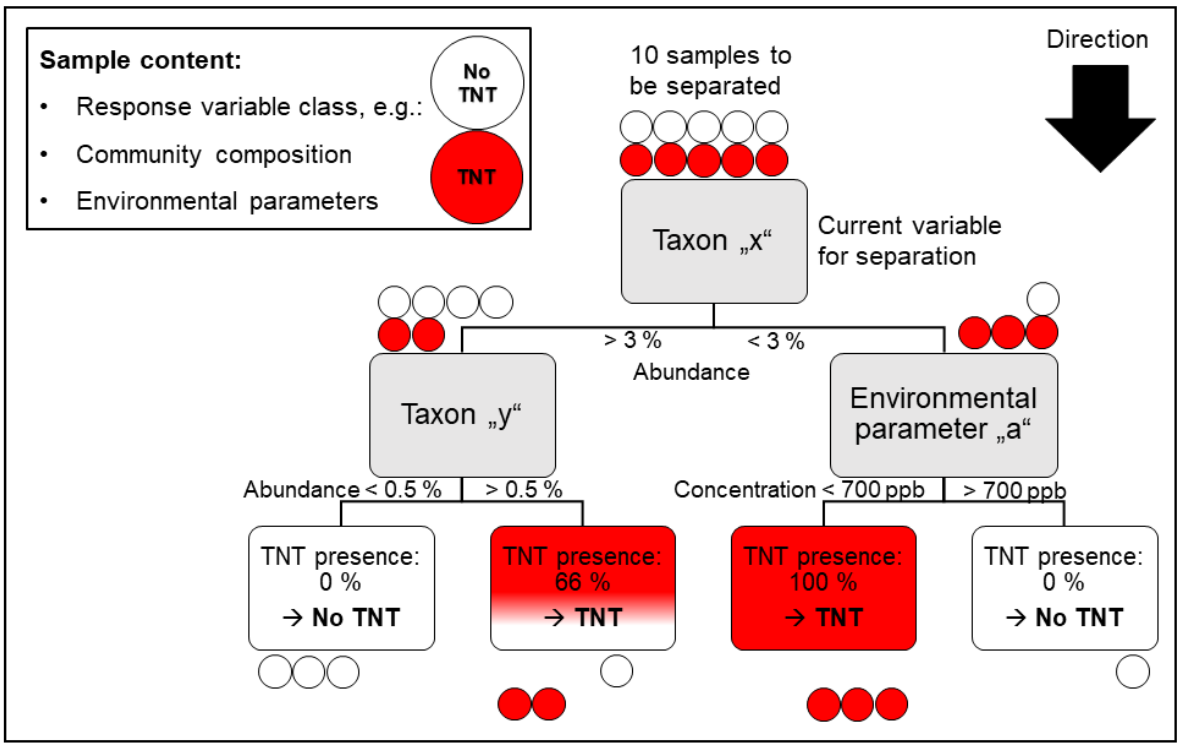


Figure B: An exemplified decision tree attempting to classify samples containing microbial community and environmental information as independent variables and the presence of TNT as response variable. The split rules are derived from training, where the separation capability for each variable is measured, for example by increase in the Gini index. The variable with the most efficient separation of classes is considered first (here “Taxon x”). A single decision tree with sufficient variables is always able to perfectly separate the response classes by overfitting (in this example, the bottom second node could be further split by another variable). To prevent this, Random Forest uses multiple decision trees, each grown on different subsets of samples and variables (see text and Chapter I for more information).

The shallow Random Forest (RF; Breiman, 2001a) algorithm uses an ensemble of (the typically weak classifiers) decision trees (Figure B) to constitute a so-called forest (described in detail in Chapter I). Each tree is based on a different subset of the variables/features and observations of the data. This process is called bootstrap aggregating or in short, bagging. Bagging reduces variance and avoids overfitting, furthermore, it increases the robustness. A majority vote based on all bagged decision trees

eventually classifies the data. Bagging results in the use of approximately 2/3 of the available data to generate the tree. The remaining third, which was “out-of-bag” is then predicted by the newly generated tree and therefore, provides a validation set. This process provides the out-of-bag error estimate. For each tree, different samples are used for tree-generation and tree-validation. Therefore, it is still required to have a separate holdout test set, completely uninvolved in training the models.

RF only possesses two important hyperparameters. These are settings to choose and optimize; named in contrast to regular parameters which are adjusted automatically by the training process. In RF these are the number of trees and the number of randomly selected variables (“mtry”) at each node split. Random forests technically allow for reconstructing their decisions, however, in reality it is not feasible to disentangle all split events in all trees.

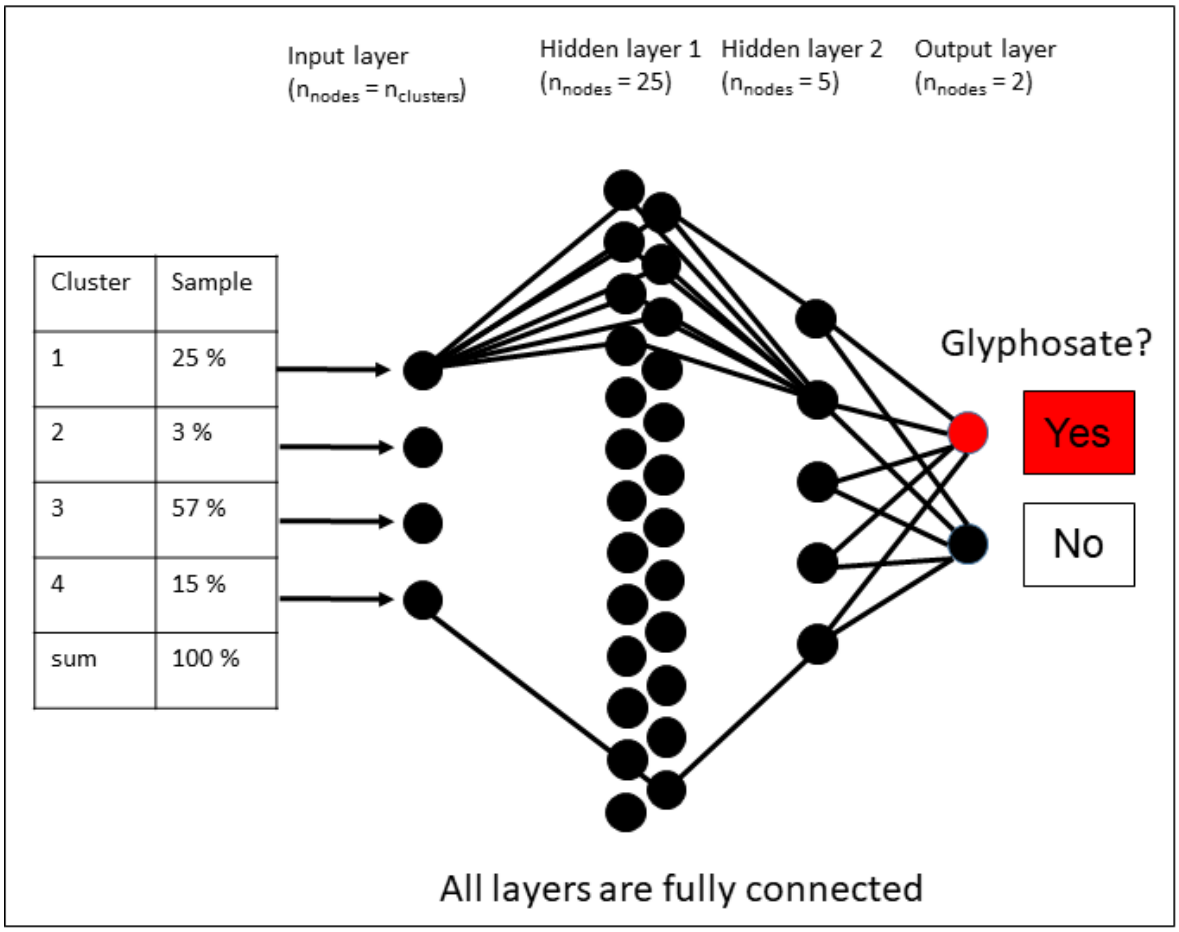


Figure C: Architecture of the ANN used in Chapter I with exemplary input data. Relative abundances of taxonomic clusters per sample are provided to the input layer, consisting of as many nodes as variables/features. The combination and feature engineering (see text) occurs in the hidden layers, were each node is connected (only a few connections are displayed to maintain clarity) to all nodes of the previous and next layer. The signal passed from node to node is adjusted along the connecting path (edge) to map the input to the output layer.

In contrast to RF, ANNs comprise a large class of deep learning algorithms with architectures reaching increasing levels of complexity and depth. Essential are an input layer for the raw data, various amount of hidden layers, number of nodes per hidden layer,

the corresponding processing mechanisms and the output layer (Figure C; more detailed explained in Chapter I). All types of neural networks have in common that the raw data is processed, manipulated and/or combined passing through each hidden layer. ANNs provide a wealth of functionalities like feed forward multilayer perceptrons, convolutional neural networks (CNN) for image and spectral data or recurrent neural networks and long short-term memory for sequential data. However, they come with a large number of hyperparameters to tune. Due to the abstraction of data and its manipulation, the inner workings of deep learning models are considered intransparent; hence being described as black boxes. Great effort is directed in developing more conceivable and interpretable models (Lapuschkin et al., 2019).

The demand for interpretability is best exemplified by the most efficient class of deep learning algorithm for pattern recognition in audio and visual data: CNN utilize automatic feature generation (reviewed in LeCun et al., 2015). It is possible to extract which features they recognized for pattern recognition across several steps of abstraction. The extracted features have frequently identified the “Clever Hans” problem, where accidental or confounding features are used to classify or recognize classes (Samhita and Gross, 2013; Lapuschkin et al., 2019). An example is the use of background information when animals should be classified in images. Hence, it became clear that a certain level of interpretability of a ML model is required for oversight and error handling. Rudin strongly expresses her opinion to use interpretable models in the first place instead of attempting to partially explain black box algorithms (Rudin, 2019). Furthermore, I want to mention the existence of adversarial examples. Such previously accurately classified images have received an imperceptible amount of perturbation to them which causes misclassification (Szegedy et al., 2014). Advanced methods to gain insights into CNNs have been described for example by Montavon et al. (2018). They defended deep learning algorithms as being interpretable.

In summary, both algorithms are capable of deriving patterns from data involving non-linear relations, but RF uses only the provided independent variables, whereas ANN can further combine the input data to engineer more valuable features. In contrast to formats for visual or acoustic data, this thesis only provided structured data as input for ML, referring to tabular data. The microbial community compositions were prepared in the form of relative abundance per taxa. With reference to the term “artificial intelligence”, it is emphasized that none of the algorithms possess a concept about what a bacterial taxa or microbial community composition represents at that point.

Variable importance and interpretable models

Certain ML methods provide a variable importance measure, which reports the relevant features for the model. Linear support vector machines or logistic regression allow for specifically reporting the contribution of each variable due to their linear function space (Topçuoğlu et al., 2020). RF has several measures to quantify the importance and calculate the significance of variables (Altmann et al., 2010; Janitza et al., 2018; Nembrini et al., 2018). However, as usual for non-linear classifiers, it is not feasible to exactly trace back every decision of even a small model consisting of e.g. 50 trees (Breiman, 2001a). In general, however, the variable importance still demands interpretation and more specific to this thesis, it does not automatically represent those variables that are indicative for a contamination event.

Although the variable importance by RF is not fully transparent, it is still meaningful, because RF leaves each variable unaltered during training and prediction. In contrast, the inherent problem with deep learning structures is that a given variable, i.e. taxon, does not exist individually anymore after the input layer (LeCun et al., 2015). An alternative is to extract and analyze the "flow" of data. However, again due to non-linear activation functions, even a small network is complicated to interpret, but it may provide interesting insights into which variables were combined.

Variable importance is limited here to the impact of the input variable on the final prediction outcome. It can be estimated e.g. by stepwise removal or addition of variables, followed by training the model and logging the prediction outcome. Another way is to permute variables, until they no longer contain useful information, and compare the outcome with unpermuted variables. This method is commonly applied, also by RF.

Clustering, classification and regression tasks for models

The most common applications for ML involve tasks such as classification, regression and clustering, stemming from the closely related field of data mining (Fayyad et al., 1996). Classification and regression are so-called supervised methods; the data contains a response variable in form of class labels or a continuous value. The model tries to map the independent variables to the discrete classes (Classification), or for regression, to the continuous value. Discretizing continuous values into intervals also enables their classification. Clustering is the process of finding similarities between observations and underlying patterns without additional info being provided, therefore, analyzing the data "unsupervised" (Angermueller et al., 2016). It is thereby similar to exploratory ordination methods for multivariate data. Clusters detected in unlabeled data sets can be assigned with classes for subsequent supervised classification. In this thesis, ML models have been

used to classify Baltic Sea microbial community compositions with regard to the presence of a contaminant. Furthermore, unsupervised clustering was used to identify the main environmental drivers of microbial communities.

Pollution of the Baltic Sea

The Baltic Sea was chosen as a research area to examine the impact of contamination on bacterial life as it has a long history of pollution and eutrophication. The brackish microbial communities have been analyzed in great detail, although foremost regarding the Baltic Sea key characteristics: the spatial distribution along the salinity gradient (Herlemann et al., 2011), across the redoxcline (Grote et al., 2007) or the temporal gradient during algal blooms (reviewed in Lindh and Pinhassi, 2018). However, much effort went into the investigation of the influence of specific pollutants such as heavy metal concentration and persistent organic pollutants (Edlund, 2007; Thureborn et al., 2013; Rodríguez, 2020).

The Baltic Sea is particularly susceptible to pollution due to the fact that it is rather shallow. The only exit to the North Sea and the Atlantic Ocean is via Kattegat and Skagerrak, respectively, therefore, the water residence can be as high as 30 years in the central Baltic (Rheinheimer, 1998). As a consequence, pollutants do not get flushed out and do not become as diluted as in the oceans, despite the Baltic Sea being the largest inland brackish sea (Snoeijs-Leijonmalm and Andrén, 2017). It is bordered by nine states. The drainage area is inhabited by 85 million people and is 4 times larger than its sea surface of 415,000 km² (Sweitzer et al., 1996). Agricultural runoff, (historical) industry, marine traffic and the water discharge of large estuaries result in the input of various pollutants and provide an oversupply of nutrients (mainly nitrogen and phosphorus compounds). The resulting eutrophication (Andersen et al., 2017) allows for strong growth of algal and microbial biomass (“algal blooms”). The breakdown of this biomass consumes oxygen, leading to its depletion. In 2006, the hypoxic bottom regions, defined as > 2 $ml \cdot L^{-1}$ dissolved oxygen, covered 67,700 km² (Conley et al., 2009). The influx of nitrogen (N) and phosphorus (P) to the Baltic Sea is therefore monitored and part of ongoing research (Ahtiainen et al., 2014). One of the agricultural run-off substances, despite its proposed soil adsorption characteristics (Bergström et al., 2011; Myers et al., 2016), is the herbicide glyphosate (Skeff et al., 2015). It is the most-applied herbicide globally since the 1970s, and can be found in soil and groundwater (Battaglin et al., 2014), marine and freshwater systems (Van Bruggen et al., 2018; Carles et al., 2019) and the Baltic Sea (Skeff et al., 2015; Wirth et al., 2021). Glyphosate has been shown to disturb microbial communities (Stachowski-Haberkorn et al., 2008). It furthermore provides carbon (C), N and P for bacteria and fungi (Lipok et al., 2007; Duke and Powles, 2008). The most common means of glyphosate

biodegradation are towards sarcosine utilizing the *phn* operon and towards aminomethylphosphonic acid (AMPA), enabled by the *gox* gene (Sviridov et al., 2015).

Whereas glyphosate has been detected in the water column, the Baltic Sea sediments have been also described to accumulate toxic substances such as polychlorinated biphenyls and polycyclic aromatic hydrocarbons (Edlund, 2007) from industrial use. After World War II, regions like the Landsort and Gotland Deep were used to dispose of chemical warfare agents (Bełdowski et al., 2016a). About 300,000 tons of conventional munition and 5000 tons of chemical warfare have been estimated to still be present in both the North and Baltic Sea (Böttcher et al., 2011). The munitions dumpsite Kolberger Heide is located close to the German city of Kiel in the Kiel Bight. It is about 2 km off the beach, about 1260 ha large and 10–15 m deep (location included in Figure D). The dumpsite comprises conventional, mostly defused, munition and the metal containments display various states of progressing corrosion. They contain mainly 2,4,6-trinitrotolouene (TNT) and 1,3,5-trinitroperhydro-1,3,5-triazine (RDX) as munition compounds (MC). Among other explosives, TNT and its degradation products have been detected in water samples and biota collected at Kolberger Heide (Gledhill et al., 2019). Little is known about the MC concentrations in sediments. For a detailed description of the site including maps, images of detonation craters and scattered, bare munition chunks the interested reader is referred to Kampmeier et al. (2020), the UDEMM project analyzing the dumpsite is summarized by Greinert (2019).

Monitoring the environmental state of the Baltic Sea

The examples of eutrophication, oxygen depletion and contamination detailed above visualize the importance of environmental monitoring to assess the quality and state of the Baltic Sea. The HELCOM members (Baltic Marine Environment Protection Commission or Helsinki commission: Denmark, Estonia, European Union, Finland, Germany, Latvia, Lithuania, Poland, Russia and Sweden) cooperate to manage and monitor the environmental state of the Baltic Sea (HELCOM, 2018). The member states are required to implement monitoring programs according to the EU Marine Strategy Framework Directive. As a consequence, the Baltic Sea action plan has been developed in 2007 to achieve good environmental status by 2021 (Backer et al., 2010). The program has been prolonged, and the plan has been updated (“Strategic plan for the BSAP update”), as the goals are unlikely to be reached. According to the HELCOM Monitoring Manual, the following organisms are currently included: birds, mammals, fish, zoo- and phytoplankton, non-indigenous species, the distribution of fauna and flora species and the abundance of the benthic community. Furthermore, the inputs and concentrations of contaminants, as well as their biological

effects, are monitored. Microorganisms are not involved except for the abundance and species composition of phytoplankton. It is true that the investigation of organisms which are not readily seen, sampled, counted, or taxonomically classified poses a challenge for their implementation in regular monitoring activities. Yet, research applying NGS for monitoring purposes to acquire information on microbial communities, even their functional potentials and expression profiles has been undertaken with promising results (e.g. by Ininbergs et al., 2015 and by the EU Bonus project BLUEPRINT).

Description of research aims

ML could prove to be a powerful tool for investigating the environmental state of a given habitat disturbed by a contamination event. This is because, in contrast to classical prior selected statistical models, ML derives the models from data. The data, in this case microbial community compositions, is obtained via NGS. To investigate this potential, microbial communities from the Baltic Sea were first investigated in the laboratory and then *in situ* for their reactions to contaminants. The Baltic Sea is well suited as a research area for contamination effects due to a) a multitude of diffuse anthropogenic influences e.g. from rivers, agricultural run-off and the atmosphere as well as b) the presence of specific contamination events such as point sources and munition dumping and c) due to its higher sensitivity towards pollution compared to open oceans. The integration of our approach with environmental monitoring to detect, investigate, and manage contaminations is socially significant due to 85 million people living close to the Baltic Sea. A taxonomically- or functionally-described portion of the community may represent a fingerprint which is indicative for an environmental condition. To automatically extract and identify such fingerprints using e.g. variable importance measures, machine learning can be of great support. These fingerprints then allow to predict the environmental condition of a habitat solely based on the community composition. The prediction of ecological niches (by salinity, water depth) and lifestyles (free-living or particle associated) using phylogenetic and functional information has been demonstrated in Alneberg et al. (2020). Similarly, ML models trained on coral reef microbiomes diagnosed shifts in the reef environment, the microbiome acted as indicators for temperature, chlorophyll and eutrophication status (Glasl et al., 2019). The studies in this thesis describe the training of ML models with Baltic Sea microbial communities to predict to the presence of glyphosate or TNT contamination events.

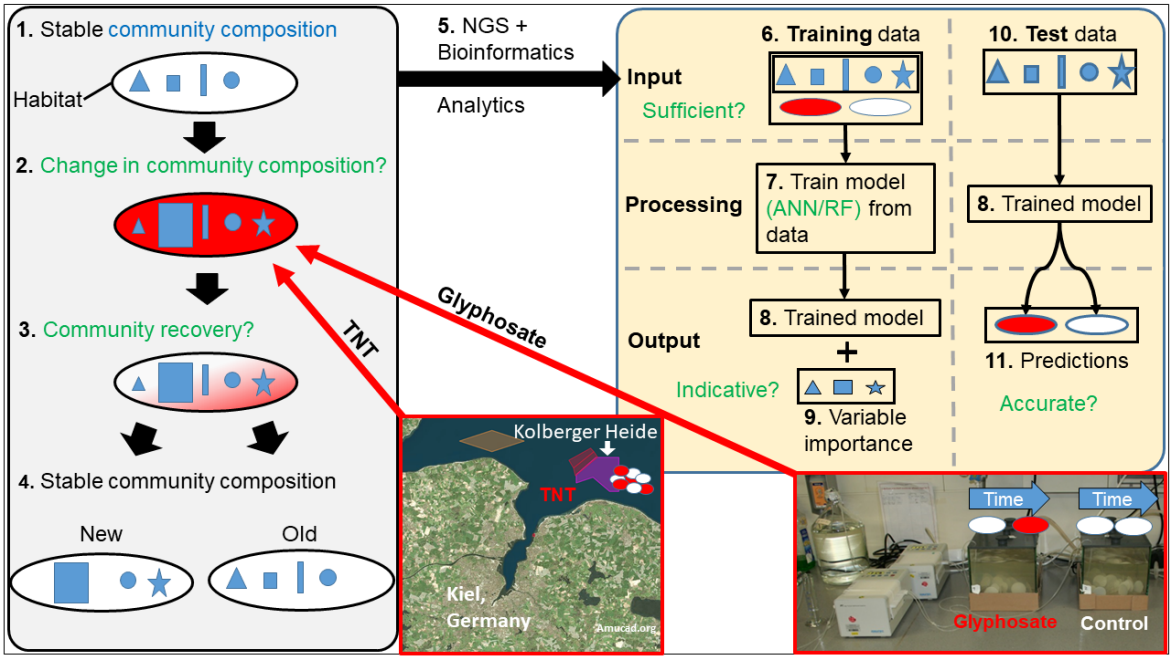


Figure D: Thesis concept overview with research questions (in green): Microbial community compositions from 140 days laboratory microcosms and from the Kolberger Heide (purple area) were investigated. 1. Community composition (shape = taxon, size = abundance) in an uncontaminated habitat (white ellipse). 2. Contamination of the habitat (by TNT or glyphosate) could alter the community composition. 3. After the contamination event, the altered composition may require time for recovery. 4. The next stable state in a formerly contaminated habitat may be the original state (1.) or a new one. 5. Community composition data from all stages were obtained via NGS; the contaminant was determined. 6. The input data for supervised ML model training were the community information and the presence of a contaminant. 7. ANN and RF models were trained for comparison of accuracy. 8. The output consisted of the trained model (for further prediction) and 9. the important variables (for interpretation), possibly indicative for the contaminant. 10. Another data set only consisting of community data could then be classified by the trained model (8.), to 11. predict the presence of a contaminant.

The conceptual approach (Figure D), executed on two different data sets, is summarized as follows: Data regarding the reactions of microbial community compositions in the presence of contamination were collected by a) conducting a laboratory microcosm experiment, where the herbicide glyphosate was added after 69 days and b) sampling in the Kolberger Heide dumpsite, allowing me to compare community compositions from TNT-contaminated and uncontaminated sediments. The microbial succession after contamination was also analyzed for a potential return to the pre-disturbed state and if, for the specific contaminant, community recovery (resilience) as a temporary state exists. Information on contaminant presence and further environmental or contextual data was collected. Community composition and contamination data was subsequently provided to train ML models and compare the aptitude of shallow and deep learning on predicting contaminations. The prediction results were analyzed with regard to accuracy and generalizability. The reasons for misclassification were also investigated. The results enabled to conclude whether the required sample size and independence of variables were met. Causal relations between important taxa and contamination were analyzed as a requirement for determining a microbial indicative fingerprint. Bioinformatics allowed me to

assess whether the findings reported by ML were biologically and ecologically meaningful and logical.

Conducted experiments and analyses

For the first two studies, I set up a one hundred and forty days long laboratory experiment involving two chemostat-like microcosms, containing brackish microbial communities. I added 82.45 µmol L^{-1} glyphosate as a pulsed stressor to one microcosm and monitored the planktonic and biofilm microbial succession with high temporal resolution. The goal was to assess whether it is feasible to analyze 16S rRNA (gene) amplicon data from free-living bacteria by ML to detect glyphosate contaminations. A stochastic subsetting approach for deep ANNs was compared against Random forest as a shallow and more interpretable ML method, which readily provides measures for variable importance. The ML task was to automatically differentiate glyphosate-treated from untreated control communities. As a function of this, the amount of taxonomic information required for a reliable classification was also investigated. The results are described in Chapter I.

In Chapter II, the detected correlations determined via ML in Chapter I, which hinted at organisms potentially involved in glyphosate biodegradation, were evaluated for their plausibility in a biological context. Using a newly developed analytical method for the detection of the glyphosate metabolite sarcosine, in combination with metagenomic information of free-living organisms, the potential pathways of biodegradation were reconstructed. Furthermore, biofilm community compositions were taken into account to compare the impact of glyphosate on their state and succession as well as their resilience and resistance with the free-living communities. The results enabled to hypothesize whether glyphosate entering the Baltic Sea will be degraded.

Chapter I and II were designed to assess both the potentials and limitations of ANN and RF models in a controlled laboratory experiment and validate the biological meaningfulness of the predictions. Following this, as described in Chapter III, the proof of principle under environmental conditions could be undertaken. ML models were tasked to predict the presence of TNT in the sediments at Kolberger Heide. In comparison to the glyphosate microcosm experiment, the concentration of TNT was in the range of pmol·g^{-1}. The microbial communities for model training came from 150 different sediment samples, which all showed varying physical and chemical attributes in addition to their individual MC contamination. Firstly, it was investigated whether prediction was possible under these conditions at all. Secondly, it was of particular interest which taxa contributed to the classification as part of a potential TNT-indicative microbial fingerprint. Given the variability of the samples, the robustness of predictions was evaluated to identify the factors which

influence the model's performance. This included assessing the false positive predictions (predicted as "TNT present", but actually "TNT absent") specifically, as they may be caused by TNT resilient community compositions, still recovering from a former TNT contamination. Depending on the community recovery time, such a phenomenon would enable the identifying of TNT contaminations when TNT itself is no longer present. Chapter I and III allowed comparing the feasibility of ML under laboratory and environmental conditions. Furthermore, using microbial communities collected from the munitions dumpsite enabled a realistic determination of the benefits of including community data and ML into environmental assessment analyses. Potential issues concerning the integration into regular monitoring activities were also discussed.

Summary of published papers

In Chapter I (Janßen et al., 2019b), both ANN and RF correctly predicted the presence of glyphosate with > 99 % accuracy by using only microbial community compositions derived from 16S rRNA (gene) amplicon sequencing. A stochastic variable subsetting approach and the RF variable importance measure showed consistently that the interaction of a few specific taxa, and even a single one (*Parvibaculum*), were sufficient to predict the presence of glyphosate. Several of these important taxa were characterized by an increase in relative abundance after the addition of glyphosate, presumably due to them degrading glyphosate or indirectly profiting from the degradation. Using DNA or RNA-derived compositions only, the sample size was likely too small for meaningful interpretation. If the technical replicates were not averaged, but provided as individual data points, the validation sample prediction reached near-perfect accuracy due to confounding variables.

Chapter II (Janßen et al., 2019a) analyzed the implications of glyphosate addition to the simulated Baltic Sea environment of the microcosms. The total cell counts increased after the addition and glyphosate degradation was ultimately determined by the presence of AMPA. These results were combined with shotgun DNA sequencing data, detecting the *gox* gene, which encodes the AMPA-producing glyphosate oxidoreductase. Analysis of the microbial succession revealed that the glyphosate pulse was sufficient to be traceable in the non-metric multidimensional scaling (nMDS) ordination. The biofilm was not as often affected as free-living bacteria, but the few affected biofilm taxa responded over a longer time period. Using a statistical model, similar taxa were identified responding to glyphosate as in Chapter I by machine learning. These taxa could also be connected to the phylogeny of the glyphosate degradation genes. Although glyphosate was still present in the range of 1 µM, the microbial responses ceased, thus, we concluded that it might persist in the Baltic Sea at prevailing concentrations of about 10 nM (Skeff et al., 2015).

Chapter III (Janßen et al., submitted) dealt with the prediction of TNT presence in sediments from the munitions dumpsite Kolberger Heide. The sediments were diverse, contained complex microbial communities and were contaminated with only pmol·g^{-1} concentrations of TNT. Yet, it was possible for RF and ANN to predict TNT using community compositions with > 80 % balanced accuracy, although TNT was not identified as a community driver. TNT was to a lesser extent also successfully predicted using geochemical and sedimentological parameters. Interestingly, the combination of both data sets revealed that the community composition already contained the relevant information. A microbial fingerprint of 25 genera was discovered as potentially indicative for the presence of TNT. The robustness and underlying factors of the prediction were thoroughly investigated to separate the spurious from the TNT-caused relationships. It was determined that the sample size has to be increased. In this regard, the analyses informed us that in particular training data from samples surrounding a mine mound was insufficient. A potential effect of resilient microbial communities was also described based on samples where TNT was not detected, but its metabolites. Finally, the implementation of this approach into regular monitoring was suggested, specifically with the current limitation of sample size in mind.

General discussion

Machine learning algorithms utilized non-i.i.d. sequencing data for accurate predictions

In this thesis, microbial community compositions were used to accurately predict the presence of glyphosate in microcosm experiments, as well as the presence of TNT in sediment samples collected at a munitions dumpsite. The taxa being most important to achieve these predictions could be identified; glyphosate-relevant taxa likely degraded glyphosate and their importance was confirmed by statistical models. It was also found that Baltic Sea sediment community compositions may conserve information of former TNT presence for a longer period, whereas communities ceased their response to glyphosate while it was still present. The experimental setup and sample size were particularly important for ML analyses. An interpretable ML model should be preferred, as confounding variables will likely occur in ecological experiments and may distort accuracy. It was found a great potential for implementing microbial community information and their analysis using ML into environmental monitoring.

The use of machine learning to predict contaminations by analyzing solely Baltic Sea microbial community compositions was (to my knowledge) not reported previous to the publication of Chapter I. The primary goal was reached when in both Chapter I (> 99 %

accuracy, required was > 90 %, for more details see below) and Chapter III (> 81 % balanced accuracy, required was > 50 %) successful predictions were achieved. However, in both studies confounding variables were identified. To avoid confoundation, ML algorithms assume that the samples are i.i.d. (Dundar et al., 2007), as do classical statistical tests, but satisfying the i.i.d. assumption often conflicts with purposeful ecological investigations (Økland, 2007). One part of the i.i.d. assumption is that variables are independent of each other. Efforts have been reported to e.g. correct support vector machines for confounding factors in biological data classification (Li et al., 2011), or use factored spectrally-transformed linear mixed models in genome-wide association studies to correct for confounding effects by population structure, family structure or cryptic relatedness (Lippert et al., 2011). Glasl et al. (2019) removed collinear variables as redundant based on a Pearson correlation coefficient > 0.7 or < 0.7. Collinearity may be caused by a confounding variable. In Chapter I a confounding factor could have been the experimental set-up. The community compositions in both microcosms were similar, but distinguishable (Figure 1.1). One of the two microcosms acted as an undisturbed control and therefore only provided samples for the “no glyphosate” class. The other provided both a control, and – after the addition of glyphosate – also a treatment class. These confounding constraints were known and permitted to calculate an accuracy threshold, indicating if the model actually had learned glyphosate-related effects. Accuracies up to 59 % were achievable by pure guessing of the majority class (“no glyphosate”) and up to 90.6 % by separating the microcosms. This means that a fictional “microcosm”-variable alone would enable > 90 % correct classification. Therefore, the achieved accuracies of 99.9 % using RF and 95.8 % by ANN (Figure 1.5) with the 10 most important, unfiltered taxa do present the identification of glyphosate-related abundance changes, but they comprise about 5–10 % of the accuracy. However, for less controlled or environmentally biological experiments, the degree of variable (in-)dependence which has to be factored in, is often not known. In Chapter III potential confounding variables identified were e.g. the sampling season and the sample area (Supplementary Material 3.1), approximated by the grain size distribution of the sediment as well as the sampling method. TNT classes (68 x present, 82 x absent [or 55 %]) were more balanced than glyphosate classes (26 x treated, 38 x control [or 59 %]), but still the balanced accuracy measure was applied to correct for imbalances (Brodersen et al., 2010). Furthermore, data sets mostly of smaller sample sizes can contain incidentally useful variables, which again are not related to the response variable. Confounding and coincidentally useful variables are a particular obstacle when investigating a potential indicative fingerprint such as in Chapter III. The problem is specifically discussed e.g. by Darrell et al. (2015), but often not mentioned in literature when deploying models

e.g. in the ML framework for microbiome analyses. Topçuoğlu et al. (2020) mentioned in general that the correlation structures of a data set should be understood.

The second precondition described by i.i.d assumes that the data comes from the same (or an identical) distribution (Darrell et al., 2015). In biological and environmental samples, often the true distribution is unknown and multiple distributions may be involved even for one distinct sampling campaign. It furthermore depends on the scope of the experiment: when investigating surface sediments, samples from a deep layer of a sediment core may not belong to the same distribution. Yet when investigating sediments from a munitions dumpsite altogether is the goal, both surface and core samples are included.

The i.i.d. assumption has been criticized across various research fields (Darrell et al., 2015). As Økland (2007) puts it with regard to statistical models explicitly referring to the nature of ecological samples: "[…] that samples with statistically desirable properties will be ecologically irrelevant [...] because natural phenomena are spatially and temporally nonrandom". However, this should not result in the rejection of ML strategies in ecology. The consequences of not being able to satisfy the i.i.d. assumption to maintain ecologically relevant analyses are three-fold (derived from Chapter III): a) careful selection is required with regard to where and how samples are taken and processed when designing an experiment. This meta data must be recorded and considered during analysis, as exemplified by the calculation of an accuracy threshold in Chapter I; b) a large number of samples is required to avoid incidentally useful variables and identify confounding ones; and c) the model must be sufficiently interpretable, so that e.g. the variable importance can be analyzed for spurious correlations. Undetected confounding variables can render a model useless to the prediction of unseen data or worse, can provide seemingly useful predictions which lead to wrong conclusions (Lapuschkin et al., 2019; Rudin, 2019). As demonstrated in Chapter III, ML can be applied to environmental data if at least the accompanying sample data and variable importance is provided for assessment.

A step towards comprehending variable importance is to analyze variable values dependent on response class and contextual data, thereby potentially identifying their relevance to the model, i.e. understanding the model's decision. It is also a recommendable approach to distinguish indicative from spurious correlations. In Chapter I, the relative abundance of important variables over time in each microcosm was examined (Figure 1.4) to identify an effect of glyphosate presence. Likewise, in Chapter III the relative abundance per area and similarity of community composition was investigated (Supplementary Material 3.8). To provide further inside into the decision-making, various algorithms allow for the extraction

of the decision boundaries for a given model, often mapped to a two-dimensional space (see for example Menze et al., 2011).

In Chapter I, a stochastic subsetting approach was chosen to analyze the variable importance in ANN models. More than 2000 sets of 20 or 30 variables were randomly selected to train the model, and the resulting accuracy was monitored. Subsetting and permutation approaches do not scale well with regard to computational effort, especially if more than one variable should be permuted at the same time to identify combinatory effects. Therefore, subsetting was feasible for the data set in Chapter I which contained (at maximum) 687 taxa; it is not recommended to be applied to a data set of e.g. 80,000 amplicon sequence variants (ASVs), the total unfiltered data set of Chapter III. To make a model more interpretable, it is often recommended to reduce the number of variables. Depending on the method, it may be required to have fewer variables than samples ($n > p$) to prevent overfitting. In fact, in both Chapter I and Chapter III, variable selection (10 genera and 25 genera, respectively) achieved more accurate predictions than by using the full community data. Breiman (2001b) has claimed that overfitting does not occur in RF due to bagging and promoted the use of more variables. The author further on cited the interesting concept of the “Rashomon effect”; named after the Japanese movie Rashomon, where a murder is described in four contradictory ways by four witnesses. Breiman transferred the concept with regard to neural nets and decision trees and described it as the “multiplicity of good models”. In short, it refers to the phenomenon that several models consisting of very few important variables (selected from the same data set) may predict similarly accurate. Yet the resulting variable importance leads to different conclusions, thus making the inference instable. Bagging was invented to avoid such behavior. In Chapter I, I analyzed the accuracy achieved by different variable subsets and identified taxa which had to be included for a stable prediction of glyphosate presence, including *Parvibaculum* and *Gallaecimonas* (Figure 1.3, Filtered data set).

Variable importance is a precondition to determine indicative microbial fingerprints

To ensure that a microbial fingerprint, as determined by variable importance, is actually indicative for a contaminant, they have to be causally related (further discussed in Chapter III). However, using 16S rRNA gene data alone, it is unlikely to identify causal relationships. Yet, due to the simple experimental design and high resolution sampling used in Chapter I, the distinct increase in abundance by several taxa following the glyphosate pulse was a reliable hint. It became evidence when in Chapter II glyphosate degradation to AMPA was analytically measured. Important variables included foremost *Parvibaculum* (Figure 1.4), a taxon by itself sufficient for classification. Furthermore, *Gallaecimonas* (Figure 2.5, Free-living) provided valuable information. Both increased in abundance after the addition of

glyphosate, but *Gallaecimonas* was also abundant at untreated time points, whereas *Parvibaculum* was only present at higher concentrations in the treated microcosm after addition of glyphosate. In contrast, *Massilia* was likely considered important for another reason: it provided information to separate the microcosms from each other – thereby representing the fictional "microcosm variable" imagined above as example of a confounding variable – regardless of glyphosate treatment, and should therefore not be considered as part of an indicative fingerprint (Figure 1.4). In line with this reasoning, prediction accuracy dropped to a maximum of exactly 90 % (the microcosm separation threshold) when *Massilia* was the only training variable (Figure 1.5).

In conclusion, this variable selection does not likely apply to Baltic Sea communities, as the laboratory microcosm conditions were different in several ways. Such were the glyphosate concentration, temperature, nutrient availability and dispersal from the environmental conditions. This is obviously common for a laboratory experiment, but highlights well the issue of transferability and purview of an indicative fingerprint with regard to different regions, habitats or ecological niches. This problematic nature will be further highlighted by an example from Chapter III, but is of general importance. The dumpsite sediment samples analyzed therein were diverse in comparison to the microcosm system. They included varying grain size distributions across kilometers in the Baltic Sea and changing redox regimes in surface samples to multicorer samples 22 cm deep (Supplementary Material 3.1). The arising interesting question is: how many indicative fingerprints do we expect? For example, muddy sediment at 20 cm depth could comprise an indicative microbial fingerprint that differs significantly from those of a coarse, oxygenated surface sediment sample. I consider this primarily as a problem of goal definition: is a model predicting for both depths (i.e. 0 cm and 20 cm) desired, or are the fingerprints, specific to each depth, requested? ML models could include both cases to map community composition to the presence of TNT (unless a given taxon behaves conflicting across habitats), however, the important variables were determined with regard to both depths, resulting in intermingled fingerprints (further discussed in Chapter III). The advantage in training separate models for each depth would be the elimination of confounding variables, in particular the grain size distribution and the redox potential in this example. It is secondly a problem of feasibility, as in this study, there were not sufficient samples available to train a model for each habitat; and it is not trivial to determine habitat borders based on DNA-derived data (the process of niche separation in freshwater systems was reviewed by Pernthaler, 2017). To address this issue and account for potentially multiple distributions and confounding variables within the community compositions, six different training/test data splits were analyzed to identify a generalizable fingerprint (Figure 3.3). Topçuoğlu et al. (2020) also recommended this strategy for

microbiological data sets, using up to 100 splits. They however did not mention the resulting information leakage. Leakage occurs when training data includes theoretically unavailable information (Kaufman et al., 2011). Applying the proposed approach, holdout test samples in one split will inevitably be the training samples in another data split, thereby informing the choice of hyperparameters. However, it is of reduced importance in comparison to having a single training/test split. For a single split, the information leaks only from one test set to one training set. Using multiple splits, the flow of information between training and test cannot be back traced. The leaked information cannot be identified, but still informs the hyperparameter selection e.g. if the in-average best hyperparameters are chosen. Thus, no truly detached holdout set was available, as discussed in Chapter III. In the Kolberger Heide data, the data splits indeed unveiled multiple distributions and identified 25 genera of importance to all data splits, which could be involved in biodegradation (Figure 3.5). The presence of TNT metabolites indicated biologically mediated transformation processes (Bernstein and Ronen, 2011). However, due to the low concentration of TNT and as no times series data was on-hand, it was not clear if TNT degradation took place *in situ* or whether it occurred e.g. in the water column and the metabolites then adsorbed to the sediment (Brannon et al., 2005). Next to potential TNT-degrading bacteria, the important variables likely also included taxa such as *Cobetia* and *Colwelliaceae* affected by confounders such as the grain size distribution (Supplementary Material 3.8, C, G). Additionally, the variables included taxa like the clade TA06, which was only present in 12 samples and therefore rather coincidentally useful in separating a small subset of samples (Supplementary Material 3.8, Y). Several of the examined 25 important genera are likely impacted by TNT, but the number of samples did not yet enable the determination of a truly indicative fingerprint.

The results from the laboratory and the environmental ML analysis raised an interesting question involving the specificity of indicative microbial fingerprints. For example, did the models in Chapter III learn to predict TNT exactly or does such a model have the ability to predict several nitroaromatic compounds, as they are similar in structure and likely cause similar reactions for a bacterial taxon (Spain, 1995)? Classification of also the metabolites was initially attempted, but unfortunately the classes were either so imbalanced or distributed corresponding to confounding variables that a prediction was not meaningful; e.g. 2- and 4-ADNT were present in 127 and 133 of the 150 selected samples. The glyphosate results resembled this theme: several *phn* operons, where the encoded enzymes potentially degrade glyphosate to sarcosine (Sviridov et al., 2012), were detected in the metagenomes. However, *phn* operons contain genetic information to degrade a variety of phosphonates (White and Metcalf, 2004), a common class of phosphorus-

containing compounds in the environment (Martinez et al., 2010). Yet another microcosm disturbed by a phosphonate similar to glyphosate would have enabled to distinguish phosphonate-shared from glyphosate-specific reactions.

Shotgun sequencing requires careful experimental design to support analyses

The ten metagenomes analyzed in Chapter II ensured that the genetic functions for glyphosate degradation in form of *gox* genes, *phn* operons and *thiO* genes was factual. It was furthermore possible to connect the abundance and phylogeny of several glyphosate degradation gene instances with taxa identified as important (Figure 2.6–8). Ultimately, it was made possible to synchronize abundance shifts for taxonomically related 16S rRNA genes and degradation-related genes to important variables detected by either ML or the statistical model applied by R package DESeq2. However, it was necessary to measure parameters such as glyphosate, AMPA, sarcosine, dissolved inorganic phosphate and glycine to estimate the utilized degradation pathway. For example, *phn* operons, theoretically providing the capability to degrade glyphosate and extract P, were detected. Yet the amount of sarcosine – indistinguishable from *L*-alanine in the applied HPLC-MS/MS method – did not change after the addition of glyphosate, additionally, sarcosine was detected in both microcosms. These results indicated, that glyphosate was only degraded via the AMPA pathway and the measured substance was probably *L*-alanine, as part of the medium (Chapter II). Five sediments samples from Kolberger Heide were subjected to metagenomic analysis as well. Martin processed and analyzed the metagenomes as part of her Bachelor Thesis (Martin, 2020), focusing on MC-degradation related, mostly nitroreductase-encoding, genes. The analyses were considered challenging, due to the complexity of the environmental sediments sampled from five distinct locations (Chapter III). Nevertheless, it was assumed that MC had a significant effect on the community composition at detonation site Mo7 (TNT: 1,600 pmol·g^{-1} wet sediment; summed MC: 5,700 pmol·g^{-1}), resulting in a difference between Mo7 and the 4 other investigated sites (0–2 pmol·g^{-1}, summed MC 1–60 pmol·g^{-1}). The author's findings showed that the abundance and diversity of such genes was indistinctive across samples. It was concluded that the impact of MC on degradation-related genes in the metagenomes was not sufficient to surpass the different main drivers of the community such as grain size distribution (Chapter III).

In summary, metagenomic data was of limited use in describing the environmental state with regard to the specific contaminant or to estimate biodegradation/transformation processes, despite the amount of data they provided. These findings demonstrate that the usability of metagenomics strongly depends on the research questions and experimental conditions, e.g. in anoxic sediments metagenomes may include conserved DNA originating

from dead cells (Thureborn et al., 2016). Nevertheless, it was possible to generate large contigs and nearly complete bins using concoct (Alneberg et al., 2014) from the microcosms. However, the responsive taxa identified via amplicon sequencing were mostly low in abundance, thus, their genomes are rarely covered by shotgun sequencing (Ni et al., 2013). This was also true with regard to the question of MC degradation/transformation in Chapter III. The genes (e.g. encoding for the nitroreductases, catalyzing the reduction of nitro-moieties as the first step in TNT reduction), are virtually ubiquitous, but do not necessarily enable the organism to degrade specifically TNT (Roldán et al., 2008).

Metatranscriptomic analyses were initially planned as part of the glyphosate degradation analysis in Chapter II, but were not conducted due to limitations in time, workforce and funding. The sediments from Chapter III were originally sampled for MC analysis by divers and hence have not been conserved appropriately for total RNA sequencing. My conclusion is to include a few, carefully selected metagenomes as mapping backbone, for bin assembly and to assess the functional potential of a habitat. In the case of Baltic Sea pelagic shotgun sequences, the assembled and functionally-annotated Baltic Sea Reference Metagenome (BARM) provides the required mapping backbone (Alneberg et al., 2018) and reduces the computational demands. However, the focus should be on metatranscriptomic analyses as a measure of functional activity. Speculating based on the findings of Chapter II, metatranscriptomic data could have indicated the activity of the glyphosate degradation pathway, among other reactions to the addition of glyphosate. With regard to Chapter III, and assuming the required sample conservation and sequencing depth was met, it could have been investigated whether MC degradation-related genes were transcribed at all. This could have helped to clarify whether TNT metabolites were formed in the sediment or originate from the water column (Chapter III) and which bacteria were involved in degradation.

Disturbed communities displayed resistance and resilience

The prediction of contamination relies on a composition-altering impact by the contaminant towards the microbial community. The general capability of glyphosate (e.g. Stachowski-Haberkorn et al., 2008) and TNT (e.g. Esteve-Núñez et al., 2001) to do so has been described. In the experimental realizations of this thesis, glyphosate could be classified as a pulse disturbance and the TNT contamination as press disturbance (Shade et al., 2012). The glyphosate addition caused an increase in the abundance of specific taxa outside of their normal operating range (Orwin and Wardle, 2004). Most water column taxa returned to original abundance levels or re-aligned with the prevailing succession before the glyphosate pulse, such as the increasing dominance of α-*Proteobacteria* (Figure 2.2). Put in other words, succession led to a changing community composition at all times, but the

change was temporarily dominated by the reaction to glyphosate. Similar behavior was identified in the metagenomic data and for the cell counts. It can be concluded that the high temporal sampling resolution allowed the observation of microbial communities being sensitive to higher and recovering at lower concentrations of glyphosate. In the biofilms, fewer taxa were identified as responsive, but their reactions were prolonged, partially until the end of the experiment (Table 2.1). In light of the fact that glyphosate was applied only once and rather served as a nutrient source than a toxic or otherwise negative stressor, this type of community reaction seems plausible. It should be noted, that possible sorption and desorption of glyphosate on surfaces was investigated in Chapter II and was found to be negligible if the microcosm was inoculated days prior to the addition of glyphosate (Supplementary Material 2.1).

Resilience could only be speculated about in Chapter III, as the samples stem from single points in time. Further information about past contaminations was drawn from the presence of TNT metabolites. Resilience as a common ecological phenomenon (Baho et al., 2012; Shade et al., 2012; Meredith et al., 2018) was considered as one explanation for false positive classifications (i.e. samples without TNT, but classified as TNT present). In case these false positives contained metabolites, they were more likely to be misclassified, although without statistical significance due to small sample size (Figure 3.6). Two interpretations (or a combination of both) were conceived: a) that the metabolites indicate that TNT was once there and had an impact on the community. Subsequently TNT was degraded or dissipated, but the resilient community still represents the impact and leads to a false positive prediction; or b) the impact of TNT compared to those of its metabolites were very similar and were therefore misinterpreted as TNT contamination (Chapter III). Resilient microbial communities can act as event recorders and offer a great potential to identify disturbances which have already passed, e.g. demonstrated by Smith et al. (2015). They detected former hydrocarbon contaminations using ML with microbial community compositions, although the hydrocarbon levels had already returned to background levels. More research is required to examine long-term reactions of community compositions to disturbances (Lindh and Pinhassi, 2018), which include alternative stable states (Allison and Martiny, 2008), or cycling through multiple states according to e.g. seasons (Lindh et al., 2015).

Random Forest is preferable to Artificial Neural Networks

Throughout my thesis I compared RF and ANN for their suitability to analyze community composition data. ANNs were included in Chapter I to reveal abstract interactions using deep learning. Yet, comparing the predictions using microbial community composition in Chapter I and III, RF proved to predict contaminations virtually always more accurately than

ANN (Figure , modified from Figure 3.1, amended by ANN scores). It was assumed that the prediction of low level TNT would be particularly hard to evaluate. Hence, both methods were applied to determine the achievable prediction rates and investigate whether misclassifications occurred algorithm- or sample-specific (Figure 3.4, B). The higher variance of the accuracy of ANN predictions compared to RF is partially attributable to RF being an ensemble classifier (discussed in Chapter III).

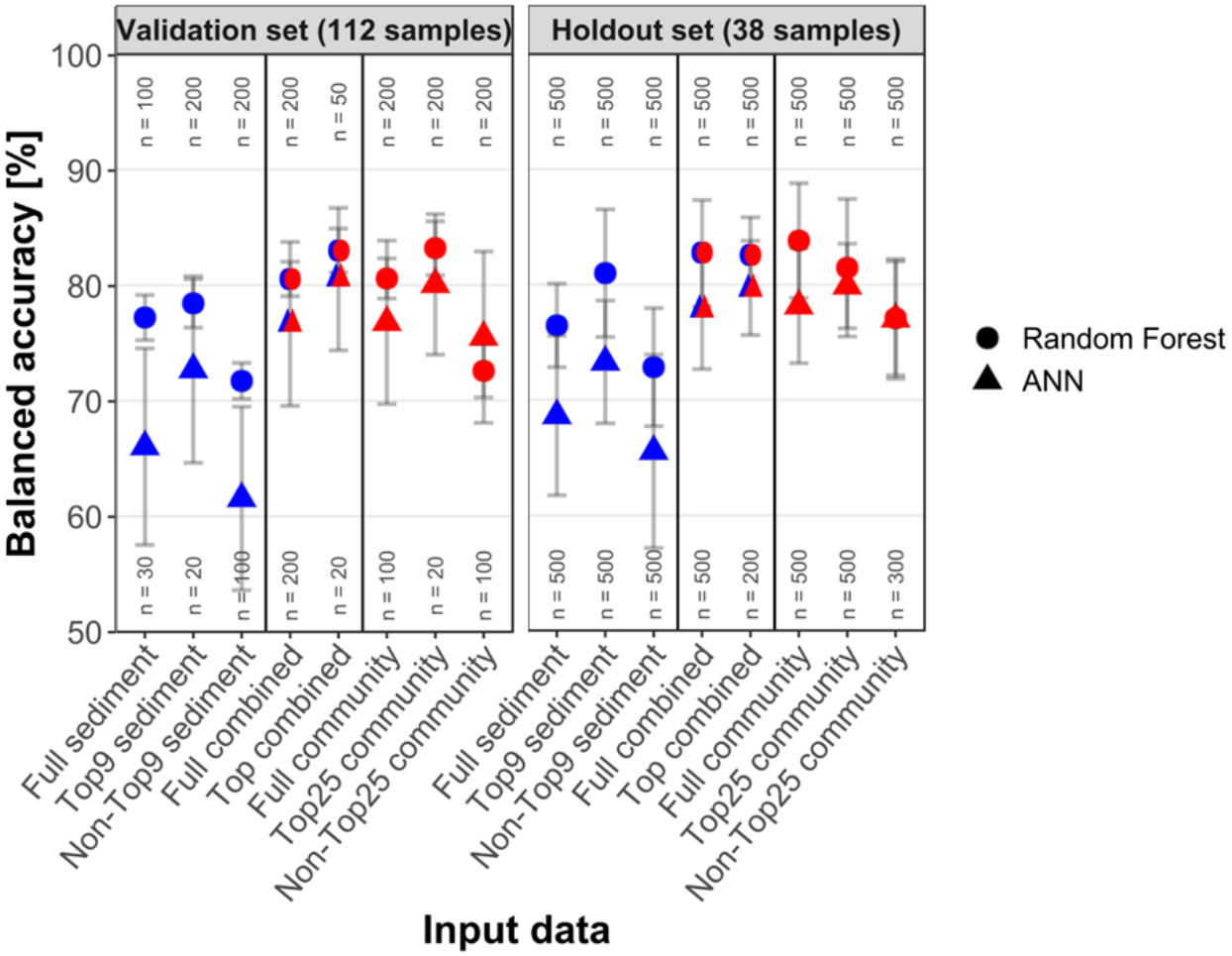


Figure D: Correct TNT classifications (Chapter III) per input data in the validation and hold out test set for RF (dot) and ANN (triangle). Red indicates community data, blue symbolizes sediment data and red-blue combined variables. Of each data type, either all variables were utilized by the model ("Full"), or only the best variables based on variable importance ("Top") or all variables except Top ("Non-Top"). Classification performance is displayed as mean and standard deviation of balanced accuracy, the classification results of the six different data set splits were averaged. The validation values are out-of-bag estimates. n indicates the number of RF (top) and ANN (bottom) models calculated.

Significant effort with regard to the application of RF and ANN went into optimization, which includes hyperparameter tuning as well as manipulating and selecting the input data. Optimization was found to increase the mean accuracy by more than 10 % for both methods when classifying glyphosate or TNT presence (Chapter I, III). Furthermore, the results in Chapter III showed that an optimized model displayed reduced variance (Supplementary Material 3.4) and confirmed that increasing the default *mtry* value – describing the number

of variables for each node split in RF, the default for classification is the square root of the number of all independent variables – is important for sparse tables with few relevant variables, because it heightens the chance detecting those relevant ones (Hastie et al., 2009).

In the literature, it was not always clear whether studies investigated variable selection or applied an approach similar to an abundance threshold as performed in Chapter I and III. Abundance cutoff values mentioned by e.g. Smith et al. (2015) or Glasl et al. (2019) seem to stem from the bioinformatics analysis. Moitinho-Silva et al. (2017) did not report a specific threshold, but used composition data on phylum, class and operational taxonomic unit (OTU) rank to predict sponges as of high or low microbial abundance. In general, higher taxonomic ranks remove the lower intra-rank variation, causing a potential loss of information. In contrast, comparing predictions based on various ranks includes an additional dimension of information. In Chapter II, *Pseudomonas* OTUs were detected, which distinctively responded to the glyphosate pulse (Table 2.1), whereas the genus *Pseudomonas* was not identified as an important variable in Chapter I. Therefore, it could be beneficial to reduce the number of variables by selecting important variables on a lower rank, instead of agglomerating the lineages. TNT predictions however still worked on class rank with 78.8 % mean balanced accuracy (Figure 3.2). Depending on the analysis, higher taxonomic ranks may be required, e.g. to predict global patterns of port microbial communities (Ghannam et al., 2020) or ballast water discharge (Gerhard and Gunsch, 2019).

The hyperparameter tuning and input data selection process is specific to the individual experiment. In Chapter III, combinations of relative abundance thresholds and hyperparameters were investigated at the same time, as both depend on each other. The hyperparameter tuning was performed for all available taxonomic ranks. A Cartesian grid search describes the process to test all combinations of values for e.g. hyperparameter 1 and hyperparameter 2. For ANN optimization, this becomes tedious and potentially unfeasible, as there are too many hyperparameters to investigate at the same time (also called combinatorial explosion). Yet e.g. the number of nodes obviously has to be investigated dependent on the size of the data set. Therefore, the number of nodes (values attempted ranged from 4 to > 1000) in both hidden layers depending on input data sets were initially determined. Fifty nodes in the first and 40 nodes in the second hidden layer showed the best results. Drop out regularization from 10 % up to 50 % of the nodes to prevent overfitting did not improve the prediction accuracy and the Adaptive Moment Estimation optimizer function outperformed Root Mean Square Propagation slightly. Two hidden layers were deemed sufficient, as they are capable of approximating virtually every

non-linear function (Schmidhuber, 2015). More advanced deep learning functions and other architectures were not explored, as basic ANNs have been reported as being the most successful ML approach for omics data sets (Yu et al., 2019). In conclusion, ANN may have achieved better predictions in Chapter I and III using other hyperparameters, however, the effort (after pre-conducted optimization) to determine such hyperparameter settings outweighs the benefits of a slightly increased accuracy.

For the ANN models deployed in Chapter I no holdout test set was set aside due to the small sample size. Instead, "leave one out cross validation" was used to calculate the prediction error: all samples were involved in training the model except one, which is "left out", and has to be classified by the trained model. The repeated cross validation combined with multiple train/test data splits was applied in Chapter III, which is better suited to address the level of generalization (Topçuoğlu et al., 2020). A holdout test set should always be included in the analysis, even if it only resembles 10 % of the samples. Such a test set is also required for precise comparison between ML methods.

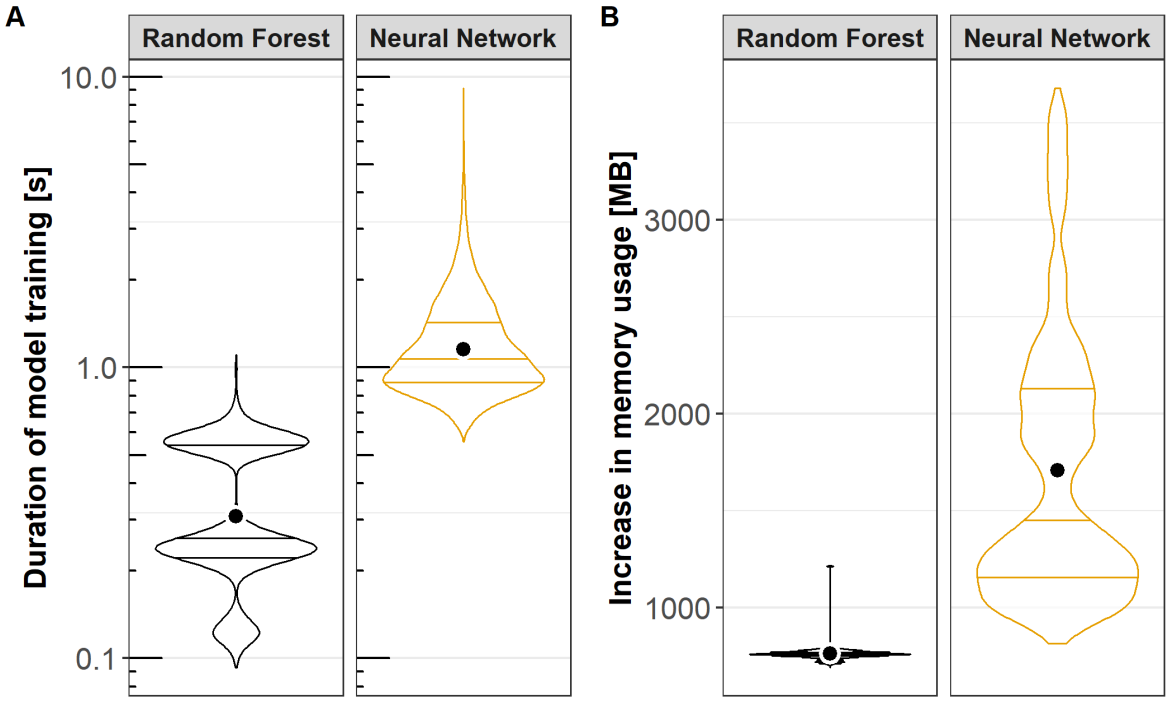


Figure E: Violin plots displaying A) the time to train a model (log transformed) and B) the increase in used memory for both algorithms. The information was logged during the optimization phase and is based on the models whose classification is shown in Figure A. The mean is represented as black dot.

Below I want to provide a short comparison between RF and ANN with regard to aspects which are not a major part of thesis. The comparison is mostly based on the extensive ML analyses for Chapter III. Including logged information from involved hyperparameters and input data selections analyzed in Figure A, it was most notable, that RF was magnitudes faster and demanded less memory (Figure F). To put the difference into perspective, the reported training time was required for a single Random Forest (0.3 s in average) or ANN

model (1 s in average). However, RF is an ensemble learner, hence consisting of (in this case) 1,000–10,000 decision trees, resulting in the same number of predictions. RF only publishes the majority prediction, e.g. if 400 trees classify a sample as “TNT present” and 600 as “TNT absent”, the output is simply “TNT absent”. However, in the background many more weak models have been trained. The training speed of ANN could be significantly improved by utilizing graphics processing units (Schmidhuber, 2015). The memory usage was affordable for RF, whereas the ANN could crash during training by reaching the memory limitation. This is one reason why training models in small sample batches has been invented (Chollet and Allaire, 2018). The speed and memory usage depend on the implementation of the algorithms. To conduct ANN analysis in Chapter III, a combination of R (R Core Team, 2017), R package Keras (Allaire and Chollet, 2020), R package TensorFlow (Allaire and Tang, 2020) and the actual TensorFlow software were used. In contrast, Random Forest required mainly R, C++ and R package ranger (Wright and Ziegler, 2017) to perform analyses. However, this comparison is skewed as TensorFlow and Keras programs provide access to virtually all deep learning variations, plus visualizations and very advanced interfaces for analysis. With regard to input data preparation, no transformation or normalization was required for RF analyses. It worked off-the-shelf with both continuous and categorical data. ANN input data transformation accelerates the convergence of the model during training significantly (Chollet and Allaire, 2018). Categorical data has to be one-hot-encoded as dummy variables. To optimize predictions, RF only possesses two relevant hyperparameters compared to an unknown greater number for ANN. Additionally, RF provides a proximity matrix. It is a distance measure, based on similarly classified samples, generated during supervised and unsupervised tasks. The matrix can be used to perform e.g. PCA (Chapter I, more detailed explained in III). This combination was the method of choice for community and sediment data ordination, including the fitting of environmental parameters with community compositions (Chapter III). It can be concluded that RF comprises a class of machine learning that is well suited to predict (classification/regression) or detect similarities within microbial community composition and environmental data, in a fast and efficient manner.

As the importance of transparent models was stated, linear ML algorithms and other tree-based methods such as gradient boosting should be considered first, even if deep learning would provide slightly better predictions (Topçuoğlu et al., 2020). It has been reported that for various typical deep-learning use cases, interpretable models achieved similar prediction scores (Rudin, 2019).

Phyloseq2ML: an R package which facilitates machine learning with microbial communities

Since the most time-consuming step during ML analyses involves the selection and preparation of data, as well as identifying the optimal hyperparameter settings, there is a large quantity of software, such as caret for R (https://topepo.github.io/caret/index.html) and scikit-learn for Python (https://scikit-learn.org/stable/), as well as programs such as DeepLearning4Java (https://deeplearning4j.org/) and platform independent high-level interfaces such as Keras (https://keras.io/) to enable ML. My motivation was to write an R package specific to the characteristics and requirements of 16S rRNA (gene) amplicon data. The purpose of the provided functions by phyloseq2ML is to connect two analysis environments with each other (https://github.com/RJ333/phyloseq2ML). On the input or source side, the frequently used R package phyloseq (McMurdie and Holmes, 2013) contains a large toolkit to analyze and manipulate microbial community data. It allows for the linking up of up to five different data sets with each other; the abundance per taxa in the community table, the taxonomic annotation, context data such as sampling information or measured environmental parameters, the reference sequences representing a taxon and a phylogenetic tree. The removal of a taxon or sample in one of these tables prompts phyloseq to update the linked data sets. Additionally, the data sets are stored in a defined format, regardless of the bioinformatic pipeline used to process the raw sequences. At this point, phyloseq2ML is designed to provide a connection to the second environment: the machine learning. More specifically, phyloseq2ML currently supports the Keras and TensorFlow interfaces for deep learning and the R package ranger for Random Forest. Phyloseq2ML functions enable the extraction, manipulation, combination and arrangement of data sets conveniently from phyloseq-class objects so they meet the formatting requirements of such machine learning implementations. It furthermore calculates a variety of performance metrics to evaluate the predictions. It was developed and extensively used during the work on Chapter III and hopefully enables other researchers, who rather want to focus on the interpretation of their data to use machine learning analysis.

Applying sequencing data and machine learning analysis to monitoring

Microbial communities may inform additionally to parameter prediction

Bacteria are currently not involved in the Baltic Sea environmental monitoring efforts, despite being essential to virtually all biogeochemical processes (Backer et al., 2010). The knowledge about microbial community compositions, their functions (reviewed with regard metagenomic based monitoring in the Baltic Sea by Ininbergs et al., 2015) and expression profiles provide insights into essential processes such as nutrient and element cycling, indeed bacteria comprise the foundations of the food web.

Table A: Costs and workload for the instrumental analysis of elements and compounds compared with next generation sequencing.

Analyte	Net cost per sample (€)	Workload in days (100 samples)	Method/ Instrument	Source for cost and workload estimation
Mercury[a]	2	2	DMA	Ines Scherff, personal communication
Total inorganic carbon[a]	2	2–3	Elemental analyzer	Ines Scherff, personal communication
Total carbon, nitrogen, sulfur[a]	2	3–4	Elemental analyzer	Ines Scherff, personal communication
Element composition, water, 500 samples[b,c]	8	10	ICP-OES	Anne Köhler, personal communication
Element composition, sediment, HCl-extracted, 500 samples[c]	9	20	ICP-OES	Anne Köhler, personal communication
Glyphosate and AMPA, water[d]	28	45	HPLC-MS/MS	Marisa A. Wirth, personal communication
Element composition, water, 100 samples[b]	33	2	ICP-OES	Anne Köhler, personal communication
Munition compounds, water and sediment, 100 samples[e]	35	18	UHPLC-ESI-MS	Aaron J. Beck, personal communication
Element composition, sediment, HCl-extracted, 100 samples[c]	36	4	ICP-OES	Anne Köhler, personal communication
16S rRNA (gene) amplicon sequencing[f]	80	10–14[g]	Illumina Sequenzing	Bonus Projekt BLUEPRINT BCC Report Deliverable 6_2
Metagenomic sequencing[f]	300	10–14 [h]	Illumina Sequenzing	Bonus Projekt BLUEPRINT BCC Report Deliverable 6_2
Metatranscriptomic sequencing[f]	350	10–14 [h]	Illumina Sequenzing	Bonus Projekt BLUEPRINT BCC Report Deliverable 6_2

a) does not include drying and homogenization of sediments
b) does not include drying and homogenization of sediments, used ICP-MS for Chapter III instead
c) workload estimate based on 100 samples
d) more sensitive method compared to Chapter II, appropriate for Baltic Sea samples, does not include sarcosine and glycine
f) Includes filtration and size fractionation, does not involve bioinformatics, Sequencing offers from LGC Genomics, Berlin, 2018; and estimated for total DNA and RNA
g) workload based on own work
h) workload estimated based on own work on 16S rRNA gene amplicon sequencing

However, the specific use of microbial community compositions to predict contaminants and environmental conditions using ML raises the question of why not to directly measure the variable. In many cases, instrumental analytics are the best approach with regard to costs and work efficiency, compared to the prediction of temperature, pH or other physicochemical parameters by microbial communities (Glasl et al., 2019; Alneberg et al., 2020). To assess the actual magnitude of difference, the (estimated) costs were compiled for some of the parameters included in Chapter I–III (Table A). The costs due not include personnel, therefore, the required working days for 100 samples (plus calibration and reference standards) are given. Omitted were variables such as temperature, conductivity, chlorophyll *a* or pH, which can be measured continuously by online sensors (e.g. used throughout public transportation such as ferries equipped with the FerryBox to monitor algal blooms in the Baltic Sea; Rantajärvi et al., 2003). The cheapest methods which involve laboratory work include sum parameters such as total nitrogen or the individual mercury determination. The costs and the workload to measure the elemental composition via ICP-OES depend on whether an extraction step has to be performed. Regardless, this method measures a double-digit number of elements at once. The most expensive and workload-intensive methods involved the analysis of glyphosate and AMPA (28 €, 45 days for 100 samples) and the suite of MC described in Chapter III (35 €, 18 days for 100 samples), returning 2 and about 10 variables, respectively. The methods are sufficiently sensitive to determine their respective analytes even strongly diluted in Baltic Sea samples (Gledhill et al., 2019; Wirth et al., 2021). Costs for NGS have been constantly decreasing, still, amplicon and shotgun sequencing rank as the most expensive methods. Costs for amplicon sequencing, however, are of the same order as elemental compositions, MC or glyphosate analytics. Shotgun sequencing is the most expensive method, but generates a great amount of primer-independent sequence data valuable for monitoring (Ininbergs et al., 2015). As more institutions such as the National Genomics Infrastructure at the Science for Life Laboratory in Stockholm, Sweden, offer sequencing services, the costs decrease further (Anders F Andersson, personal communication). The other important factor is the hands-on and analysis time. Sequencing (including sample filtration, nucleic acid extraction and library preparation) requires fewer working days compared to the MC or glyphosate analyses. In research, the subsequent bioinformatic analysis is a major time consuming step. For monitoring purposes specific indicators are targeted. This enables streamlined and automated processing and analysis, including the ML prediction. In the long term, *in situ* library preparation and sequencing (e.g. at a MARNET monitoring station), data upload and fully automated processing and analysis will be possible.

Still, it seems unnecessarily complicated to integrate NGS of microbes solely to predict another parameter by it. Yet, even more and significant advantages originate from information uniquely accessible via sequenced microbial communities analyzed by ML:

a) The same community data points can be used to predict several parameters, such as glyphosate, AMPA and TNT as well as mercury or nitrate and uranium (He et al., 2018) or water depth and salinity (Alneberg et al., 2020), rendering the approach more resource-efficient.

b) Communities may exhibit resilience towards a disturbance, substance or contaminant, which means, that the effect of such is still represented by the community composition, although the disturbance itself is over (Shade et al., 2012; Smith et al., 2015). Such information is evasive to direct instrumental measurements.

c) Disturbances of contaminants can only be identified if they actually impact the community. In return, a distinguishable community composition supports the determination of an impact threshold on microbial ecology (effective concentration) in real world settings. To display the sensitivity of this approach, Wood (2019) has shown that bacteria decide precisely when building certain enzymes is worthwhile; responses to antibiotics were initiated when a given concentration was surpassed, but long before inhibitory concentrations were reached. Similarly, glyphosate degradation was likely not worthwhile anymore when concentration fell below 1 µM, as demonstrated by the cease in reaction towards glyphosate (Chapter II).

d) Each community composition, together with a set of meta data, is, in itself, a fully valid sample set. The data set becomes even more valuable if not only the composition, but the metagenome or metatranscriptome were also sequenced. It needs to be made publicly available to enable in-depth data mining/analyzes, microbiological and ecological research. A central data base, comparable to the BalticMicrobeDB (Alneberg et al., 2018) should be provided for organized and accessible data storage, and methods should be standardized to reduce the chance of confounding batch effects, such as described by Soneson et al. (2014) for publicly available gene expression data sets.

These points illustrate the significant improvement of environmental monitoring efforts that is possible by the inclusion of community composition data and ML analyses. It should be stressed that the time when e.g. glyphosate is solely determined by ML prediction has not yet arrived; the models have to be trained in a supervised manner, and therefore the parameters still have to be analytically measured. Ongoing measurements to ensure and calibrate the prediction quality are indispensable. Yet now is a perfect time to start collecting

samples for this approach, as all preconditions in terms of equipment, instrumentation, software and knowledge are available.

Microbial monitoring requires specific collection and storage of samples

Regular monitoring already provides a framework optimal for data analysis, including a set of indicators to assess and a set of sample data to record, a system of monitoring stations and routes for cruises to cover the ecologically- or socially-relevant areas of the Baltic Sea as well as an established infrastructure and the required logistics to store, process and analyze the acquired data in a standardized methodology. To integrate NGS into monitoring, protocols for molecular biology-suited sampling need to be implemented. DNA is more robust; in contrast expression levels captured by metatranscriptomes possess turn-over time in the seconds' range. Therefore, DNA samples are advised to be, and RNA samples *must* be conserved *in situ* (Charvet et al., 2019). The details of integrating NGS have been explored in the Bonus Project BLUEPRINT, which I want to refer interested readers to.

One of the advantages of using ML with data generated by NGS comes from the number of samples and the breadth of potential response variables collected. However, supervised learning demands training data including the response variable, which could be a physicochemical parameter or as yet unknown contaminants. In the future, the Baltic Sea may be affected by contaminants we currently do not know much about, the so-called emerging contaminants (de Wit et al., 2020). It should be avoided starting with zero training samples when a new response variable is added on to the list of monitored substances. On that account, it should be investigated if a certain amount of retained samples could be e.g. deep frozen (–80° C) or otherwise preserved for complicated and sensitive future analyses. This furthermore would allow the use of the same community composition for all response variables, enabling comparisons between them. Structurally related substances should be included to identify the specificity of a fingerprint, e.g. whether a TNT-trained model will also report 2,4-DANT-impacted communities. Finally, long-term monitoring could fill a gap of statistically powerful investigations on resilience in natural ecosystems (Lindh and Pinhassi, 2018). The selection of appropriate sampling locations with the desired ecological relevance or statistical independence (i.e. determining the habitat borders for individual indicative fingerprints) could be based on the environmental parameters, for example above and below the redoxcline. In return, an assessment of the model's prediction robustness can inform upcoming sampling campaigns about where additional training data is required (Chapter III).

Selecting appropriate algorithms for integration with environmental monitoring

The ML step in monitoring is advised to focus less on most accurate predictions and more on the interpretability of the model (Topçuoğlu et al., 2020). One reason is that the i.i.d. assumption will likely be violated and the effects of such have to be considered (Økland, 2007). As a consequence, a meaningful variable importance measure needs to be reported. This measure is required to validate that model-relevant variables are actually monitoring-relevant (like an indicative fingerprint), too. Breiman (2001b), on the contrary, states the Occam Dilemma: “Accuracy generally requires more complex prediction methods. Simple and interpretable functions do not make the most accurate predictors. Using complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy first, then try to understand why.” He furthermore mentioned: “The goal is not interpretability, but accurate information” and that asking for interpretability is misled. However, Breiman (2001b) seemed to already consider Random Forest as uninterpretable, which at least readily reports the variable importance, and additionally may not have had ecological data sets in mind when making these statements. Rudin (2019) requests the exact opposite and strongly promotes the use of interpretable models. This conflict need not be of great concern, as it has become the standard to compare various ML methods (Topçuoğlu et al., 2020). Thus, an assessment on how much more accurate a more complex model can become and whether it is worth the loss of interpretability should be, and typically is, conducted. However, the community data collected by monitoring cruises likely does not require complex deep learning approaches (Yu et al., 2019), such as reviewed by Cao et al. (2020). Ultimately, the training speed of the model should not be of primary concern; after the optimization phase the model will only be re-trained with new incoming monitoring samples. I would gladly witness these assessments of mine refuted with the first 100,000 samples analyzed by a month-long trained deep learning model, uncovering microbial and ecological relationships of unexpected complexity.

Conclusion and outlook

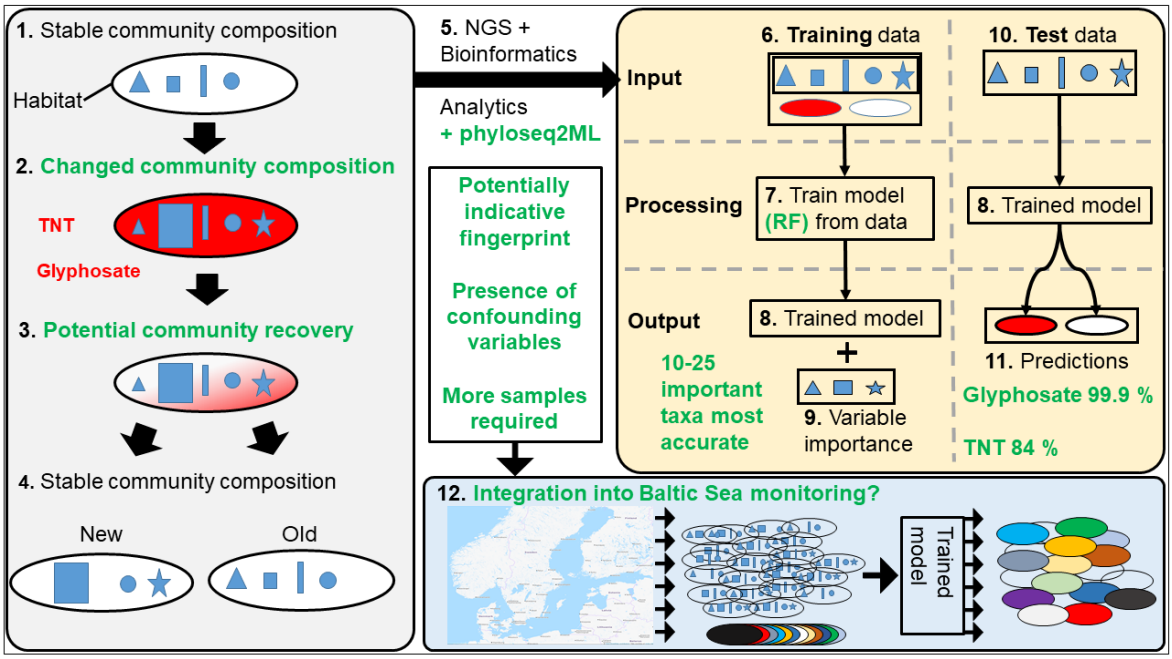


Figure G: The findings of this thesis are colored in green. Integration with 12. environmental monitoring would enable the required increase in sample size of community compositions and potential response variables, allowing for truly indicative fingerprints.

The contribution of this thesis to the scientific field, with regard to the research questions presented at the beginning (Figure D) can be described as follows (Figure G): It was found that microbial community compositions were altered to a certain extent by the presence of TNT and glyphosate. Machine learning with Baltic Sea community data worked to predict contaminations in the laboratory (up to 99.9 % accuracy) and in the environment (up to 84 % balanced accuracy). It can be concluded that the presence of TNT and glyphosate (at environmentally detected concentrations) do not shape the microbial communities as a main environmental driver would, and therefore it is not assumed that microbial ecosystem functioning is altered or impaired. An R package was written to facilitate the use of microbial community composition data for Random Forest and ANN, and its applicability to query microbial community composition for information about their habitat and contamination status was demonstrated. Random Forest was consistently more accurate and predictions varied less compared to ANN. It was also faster, less computationally demanding, had fewer hyperparameters to tune, was more interpretable and provided several variable importance measures. In general, a partially interpretable model should be included in the ML analysis, because confounding variables in ecology most often cannot be avoided and therefore need to be monitored. The sample size was not sufficient to unambiguously identify an indicative microbial fingerprint for TNT and to capture the underlying data structure completely. Glyphosate-disturbed communities demonstrated short community recovery time while glyphosate was still present. Potentially resilient communities after the dissipation of TNT

were detected based on false positive classification. They could alternatively be explained by the structural similarity of nitroaromatic compounds. Current monitoring describes insufficient progress on improving the environmental state of the Baltic Sea, but they also ignore the bacteria as source of information (the main drivers of the basic processes). Integrating microbial community compositions and ML into regular monitoring could provide information for more efficient actions, e.g. to detect intermittently occurring contamination which is not discovered by analytics at the time of the monitoring cruise. It allows to assess environmental state-relevant processes at their origin and could be an important step towards holistic modelling of the Baltic Sea. It was also shown that NGS, though being relatively expensive, is in the same cost range of analytical methods such as glyphosate or MC detection. Monitoring, in return, would provide the urgently required training data, including a plethora of environmental parameters and contaminants for supervised learning. In fact, limited sample size was identified as main constraint of the analyses.

With regard to the future, I do believe that the way to unlock the full potential of ML to analyze microbial community compositions and predict environmental conditions involves a strong increase in the number of samples. This would enable the determination of truly generalized patterns and is the distinction to statistical models, which of course also profit from increased sample sizes. Studies involving ~ 100–500 samples demonstrate the potential and importance of ML in microbial ecology, as I did with this thesis with regard to contaminations in the Baltic Sea. Yet they cannot transcend the limitations imposed by the sample size. Thus, after having these ~ 100 sample data sets analyzed, new experiments or studies are started from scratch, which cannot make use of the existing data. Nowadays, interesting studies with more potential include at least one order of magnitude greater sample size, such as in Ghannam et al. (2020) with > 1200 samples. In comparison, for image classification, it is possible to access or download huge models online pre-trained with millions of images and categories: ImageNet contains about 15,000,000 images of 20,000 classes (Fei-Fei et al., 2010), the pretrained EfficientNet model achieved 84.3 % accuracy (Tan and Le, 2019). The BARM and the BalticMicrobeDB (Alneberg et al., 2018) already exist as a starting infrastructure to collect community composition data. Environmental monitoring should work as a prototype of constantly retraining ML models, integrating the newest information, just like physics and climate modelers – and now image classifiers – have done it for a long time.

Chapter I

An artificial neural network and Random Forest identify glyphosate-impacted brackish communities based on 16S rRNA amplicon MiSeq read counts

The following chapter was published in the journal *Marine Pollution Bulletin* as:

René Janßen, Jakob Zabel, Uwe von Lukas, and Matthias Labrenz (2019). An artificial neural network and Random Forest identify glyphosate-impacted brackish communities based on 16S rRNA amplicon MiSeq read counts. *Mar. Pollut. Bull*. 149:110530. doi: 10.1016/j.marpolbul.2019.110530

Declaration of author contributions:

René Janßen designed and conducted the experiment, performed laboratory work and bioinformatically processed and analyzed the data. He performed the Random Forest analysis, further analyzed the Artificial Neural network results and rewrote the entire manuscript.

Jakob Zabel performed the Artificial Neural network analysis and wrote the first draft of the manuscript.

René Janßen, Jakob Zabel and Matthias Labrenz discussed the data.

Matthias Labrenz conceived of the idea and critically commented on the manuscript and redrafted parts of it, Uwe von Lukas critically commented on the manuscript.

René Janßen's contribution to the written manuscript was ~ 80 %.

Abstract

Machine learning algorithms can be trained on complex data sets to detect, predict, or model specific aspects. Aim of this study was to train an artificial neural network in comparison to a Random Forest model to detect induced changes in microbial communities, in order to support environmental monitoring efforts of contamination events. Models were trained on taxon count tables obtained via next-generation amplicon sequencing of water column samples originating from a lab microcosm incubation experiment conducted over 140 days to determine the effects of glyphosate on succession within brackish-water microbial communities. Glyphosate- treated assemblages were classified correctly; a subsetting approach identified the taxa primarily responsible for this, permitting the reduction of input features. This study demonstrates the potential of artificial neural networks to predict indicator species for glyphosate contamination. The results could empower the development of environmental monitoring strategies with applications limited to neither glyphosate nor amplicon sequence data.

1.1 Introduction

Monitoring the environmental status of the Baltic Sea is required by law as part of the HELCOM agreement (Backer et al., 2010), including distinct events such as contamination. Based on molecular methods, such as 16S/18S rRNA next generation sequencing (NGS) and meta-genomics/-transcriptomics, microbial reactions during known contamination events can be identified. With sufficient knowledge about such reactions contamination events potentially can be, vice versa, discovered by molecular methods without information about the contamination event. Thus, NGS and -omics have the potential to support environmental monitoring of the Baltic Sea in the future. However, these methods collect such a large amount of data that the data cannot be evaluated manually.

Machine learning algorithms are important tools to support data analysis and decision-making because they are capable of performing regression and classification tasks on complex data sets and solving non-trivial tasks. Classification resembles the decision between discrete variables, e.g., “yes” or “no”, whereas a regression fits the provided data within the range of a continuous variable (Bourdès et al., 2010). The term supervised learning refers to the practice that a machine learning algorithm is provided the input data and the correct output and adjusts its specific parameters to correlate both (Angermueller et al., 2016). Random Forest (RF) is an established machine learning ensemble classifier (Breiman, 2001a). RF makes use of decision trees, which on their own are weak classifiers, prone to low robustness and overfitting. However, RF as ensemble classifier builds a forest of decision trees, each tree based on a different subset of the features and observations of the data, thereby reducing the variance and increasing the robustness. A majority vote based on all decision trees eventually classifies the data. RF has been applied in many fields of data science with great success (Fernández-Delgado et al., 2014). RF is also used for analyzing NGS and environmental data because RF “off-the-shelf” can process continuous, discrete and logical values as input. An overview for supervised learning on microbial community composition data regarding, for instance, the classification of the human microbiome, is given in Knights et al. (2011). More focused towards contamination events, Smith et al. (2015) used microbial community compositions from a nuclear waste site to predict uranium and nitrate levels and from the Deepwater Horizon oil spill to classify for hydrocarbon contamination. Similarly, He et al. (2018) analyzed groundwater microbiomes for their functional gene richness and diversity using microarrays and found that increasing uranium levels led to generally decreased functional richness and diversity, while specific functional guilds related to uranium increased.

The general potential of another common machine learning technique, the artificial neural network (ANN), is illustrated by the correct prediction of a XOR-logic gate output, which is 1 only if exactly one of the inputs is 1. This cannot be achieved by a linear decision boundary

(Rosenblatt, 1958) but rather by an ANN with a hidden layer (Sprinkhuizen-Kuyper and Boers, 1996). Attempts have been made to implement ANNs in NGS data analyses, as NGS is a well-established method in medicine, environmental microbiology, biotechnology and related fields. NGS generates a high number of sequencing reads of DNA and reverse-transcribed RNA. Therefore, an appropriate data format to supply the ANN with the information is essential to link the methods. Nguyen et al. (2016) applied a convolutional neural network (CNN) to treat DNA sequences as a string input and store the position of the nucleotides in the sequence. Another option is to use sequencing-derived or processed data, not the raw sequences themselves. Larsen et al. (2012) used microbial community composition and environmental data to calculate an environmental interaction network, which identified significant relationships. This interaction information was used to generate ANNs predicting the abundance of microbial taxa depending on changes of environmental factors as a bioclimatic model. Using microbial community composition and phylogenetic trees to incorporate similarity information in a CNN model, Fioravanti et al. (2018) classified microbial communities associated with Inflammatory Bowel Disease.

ANNs in its various architectures are known to require more information for training than RF, but perform better on more complex data sets. Additionally, ANNs may continuously gain performance with growing data amounts, which could be provided along a monitoring effort (Fernández-Delgado et al., 2014). The aim of our study was to check whether an ANN analysis of 16S rRNA NGS data is suitable to detect glyphosate contaminations in the Baltic Sea and potentially support environmental Baltic Sea monitoring.

Glyphosate is the most-applied herbicide globally since the 1970 and acts as a potentially harmful herbicide (Van Bruggen et al., 2018), as well as a phosphorus-providing substrate (Hove-Jensen et al., 2014). Recent studies have proven that glyphosate is mobile despite its soil adsorption characteristics (Bergström et al., 2011; Kwiatkowska et al., 2016; Myers et al., 2016). Due to its intensive use in agriculture around the world, glyphosate is present in significant quantities in soil and groundwater (Battaglin et al., 2014), and has entered the brackish Baltic Sea (Skeff et al., 2015). A laboratory microcosm experiment was set up in which the herbicide was added as a stressor to a brackish-water microbial community. To assess the impact of glyphosate independently of specific glyphosate detection methods, a combined approach of artificial neural networks and 16S rRNA and rRNA gene NGS was applied. An ANN was trained on compositions, which were declared as glyphosate-impacted or not. The ANN was then challenged to classify a previously unknown sample with regard to the presence of glyphosate. The aim was to automatically differentiate glyphosate-treated from untreated control communities. The robustness of the ANN setup and the amount of taxonomic information required for a reliable classification were investigated and, as control, compared to Random Forest analysis.

1.2 Material and methods

1.2.1 Laboratory & sampling

1.2.1.1 Overview of the experimental setup

Two 12-L (20×30×20 cm) microcosms comprising float glass and silicone glue were obtained from Rebie Aquaristik (Bielefeld, Germany). The experiment lasted for 140 days, starting at day −69 for an equilibration period until day 0, when a glyphosate pulse introduced the incubation period in the treatment microcosm until day 71 (Supplementary Material 1.1). On a total of 16 time points (days −25, −7, 0, 3, 7, 10, 14, 17, 22, 29, 36, 43, 50, 57, 64, 71) water samples from both microcosms were sampled. In three technical replicates each, 16S rRNA and 16S rRNA gene based community compositions were generated, summing up to 12 communities per time point and a total of 187 communities. For averaged technical replicates, 64 abundance tables were yielded.

1.2.1.2 Microcosm setup

The microcosms were cleaned with EtOH (70 %) and rinsed with sterile, filtered MilliQ water before they were filled with the sterilized substrates. Surface brackish water for inoculation was collected 2.5 km north of Warnemünde, Germany (54.199412, 12.042317). Five hundred milliliters of water was sterile filtered per GVWP filter (0.22 µm, Merck Millipore, Darmstadt, Germany) until the collected volume was filtered. The filters were immediately shock frozen in liquid nitrogen and stored at −80 °C. Modified artificial brackish water (ABW, Bruns et al., 2002) served as the substrate, containing double the amount of KH2PO4 to prevent phosphate limitation. A stock solution of 20 g casein hydrolysate (Merck, Darmstadt, Germany)·L^{-1} dissolved in MilliQ (Merck Millipore) served as the carbon and nitrogen source. The solution was sterile filtered (0.22 µm, Sartorius, Göttingen, Germany) and stored at 15 °C. The casein hydrolysate was added to the ABW after autoclaving to a final concentration of 2.5 mL·L^{-1}. Fire-dried quartz sand (0.1–0.4 mm, Quarzwerke, Frechen, Germany) was combusted for at least 4 h at 500 °C in aluminium trays (Alcan, Brazil) and served as an artificial, carbon-free hard substrate. The microcosms were filled with 2 kg of quartz sand (~1.6 L) and 8 L of ABW. Combusted GF/F microfibre filters (Ø 47 mm, Whatman, Little Chalfont, UK) were placed into the hard substrate to provide easily collectible biofilm-overgrown material. Air pumps (2×200 L·h^{-1}, 4 W, EHEIM GmbH, Deizisau, Germany) delivered sterile-filtered air (0.2 µm, Midisart 2000, Sartorius Stedim, Göttingen, Germany). Three thawed GVWP inoculum filters were cut in half, with one half placed in each microcosm overnight. ABW was refilled on days −56 and −34; the batch mode lasted from day −69 to day −31 to ensure that the bacteria formed biofilms on all surfaces, thereby preventing glyphosate adsorption (Supplementary Material 1.1). Beginning on day −31, stable nutrient conditions were provided by changing the cultivation mode to a chemostat-like continuous culture to prevent substrate depletion and product

accumulation. A peristaltic pump (Ismatec IPC 8, Cole Palmer, Wertheim, Germany) transported ABW from a sealed, autoclaved 5-L Schott bottle through clean, sterile tubing (Ø 1.02 mm (ID), silicone peroxide, Ismatec) at a flow rate of 0.37–0.38 mL·min^{-1} (537–548 mL·day^{-1}) into the microcosms representing the water column. A second peristaltic pump with a flow rate of 0.33–0.34 mL·min^{-1} (475–489 mL·day^{-1}) removed water from the opposite end of the microcosms such that excess volume was available for sampling. The 5-L Schott bottle was regularly exchanged together with the inlet tubing. On day 0, a pulse of sterile filtrated glyphosate (13.49 mg·L^{-1} final concentration) was added to the water column of the treatment microcosm and mixed by stirring.

1.2.1.3 Sampling procedure

Samples were taken for the determination of total cell counts and glyphosate and aminomethylphosphonic acid (AMPA) concentrations, respectively. One-hundred-millilitre water column samples were sterile filtered in three replicates; for the analysis of biofilm communities, three overgrown GF/F filters were picked with sterile tweezers. These filters were used for nucleic acid extraction. Samples for the DNA/RNA extraction were shock frozen in liquid nitrogen and stored at −80 °C. Five-millilitre samples for the determination of glyphosate and AMPA concentrations were stored at 20 °C without further treatment.

1.2.1.4 Nucleic acid extraction and sequencing

Nucleic acid extraction and DNA digestion were performed according to Bennke et al. (2018) for the filtered water samples. Biofilm samples were extracted using the phenol-chloroform method described in Weinbauer et al. (2002). cDNA synthesis was performed using 20 ng DNA-free total RNA as the input for the MultiScribe (Fisher Scientific GmbH, Germany) Reverse Transcriptase system with reverse primer 1492r (5′ TACGGYTACCTTGTTACGACTT (Lane, 1991)). Illumina amplicon sequencing was prepared as described in Bennke et al. (2018). The V3–V4 region on the 16S rRNA gene was targeted with the primer set 341f-805r (forward: CCTACGGGNGGCWGCAG, reverse: GACTACHVGGGTATCTAATCC (Herlemann et al., 2011)). Indexed amplicon libraries were pooled to a concentration of 4 µM. The PhiX control was spiked into the library pools at a concentration of 10%. Each final library pool (4 pM) was subjected to one of two consecutive individual paired-end sequencing runs for water column samples using 600 cycle V3 chemistry kits on an Illumina MiSeq. During the 16S rRNA gene amplicons run, 706 K·mm^{-2} clusters were sequenced; generating 17.6 million reads that passed the filter specifications. Over 70 % of the sequencing and index reads were found to have a Qscore ≥30. During the 16S rRNA amplicons run, 555 K·mm^{-2} clusters were sequenced. This generated 13.9 million reads passing filter specifications. Over 74 % of the sequencing and index reads had a Qscore ≥30.

1.2.1.5 Bioinformatic and statistical analysis of amplicon data

Sequence data preparation for the SILVAngs pipeline was performed as described previously (Bennke et al., 2018). The SILVAngs pipeline dereplicated 100 % identical sequences. Of the remaining unique 16S rRNA sequences, OTUs with a similarity threshold of 98 % were selected. The representative sequence per OTU was taxonomically annotated using the ARB-SILVA database (SILVA release 128). Identical annotations for different OTUs were merged into clusters on the genus level; thus, the term “clusters” is used instead of OTUs. From the resulting taxonomy file containing reads per sample per cluster the clusters annotated as “No relative” were discarded. The relative abundance per cluster in % was calculated from the read fraction of the cluster of the library size of the sample. The unfiltered data set underwent no further quality check; for the filtered data set clusters with fewer than five reads were excluded. The sequences were deposited in the NCBI database under BioProject ID PRJNA434253 and SRA accession SRP151042.

1.2.1.6 Non-metrical multidimensional scaling (nMDS) and Principal Coordinates Analysis (PCoA)

Ordination methods belong to the first steps of exploratory analysis. nMDS and PCoAs, both commonly used in ecology, are approaches to find similar samples and possible patterns in a data set. While an unsupervised approach was used to produce a nMDS from our dataset, a PCoA was produced on a supervised Random Forest Model that displays patterns generated due to class labels. The nMDS was generated by phyloseq v. 1.26.0 (McMurdie and Holmes, 2013) within R v. 3.5.1 (R Core Team et al., 2017). The distance matrixes required for ordination where calculated as follows: the table with the unfiltered relative abundances of clusters were square root transformed and the dissimilarity based on Bray-Curtis was calculated for nMDS. 100 Random starts were performed to reach the ordination with the lowest stress. To visualize the similarity between samples analyzed by a Random Forest model, the proximity matrix (see Material and methods: Random Forest) was converted into a distance matrix. The R base function cmdscale() was called to perform classic multidimensional scaling (MDS) or Principal Coordinates Analysis (PCoA) with a Euclidean distance. All plots were created using ggplot2 v. 3.1.0 (Wickham, 2016).

1.2.2 Machine learning

1.2.2.1 Neural network architecture

The Java library for the feed-forward backpropagation ANN was Deep Learning for Java (DL4J), with N-dimensional arrays (Patterson and Gibson, 2017) and default values for most parameters. ANNs receive data via an input layer with a number of input nodes representing the dimensions of the input data; therefore, the input layer contained one neuron per feature (taxonomic cluster) in the respective data set, ranging from 1 to 687.

The signal is transported from one node to the next (layer), and the strength of the signal is altered by the weight of the connection, allowing for separation between more and less impactful nodes. The weights were Xavier initialized (default). Scaling of the signal between 0 and 1 was achieved using Softmax for output normalization. On each node or neuron, a threshold must be met by the incoming signal to activate the forwarding to the next layer. The neuron activation function was tanh (the default).

By employing further connected layers, the hidden layers, more complex interactions are enabled due to more combinations of input signals. Hidden layer 1 comprised 25 neurons, hidden layer 2 comprised 5 neurons and the output layer comprised 2 neurons, one for each class “treatment” and “control”. The output layer showed the aggregated result of the signals channeled through the preceding nodes. Each layer was fully connected to the next. The ANN had to be trained beforehand to classify the microbial community compositions. The option used was providing training data of a given format and amount as well as the expected classification, consequently the so called supervised learning. Using backpropagation as an iterative learning process, the ANN adjusts the weights of the connections between the nodes to yield the provided classification. For this, the loss function was the negative log likelihood (Glorot and Bengio, 2010). Every experiment used 2000 epochs, with a learning rate of 0.1 for each repetition. Prior unknown data of the same format might then be classified by the trained ANN. To do so, the initial data set must be split into a training quota and a test quota. A third quota is required if, e.g., several ANN setups are to be compared and validated before the actual testing. It is therefore only feasible if sufficient data is available (Wu et al., 2013). As sequencing is still comparatively expensive, the amount of samples processed for this experiment was large but limited. Therefore, the data were split into training and test data sets, with the largest portion being training data.

1.2.2.2 Random Forest

The RF algorithm selects randomly a set of 16S rRNA (gene) community compositions and builds a decision tree, where every decision is a node splitting the observations. The final outcomes or ends of a decision tree are the leaves and depending on the values of the feature, the respective leaf votes for a class. On each node, a random set of taxonomic clusters is selected. The Gini impurity measure describes which of the selected clusters performs best to split the samples according to the provided classification and is therefore used. This process is repeated for the appointed number of trees; the Random Forest is “grown”. Thus, the Random Forest model is based only on the randomly sampled observations, which were “in bag”, whereas those not involved in building the Random Forest were “out of bag”. Therefore, such samples can be used as testing set to evaluate the out of bag error (OOB). Practically, to evaluate the OOB error, a decision tree in a grown Forest asks on a given node for the OOB observations, whether the relative abundance e.g. of *Parvibaculum* spp. is above or below a

given threshold. The respective threshold was determined when growing the forest, based on the best Gini impurity. Consequently, it begins to separate between samples classified as glyphosate treated and nontreated. As one node is usually not sufficient, it proceeds to the next higher node and asks for the relative abundances of, e.g., *Massilia* spp. According to the respective values, the remaining not correctly split samples are further divided. This continues until no more features are available or no improvement of Gini impurity is observed, the node becomes a leaf.

Due to the Gini impurity measure, important features are placed as first node (root node) or early on in the decision tree as they contribute to the classification. The mean minimum depth of features estimates the importance across the whole Forest. Therefore, feature sampling size and number of trees are important parameters to tune. For classification the default feature sample size is n with n = total number of features. In noisy data sets with many unimportant or sparsely distributed features, it might improve the classification performance if a higher number of features are evaluated at every node, thereby increasing the chance of sampling a valuable feature. Increasing feature sample size, however, significantly affects the computational efforts. The number of trees per Random Forest should be increased until the OOB error stabilizes.

The frequency of individual samples ending up in the same terminal node of a tree can be reported in a proximity matrix. If e.g. sample A and sample B both land in the same end node, the proximity between A and B is increased. By this means, the proximity matrix can be used as a measure of similarity.

1.2.2.3 Application of the Random Forest

To perform the RF analysis on the data sets, the community compositions and the corresponding classifications were retrieved from phyloseq to use with the randomForest package v. 4.6-14 (Liaw and Wiener, 2002). Mean minimum depth of features were extracted by the randomForestExplainer v. 0.9 (Paluszynska and Biecek, 2017). The randomForest function was called using ntree = 5000, and the parameter mtry = 40 for the main data sets and 16S rRNA and 16S rRNA gene subsets; mtry = default (3) for the various top 10 selections; and mtry = default (1) for single clusters. 100 Random Forests were built and evaluated to receive a distribution of OOB errors, presented as percentage of correct classification. The same input tables as for the ANN were used and, since the OOB error evaluation was used, the data was not divided into training and test sets. For the same reason, the remaining parallels of a given test sample in the filtered data set were not removed, resulting in one or two training sample parallels being very similar to the respective test sample.

1.2.2.4 Format of the main data sets

The taxonomy tables contained the relative abundance of a taxonomic cluster as the input feature for a given sample (Supplementary Material 1.1). Each sample represented a unique combination of time point, nucleic acid, microcosm, habitat, and – in case of the filtered data sets – technical replicate (2 or 3).

Main data set 1 was the “unfiltered data” set, which consisted of all 687 clusters and the averaged technical replicates, resulting in 64 observations. Note that from the original experimental data containing information on water column and biofilm, the relative abundances of clusters within the biofilm samples were removed in this study at a later step but remained a feature of the taxonomy file format and were considered with an input node. Effectively, it represents setting all biofilm-originated clusters to 0. Therefore, the unfiltered data set contained 687 taxonomic clusters with 213 being biofilm-originated, resulting in 474 clusters different from 0. One observation was randomly selected to test the classification performance of the ANN; the remaining tables comprised the training data.

Main data set 2 was labelled “filtered data”. The tables were filtered before the relative abundances were calculated by removing clusters with less than five counts in a sample. Additionally, the replications were not averaged but rather used as separate observations, yielding 187 observations (111 × “control”, 76 × “treated”). Consequently, the filtered data set contained 360 taxonomic clusters with 86 clusters being biofilm-originated, resulting in 274 clusters different from 0. One observation was randomly selected for testing; the remaining tables comprised the training data. Excluded were the remaining tables of the replicate as they were very similar to the test observation. Which observations (tables or samples) and which features (taxonomic clusters) were additionally used in the various classification setups is illustrated in Supplementary Material 1.2.

1.2.2.5 Note on classification thresholds

As only the samples from the treatment microcosm after day 0 were in contact with glyphosate (denoted as “treated”), those samples were to be separated from both the samples before the glyphosate addition and all samples from the second microcosm (denoted as “control”, Supplementary Material 1.1). Hence, the unfiltered data set with averaged replicates consisted of 38 × “control” and 26 × “treated” tables. Consequently, purely guessing “control” as classification would be correct for ~59 % of the tables. Moreover, the 38 control tables combine 32 tables from the glyphosate-unimpacted microcosm as well as 6 tables from the treatment microcosms before the addition of glyphosate. Therefore, a classifier that is able to distinguish the microcosms and votes for “control” would be wrong only for the 6 tables which originate from the treatment microcosm. Therefore, a classification rate of 1 – (6/64) = 90.625 % could be achieved without learning to classify before and after the addition of glyphosate. The

corresponding threshold for the filtered data set is 1 – (18/187) = 90.374 %. A classification rate superseding those thresholds must be accomplished to evaluate the model as having learned more than solely separating the microcosm communities, a task otherwise no machine learning is required for (Figure 1.1). We did not duplicate existing “treated” tables to generate a 1:1 ratio of “treated” and “control” as the deviation was not considered to be problematic.

1.2.2.6 Test ANN classification

The general applicability of ANNs in classifying community composition data was tested using the unfiltered data set as the input. The classification was repeated 2000 times.

1.2.2.7 Identifying clusters present in a successful classification

To identify the taxonomic clusters participating in successful classifications, a subsetting approach was applied. For the unfiltered data set, 30 clusters were chosen randomly, and the network was trained with those 30 clusters. For each subset, the classification was repeated 256 times (so with 64 tables each covering 4 times the test table), and the number of correct classifications and the chosen clusters were documented. The order of averages stabilized after ~1000 subsets, and the process was stopped at 1220 subsets. Each subset classification required approximately 1 h on a virtual machine with 4 CPU cores and 16 GB RAM. Comparably for the filtered data set, the top 10 ranking clusters were calculated by the random subsetting approach based on a sample size of 20 clusters. Subsetting was performed 1066 times, and on each subset 1000 classification runs were performed. Four specific clusters appearing in the unfiltered and filtered subsets were compared regarding their participation in correct classifications.

1.2.2.8 Determine required feature amount for classification

The previous experiment ranked each cluster based on the number of times it was present in successful classification subsets. From this order, the 10 top-ranked clusters were selected to determine whether a classification was possible with a significantly reduced number of features. Furthermore, the limitations of cluster reduction were explored by sequentially removing the lowest-ranked cluster of the top 10.

1.2.2.9 Determine required number of observations for classification

To examine whether 16S rRNA or 16S rRNA gene data alone were sufficient input for classification and which is better suited, the unfiltered data set was split into 32 16S rRNA gene and 32 16S rRNA tables, with 31 tables used for training and the remainder for testing. We conducted 13568 repetitions for classification based on 16S rRNA gene and 2048 repetitions based on 16S rRNA. Additionally, to investigate the required sampling resolution, half of the time points from the unfiltered data set were evenly distributed removed from the data set, and the classification was tested 6656 times.

1.2.2.10 Machine learning algorithms from other tool kits tested

In the process of analyzing the community assemblage data, Weka toolkit implementations of Decision Table, Random Forest and ANNs were tested (Hall et al., 2009). None of the WEKA approaches provided correct classifications with standard parameters, and the computation times of the respective ANNs were excessive.

1.3 Results

1.3.1 ANN identifies glyphosate-treated microbial communities

The unfiltered community composition data was displayed by PCoA (data not shown) and nMDS ordination (Figure 1.1). The nMDS provided substantially better clustering, outlining the community dissimilarities across nucleic acids and treatments. Applying an ANN to the same data achieved 1905 correct classifications out of 2000 repetitions (95.25 %). As explained above in the methods, a classification based purely on guessing would have been correct ~59 % of the time and if only microcosms were separated, a 90.625 % correct classification could theoretically be reached at best. It is therefore generally possible to use an ANN to separate community composition data. The Random Forest classification was used as the reference machine learning algorithm and evaluated on the OOB error, classified 97.1 % correctly.

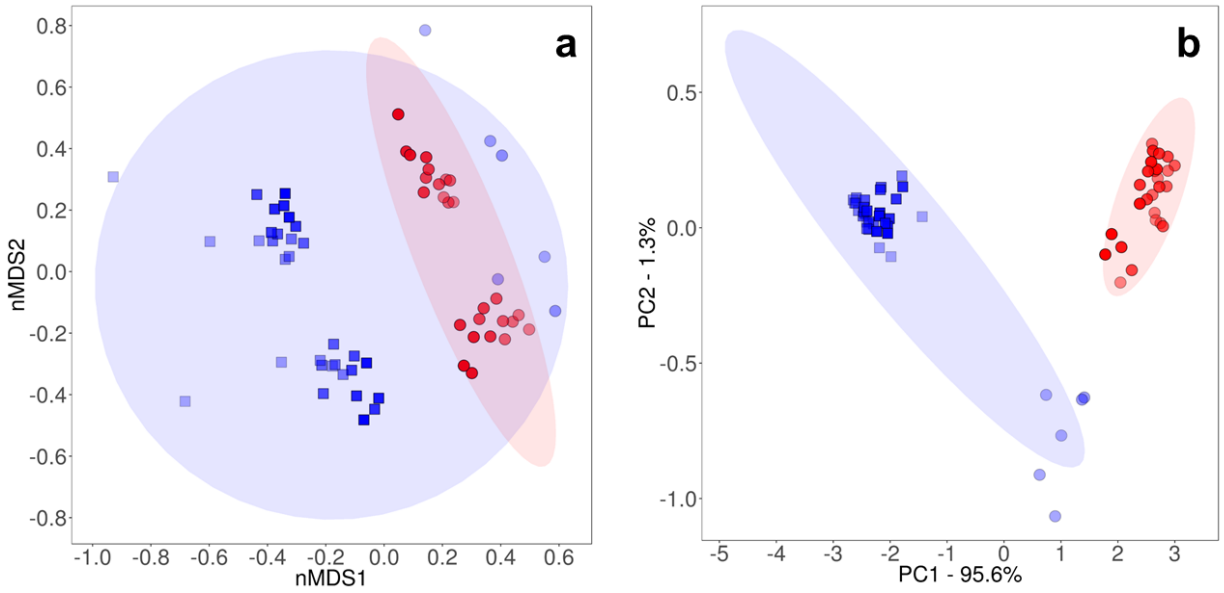


Figure 1.1: Reduction of multidimensional community composition data using a) the Bray-Curtis dissimilarity for nMDS and b) the Random Forest-generated proximity matrix for PCoA with Euclidean distance. The nMDS showed four clusters, one for each combination of microcosm and nucleic acid. Random Forest classification generated three clusters, regardless of nucleic acid: communities from the control microcosm (blue squares), communities from the glyphosate microcosm prior to the addition of glyphosate (blue dots) and after the addition of glyphosate (red dots) and was able to isolate the treated communities. The separation took mainly place on the PC1 axis with 95.6% variance explained.

The resulting proximity matrix of sample classification was able to display the separation achieved, which is mainly shown along PC1, with 95.6 % of variation being explained. The few blue samples in the lower cluster of the MDS plot originate from the treatment microcosm before the glyphosate addition and were not separated by nMDS.

1.3.2 Identification of clusters present in successful classifications by the ANN

To understand which clusters participate in correct classification and hence are possibly important, subsets comprising randomly chosen clusters were tested.

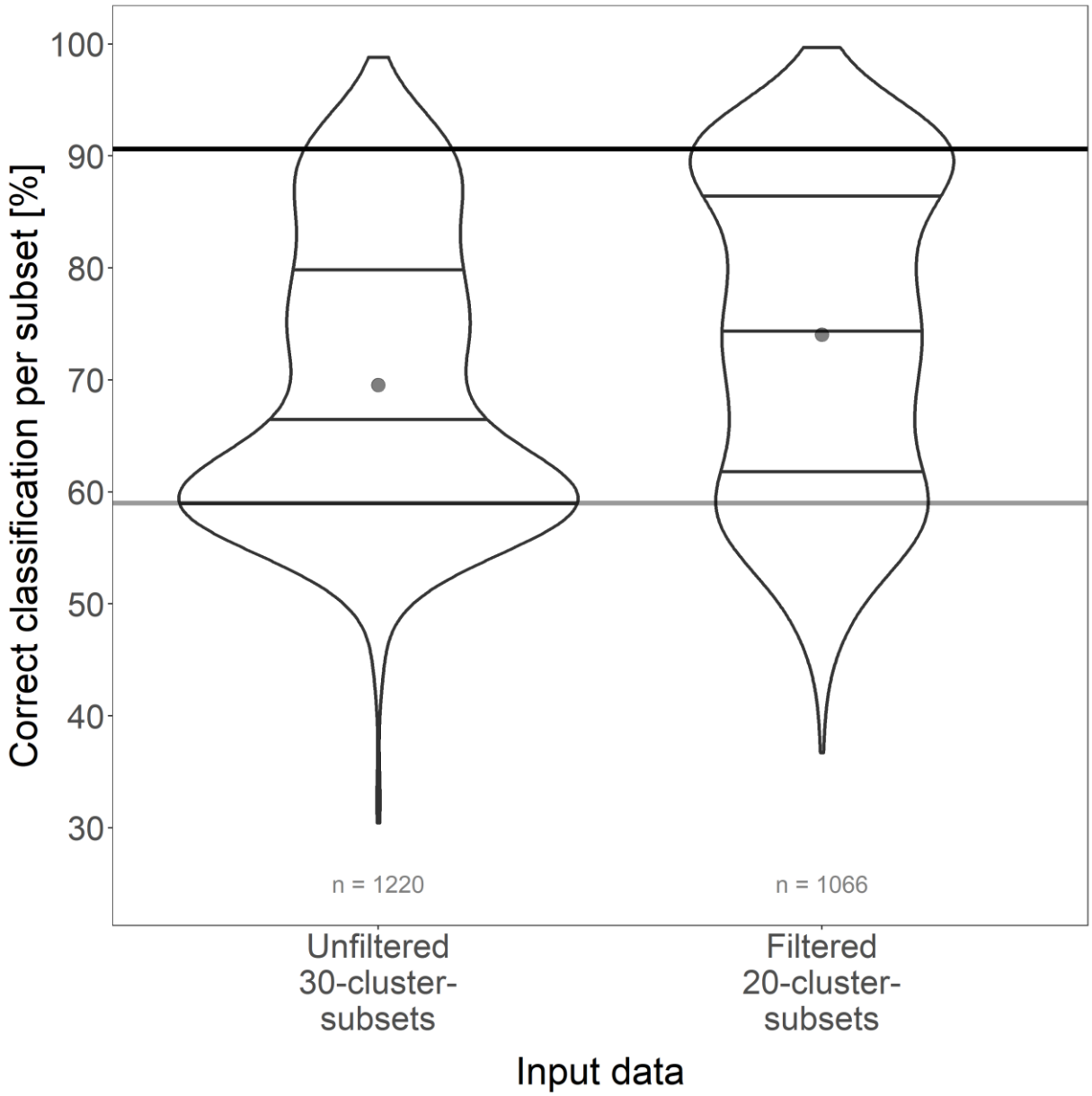


Figure 1.2: Violin plots of correct classification rates by random subsets of size 30 for the unfiltered data set and size 20 for the filtered data set, respectively. *n* is the number of subsets that were generated for the respective data set. The dot represents the average; the three horizontal lines within the violin plot depicture the 25%, 50% and 75% quantiles. The horizontal bar at 59% displays the classification achievable by pure guessing, the upper bar marks the threshold for a classification which both separates the microcosms and before and after glyphosate addition. A shift from many subsets classifying around the guessing level for the unfiltered data set to improved classification rates in the filtered data set is shown.

For the unfiltered data set, the subset size was 30 clusters, and 1220 subsets were generated. Of 256 classification repetitions for each subset, the number of correct classification ranged from 30.1 to 98.4 % (Figure 1.2). For the filtered data set, 1066 subsets with a size of 20 clusters were selected, and on each subset 1000 classifications were performed, with the correct classification per subset ranging from 36.7 to 99.7 %. The range is comparable; the distribution of the classification, however, displayed an increase in valuable subsets for the filtered data set. Fifty percent of the subsets of unfiltered data achieved a classification rate centered around the guessing level, and only a small fraction of subsets was above the 90.625 % threshold compared with a significantly reduced fraction for the filtered data set at guessing level and increased fraction at all higher classification rates, especially around the critical 90.625 % threshold. Furthermore, an increased average of the classifications was observed, as well as reduced computational efforts, as the distribution of classifications stabilized after 300 subsets for the filtered data set, compared with 1000 subsets for the unfiltered data set.

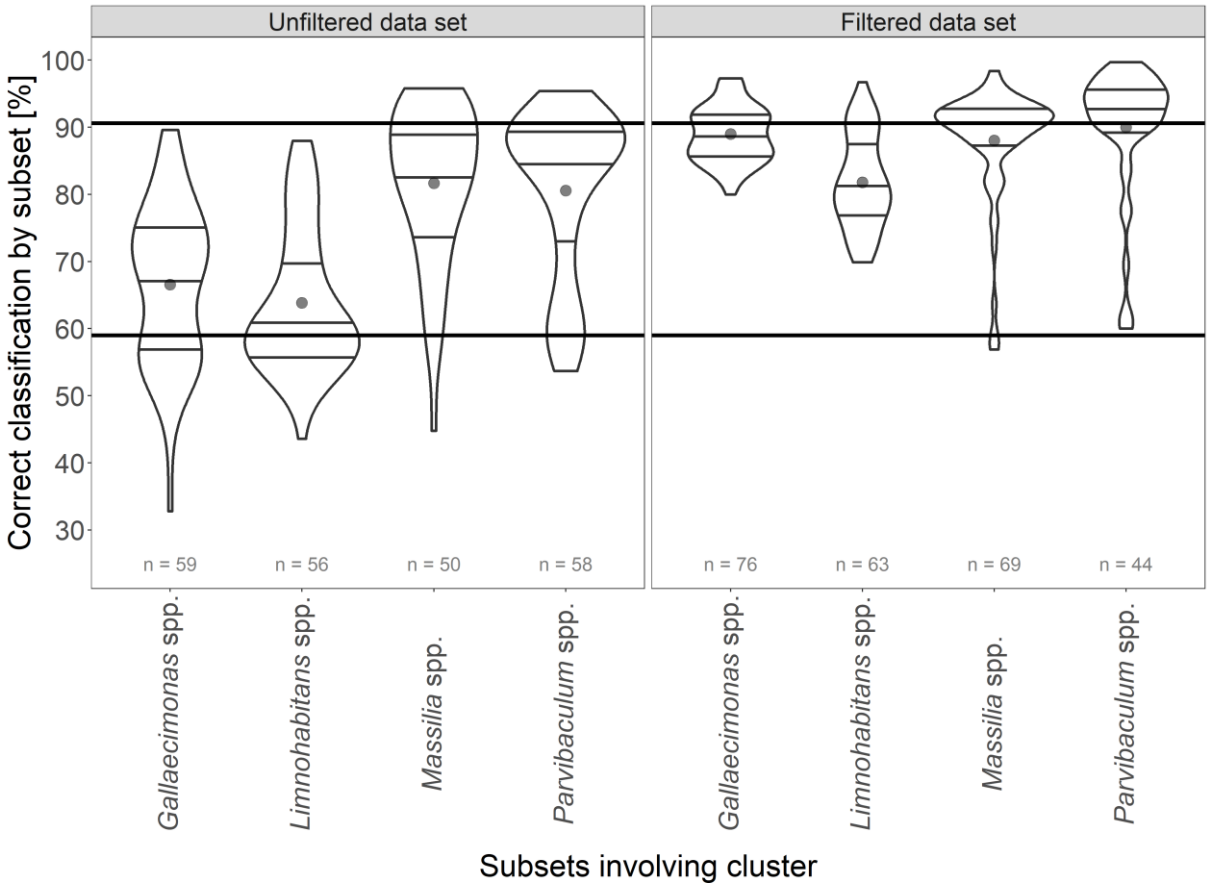


Figure 1.3: Violin plots of correct classification by subsets containing specific taxonomic clusters for both data sets. *n* is the number of subsets that included the respective cluster. The dot represents the average classification, the three horizontal lines within the violin plot depicture the 25%, 50% and 75% quantiles. The horizontal bar at 59% displays the classification achievable by pure guessing, the upper bar marks the threshold for a classification which both separates the microcosms and before and after glyphosate addition. The filtering step improved the ANN's performance by reducing the range and frequency of less good classifications towards a higher number of better classifications. *Gallaecimonas* spp. containing subsets drastically improved classification rates.

The ranking, based on the average classification from all subsets a cluster was part of, revealed which clusters were frequently part of the correctly classified subsets. The

classification of subsets containing the clusters *Massilia* spp., *Parvibaculum* spp., *Gallaecimonas* spp. and *Limnohabitans* spp. were compared (Figure 1.3; relative abundances in Figure 1.4 and Supplementary Material 1.3l and r). They appeared in both data sets. In particular, *Gallaecimonas* spp. and *Parvibaculum* spp. displayed a distinct increased relative abundance following the glyphosate addition, which could be useful for the classification by the ANN. *Limnohabitans* spp. increased in abundance in the control treatment after day 0. For *Massilia* spp. the temporal abundance course did not reveal a response to glyphosate but rather differed between the two microcosms. *Massilia* spp. in both data sets were part of the well-performing subsets and identified as a very valuable cluster for classification in the setup described below.

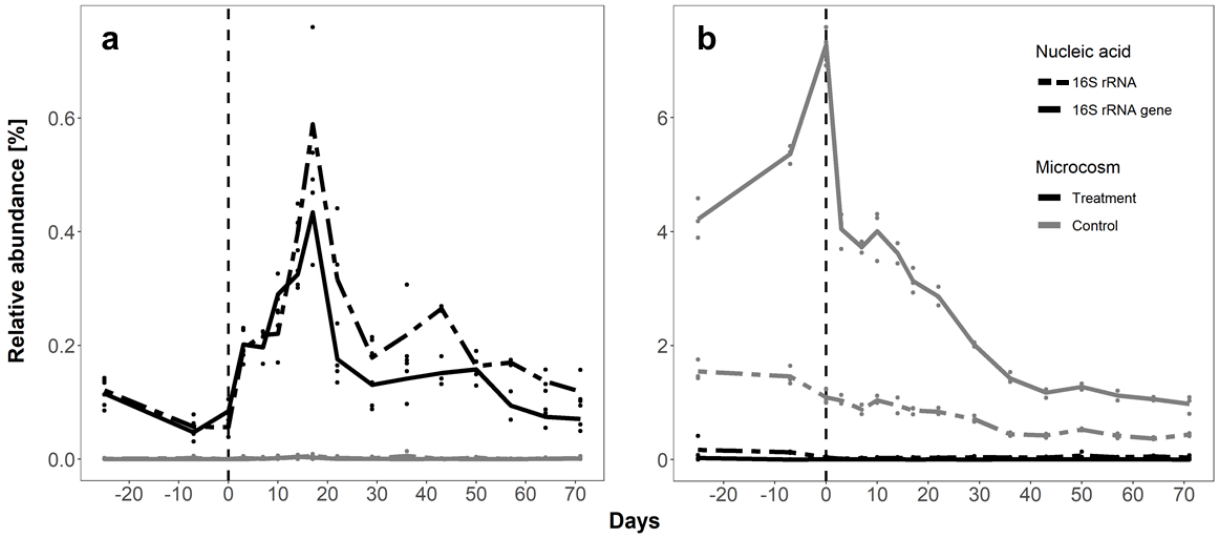


Figure 1.4: Relative abundances of the taxonomic clusters *Parvibaculum* spp. and *Massilia* spp. a) *Parvibaculum* spp. differed between treatment (black) and control (grey) microcosms as well as before and after glyphosate addition (dashed vertical line). b) *Massilia* spp. displayed consistent differences between the microcosms.

As shown in Figure 1.3, subsets containing *Gallaecimonas* spp. and *Limnohabitans* spp. did perform poorly on the unfiltered data set, and the classification rates for *Gallaecimonas* spp. containing subsets ranging from 34 to 89 %. This changed for subsets of the filtered data set containing *Gallaecimonas* spp., as a high fraction of the subsets were classified ~90 % correctly; however, subsets with *Limnohabitans* spp. improved, too. *Parvibaculum* spp. - containing subsets showed a good performance in both data sets but improved still in the filtered data as they reached the upper threshold for 50 % correctness in the subsets, similar to subsets containing *Massilia* spp. For both data sets it should be noted that performing the actual classification with an ANN trained on random subsets is not effective. The averaged classification rate over all generated subsets from filtered data is 74.1 % and 69.5 % from unfiltered data.

1.3.3 Exploring the limits of the required cluster features and assembling a highly indicative selection

The random subsets were expected to be unable to perform a sufficient classification contrary to the full data set. But the results indicate that some clusters are more decisive than others. Thus, the 10 clusters with the best average classification rate from the subsetting approaches were selected as the sole training data for the ANN. This yielded a classification rate of 94.4 % for the unfiltered data. 95.8 % for the filtered data was achieved if the highly similar technical replicates of the test sample were removed; otherwise the ANN classified 97 % correctly (Figure 1.5). The two top 10 cluster lists contained 4 shared clusters (Table 1.1). The difference is partially due to the removal of low abundance clusters during the filtering step. The maximal relative abundance per top 10 clusters ranged from 0.07 % (*Dokdonella* spp., only unfiltered data) and over 0.76 % (*Parvibaculum* spp., both sets) to 9.27 % (*Gallaecimonas* spp., both sets). Essentially, both top 10 data sets yielded a classification as good as the full unfiltered data set (95.25 %).

In the Random Forest models for the unfiltered and filtered data, the mean minimum depth measure was assessed to identify the 10 clusters most important for classification (Table 1.1). At minimum, 8 of 10 clusters were identical between the RF selection and the filtered ANN selection, thus, the filtering step marginally altered the top 10 RF selection. Random Forest classified 99.9 % correctly based on its unfiltered top 10 clusters, the filtered data set performed almost as good with 98.9 %. Further reducing the number of features revealed that using at least the six best-ranked clusters as the input for the ANN was required to yield a classification rate > 90.625 % for filtered data, whereas the unfiltered data sets kept meeting the threshold using as few as two clusters. A further stepwise reduction of the filtered data to only the top two clusters lowered the classification rate to 88.8 %. Interestingly, the classification rate with unfiltered data decreased to near the guessing level (63.5 %) when only the best ranked cluster was used (*Massilia* spp., Figure 1.4b). *Massilia* spp. comprised a cluster only abundant in the control microcosm. In the filtered data set, it achieved 90.32 %, within the reach of the microcosm-separating threshold 90.374 %. The second best unfiltered cluster was *Parvibaculum* spp. (84 %), and it was determined to be the best-ranked cluster for the filtered data set and, in contrast, performed well on its own (91.7 %). The relative abundance of *Parvibaculum* spp. was different between both microcosms as well as before and after the glyphosate addition (Figure 1.4a). It was also observed that 2 clusters yielded a better classification than 3 or 4 for filtered data, and the decrease was not linear for the unfiltered data.

Table 1.1: Listing of the 10 most important taxonomic clusters for classification for both data sets revealed by ANN random subsetting and Random Forest, compared with results of bioinformatic analysis achieved by applying R package DESeq2. Lineage (SILVA release 128), maximum abundance per 16S rRNA gene or 16S rRNA targeted approach and the DESeq2 *p* value, if available, were included.

Listing and comparison of all taxonomic clusters revealed by random subsetting or bioinformatic analysis applying R package DESeq2. Includes the rank, maximum relative abundance per DNA or cDNA targeted approach and the p-value, if available.

Genus	Family	Order	Class	Phylum	Unfiltered data top 10 by ANN	Filtered data top 10 by ANN	Unfiltered data top 10 by RF	Filtered data top 10 by RF	DESeq2 Wald test adjusted *p*-value	Maximal relative abundance of 16S rRNA (gene) [%]
Massilia	*Oxalobacteraceae*	*Burkholderiales*	*Betaproteobacteria*	*Proteobacteria*	x	x	x	x		7.6
Parvibaculum	*Rhodobiaceae*	*Rhizobiales*	*Alphaproteobacteria*	*Proteobacteria*	x	x	x	x	***	0.76
Brevundimonas	*Caulobacteraceae*	*Caulobacterales*	*Alphaproteobacteria*	*Proteobacteria*		x	x	x	**	0.37
Gallaecimonas	Unknown Family	*Gammaproteobacteria* Incertae Sedis	*Gammaproteobacteria*	*Proteobacteria*		x	x	x	***	9.27
Hyphomonas	*Hyphomonadaceae*	*Caulobacterales*	*Alphaproteobacteria*	*Proteobacteria*		x	x	x	***	0.66
Sphingopyxis	*Sphingomonadaceae*	*Sphingomonadales*	*Alphaproteobacteria*	*Proteobacteria*		x	x	x		2.75
Thalassobaculum	*Rhodospirillaceae*	*Rhodospirillales*	*Alphaproteobacteria*	*Proteobacteria*		x	x	x	***	7.66
Caulobacter	*Caulobacteraceae*	*Caulobacterales*	*Alphaproteobacteria*	*Proteobacteria*	x	x		x		0.37
Aminobacter	*Phyllobacteriaceae*	*Rhizobiales*	*Alphaproteobacteria*	*Proteobacteria*	x		x	x		0.15
Rhizobium	*Rhizobiaceae*	*Rhizobiales*	*Alphaproteobacteria*	*Proteobacteria*		x				6.22
Sphingomonas	*Sphingomonadaceae*	*Sphingomonadales*	*Alphaproteobacteria*	*Proteobacteria*		x				1.75
Ahrensia	*Phyllobacteriaceae*	*Rhizobiales*	*Alphaproteobacteria*	*Proteobacteria*			x			0.8
uncultured	*Rhodobiaceae*	*Rhizobiales*	*Alphaproteobacteria*	*Proteobacteria*			x			0.04
Dokdonella	*Xanthomonadaceae*	*Xanthomonadales*	*Gammaproteobacteria*	*Proteobacteria*	x					0.07
-	-	B38	*Gammaproteobacteria*	*Proteobacteria*	x					0.12
Idiomarina	*Idiomarinaceae*	*Alteromonadales*	*Gammaproteobacteria*	*Proteobacteria*	x				**	0.62
Loktanella	*Rhodobacteraceae*	*Rhodobacterales*	*Alphaproteobacteria*	*Proteobacteria*	x				**	0.48
Nesiotobacter	*Rhodobacteraceae*	*Rhodobacterales*	*Alphaproteobacteria*	*Proteobacteria*	x					0
Reyranella	*Rhodospirillales* Incertae Sedis	*Rhodospirillales*	*Alphaproteobacteria*	*Proteobacteria*	x					0.09
Ferrovibrio	*Rhodospirillaceae*	*Rhodospirillales*	*Alphaproteobacteria*	*Proteobacteria*				x		49.7
Limnohabitans	*Comamonadaceae*	*Burkholderiales*	*Betaproteobacteria*	*Proteobacteria*						20.97

Column I: excerpt of statistically significant abundant taxonomic clusters after addition of glyphosate in the treatment microcosm. Statistical significance (*** < 0.001, ** < 0.01, *<0.1) is based on Benjamini-Hochberg corrected *p*-values , tested with Wald test using R package DESeq2

Random Forest classifications were conducted similarly. *Parvibaculum* spp. performed well for the unfiltered data set with 92.2 % and was close to the 90.374 % for the filtered data set with a classification rate of 90.38 %, below the ANN's rate. *Massilia* spp. gained a correct classification of 89.1 % for the unfiltered and 86.1 % for the filtered data set, respectively. RF on single clusters performed better for the unfiltered data set.

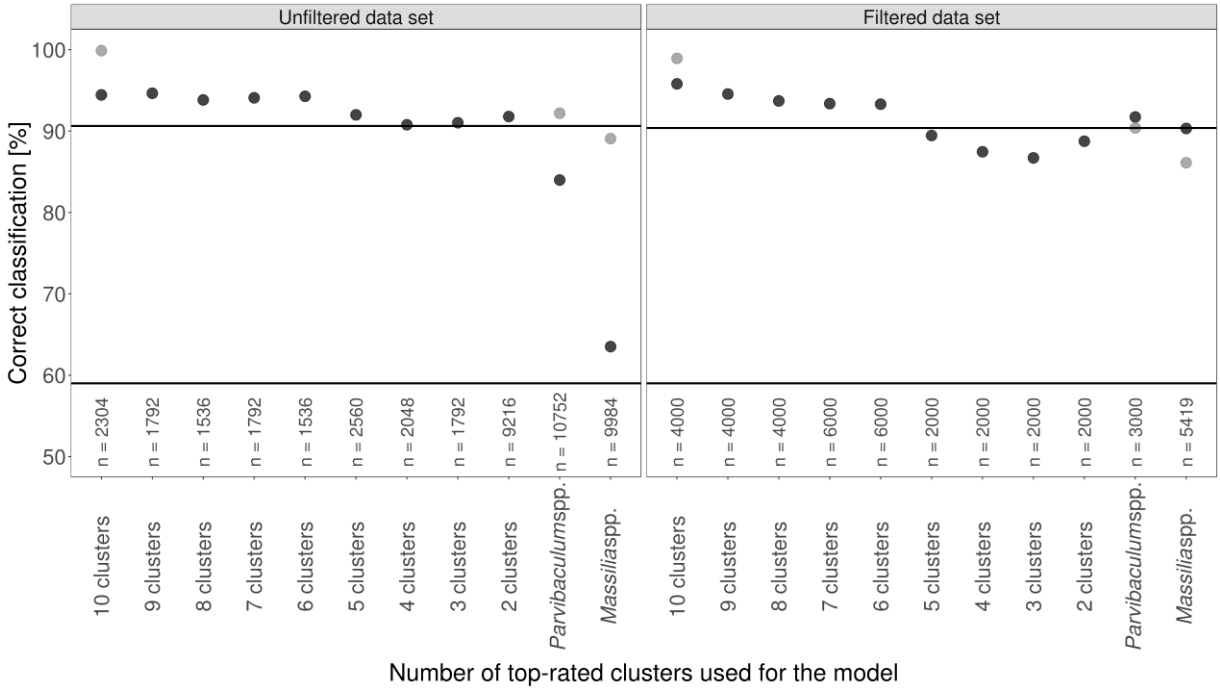


Figure 1.5: Classification rates achieved by using a top ranked selection of clusters. In black the ANN classifications, in grey the RF values. *n* is the number of classifications performed with the respective clusters by the ANN. The horizontal bar at 59% displays the classification achievable by pure guessing, the upper bar marks the threshold for a classification which both separates the microcosms and before and after glyphosate addition. Information on 10 clusters is sufficient to classify as well as using the full data set. Removing one cluster at a time from the input did not result in a linear decrease for the ANN. Depending on the cluster, one can provide sufficient data (*Parvibaculum* spp.) for the ANN and the classification was improved by the filtering step. RF was able to classify using *Parvibaculum* spp. in the unfiltered data set, but performance decreased using the filtered data.

1.3.4 Comparing the use of 16S rRNA gene - vs. 16S rRNA-derived data

After investigating the number of features used for classification, these approaches targeted the number of observations required. The unfiltered data set was used. The classification rate of the ANN decreased to 82.2 % if only 16S rRNA gene data was used and to 84.8 % for the 16S rRNA-derived data (Supplementary Material 1.4). Both values were within the range needed to distinguish between the two microcosms regardless of the glyphosate addition. Random Forest models showed that 16S rRNA gene data performed successful (96.6 %) and better than the 16S rRNA-derived data (92.5 %). Excluding half of the sampling time points, which was tested for the ANN, resulted in a classification of 82.1 %.

1.4 Discussion

Information collected over 16 time points from a microbial community assemblage obtained in a lab microcosm experiment in which glyphosate was applied as a disturbant was used to train an ANN for the classification of treated and control communities. Glyphosate is not considered

as strong a microbial stressor as, for example, toxic or antibiotic substances and, in fact, can be utilized as a nutrient or energy source by many microbes such that positively reacting clusters could be used for classification (McGrath et al., 1997; Hove-Jensen et al., 2014). The artificial neural network successfully distinguished between treated and untreated communities and demonstrated the general feasibility of the combined NGS-ANN approach. In particular, the ANN was required to separate community compositions of two independent microcosms with slightly different assemblages from the beginning of the experiment, which was easily achieved by standard ordination methods such as non-metric multidimensional scaling (Figure 1.1). In addition, the experimental design also demanded the ANN to identify traits present in the control-labelled samples from both microcosms and to separate those from the treated samples, which it successfully accomplished. However, the RF model employed as reference for machine learning performed better on this task.

1.4.1 A statistical approach to identifying decision-important clusters improved with fewer features

The subsequently applied random subsetting method for the ANN was developed because a systematic approach was not considered feasible for this study; any expedient selection of factors resulted in combinatorial explosion. The subsetting led to two important conclusions:

1) It was possible to stochastically identify and rank input features. This involved the identification of important taxonomic clusters to differentiate between the samples. Therefore, it could help determine indicator candidates for environmental monitoring purposes.

2) The required number of features could be significantly reduced, as demonstrated by the equally successful classification by only the top 10 ranked clusters of each data set. Using less than approximately 10 clusters might result in a loss of required information (Figure 1.3). This was shown by the non-linear decrease in classification rates. This indicated that each cluster may contribute a certain dimension of information; the conducted removal of clusters was based on their average classification within a subset, which may not reflect the value of these pieces of information. To conclude, it is presumably not worth the reduced computational costs to base classification on a single-digit number of features. Testing classifications based on a single node was a rather theoretical approach, as reducing the feature amount to 1 renders the node interaction of the ANN useless, could cause problems with the calculation of a split value for the RF, not to speak of the impossibility of bagging. The ANN using filtered single clusters was the test where the ANN outperformed the RF.

RF models provide several metrics as the decrease in Gini impurity; the numbers of times a feature was a root node or the mean minimum depth per feature to identify classification-relevant clusters. It was observed (data not shown) that the top 10 selection from the ANN provided as input to the RF slightly reduced the RF's performance. This hints towards a

differential utilization of information by RF and ANN. Overall, the selected features by both methods were largely the same and the RF classification was significantly improved.

The filtering of the data shortened the time needed for the ANN's subset ranking to stabilize, which might also depend on the size of the subset. In all comparisons between the two data sets, e.g., of subsets (Figure 1.2) containing specific clusters (Figure 1.3), the top 10 selections and the classification trained solely on *Massilia* spp. or *Parvibaculum* spp. abundance (Figure 1.5), proved a significant increase in the contribution to a successful classification if the data was filtered from low abundant clusters. We encourage the application of a filtering step on microbial community composition data sets for similar approaches if applying an ANN. All RF based models, however, performed slightly worse when processing the filtered data compared to their accuracy on unfiltered data. The added noise of the technical replicates might be exactly what the ANN requires for generalization, the reduced feature noise possibly supported identifying important clusters. The RF did not gain significant improvements by the removal of low abundant clusters, as it was shown that the filtering step altered the top 10 selection of the RF only marginally and its performance decreased. This is another hint that the RF processes the community composition differently than ANN. Solely random subsampling led on average to unsatisfactory classification results by the ANN, which indicated that each participating cluster may also contribute its information in a misleading way, e.g., clusters that were unresponsive to glyphosate or empty clusters. However, a few subsets in both data sets breached the classification threshold of 90.625 %. In Figure 1.3, the contribution of certain clusters towards the classification performance of their subset is displayed. It can be assumed that evenly distributed classifications ranging between many percentages (*Gallaecimonas* spp. in unfiltered data) indicate that the cluster within this subset is not a dominant contributor of information; hence, the classification success rather depends on the other members of the subset. A distinct range of classification within a subset (filtered data, *Parvibaculum* spp.) might rather indicate a decisive cluster of a subset. The clusters *Parvibaculum* spp. and *Massilia* spp. were part of both top 10 lists. How *Massilia* spp. supported a classification, while being present in only one microcosm, is a matter of speculation. It may be that the abundance of *Massilia* spp. separates the microcosms while another cluster contributes the information needed to distinguish between before and after the glyphosate addition (Figure 1.4). This is supported by the sharp fall in classification rate when *Massilia* spp. were used as a single input feature, whereas the ANN solely trained on *Parvibaculum* spp. succeeded. It however does not explain the increase in classification rate increase from the unfiltered to the filtered data set. The bioclimatic model of Larsen et al. (2012) predicted bacterial community assemblages on order level. It incorporated 16S rRNA gene pyrosequencing data in a preceding data analysis step before applying an ANN to spatially and temporally extrapolate microbial diversity. Their findings suggested that the strength of the ANNs is to combine the information on abundances

of multiple taxa. Another interesting finding of our study was the identification of the practically "empty" cluster *Nesiotobacter* spp. as a member of the top 10 clusters (Supplementary Material 1.3i). Its appearance in only the unfiltered data suggests that it represented the many "empty" clusters that were also part of the "treated" tables. It could also be coincidentally part of the well-performing subsets. It points out to the possible issue of finding an appropriate subset size and the required amount of subset samplings for a given data set. Environmental communities resolved onto OTU level would harbor even more features, making this an important computational issue. The abovementioned Gini impurity to identify important features in RF is of advantage here, as they directly assess the information value of a feature per split in a decision tree. The ANN instead immediately combines information from the features to find more generalized, abstract interactions and therefore, after the first fully connected layer, the importance of a specific input features is hard to assess. Currently, much research effort is targeted towards understanding and visualizing why a neural net decides or recognizes as it does. It has to be stressed that both machine learning techniques purely correlate the provided data with the provided output, hence, no causal relationship can be concluded. This was displayed by ANN findings about *Limnohabitans* spp. (Supplementary Material 1.3r), a feature increasing in the control microcosm after day 0 and therefore helpful for classification, but probably not linked to glyphosate treatment. The advantage of community composition data (or OMICS data in general) is that each feature has an intrinsic information value which is independent from the context. It is more helpful to find that, e.g., *Parvibaculum* spp. or gene *phnJ* is important for classification compared to "the pixel at position 2,2" in visual pattern recognition. However, the most prominent clusters increasing in abundance specifically after glyphosate addition - *Parvibaculum* spp., *Gallaecimonas* spp. and *Hyphomonas* spp. - were so far not mentioned in literature to be related to glyphosate degradation. In contrast, Smith et al. (2015) detected taxa important for the prediction of uranium and oil, which were known to interact with these contaminants. It should be considered that the abundance changes might not be directly caused by glyphosate utilization of the same cluster, but e.g. by a metabolism product or the suppression of a competing taxon.

1.4.2 More observations should be generated

In general, more complex data sets require more general models to fit the data. If the data sets are also noisy, characterized by a larger variance, even more training data and - more specific - more observations to adjust the weights of the ANN are necessary. This was demonstrated in experiments with decreased numbers of taxonomic tables. The use of only 16S rRNA- or 16S rRNA gene-derived data as well as only half of the time points reduced the classification rate to below 90.625 % (Supplementary Material 1.4), with 16S rRNA gene data performing better than 16S rRNA data. It is possible that the decrease was due to the small sample size. RF was not limited by the sample size and accurately classified the microbial communities

based on each 32 samples. The RF findings indicated contrastingly that 16S rRNA, the expressed 16S rRNA gene as potential activity measure, is a better proxy of glyphosate response. These findings indicate that the present number of samples is close to the limit for maintaining a correct classification for the ANN. Fortunately, if such an ANN could be implemented in monitoring programs, additional data would be generated at each monitoring event, which can be progressively included into the model such that the observations-to-features ratio is continuously improved.

Different types of neural networks could be explored to compare which architecture achieves the most for a given data set. A more sophisticated CNN model including a customized layer for phylogenetic similarity was presented in Fioravanti et al. (2018). It should be mentioned that Yu et al. (2019) reviewed neural nets in various omics applications and found that basic architectures performed better.

1.4.3 The outcome of the ANN was confirmed by bioinformatic analysis

The samples from the same glyphosate incubation experiment were also examined with bioinformatics tools for a second manuscript, guided by slightly different hypotheses. From the 20 unique clusters in the unfiltered and filtered top 10 clusters established by ANN and RF (Table 1.1), seven were also identified by the R package DESeq2, which tests for statistically significant differences in abundance and was developed for NGS data (Love et al., 2014). It was applied to compare the cluster abundances before and after the glyphosate pulse. Subjecting the DESeq2 input to a filtering step excluded some of the clusters identified by the ANN. This step was thought to improve the reliability of statistics. The data suggested that the ANN can profit from low abundant clusters as well (Table 1.1). A combination of traditional bioinformatic or molecular ecology approaches and ANN technologies seems practical.

1.4.4 Concluding further steps in the application of ANN with NGS

While the results presented by this study are promising, the community assemblage data were still low-dimensional, containing information on the relative abundance per cluster, time point, glyphosate treatment, technical replication, and nucleic acid analyzed. The samples were treated as independent observations. To make use of the capacity and potential of machine learning technologies, various aspects can be targeted for improvement. For example, the number of dimensions could be increased by adding meta data, often available from standardized monitoring campaigns, to the input. Temporal or spatial information should be included. At this step, a shift from classification to regression could be appropriate.

Both techniques proved to be powerful and should not be seen as competitive, but as two different means to process the same information. Therefore, the hypothesis and the data characteristics should inform the choice of which technique to use. As a rule of thumb, we suggest starting with Random Forest models, which worked off-the-shelf and provided rapid

results. If the data sets grow larger, more complex and noisier, basic ANN models can be tested. For further tuning, the variety of neural network architectures provide all means of hyperparameter control and abstraction rate. However, the first aim here was a robust ANN that can achieve a correct classification based only on sequencing-derived data. OTUs, or in this case, clusters, inherit a vast amount of functions, and their predictive ability is therefore limited to specific scenarios and environments. This can be leveraged by using available 16S rRNA gene amplicon data sets, e.g., from the Baltic Sea with known meta data. ANNs could be trained on these data as “standard”, and if a new sample is not classified as “standard”, it should be investigated to identify the reason for the deviation. The monitoring would not only help to survey the environmental state of the Baltic Sea but would also serve as a steadily growing data resource. It was just demonstrated herein that the knowledge about degradation abilities of isolates is sparse relative to the number of strains and pollutants. This data resource would support the identification of taxa linked to a contamination. Since taxonomic resolution achieved by amplicon sequencing is limited, whereas functions can be strain-specific, the next logical step is to use data from metagenomic and metatranscriptomic sequencing. This would complete the efforts undergone by He et al. (2018) based on microarray data but still include the phylogenetic and taxonomic dimensions. With a function- versus phylogeny-targeted approach, the features would be the abundance of genes or their transcripts. The general principle was already demonstrated by Lin et al. (2017) who, employing a CNN, improved the assignment of single-cell RNAseq reads to their cell types of origin. Although our suggested approach would necessitate more training data, the approach is feasible, as sequencing costs are decreasing, and many suited data sets for training, validation and testing are publicly available. It must be stated, that the herein discussed models are based on microcosm data, which intrinsically is an abstraction with regard to the environmental situation. However, this use of microcosm data allowed us to isolate the influence of glyphosate on a microbial community. To transfer a supervised machine learning approach to environmental monitoring, the training data set must be sufficient to explain or at least correlate observed changes in the microbial communities with the vast spectrum of variables such as salinity, temperature, nutrient concentrations, pH, anthropogenic influences and so on. Machine learning offers the ability to detect links between features that otherwise might have gone unrecognized, but the demand for contextual data is even more essential to utilize such models. There are numerous fields of application for our findings and include the monitoring of specific events by classifying (e.g., contamination events or algal blooms), as well as more generally fitting microbial community compositions via regressions (e.g., a salinity, temperature or temporal gradients) and describing a set of indicative organisms for a given classification. Supplementary data to this article can be found online at https:// doi.org/10.1016/j.marpolbul.2019.110530

Chapter II

A glyphosate pulse to brackish long-term microcosms has a greater impact on the microbial diversity and abundance of planktonic than of biofilm assemblages

The following chapter was published in the journal *Frontiers in Marine Science* as:

René Janßen, Wael Skeff, Johannes Werner, Marisa A. Wirth, Bernd Kreikemeyer, Detlef Schulz-Bull, and Matthias Labrenz (2019). A Glyphosate Pulse to Brackish Long-Term Microcosms Has a Greater Impact on the Microbial Diversity and Abundance of Planktonic Than of Biofilm Assemblages. Front. Mar. Sci. 6:758. doi: 10.3389/fmars.2019.00758

Declaration of author contributions:

René Janßen designed and conducted the microcosm experiment, performed laboratory work and bioinformatically processed and analyzed the data.

Wael Skeff designed and conducted the adsorption experiment and measured glyphosate and AMPA for both experiments.

Marisa A. Wirth developed a method to measure sarcosine and re-measured samples.

Johannes Werner reviewed code and performed the metagenomic analyses.

René Janßen, Johannes Werner, Wael Skeff and Matthias Labrenz discussed the data.

René Janßen drafted the manuscript, Matthias Labrenz critically commented on the manuscript and redrafted parts of it, Johannes Werner, Detlef Schulz-Bull and Bernd Kreikemeyer critically commented on the manuscript.

René Janßen’s contribution to the written manuscript was ~ 80 %.

Abstract

The widespread herbicide glyphosate has been detected in aquatic coastal zones of the southern Baltic Sea. We monitored community dynamics in glyphosate-impacted chemostats for 20 weeks to evaluate the potential impact of the herbicide on free-living and biofilm-associated bacterial community assemblages in a brackish ecosystem. A HPLC-MS/MS method was developed to measure glyphosate, aminomethylphosphonic acid and sarcosine concentrations within a brackish matrix. These concentrations were analyzed weekly, together with prokaryotic succession, determined by total cell counts and next generation 16S rRNA (gene) amplicon sequencing. Shotgun metagenomics provided insights into the glyphosate degradation potential of the microbial communities. Temporal increases in total cell counts, bacterial diversity and the abundances of distinct bacterial operational taxonomic units were identified in the water column. Biofilm communities proved to be less affected than pelagic ones, but their responses were of longer duration. The increase of glyphosate oxidoreductase (*gox*) and *thiO* gene as well as the *phn* operon abundance indicated glyphosate degradation by first the aminomethylphosphonic acid pathway and possibly a subsequent cleavage of the C-P bond. However, although glyphosate concentrations were reduced by 99 %, 1 µM of the herbicide remained until the end of the experiment. Thus, when present at low concentrations, glyphosate may evade bacterial degradation and persist in Baltic Sea waters.

2.1 Introduction

Microorganisms are ubiquitous on Earth and respond rapidly to environmental changes. The majority of microorganisms live within biofilms, which promote high cell abundances and activities (Costerton et al., 1995). In mature biofilms, extracellular polymeric substances produced by resident species give rise to a distinct three dimensional structure. That way microorganisms are protected from disturbances that for planktonic cells or even higher organisms induce toxicity and other forms of stress (Davey and O'Toole, 2000; Reese et al., 2016). However, biofilms are not completely invulnerable (Qu et al., 2017), as evidenced by changes in their assemblages in response to a wide range of disturbances.

A potential environmental stressor is glyphosate, which has been in use since the 1970s. Following assessments demonstrating its relatively low environmental toxicity, it has become the most widely produced and sold herbicide worldwide. However, as a synthetic combination of glycine and a phosphate residue, coupled to form a stable phosphonate, glyphosate provides carbon (C), nitrogen (N), and phosphorus (P) for bacteria and fungi (Lipok et al., 2007; Duke and Powles, 2008). Two major routes of glyphosate biodegradation have been described according to their first respective intermediate: the sarcosine pathway and the aminomethylphosphonic acid (AMPA) pathway, encoded mainly by the *phn* operon and the glyphosate oxidoreductase (*gox*) gene, respectively. The *phn* operon encodes a C-P lyase, whose activity makes the P component of phosphonate bioavailable. In the AMPA pathway, glyphosate is cleaved at the C-N bond, resulting in AMPA and glyoxylate. An alternative pathway to yield AMPA from glyphosate was discovered with the enzyme glycine oxidase encoded by *thiO*. However, this enzyme possesses an unspecific Km of 87 mM for glyphosate, compared to 0.6 mM for glycine (Pedotti et al., 2009).

Glyphosate has been detected in marine and freshwater systems (Van Bruggen et al., 2018; Carles et al., 2019), representing a disturbance to microbial communities at concentrations upwards of 5.92 nM (Stachowski-Haberkorn et al., 2008). Moreover, its dissipation is enhanced by biofilms, probably due to their adsorption capacities (Klátyik et al., 2017). The presence of glyphosate in the brackish Baltic Sea from agricultural runoff has been reported (Skeff et al., 2015), but the effects of the herbicide on its ecosystems are as yet unknown. The Baltic Sea is known for elevated contamination levels and monitoring of the environmental state is mandatory (HELCOM, 2018). Thus, the aim of this study was to investigate the impact of glyphosate on the state and succession of bacterial community assemblages in a Baltic-Sea-like environment. Potential effects were compared between free-living and biofilm communities, as biofilm communities are expected to be more resilient. Furthermore, the potential for and means of biodegradation, as well as the possibly involved OTUs, were analyzed to evaluate the fate of glyphosate entering the Baltic Sea.

2.2 Material and methods

2.2.1 Experimental setup

2.2.1.1 Microcosm experiment

The experiment was conducted in two 12 L (20 × 30 × 20 cm) microcosms (Rebie Aquaristik, Bielefeld, Germany) made of float glass plates sealed with silicone glue. The microcosms were filled with 2 kg of combusted quartz sand as hard substrate, 8 L artificial brackish water (ABW) amended with casamino acids as liquid medium (modified after Bruns et al. (2002) and combusted GF/F microfiber filters (Ø 47 mm, Whatman, Little Chalfont, United Kingdom) as collectible, inert biofilm substrate. An air pump aerated and mixed the system continuously. The microcosms were incubated with a Baltic Sea-derived inoculum and the 140-day experiment started with an equilibration period from day −69 until day 0 to allow biofilm to form and mature. On day −31 the system switched from batch to continuous cultivation mode with an average efflux rate of 475–489 $mL \cdot d^{-1}$. During the whole period microbial succession in both microcosms was monitored. On day 0, a sterile-filtrated glyphosate solution (final concentration of 82.45 µM; Dr. Ehrenstorfer, Augsburg, Germany) was syringe-injected into the water column of the treatment microcosm and dispersed throughout by manual stirring. Monitoring went on until day +71. For further details on experimental procedures see Janßen et al. (2019).

2.2.1.2 Prevention of glyphosate adsorption to abiotic surfaces

Glyphosate can adsorb to glass or sediment surfaces (Bergström et al., 2011; Huang and Zhang, 2011) and might also adhere to biofilms. Adsorption may affect not only glyphosate degradation in the liquid phase but also act as a glyphosate reservoir during incubations. However, a surface adsorption test performed prior to the start of our experiment showed stable glyphosate concentrations in the water column of the glyphosate-containing microcosms (Supplementary Material 2.1).

2.2.2 Sampling procedure

Five-mL water samples for glyphosate, AMPA, sarcosine/*L*-alanine and nutrient analyses were stored at −20°C without further treatment. For nucleic acid extraction and subsequent next-generation sequencing (NGS) of planktonic cells, 100 mL of water was filtered through 0.22-µM GVWP filters in three replicates. For the analysis of biofilm communities, three overgrown GF/F filters were selected with sterile tweezers. The total data set consisted of 287 samples, with water samples covering 16 time points (days −25, −7, 0, +3, +7, +10, +14, +17, +22, +29, +36, +43, +50, +57, +64, +71) and biofilm filters eight time points (days −7, 0, +7, +17, +29, +43, +57, +71). Detailed meta-information describing the samples is provided in Supplementary Material 2.2. The filters were shock frozen in liquid nitrogen and stored at −80°C until their use for DNA/RNA extractions. Planktonic cell counts were determined in 1-

mL water samples fixed with 1/10 $v \cdot v^{-1}$ formol (37 %, sterile filtered, Rotipuran p.a. ACS, Carl Roth GmbH, Karlsruhe, Germany), incubated for at least 2 h at room temperature or overnight at 4°C and processed within 24 h. For C and N analyses, 100 mL of water was collected on day +71.

2.2.3 Determination of total cell counts

Water column cell counts were determined by 4′,6-diamidino-2-phenylindole (DAPI; Applichem GmbH, Darmstadt, Germany) staining according to Porter and Feig (1980). To ensure that cells on the filter surfaces were evenly distributed, the cells on the filter were diluted, if necessary, using sterile ABW. The cells obtained by filtering 50–500 µL of water on a Cyclopore filter (PC BLK, 25 mm, 0.2 µm, Whatman, Maidstone, United Kingdom) were stained with 10 mg $DAPI \cdot L^{-1}$ for 3 min and embedded using AF1 (Citifluor Ltd, London, United Kingdom) and Vectashield (H 1000, Vector Laboratories, Burlington, CA, United States) at a 7:1 ratio. Total cell counts were determined in triplicate samples using an Axio Lab. A1 equipped with a N-Achroplan 100x oil dispersion objective (both Carl Zeiss AG, Göttingen, Germany). Twenty small quadrats were counted in 25 different fields of view per filter.

2.2.4 Significance testing applied to total cell counts

To test for a statistically significant change in total cell counts after the addition of glyphosate, the cell counts prior to (days −7 to +3) and after (days +28 to +36) the cell number increase were combined and compared with the counts from days in which cell numbers increased (days +7 to +22). A second comparison was performed between treatment and control microcosms for the cell counts from day +7 to +22 only. Total cell counts were analyzed in triplicate samples using a two tailed t-test for two heteroscedastic samples. Significant changes ($p < 0.05$) are marked with * in Figure 2.1A.

2.2.5 Nutrient analysis

To understand the nutritional relevance of glyphosate, particulate organic nitrogen and carbon (POC/PON) concentrations were analyzed using an vario Micro element analyzer (Elementar Analysensysteme GmbH, Langenselbold, Germany), and dissolved organic carbon and nitrogen (DOC/DON) concentrations using a Shimadzu TOC-V + TNM1 analyzer (Duisburg, Germany). Dissolved inorganic phosphorus (DIP) was measured following the method of Grasshoff et al. (1999).

2.2.6 Glyphosate and AMPA analysis

Glyphosate and AMPA analyses followed the procedure of Skeff et al. (2015, 2016). Internal standards of glyphosate (1-2-$^{13}C_2$ ^{15}N glyphosate) and AMPA (^{13}C ^{15}N AMPA) were prepared in the same sample matrices and added to the samples. The samples were adjusted to pH 9 by the addition of 100 µL of borate buffer and then derivatized by treatment with 100 µL of 19.8 mM FMOC-Cl in acetonitrile. After 4 h of incubation at 21°C, the derivatized samples were

filtered through a 0.45-µm Phenex-RC 15-mm syringe filter and subjected to LC–MS/MS. The target compounds were analyzed using an Accela HPLC system connected to a TSQ Vantage triple quadrupole mass analyzer with a heated electrospray ionization source interface. Chromatographic reversed-phase separation was achieved on a Gemini-NX C18 column coupled to a Gemini-NX Security Guard cartridge. The samples were eluted gradually from the column with (a) a 2 mM ammonium hydrogen carbonate buffer and ammonia solution (32 %, $v \cdot v^{-1}$) at pH 9 and (b) acetonitrile. Before the analysis, the instrument was calibrated for the target substances using the same sample matrices. Each compound, including the internal standard, was scanned for two transitions in selected reaction monitoring mode. The most abundant transition was used for quantification and the other transition for confirmation.

Additional measurements for AMPA and sarcosine were carried out after an initial evaluation of the data. The applied method generally followed the procedure described above, with the following differences: After derivatization of the samples, 1 mL of dichloromethane was added to the mixture to extract the remaining FMOC-Cl. Samples were shaken and then centrifuged for 10 min at 1000 rpm. The supernatant was removed and transferred into a vial for analysis. Chromatographic separation and mass spectrometric detection was carried out as described above, but with a LC-2040C Nexera-I and a triple quadrupole mass spectrometer LCMS-8060 as also described in Wirth et al. (2019). Compounds were detected through SRM events, as described above. Sarcosine has the same MS fragments and retention time as *L*-alanine, since the two compounds are isomers. Thus, they could not be differentiated with the utilized method. To acquire evidence for the presence or absence of sarcosine in the samples, comparative measurements between samples from both microcosms were conducted, since the *L*-alanine concentration should be identical.

2.2.7 Nucleic acid extraction and sequencing

The kit-based extraction of nucleic acids from free-living bacteria and subsequent DNAse digestion of the RNA extracts were performed according to Bennke et al. (2018). Biofilm samples were extracted using the phenol-chloroform method described in Weinbauer et al. (2002). cDNA synthesis was performed using 20 ng of DNA-free total RNA as the input for the MultiScribe (Fisher Scientific GmbH, Germany) reverse transcriptase system using the reverse primer 1492r (5′ TACGGYTACCTTGTTACGACTT, Lane, 1991). Illumina amplicon sequencing was prepared as described in Bennke et al. (2018). The V3-V4 region of the 16S rRNA gene was targeted using the primer set 341f-805r (forward: CCTACGGGNGGCWGCAG, reverse: GACTACHVGGGTATCTAATCC, (Herlemann et al., 2011)). Indexed amplicon libraries were pooled to a concentration of 4 µM. The PhiX control was spiked into the library pools at a concentration of 10 %. Each final library pool (4 pM) was subjected to one of three consecutive individual paired-end sequencing runs using 600 cycle V3 chemistry kits on an Illumina MiSeq.

2.2.8 Bioinformatic and statistical analysis of the amplicon data

Amplicon read processing and annotation were conducted using Mothur v. 1.39.5 (Schloss et al., 2009). Sequences were combined in a pre-cluster step if there were less than 2 mismatches. Chimeras were removed using VSEARCH (Rognes et al., 2016). OTUs were picked based on a 98 % similarity threshold. When counting the number of OTUs, singletons were ignored, but not removed from the data set. OTUs were only removed where mentioned and all parameters are deposited in the Github repository listed in the data availability statement.

The operational taxonomic unit (OTU) and taxonomy table were imported into R v. 3.5.1 (R Core Team, 2018) and analyzed using phyloseq v. 1.26.0 (McMurdie and Holmes, 2013), ggplot2 v. 3.1.0 (Wickham, 2016) and DESeq2 v. 1.22.1 (Love et al., 2014). Taxonomic annotation of the data presented herein was accomplished using the Silva release 132 (Yilmaz et al., 2014), including the taxonomic changes proposed by Parks et al. (2018).

Basic information on the amplicon sequencing-based approaches is provided in Supplementary Material 2.3, including the MiSeq run statistics, sequencing depth and average sequence length in the 16S complementary rRNA and 16S rRNA gene libraries.

The composition of the microbial communities was plotted enforcing a relative abundance cut-off value of 0.15 % at order level to reduce the legend size. To identify OTUs whose abundance changed after glyphosate addition, unfiltered 16S rRNA gene and 16S rRNA OTU tables were used separately as input for DESeq2, as suggested by (McMurdie and Holmes, 2014). DESeq2 performed the Wald test on two time points (in three technical replicates) before glyphosate addition versus five time points directly thereafter. For the less-frequent biofilm sampling, the time span was the same, resulting in comparisons of two time points before versus two time points immediately after glyphosate addition. The abundances of selected OTUs were plotted. The relative abundances of the OTUs in treatment and control microcosms were compared manually to identify those OTUs that responded to glyphosate.

The similarity of microbial communities was visualized in non-metric multidimensional scaling (NMDS) analyses based on Bray–Curtis dissimilarities. Relative abundances were used as input, square-root-transformed and Wisconsin double-standardized. The ordination with the lowest stress was determined based on 100 runs. OTUs with at least three reads were included. OTU tables for the Chao1 richness estimate and Shannon index included singletons. A t-test was applied to analyze the significance of a change in α-diversity after glyphosate addition and was performed for all sample subsets from day −22 to day 0 vs. day +3 to day +17.

To include the concentration of glyphosate into the ordination, canonical correspondence analysis (CCA) and redundancy analysis (RDA) were performed within phyloseq using its ordinate function. The input data was as described for NMDS and glyphosate concentration was the constraint. The resulting plots are shown in Supplementary Material 2.4.

2.2.9 Metagenomic analysis

For metagenomic analyses, technical replicates of DNA extracts were pooled. Metagenomic reads of seven treatment and three control microcosm water-column samples were generated by a full run on an Illumina Nextseq500 (LGC Genomics GmbH, Berlin, Germany). Reads were quality checked using FastQC v. 0.11.71 and trimmed with Trimmomatic v. 0.38 (Bolger et al., 2014). The individual samples were merged and co-assembled using MEGAHIT v. 1.1.3 (Li et al., 2016) with the k-mer list 21, 25, 29, 33, 37, 41, 45, 49, 53, 57, 61, 65, 69, 73, 77, 81, 85, 89, 93, 97, and 99. The genes were predicted and functionally annotated using Prokka v. 1.13.0 (Seemann, 2014). For gene quantification, the reads of the individual samples were mapped on the assembled contigs using Kallisto v. 0.44.0 (Bray et al., 2016).

2.2.10 Functional tree calculation

Correlations between the abundances of OTUs and glyphosate degradation genes were identified. Protein sequences of organisms related to the OTUs identified in this study were downloaded from UniProt (Bateman et al., 2017). The corresponding genes identified in the assembled metagenome were translated and added to this sequence set. After the removal of exact duplicates using CD-Hit auxtools v.4.6.8 (Fu et al., 2012), a multiple sequence alignment was built using Mafft v. 7.407 (Katoh and Standley, 2013). A phylogenetic tree was calculated using RAxML v. 8.2.12 (Stamatakis, 2014), with “PROTCATAUTO” as the amino acid substitution model, and plotted together with the respective abundances using R package ggtree v. 1.8.2 (Yu et al., 2018). This workflow was implemented in Nextflow v. 18.10.1 (Di Tommaso et al., 2017).

No *gox* genes were annotated in the metagenomes. Instead, a sequence-based approach was used: reference sequences of *gox* were downloaded from UniProt and GenBank to create a DIAMOND database (v. 0.9; Buchfink et al., 2015). The metagenomic sequences were blasted against the DIAMOND database (e-value of $1E^{-8}$, sequence identity ≥ 40 %, query coverage ≥ 70 %) and eventually phylogenetic trees with the corresponding abundance were plotted as described above.

2.3 Results

2.3.1 Total cell counts, glyphosate and AMPA concentrations and nutrients

Total cell counts in the water column were in the range of 2–4 × 10^7 cells·mL^{-1} both in the treatment and control water samples (Figure 2.1A). Following glyphosate addition, they

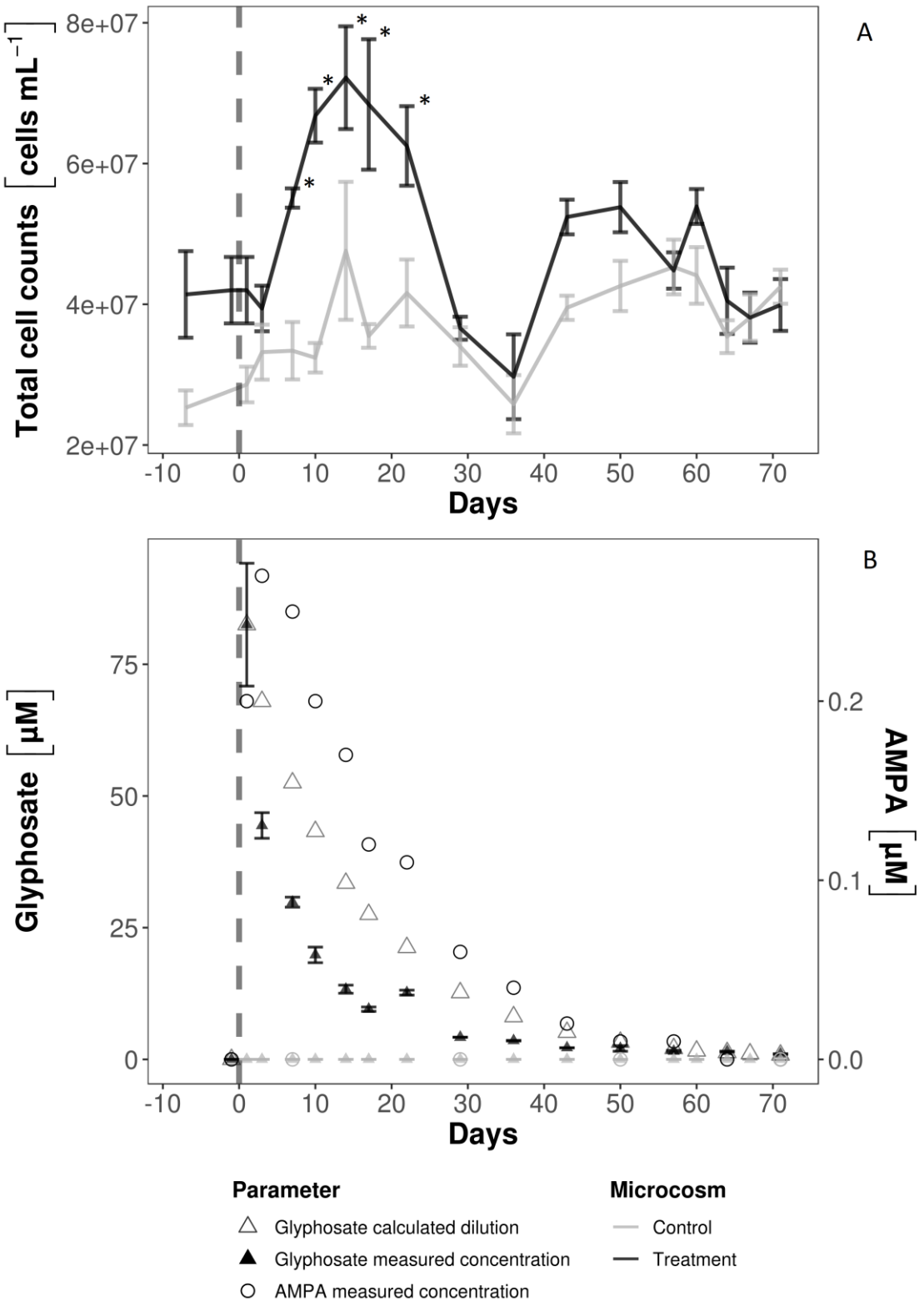


Figure 2.1: Total cell counts (A) and glyphosate and AMPA (B) concentrations in the water column of the microcosms. A: Total cell counts after glyphosate addition at day 0 increased significantly (*) compared to cell numbers before and later after herbicide addition ($p < 0.001$) and to the cell counts of the control microcosms during the same time period ($p < 0.001$). B: On day 0, the concentration was measured before and after glyphosate addition, AMPA was already present in the first sample takes 4 h afterwards, though not at its highest concentration. The decrease in the calculated glyphosate concentration was slower than the measured values. Note the different scales for glyphosate and AMPA and that at the end of the experiment glyphosate persisted at a concentration > 1 µM.

increased significantly, up to 7 × 10^7 cells·mL^{-1}, and remained elevated over the following 14-day period, during which the decrease in glyphosate was the strongest (Figure 2.1B). Based on the chemostat's volume and flow rate, the theoretical glyphosate concentration after approximately 60 days of incubation was close to zero. With a starting glyphosate concentration of 82.45 µM at day 0, the measured glyphosate concentrations, especially within the first two weeks of incubation, were 18–24 µM lower than the theoretical values. The results of the adsorption test (Supplementary Material 2.1) suggested that glyphosate was neither incorporated into biofilms nor adsorbed onto surfaces under our experimental conditions. After 71 days, glyphosate concentrations were reduced by 99 %. AMPA was detected as soon as 4 h after addition in the first sample. 3 days later AMPA concentrations ranged from 0.27 µM to below the detection limit (LOD) by day +64 and +71. The highest ratio of AMPA to glyphosate was 1.35 % on day +29. Peaks representing the isomers sarcosine and *L*-alanine could also be detected but were not reliably quantifiable, as their concentration (−0.017 to 0.016 µM) was close to the LOD. The peaks were present in both microcosms, before and after the addition of glyphosate.

The DIP concentration on day −69 was 15 µM, and on day 0 before and after glyphosate addition 23.3 and 24 µM, respectively. On day +71, at the end of the experiment, it declined to 16 µM. DOC and DN concentrations in the microcosms on day +71 were 80,000 and 20,000 µM, respectively. The resulting DOC:DN:DIP ratios were 238:56:1 for 24 µM DIP and 380:90:1 for 15 µM DIP.

2.3.2 16S rRNA and rRNA gene based community compositions

Among the 12,852 OTUs with more than one read, 10,692 originated from the water column. Two thousand nine hundred and three OTUs stem from the biofilm and 743 OTUs were present in both habitats. Planktonic 16S rRNA was roughly twice as rich in OTUs as either the planktonic 16S rRNA genes or the biofilm communities (Supplementary Material 2.2). Based on the number of reads, free-living (Figure 2.2) and biofilm (Supplementary Material 2.5) microbial community assemblages consisted almost exclusively of Proteobacteria, mainly Alpha- and *Gammaproteobacteria*. After the glyphosate pulse *Alphaproteobacteria* eventually comprised > 90 % of the bacterial assemblages in the treatment microcosm. Therein, *Rhizobiales* and *Rhodospirillales* represented large and increasing portions thereof.

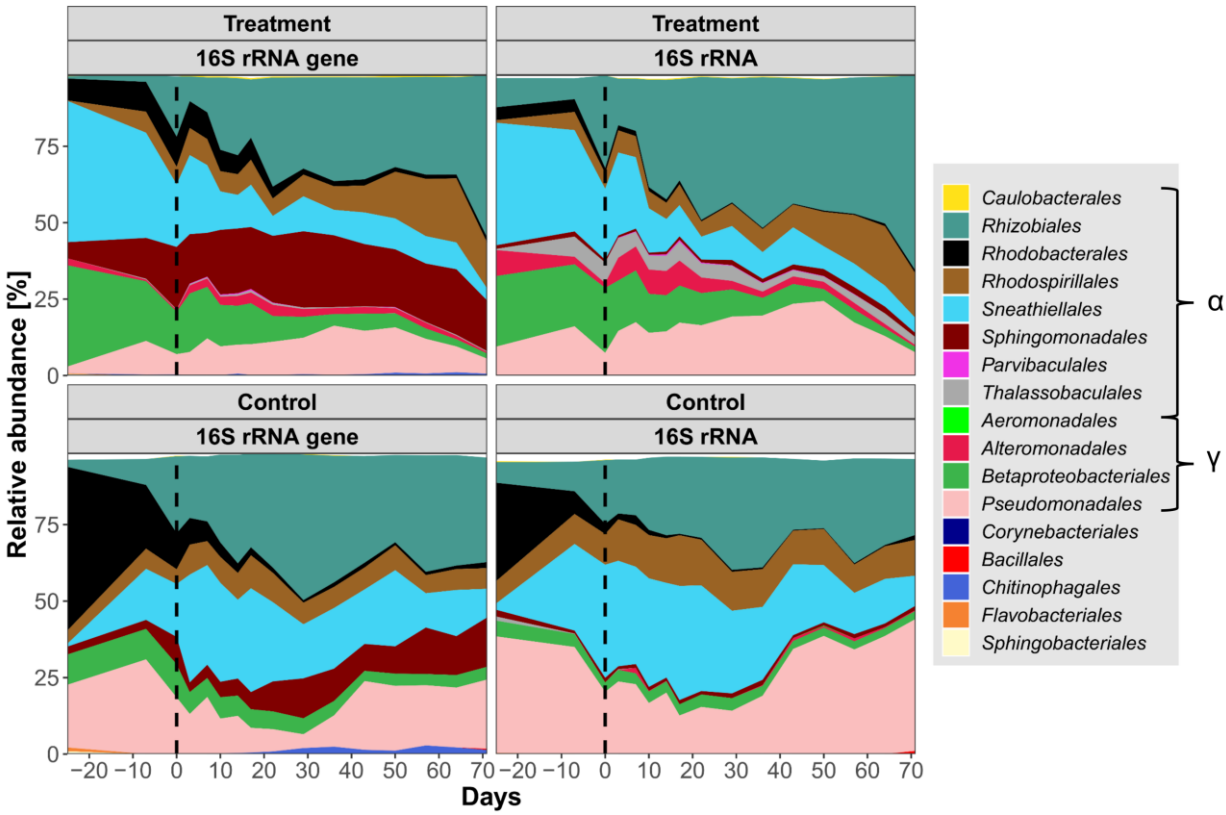


Figure 2.2: Relative planktonic community composition in the treatment and control microcosms based on 16S rRNA gene and 16S rRNA abundances. Taxa were cumulated on the order level, sorted by class. α = *Alphaproteobacteria*, γ = *Gammaproteobacteria*. All orders with a relative abundance > 0.15 % are displayed. Glyphosate addition is indicated by a vertical dashed line. Notice the dominance of *Proteobacteria* and the overall increase of planktonic *Alphaproteobacteria*.

Unclassified *Rhizobiales* OTU 1 was the most abundant OTU, up to 84 % in the biofilm 16S rRNA (Supplementary Material 2.6). However, *Pseudomonas* OTU 7 reads (Supplementary Material 2.6) represented up to 25 % of the 16S rRNA community in the water column of the treatment microcosm. *Pseudomonas* OTU 7 increased in abundance after the glyphosate pulse together with *Alteromonadales*, which includes the genus *Gallaecimonas*. In total, the OTUs covered > 320 genera, with 280 genera represented by 1–10 OTUs each. Ten very abundant genera, including *Hoeflea*, Ferrovibrio and undistinguished taxa (e.g., “unclassified” or “uncultured”), were represented by 100–318 OTUs. Based on a 0.01 % relative abundance threshold, the biofilm community consisted of 90 genera and the water column community of 75 genera, with 59 shared genera (Supplementary Material 2.2). The diversity of members of the *Gammaproteobacteria* was evidenced by the finding that 10,088 of the 12,852 OTUs belonged to *Pseudomonas*, although > 98 % of them were present at abundances of < 0.01 %.

2.3.3 NMDS ordination

Overall changes in 16S rRNA and 16S rRNA gene OTU composition were visualized via NMDS. Both 16S rRNA genes and 16S rRNA based assemblages were mainly arranged along the NMDS 2 axis, thus correlating with the sampling time (Figure 2.3). Samples from treatment and control microcosms were clearly separated. The PERMANOVA yielded *p*-values of < 0.001, with no significant differences in the dispersion of the control vs. the treatment groups

for all subsets. In general, the water column samples from the treatment microcosms were more similar along NMDS 2 than were the control microcosms. Water column 16S rRNA gene (stress 0.113) and 16S rRNA (stress 0.102) based community compositions produced similar ordinations. However, the 16S rRNA gene data formed two main clusters that were separated by the glyphosate pulse (day 0 vs. day +3). As long as glyphosate concentrations exceeded 5.92 µM (day +3 to day +22), the free-living community composition in the treatment microcosm formed a subcluster (red polygons in Figure 2.3). These observations also applied to the 16S rRNA data but the differences were less distinctive.

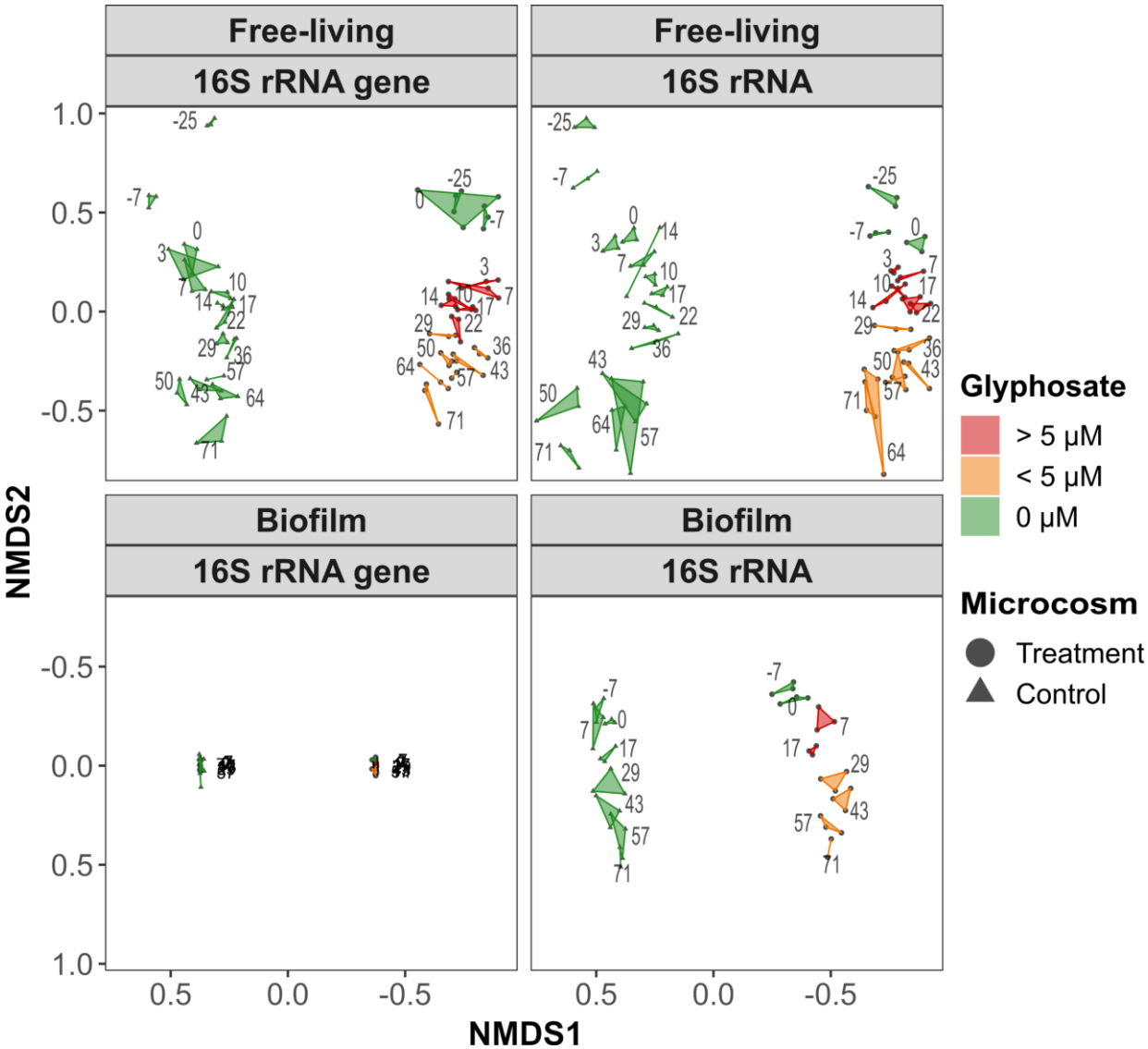


Figure 2.3: NMDS ordination plots based on the Bray-Curtis dissimilarity of the square-root-transformed and Wisconsin double-standardized 16S rRNA gene and 16S rRNA based community composition in the biofilms and water column. The axis direction for the biofilm ordinations was reversed to correspond to the pelagic orientation. The numbers refer to the sampling day; glyphosate was added before day 3. Technical replicates are connected by a polygon colored according to the measured glyphosate concentration.

Based on 16S rRNA gene data from the biofilm (stress 0.042), all communities remained stable. Biofilm community succession was generally less pronounced than that of planktonic communities, while control and treatment assemblages spanned a similar distance on NMDS2.

In contrast to the control samples, the overall biofilm 16S rRNA (stress 0.072) communities before and after glyphosate addition (days −7 to +7) formed distinguishable clusters.

2.3.4 Alpha diversity measures

The Shannon index was statistically assessed to test the impact of glyphosate on community diversity. Samples were grouped before and after day 0. For planktonic samples, the trend in the diversity of control microcosm communities was toward lower indices whereas in the treatment microcosm diversity increased temporarily after glyphosate addition, from a Shannon index of about 2.2 to > 2.5 (Figure 2.4). This development was again more pronounced for the 16S rRNA gene data, in which a significantly higher estimated richness (Chao1) after the pulse was also evident.

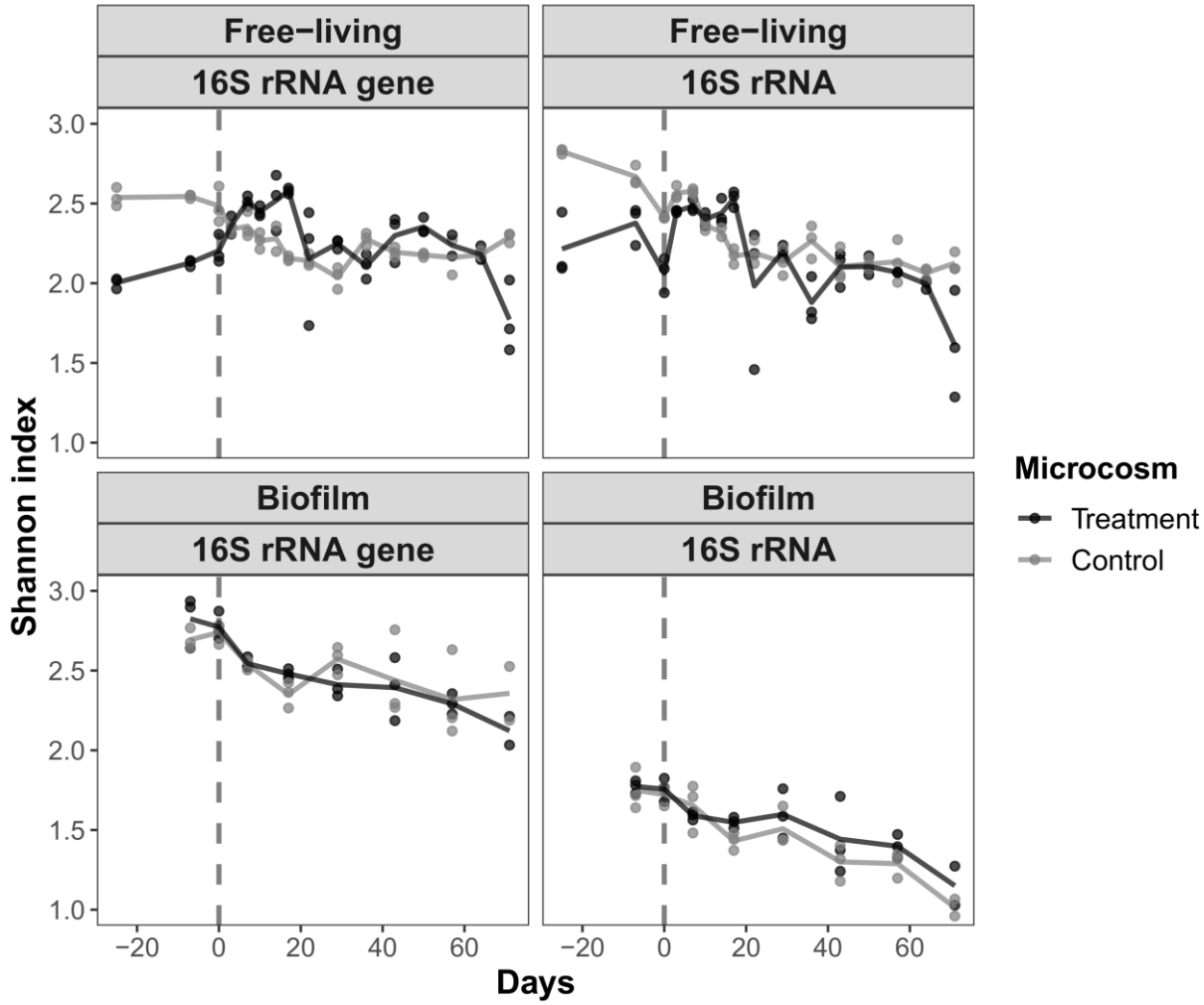


Figure 2.4: The change in α-diversity (Shannon index) of the free-living and biofilm communities according to the 16S rRNA gene and 16S rRNA data from the treatment and control microcosms. The vertical dashed line indicates the time of glyphosate addition. Samples obtained between the start of the experiment and day 0 (group 1) and from day +3 to day +17 (group 2) were compared in a t-test. The increase in the diversity of the free-living communities was significant.

By contrast, the Shannon index of the biofilm community samples decreased after day 0 regardless of the treatment, from approximately 2.7–2.3 (16S rRNA gene) and from 1.8–1.2 (16S rRNA), hence displaying a uniform mode of succession. A decrease in the diversity of the

planktonic control communities was also observed. Shannon indices between sample groups before and after day 0 were significant, ranging from a *p*-value of $1.04 \cdot 10^{-7}$ for changes in the 16S rRNA gene of the treated planktonic samples to 0.03 for changes in the control 16S rRNA of the biofilm (Supplementary Material 2.7).

2.3.5 OTUs increasing in abundance after glyphosate treatment

The succession in planktonic and biofilm community composition was analyzed based on the relative OTU abundances that increased significantly after glyphosate addition. The analysis identified 24 OTUs, assigned to seventeen genera, that responded to glyphosate in the water column; three more OTUs originated from biofilms (Table 2.1, detailed statistics are provided in Supplementary Material 2.8). Distinctive positive responses were determined for OTUs of three *Gallaecimonas* spp. (Figure 2.5 and Supplementary Material 2.6, OTU 109/129),

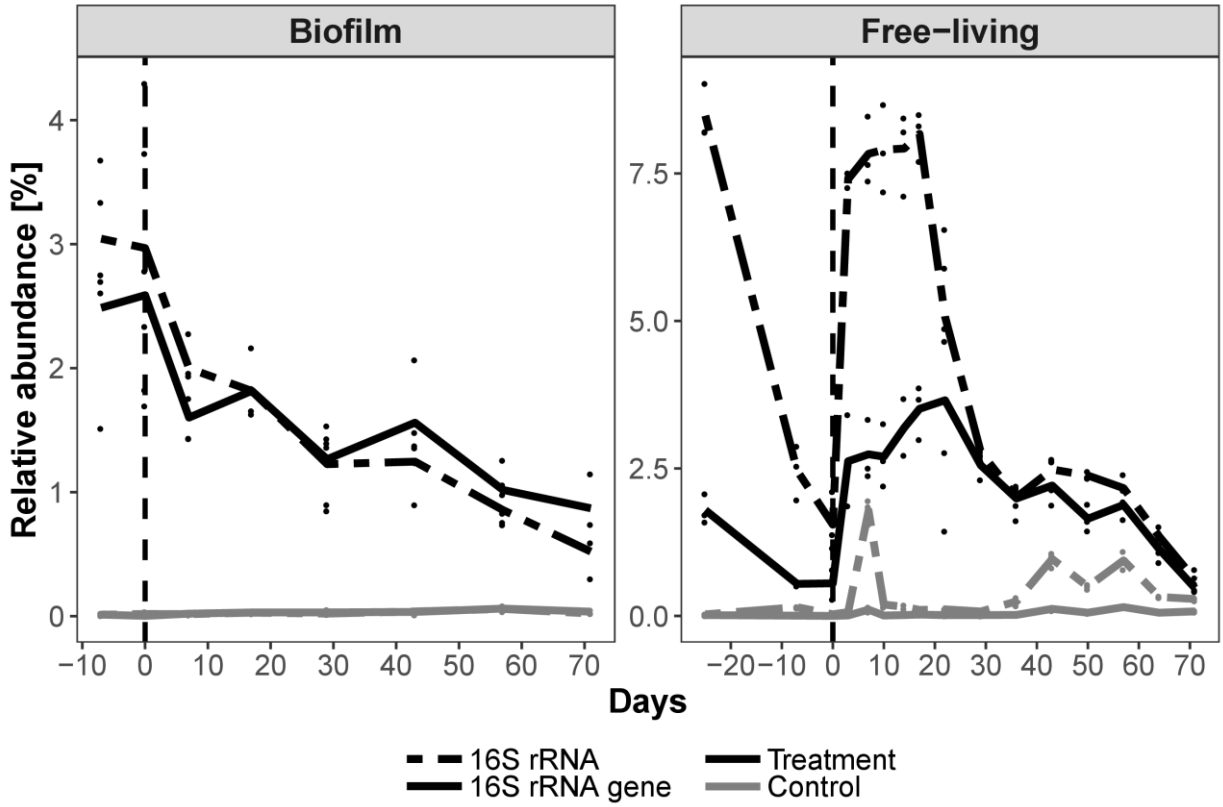


Figure 2.5: Changes in the relative abundance of *Gallaecimonas* OTU 11 in the treatment and control microcosms over time as determined by 16S rRNA gene and rRNA amplicon analyses. The vertical dashed line indicates the time of glyphosate addition. There was no evidence of an impact of glyphosate on the biofilms whereas planktonic abundance responded rapidly based on 16S rRNA gene and 16S rRNA abundances.

Methylotenera spp., *Hyphomonas* spp. and *Parvibaculum* spp., with both 16S rRNA and rRNA gene abundances increasing immediately after glyphosate addition (Supplementary Material 2.6, OTU 44/25/46). In agreement with the results reported above, the corresponding biofilm abundances for these OTUs remained stable.The genus *Pseudomonas* accounted for most of the overall diversity within the microcosms, with variable responses by individual *Pseudomonas* OTUs to glyphosate addition. Thus, while

Table 2.1: Differentially abundant OTUs in the treatment microcosms after the addition of glyphosate. OTU abundance before and after glyphosate addition were first tested using the Wald test for Benjamini-Hochberg corrected *p*-values < 0.01. From this selection, 24 OTUs in the water column and 3 in the biofilm were then identified as potentially glyphosate-responsive based on a visual comparison with the corresponding OTUs in the control microcosm.

Free-living/ Biofilm	16S rRNA gene/16S rRNA	Order	Family	Genus	OTU ID
- / X	X / X	*Caulobacterales*	*Caulobacteraceae*	*Brevundimonas*	Otu000042
X / -	X / X	*Caulobacterales*	*Hyphomonadaceae*	*Hyphomonas*	Otu000025
X / -	X / X	*Parvibaculales*	*Parvibaculaceae*	*Parvibaculum*	Otu000046
X / -	X / -	*Rhizobiales*	*Rhizobiaceae*	*Aminobacter*	Otu000072
X / -	- / X	*Rhizobiales*	*Rhizobiaceae*	*Hoeflea*	Otu000320
X / -	X / X	*Rhizobiales*	*Rhizobiaceae*	*Mesorhizobium*	Otu000056
X / -	X / -	*Rhizobiales*	*Rhizobiaceae*	*Rhizobiaceae*_unclassified	Otu000037
X / -	X / X	*Rhizobiales*	*Rhizobiaceae*	*Rhizobiaceae* unclassified	Otu000070
X / X	X / X	*Rhizobiales*	*Xanthobacteraceae*	*Pseudolabrys*	Otu000038
- / X	X / X	*Rhodobacterales*	*Rhodobacteraceae*	*Defluviimonas*	Otu000098
X / -	X / -	*Rhodobacterales*	*Rhodobacteraceae*	*Loktanella*	Otu000059
X / -	X / -	*Rhodobacterales*	*Rhodobacteraceae*	*Seohaeicola*	Otu000094
X / -	X / -	*Sphingomonadales*	*Sphingomonadaceae*	*Sphingorhabdus*	Otu000032
X / -	X / -	*Thalassobaculales*	*Thalassobaculaceae*	*Thalassobaculum*	Otu000018
X / -	X / X	*Alteromonadales*	*Gallaecimonadaceae*	*Gallaecimonas*	Otu000011
X / -	X / X	*Alteromonadales*	*Gallaecimonadaceae*	*Gallaecimonas*	Otu000109
X / -	X / X	*Alteromonadales*	*Gallaecimonadaceae*	*Gallaecimonas*	Otu000129
X / -	X / -	*Alteromonadales*	*Idiomarinaceae*	*Idiomarina*	Otu000049
X / -	X / X	*Betaproteobacteriales*	*Burkholderiaceae*	*Hydrogenophaga*	Otu000139
X / -	X / X	*Betaproteobacteriales*	*Methylophilaceae*	*Methylotenera*	Otu000044
X / -	X / X	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*	Otu000007
X / -	X / X	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*	Otu000036
X / -	X / X	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*	Otu000078
X / -	X / X	*Puniceispirillales*	uncultured	uncultured_ge	Otu000191
X / -	X / X	*Sphingomonadales*	*Sphingomonadaceae*	*Sphingobium*	Otu000176
X / -	X / -	*Planctomycetales*	*Gimesiaceae*	*Gimesia*	Otu000058

2.3.6 Duration of the detected signals

The detected microbial signals representative of free-living and biofilm communities after the glyphosate pulse differed in length and intensity. Total cell counts in the water column

increased significantly from day +7 to day +22, whereupon the glyphosate concentration remained ≤ 4.4 µM and AMPA < 0.1 µM. The Shannon index increased significantly from day +3 to day +17 for both the 16S rRNA and 16S rRNA gene based planktonic communities. The clusters in the NMDS of the 16S rRNA gene (except for one technical replicate) and 16S rRNA data from free-living bacteria indicated that the community composition from day +3 to +22 (Figure 2.3; red polygons) was more similar among these samples than in subsequent samples. The relative abundances of the responding planktonic OTUs generally increased from day +3 to day +22. Some of the detected planktonic OTUs retained elevated abundances for a longer period, until day 64, such as several *Pseudomonas* OTUs. However, this behavior was commonly observed for biofilm OTUs, and specifically for *Brevundimonas* OTU 42, *Defluviimonas* OTU 98 and *Pseudolabrys* OTU 38 (Supplementary Material 2.6). The increase in abundance began gradually and was first detected typically after day +7, but it continued until the end of the experiment. For these biofilm OTUs, the continuously high abundances were accompanied by corresponding changes in planktonic abundances. Thus, biofilm reactions were maintained whereas most planktonic reactions ended on day +22, when the glyphosate concentration was 12.7 µM.

2.3.7 Distribution of glyphosate degradation genes in metagenomic samples

Metagenomes of free-living microbial communities were analyzed to gain insights into glyphosate-related bacterial functions. All relevant glyphosate-degradation genes *gox*, *thiO* and *phnC-P* were detected. The *phn* operon might be involved in metabolization at two steps, either degrading glyphosate to sarcosine or cleaving the C-P bond in AMPA. Identifying the responsible pathway, if not all of them, was required. For the sarcosine pathway, whether a particular *phn* operon enables glyphosate degradation at all depends on the encoded substrate specificity. Therefore, we screened for sequence clusters that became more abundant after glyphosate addition, as these may also have contained sequence motifs typical of glyphosate degradation. An example is the *phnJ* gene, which codes for an essential protein within the C-P lyase multienzyme core complex. Nonetheless, in samples from the treatment microcosm, the abundance profiles proved to be complex even for closely related sequences of *phnJ* genes (Figure 2.6). Based on phylogenetic analyses, *phnJ* genes similar to that of the alphaproteobacteria *Yoonia vestfoldensis* spp. (formerly *Loktanella vestfoldensis*, UniProtKB: A0A1Y0ECC7) were most abundant on day +14, when the total cell count reached a peak. This development was similar for the *phnJ* sequences of *Ruegeria pomeroyi* strain ATCC 700808 (UniProtKB: Q5LW71), *Rhizobium meliloti* strain 1021 (UniProtKB: Q52987) and *Agrobacterium radiobacter* strain ATCC BAA-868 (UniProtKB: B9J6Q8). Moreover, *phnJ* sequence reads correlated with the abundance of 16S rRNA gene OTUs, such as those of *Yoonia* spp., based on the taxonomy of the reference genes (Supplementary Material 2.6, OTU 59).

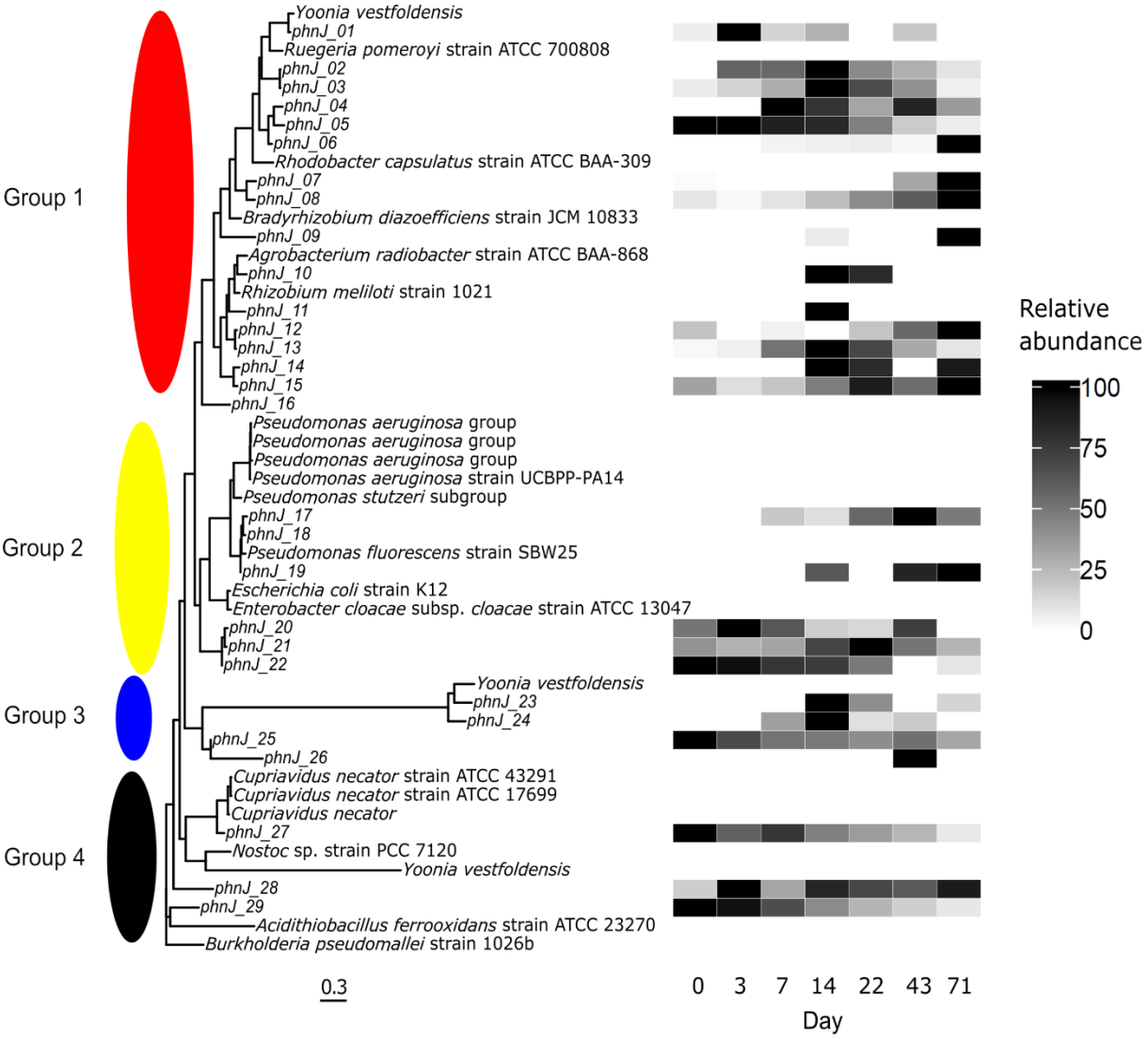


Figure 2.6: Multiple sequence alignment of protein fasta sequences of the *phnJ* gene (A) from the free-living treatment metagenomes. As a reference, the tree contains sequences from cultivated organisms related to the differentially abundant OTUs detected in this experiment. Taxonomy is presented according to UniProt. The heat map shows the abundance of a given *phnJ* gene relative to the other samples over time. Four groups are labeled, the first two consisting exclusively of *Alphaproteobacteria* and *Gammaproteobacteria*, respectively.

From the 29 *phnJ* genes annotated in the treatment microcosm, four main groups could be recognized. Based on the embedded reference genes from known organisms, the largest group consisted solely of the alphaproteobacterial *phnJ* sequences grouping with sixteen genes from the metagenomes. *Alphaproteobacteria* were by far the most abundant class inhabiting the microcosms. The second group solely contained gammaproteobacterial reference genes and six metagenomic genes. For the first two groups, phylogenetic relationships based on the 16S rRNA gene were similar to the clustering of the *phnJ* sequences, as highlighted by the subgroup of sequences from *Enterobacter cloacae* ssp. cloacae strain ATCC 13047 (UniProtKB: A0A0H3CFJ4) and *Escherichia coli* K12 (UniProtKB: P16688). Groups 3 and 4 gathered *phnJ* sequences from several less-related organisms. Also

in these groups multiple *phnJ* sequences were those of the alphaproteobacteria *Yoonia vestfoldensis* spp. and were encountered in the first, third and fourth group. In the latter two groups, there was a low similarity with their closest relatives, which even included the cyanobacterium *Nostoc* sp. strain PCC 7120. Interestingly, the highly diverse genus *Pseudomonas* was represented by only three sequences; these were most closely related to *phnJ* from P. *fluorescens* strain SBW25 (UniProtKB: C3K5L9). Comparable results, i.e., varying numbers of *Pseudomonas*-related genes (data not shown), were achieved for other *phn* and the sarcosine oxidase (*sox*) genes.

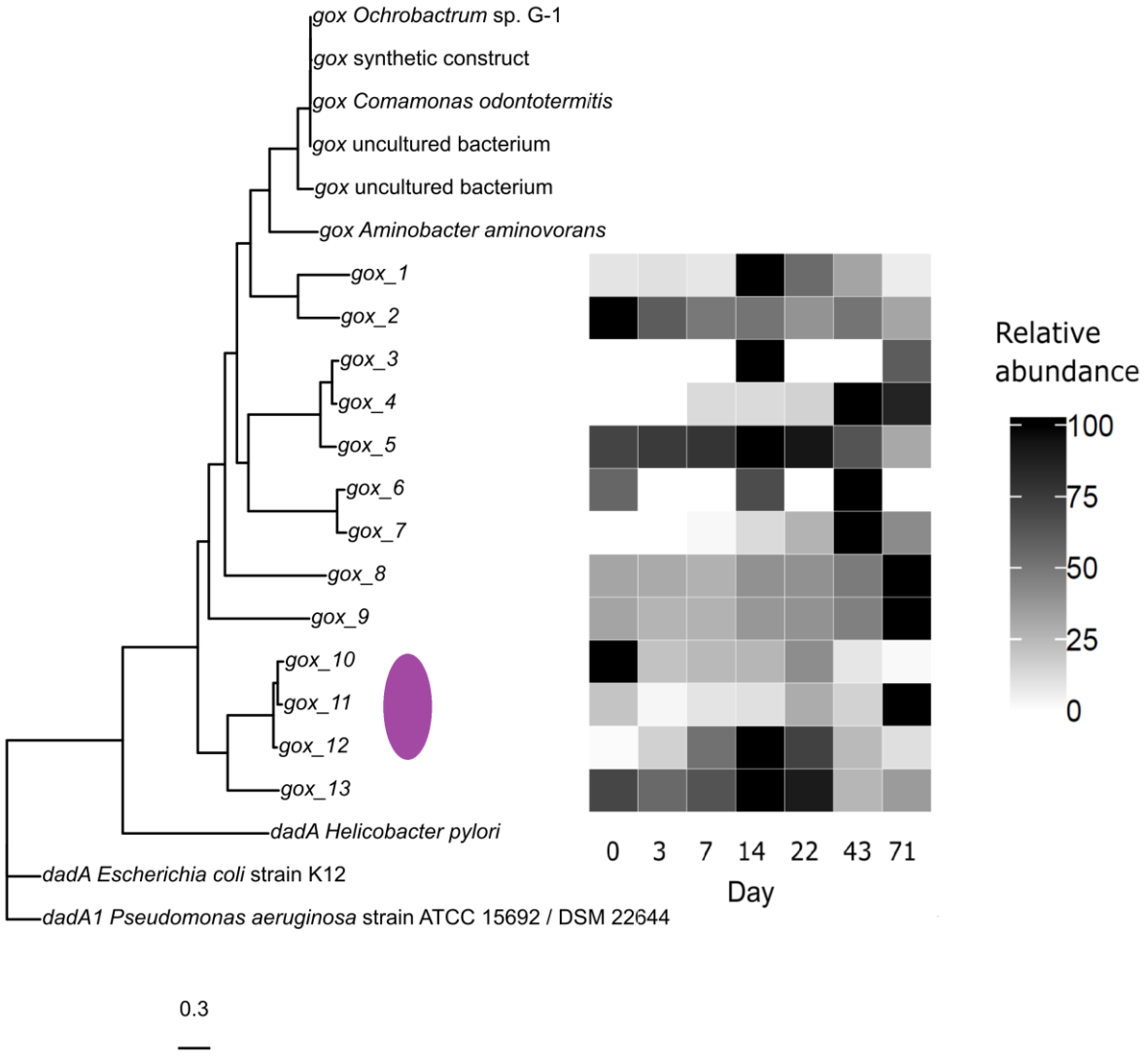


Figure 2.7: Multiple sequence alignment of protein fasta sequences of the *gox* and *dadA* genes deposited in UniProt and GenBank clustered separately with the metagenomic sequences more related to *gox*. The purple ellipse marks sequences similar to those of the genus *Hoeflea.*

The *gox* gene was not identified by Prokka in the metagenomes. A manually conducted comparison, however, detected thirteen closely related sequences which were annotated by Prokka as *dadA* (*D*-amino acid dehydrogenase), but instead appear to be more closely related to *gox* genes. The reference *gox* genes create a distinct group (UniProtKB: D2KI28,

A0A142MF04, D4NZ76, D4NZ75; GenBank: ATE50174.1, ADV58259.1), whereas the metagenomic sequences are distinguished from this group (Figure 2.7). When challenged, the annotation of these sequences, by adding the Prokka-referenced *dadA* sequences (UniProtKB: P0A6J5, Q9HTQ0, A3KEZ1), the metagenomic sequences were indeed more similar to *gox* genes. The abundance of the potential *gox* sequences gox_1, 3, 5, 12, and 13 converged with the total cell counts peak. Due to the separate clustering of the reference sequences, the taxonomic inference remains unknown. However, a basic online BLASTp analysis assigned gox_10, 11 and 12 (Figure 2.7, purple ellipse) to FAD-binding oxidoreductase from *Hoeflea marina* (UniProtKB: A0A317PMM8) and *Hoeflea* sp. BRH c9 (UniProtKB: A0A0F2P8D1) with a query coverage of 100 % and an identity > 88 %.

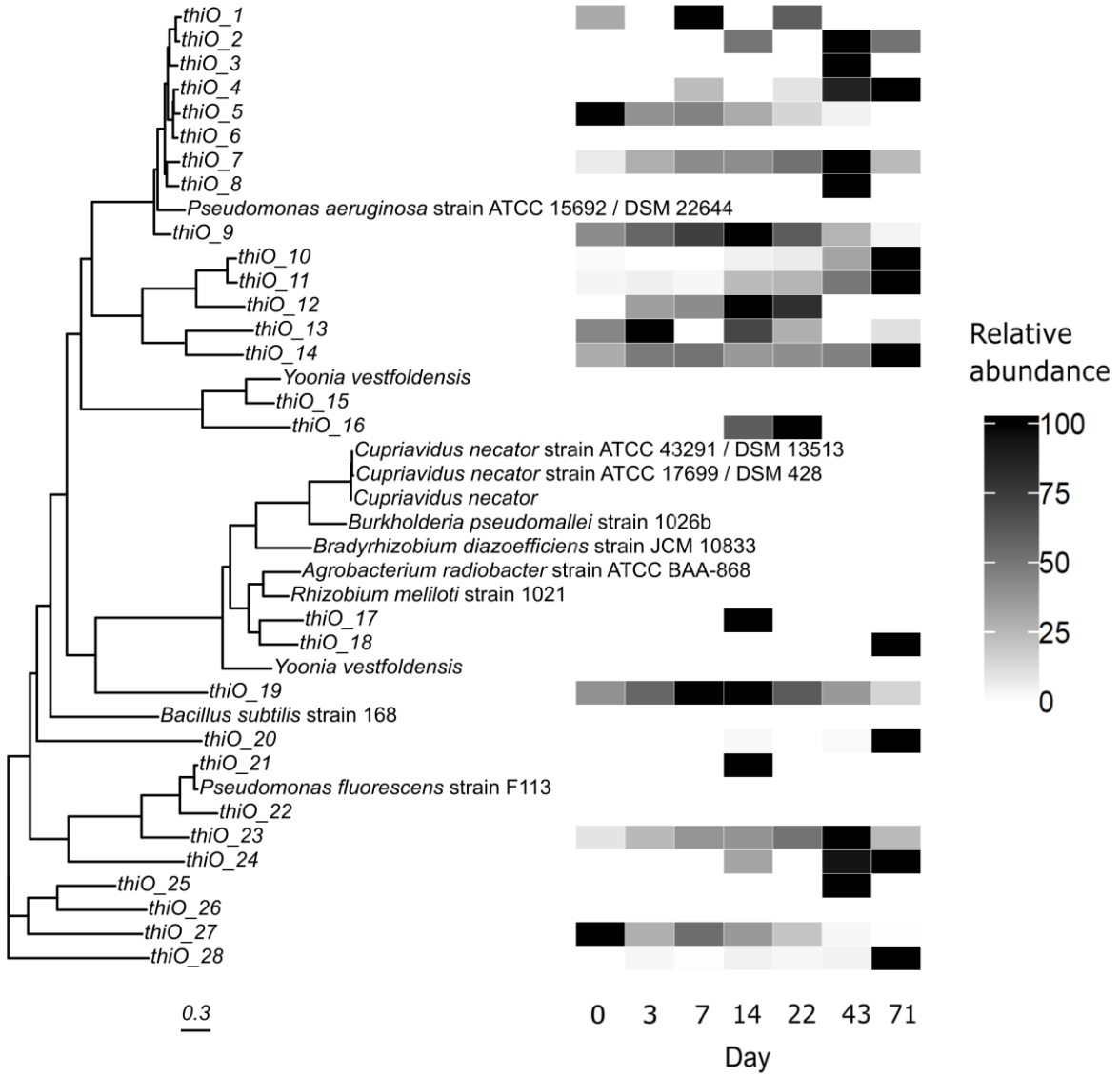


Figure 2.8: Multiple sequence alignment of protein fasta sequences of the *thiO* gene could be different between the same or closely related organisms such as *Yoonia vestfoldensis* or *Pseudomonas*. The presence of *thiO* genes over all sampled time points was less comprehensive compared to *gox* and *phnJ* genes.

thiO sequences were detected 28 times (Figure 2.8) with no clear taxonomic separation based on the positioning of the reference sequences. Three sequences were most abundant at day

+14 in time with the cell counts peak (thiO_9, 12, 19) with thiO_19 being somewhat related to *Yoonia vestfoldensis* (UniProtKB: A0A1Y0E718). thiO_1 to 9 grouped with *Pseudomonas aeruginosa* ATCC 15692 (UniProtKB: G8PX29). *Betaproteobacteria* clustered together (Cupriavidus, UniProtKB: Q0KF33, G0ETC1, A0A1K0I947 and *Burkholderia*, UniProtKB: A0A0H3HPX7), although alphabacterial sequences were also similar. Interestingly, *Yoonia vestfoldensis* (UniProtKB: A0A1Y0EFI8, A0A1Y0E718) and *Pseudomonas* (UniProtKB: P33642, G8PX29) harbored dissimilar *thiO* sequences, which might be obtained by horizontal gene transfer.

2.4 Discussion

2.4.1 Potential impacts of glyphosate on a brackish microbial ecosystem

A glyphosate incubation experiment with a brackish water community was conducted to investigate the impact of glyphosate on free-living and biofilm microbial assemblages. Following the glyphosate pulse, changes in community composition and increases in total cell counts, α-diversity and the abundances of specific 16S rRNA (gene) OTUs were detected in the water column. By contrast, with a few exceptions, the biofilm, which was 69 days old when glyphosate was added, remained undisturbed. Other studies have also shown that organisms embedded in a biofilm are less responsive to disturbances in the surrounding medium (Davey and O'Toole, 2000; Tlili et al., 2011). Similarly, in this study, compared to the water column, fewer OTUs in the biofilm were affected by glyphosate. A smaller impact of glyphosate on freshwater biofilms, and especially on phototrophic organisms, was previously reported. Khadra et al. (2018) investigated periphytic biofilms differing in age (at least 2 months) and exposed to different glyphosate concentrations (35.4, 383.5, and 3540 nM) in a lake. They concluded that glyphosate had no effect on biofilms, a finding also reported by Lozano et al. (2018), who showed that periphyton was more resistant than phytoplankton to 17.7 µM glyphosate. Although the latter study did not find a biomass-based effect on periphyton, the abundance of certain taxa decreased. Vera et al. (2010) noted a delayed increase in the biomass of periphyton exposed to 47.3 µM glyphosate in freshwater mesocosms.

In our study, from the initial 82.45 µM glyphosate added at day 0, 1 µM remained at the end of the experiment. Thus, with the decreasing availability of glyphosate, the cost-benefit ratio of producing proteins for its metabolism seems to become increasingly unfavorable. This is supported by the absence of AMPA as indication of degradation in the later samples. Given that at 4.4 µM glyphosate a response by planktonic communities was no longer detectable, then at the pM to nM concentration ranges measured in estuaries of the Baltic Sea (Skeff et al., 2015) neither biofilms (harboring the majority of microbial cells) nor planktonic bacteria in the Baltic Sea are likely to be disturbed by the herbicide. However, the nutrient regime in Baltic seawater differs from that of the rich medium provided in this study. In addition, the long-term

effects of glyphosate on microbial communities are as yet unknown. Mercurio et al. (2014) demonstrated an unexpected glyphosate persistence in seawater in the presence and absence of light. The study of Stachowski-Haberkorn et al. (2008) suggested that even low-nM glyphosate concentrations can affect natural coastal microbial communities in marine environments. According to our results, planktonic bacteria better reflect short-term disturbances, whereas the accompanying biofilm provides a reference for examining overall succession trends occurring as a result of exchange with the water column. A biofilm response would thus indicate that a threshold of disturbance tolerance had been exceeded. Expanding the shotgun sequencing to involve the biofilms could support this idea.

2.4.2 Differences in the responses of water column and biofilm communities to glyphosate addition

The response of biofilm OTUs to glyphosate addition, as measured by abundance, was minor but detectable until the end of the experiment. One interpretation of this result is that the glyphosate pulse favored these OTUs over others in the biofilm community. Glyphosate has been shown to accumulate in biofilms, including those within a Brazilian river (59.2–1806 nmol·kg^{-1}, AMPA 450.29–6033.89 nmol·kg^{-1}; Fernandes et al., 2019), or to persist at a very small percentage of the initial concentration, as demonstrated in a microcosm study of biofilms in a French river (Carles et al., 2019). Either would provide glyphosate-degrading OTUs with a nutritional advantage. Our initial tests conducted prior to the experiment showed that the biofilms did not enrich glyphosate, at least not during the first 3 days after the glyphosate pulse. However, in the few cases in which a response by biofilm OTUs was identified, the respective signal was also detected in the water column, both for a longer time and indicative of a higher abundance. Several of the abundant biofilm OTUs, however, were characterized by a concise albeit constant changes in abundance regardless of the condition, which complicated the detection of glyphosate-responding OTUs (Figure 2.5). It should be noted that the growth substrates for the biofilm were initially sterile and that all colonization occurred via the inoculated water column. This could explain the overall concordance between abundant OTUs in the water column and in the biofilms.

Microbial responses within the water column were in most cases limited to day +22, which coincided with the strongest decrease in the glyphosate concentration to ≤ 4.4 µM, the AMPA concentration fell below 0.1 µM afterward. Transient effects on microbial communities by glyphosate have been previously described. For example, Weaver et al. (2007) found small, brief (< 7 days) changes in the levels of fatty acid methyl esters and a reduction in the hydrolytic activity of a soil microbial community exposed to a glyphosate concentration of 277–828 µmol·kg^{-1}. Using Biolog assays and phospholipid fatty acids analyses, Ratcliff et al. (2006) measured a non-specific, short-term stimulation of bacteria at a high glyphosate concentration. The increased α-diversity determined in this study confirmed the findings of Lu et al. (2017),

who analyzed the rhizosphere of a glyphosate-tolerant soybean line based on 16S rRNA gene amplicon sequencing. The authors also found a higher diversity and varying OTU abundances in the rhizosphere of the treated than of the control cultivar. In a metatranscriptomic analysis, Newman et al. (2016) investigated changes in bacterial gene patterns in response to long-term glyphosate exposure. The results indicated a potential shift in bacterial community composition toward more glyphosate-tolerant bacteria. Wang et al. (2017) described the effects of two glyphosate concentrations on the microbial community associated with the dinoflagellate *Prorocentrum donghaiense* and showed that 36 µM glyphosate caused a decrease and 360 µM an increase in α-diversity. Several OTUs detected in our study belonged to genera whose abundance increased following glyphosate treatment (*Methylotenera*, *Pseudomonas*, *Sphingobium*, *Thalassobaculum*), demonstrating the ability of glyphosate to cause favorable conditions for these genera across various habitats. On a further note, the herein identified *Rhodobacteraceae* and *Rhizobiaceae* OTUs were confirmed in a novel approach using artificial neural networks and Random Forest to detect responding OTUs (Janßen et al., 2019b).

2.4.3 Glyphosate-induced changes in OTU abundance

In our study, temporally highly resolved NGS data revealed increased OTU abundances, but the mechanisms of the increases were unclear. While glyphosate can be considered as a source of C, N, or P, the microcosms were supplied with sufficient amounts of C and N (evidenced by the medium composition and end-of-experiment data points) and P from other sources.

Specific reactions to glyphosate have been described in studies of bacterial cultures, especially those of degraders (Wang et al., 2016a) and resistant *cyanobacteria* (López-Rodas et al., 2007). Within the same species, different strains may or may not be capable of degrading glyphosate and several pathways for glyphosate degradation may be present in a single strain. This is the case in *Pseudomonas* (Jacob et al., 1988; White and Metcalf, 2004; Zhao et al., 2015; Lidbury et al., 2016) and would explain why some, but not all of the *Pseudomonas* OTUs detected in our study became abundant after glyphosate addition. Thus, the pronounced diversity of *Pseudomonas* was also expressed by its reactions toward glyphosate.

2.4.4 Probability of glyphosate degradation

The responses mainly by free-living bacteria, such as the increase in cell counts and the presence of AMPA indicated that glyphosate was degraded. The amount of AMPA detected in comparison to the corresponding glyphosate concentrations suggests that only a minor fraction was metabolized, a quality associated with the glycine oxidase *thiO*. The increase in total cell counts and the discrepancy between the measured and the calculated glyphosate levels require a more complete degradation of glyphosate. Sarcosine/*L*-alanine levels do not

compensate for this difference and as they did not change after glyphosate addition and were present in both microcosms it was more likely only *L*-alanine was present due to the inclusion of the casamino acids. This implicates that glyphosate was not metabolized by the sarcosine pathway. It is possible that rapid degradation of the intermediate product could have occurred, thus rendering it hardly detectible. However, it is unlikely that the glycine oxidase would be capable of such a degradation rate due to its low specificity toward glyphosate. A possible explanation is the degradation of glyphosate by *gox* into AMPA with an immediate continuation by C-P lyase. Sviridov et al. (2015) concluded in their review that the majority of described glyphosate-degrading bacteria use the *gox* gene and consequently export AMPA into their environment, but also stated that organisms not being capable of degrading glyphosate might still metabolize AMPA.

Furthermore, the abundances of *gox* genes, *thiO* genes, the *phn* operon, *sox* genes and *aroA* genes correlated with those of the detected OTUs (via multiple sequence alignment and reference sequences; Figure 2.6). It must be noted that *phn* operons encode functions that result in the degradation of a variety of phosphonates, although not necessarily including glyphosate. The respective genes are subject to extensive lateral transfer, which complicates data interpretation (Huang et al., 2005). The results of our metagenomic analysis suggested that *phn* genes have a higher sequence similarity based on phylogeny than on substrate specificity. Sequence abundances of a *phnJ* gene correlated with OTU 59 (classified as *Yoonia*/*Loktanella* spp.). This suggested that this OTU possesses *phn* genes whose abundances' increase might be in response to the presence of glyphosate or AMPA as a nutrient source. The same reference organism, as well as *Pseudomonas aeruginosa*, correlated with the abundance of *thiO* genes. The phylogenetic comparison of *gox*, *dadA* and our metagenomic sequences (Figure 2.7) underlined the demand of properly described references and the potential of undiscovered *gox* variants. For the *Hoeflea*-related *gox* sequences, no treatment-specific abundance change could be assigned to *Hoeflea* OTUs. In conclusion, a metatranscriptomic analysis that describes the expressed *phn*, *gox*, and *thiO* mRNAs may have provided clearer evidence of the pathways used for glyphosate degradation (Martínez et al., 2013; Wang et al., 2016b) as well as the involved organisms.

However, amplicon sequences still proved to be a cost efficient and sensitive method for community analysis, as comparisons of 16S rRNA (gene) and shotgun sequencing data indicated that glyphosate-responsive low-abundance OTUs were not covered in the metagenome. Furthermore, the 16S rRNA gene amplicon counts were a better indicator of community changes than 16S rRNA, indicating that DNA is a better proxy of abundance. Field experiments or laboratory studies involving more than one determinant should further investigate the potential of using detailed community composition data as an indicator of community disturbance.

2.5 Data availability statement

The datasets generated for this study can be found in the NCBI database under BioProject ID PRJNA434253 and SRA accession SRP151042. OTU and taxonomy table as well as corresponding code to process and analyze the data are available in the GitHub repo: https://github.com/RJ333/Glyphosate_gene_richness, code for the metagenomic analysis is available under https://github.com/RJ333/calculate-functional-trees.

Chapter III

Machine learning predicts the presence of 2,4,6-trinitrotoluene in sediments of a Baltic Sea munitions dumpsite using microbial community compositions

The following chapter was submitted to the journal *Frontiers in Microbiology*'s special issue *Advancements in the Understanding of Anthropogenic Impacts on the Microbial Ecology and Function of Aquatic Environments* as:

René Janßen, Aaron J. Beck, Johannes Werner, Olaf Dellwig, Johannes Alneberg, Bernd Kreikemeyer, Edmund Maser, Claus Böttcher, Eric P. Achterberg, Anders F. Andersson, Matthias Labrenz (2020): Machine learning predicts the presence of 2,4,6-trinitrotoluene in sediments of a Baltic Sea munitions dumpsite using microbial community compositions (submitted)

Declaration of author contributions:

René Janßen performed molecular and geological lab work (except MC measuring), acquired funding, planned the data analysis, performed the bioinformatic processing, wrote the analysis scripts, wrote the R package, conducted the machine learning and analyzed the results

Aaron Beck and Eric Achterberg measured MC concentrations, designed the sampling campaigns at Kolberger Heide and commented on the manuscript.

Edmund Maser acquired funds and samples and provided toxicological assessment and commented on the manuscript revision.

Bernd Kreikemeyer provided the 16S rRNA gene amplicon sequencing device and materials and commented on the manuscript revision

Claus Böttcher connected various subprojects to make this study possible, provided historical and jurisdictional context, acquired funding, addressed monitoring requirement and commented on the manuscript.

Olaf Dellwig supervised the geochemical analysis and conducted ICP-MS measurements, interpreted the data and commented on the manuscript.

Johannes Werner JW supported the bioinformatics analysis, performed code review, supervised the R package development, provided the IT infrastructure and commented on the manuscript.

Johannes Alneberg guided the initial bioinformatic analysis and provided statistical and general coding support and commented on the manuscript.

Anders F. Andersson came up with essential ideas regarding the analysis of microbial communities by machine learning in an ecologically meaningful way, supervised bioinformatics and statistical data analysis discussed the data and commented on the manuscript.

Matthias Labrenz conceived the concept and supervised the project, acquired and provided funding and discussed the data with René Janßen.

René Janßen, Johannes Werner, Anders F. Andersson, Aaron J. Beck, Olaf Dellwig, Eric Achterberg, Claus Böttcher and Matthias Labrenz discussed the data of the initial draft.

René Janßen drafted the manuscript, Matthias Labrenz critically commented on the manuscript and redrafted parts of it. All other authors commented on the manuscript.

René Janßen’s contribution to the written manuscript was ~ 90 %.

Abstract

Bacteria are ubiquitous and live in complex microbial communities, which can react rapidly to changing environmental conditions. Their physiological variety enables communities to respond in specific ways to environmental drivers, potentially resulting in distinct microbial fingerprints for a given environmental state. Our goal was to assess the opportunities and limitations of machine learning to detect fingerprints indicating the presence of the munition compound 2,4,6-trinitrotoluene (TNT) in southwestern Baltic Sea sediments.

Over 40 environmental variables including grain size distribution, elemental composition and concentration of munition compounds (mostly at pmol g^{-1} levels) from 150 sediments collected at the near-to-shore munition dumpsite Kolberger Heide by the German city of Kiel were combined with 16S rRNA gene amplicon sequencing libraries. Prediction was achieved using Random Forests; the robustness of predictions was validated using Artificial Neural Networks. To facilitate machine learning with microbiome data we developed the R package phyloseq2ML.

Using the most classification-relevant 25 bacterial genera exclusively, potentially representing a TNT-indicative fingerprint, TNT was predicted correctly with up to 81.5 % balanced accuracy. False positive classifications indicated that this approach has also the potential to identify samples where the original TNT contamination was no longer detectable. The sensitivity of this approach can be deduced from the fact that TNT presence was neither identified among the main drivers of the microbial community composition, nor did it correlate with sediment metal content, demonstrated by decreased prediction rates using environmental variables.

Our results suggest that microbial communities can predict even minor influencing factors in complex environments, demonstrating the potential of this approach for the discovery of contamination events over an integrated period of time and for environmental monitoring in general.

3.1 Introduction

Microbes are the most diverse, abundant and ubiquitous life forms on Earth. They live in complex microbial communities, which can react rapidly to environmental changes, a result of consistent evolutionary pressures applied by fluctuating conditions (Lindh and Pinhassi, 2018). The developed variety of physiologies enables communities to respond in specific ways to environmental drivers, hence functioning as indicator for surrounding conditions. This principle was demonstrated for very different habitats: it was possible to match individual human skin microbiomes with those on the occupant's household surfaces (Wilkins et al., 2017), to associate subway microbiomes to the major cities they were located in (Ryan, 2019) or to distinguish microbial communities in the brackish Baltic Sea along the salinity gradient (Herlemann et al., 2011) and its anoxic regions (Thureborn et al., 2016). However, relevant indicative fractions of the communities, conceivably acting as microbial fingerprints, may only emerge by analyzing a sufficiently large number of communities. Next generation sequencing allows for processing such larger amounts of samples to extract this information, but it might be accompanied by a large portion of irrelevant data with regard to the particular indication.

The ensemble classifier Random Forest (RF) is capable of identifying such potential fingerprints – even if they include nonlinear relations - in large and complex data sets (Breiman, 2001a). RF is among the most popular machine learning methods and has frequently been used in biological sciences (Fernández-Delgado et al., 2014). The features relevant for the model's decisions can be assumed equivalent to an indicative fingerprint and the RF variable importance measure readily identifies them (e.g. Altmann et al., 2010; Janitza et al., 2018). Fingerprints related to community-shaping drivers are revealed by performing unsupervised classification, whereas specific influences can be targeted by the application of supervised machine learning. In microbiological studies, RF has been deployed to localize the geographic origin of port water across three continents based on dominant bacterial phyla (Ghannam et al., 2020). Moitinho-Silva et al. (2017) used RF among other classifiers to separate between sponges of high and low microbial abundance. Thompson et al. (2019) used RF and artificial neural networks (ANN) to identify important taxa for the prediction of dissolved organic carbon concentrations. In a previous study we demonstrated the identification of glyphosate-impacted free-living community compositions by ANN and RF after a 82.45 $nmol \cdot mL^{-1}$ glyphosate pulse in a lab microcosm experiment (Janßen et al., 2019b).

In this study, we are particularly interested in the question to what extent environmental microbial communities can reliably predict anthropogenic pollutants using the above algorithms. We tested this approach for a munitions dumpsite in the southwestern Baltic Sea, where sediments are contaminated with explosive compounds such as 2,4,6-trinitrotolouene (TNT). The munitions dumpsite Kolberger Heide in the Kiel Bight (Germany) is an

approximately 1260 ha large area of 10 - 15 m water depth. Conventional munition, mostly incomplete or unfused was disposed of at this site after World War II (Kampmeier et al., 2020). About 30.000 tons are estimated to be still on site, containing mainly of TNT and 1,3,5-trinitroperhydro-1,3,5-triazine (RDX) as munition compounds ([MC], Böttcher et al., 2011). The containments such as mines, shells and torpedo heads display various states of corrosion (Kampmeier et al., 2020), resulting in the leakage of MC (Beck et al., 2019). In addition, bare munition chunks are scattered across the sediment bed, potentially due to low-order, or incomplete detonation during blow-in-place clearance activities (Pfeiffer, 2009; Maser and Strehse, 2020). Dissolved TNT can be rapidly dissipated or metabolized in direct proximity to its source, complicating the quantification of TNT released into the environment (Elovitz and Weber, 1999; Beck et al., 2019). However, the presence of MC including TNT and its transformation products in the Kolberger Heide water column samples (ca. 1 – 15 $ng \cdot L^{-1}$) and biota (1 – 24000 $ng \cdot g^{-1}$) has been reported (Gledhill et al., 2019). Little is known about the MC concentrations in accordant sediments.

Sediment in the Kolberger Heide is contaminated by TNT at $pmol \cdot g^{-1}$ levels. It was our aim a) to investigate if machine learning is capable of predicting TNT in sediments and identifying indicative microbial fingerprints; b) to assess how robust the predictions are and which factors influenced the model's performance; c) to evaluate whether a microbial fingerprint is sufficiently persistent to detect a history of TNT, indicated by TNT transformation products. Finally, we discuss how the described approach could supplement and be integrated into regular monitoring activities.

3.2 Material and methods

3.2.1 Collection of sediments and determination of munition compounds

One hundred sixty-seven sediment samples were collected within the Kolberger Heide munitions dumpsite and its surroundings during the course of the UDEMM (Environmental monitoring for the delaboration of munitions on the seabed, Greinert, 2019) project. Samples were obtained during several cruises and individual sampling events. Additional sampling took place at defined distances from mines and at a site of a controlled detonation. Sediment samples within the dumpsite were collected manually by scientific divers, or using an ROV. Outside the dumpsite's restriction zone, surface sediments were collected using a Van Veen grab. Duplicate sediment cores were collected using a multi-corer at two sites east and west of the dumpsite (map provided in Supplementary Material 3.1). Sampling was conducted in December 2016 and from June to December 2017. Supplementary Material 3.2 details contextual data such as position of sample collection, cruises and experiments as well as measured parameters. "Experiments" refer to the goal of a sampling, e.g. investigating a spatial MC gradient in cardinal directions around a mine, analyzing the MC distribution across a mine

mound or along a sediment profile. Sediments were stored in sealable plastic bags (Whirl-paks; Nasco, Madison, WI, USA) at -20 °C for subsequent MC analysis using an ultra-high performance liquid chromatographic system coupled to a heated electrospray ionization source and a high resolution quadrupole/orbitrap mass analyzer (UHPLC-HESI-MS, Q Exactive, ThermoScientific) detection after thawing and extraction using LCMS-grade acetonitrile (Fisher). Munition compounds were measured according to Gledhill et al. (2019) including TNT, RDX, 2-amino-4,6-dinitrolouene (2-ADNT), 4-amino-2,6-dinitrolouene (4-ADNT), 2,4-dinitrotoluene (2,4-DNT), 2,6-dinitrotoluene (2,6-DNT), 1,3-dinitrobenzene (DNB), 1,3,5-trinitrobenzene (TNB), Octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine (HMX) and Tetryl (N-methyl-N-2,4,6-tetranitrolaniline). The TNT transformation products, 2,4-diamino-6-nitrotoluene (2,4-DANT) and 2,6-diamino-4-nitrotoluene (2,6-DANT) are not included in the Gledhill and colleagues (2019) suite of compounds, but were analyzed using the same method, and quantified after standardization using single-compound standards (AccuStandard, Connecticut, USA). For geological and molecular biology analyses sediments were slowly thawed, homogenized under a clean bench, and split into two 15 mL aliquots. The aliquots were stored at -80 °C.

3.2.2 Geochemical and sedimentological analyses

3.2.2.1 Sample preparation

The frozen (–20 °C) sediment samples were freeze-dried (Christ LOC-1M Alpha 1-4 and Christ Delta 1-24 LSCplus, Osterode am Harz, Germany) for 60 – 72 hours. Except for the grain size analyses, the dried samples were homogenized in an agate ball mill (Fritsch Pulverisette, Idar-Oberstein, Germany) at 200 rpm for 10 min.

3.2.2.2 Carbon, nitrogen, and sulfur

About 10 – 17 mg of the sediments were weighted into tin crucibles, a spatula tip of vanadium(V) oxide (Alpha Resources, Stevensville, MI, USA) was added as catalyzer and total C, total N, and total S were determined by an elemental analyzer (EuroEA, HEKAtech, Wegberg, Germany). For total inorganic carbon, 50 – 70 mg of sediment was treated with 40 % orthophosphoric acid and analyzed with an elemental analyzer (multiEA 4000, Analytik Jena, Jena, Germany). Total organic carbon was calculated by subtracting total inorganic carbon from total carbon. Precision and trueness were checked with in-house standards (MBSS, OBSS) and were <3.5% (Häusler et al., 2018).

3.2.2.3 Mercury

The sedimentary mercury content was determined by a direct mercury analyzer (DMA 80, Milestone Srl, Italy) using 100 - 120 mg per analysis (50 mg for sample “Udemm1277”, which exceeded the calibration range). Precision and trueness were checked with the certified reference material BCR-142R (Community Bureau of Reference) and an in-house standard

comprising Baltic Sea sediments (Mecklenburg Bay Sediment Standard, MBSS) and were <3% and <10%, respectively (Häusler et al., 2018). Sediments exceeding 1000 µg Hg kg^{-1} were measured three times and averaged.

3.2.2.4 Reactive iron and trace element contents

For determination of reactive element contents, about 200 mg of sediment material was weighed into pre-cleaned 11.5 mL polystyrene tubes and 10 mL of 0.5 M HCl was added. The tubes were shaken for 60 min at 175 rpm, followed by 6 min of centrifugation at 4000 x g and filtration of the solutions with 0.45 µm syringe filters. Three procedural blanks were analyzed together with the samples. The contents of Fe, P, and trace metals in the 0.5 M HCl extracts were determined by Q-ICP-MS (iCAP Q; Thermo Fisher Scientific, Germany) after automated 50-fold dilution with 2 vol% HNO_3 via a prepFAST module (Elemental Scientific, Omaha, NE, USA) and external calibration. Helium was used as collision gas (KED mode) to minimize polyatomic interferences and a Rh and Ir containing solution added online by the prepFAST module served as internal standard to compensate for matrix effects and instrument fluctuations. The calibration was checked with the international reference material SGR-1b (USGS), which underwent total acid digestion in closed PTFE vessels using a HNO_3-HF-$HClO_4$ mixture (Dellwig et al., 2019). For stable $^{206/207}$Pb isotope ratios the NIST SRM-981 was used as reference material (Dellwig et al., 2018). Precision and trueness of the measurements of the reference materials were <4.4% and 8.1%, respectively.

3.2.2.5 Grain size distributions

The grain size of the <2 mm sediment faction was measured using a Hydro EV accessory to the Mastersizer 3000 (Malvern Panalytical GmbH, Herrenberg, Germany). The samples were stirred at 2500 rpm and sonicated for 10 s. Eight measurements were performed per sample, followed by purging steps with distilled water. Outliers (values exceeding 1.5 times the interquartile range) were removed and the remaining values per sediment were averaged.

3.2.3 Molecular biology and bioinformatics

The methods described in the following were applied to the molecular biology aliquots of each sediment sample.

3.2.3.1 Extraction of nucleic acids

The sediments were collected using the appropriate collection and storage procedures for the determination of MC. To retrieve the best possible results in subsequent molecular biological analyses and due to the long term presence of TNT in the Kolberger Heide, the more robust 16S rRNA gene was preferred over the more sensitive 16S rRNA as sequencing target. DNA was extracted from 250 mg wet sediment using the Qiagen PowerSoil DNA Kits or from 2000 mg wet sediment using the Mobio PowerSoil RNA kit with the DNA elution kit (Hilden,

Germany). For each kit an extraction control without sediment was processed along with regular samples.

3.2.3.2 Sequencing 16S rRNA gene amplicons

The V4 region of the 16S rRNA gene was targeted with the primer set 515f-806r (forward 5' GTGCCAGCMGCCGCGGTAA 3', reverse 5' GGACTACHVGGGTWTCTAAT 3', Caporaso et al., 2011). Indexed amplicon libraries were pooled to a concentration of four µM. As usual for low diversity libraries, the PhiX control was spiked into the library pools at a concentration of 40 pM (10%).Each final library pool (4 pM) was subjected to 1 of 3 consecutive individual paired-end sequencing runs using 500 cycle V2 chemistry kits on an Illumina MiSeq (Berlin, Germany). Additional information with regard to the 16S rRNA gene libraries is provided in Supplementary Material 3.3.

3.2.3.3 Processing 16S rRNA gene amplicon sequences

Amplicon read processing – including the removal of primer and two-parent chimera sequences, the quality filtering step and the taxonomic annotation - was conducted using the DADA2 pipeline v 1.10.0 (Callahan et al., 2016) with R v. 3.5.1 (R Core Team, 2017). DADA2 corrected for sequencing errors during the generation of amplicon sequence variants (ASV). As recommended, such a correction was applied separately for each sequencing run. The individual tables were merged afterwards. Only ASV of length from 231 – 272 bp were kept according to the expected amplicon lengths reported in Ziesemer et al. (2015).

Taxonomic annotation of herein presented data was accomplished using the Silva release 132 (Yilmaz et al., 2014), including the taxonomic changes that were proposed by Parks et al. (2018). The ASV and taxonomy table were imported to and analyzed with phyloseq v. 1.30.0 (McMurdie and Holmes, 2013) accelerated by speedyseq v. 0.1.1 (McLaren, 2020). Plots were generated using ggplot2 v. 3.3.1 (Wickham, 2016).

ASV which were present in negative PCR or extraction controls and also found abundantly in actual samples were individually checked due to potential cross contamination directed from samples towards the controls. ASV with more than 35 reads in controls were removed from the dataset. ASV00001 was excluded from this rule because it was much more abundant in actual samples (extraction control: 75 reads, samples: > 10000 reads). ASV which were present in controls and less abundant in samples were removed. Subsequently, it was checked if any of the as important detected taxa were also present in control samples. ASV00063 belonged to the important genus *Maribacter* (4 reads in positive PCR control) and ASV00074 to *Cobetia* (5 reads in negative PCR control). As no reads were found in the extraction control and they were as abundant as up to 3000 reads in sediments, these ASV were left unaltered.

3.2.4 Machine learning analyses

Analyses were carried out on six virtual machines provided by the German Network for Bioinformatics Infrastructure (de.NBI Cloud). The virtual machines ran Ubuntu 18.04.4 LTS as operating system on 28 Intel Xeon Gold 6140s cores with 256 - 512 GB memory available. RF analyses were performed utilizing R package ranger v. 0.12.1 (Wright and Ziegler, 2017). ANNs were generated with the R Keras framework v.2.3.0.0 (Allaire and Chollet, 2020) and the TensorFlow back end v 2.2.0 (Allaire and Tang, 2020). Our efforts to extract abundance, taxonomical and contextual data from phyloseq objects and subject those to machine learning led to the development of the R package phyloseq2ML v. 0.5.1 (https://github.com/RJ333/phyloseq2ML). It facilitates modification and combining such data sets as needed – using objects of class “phyloseq” as source - and formats the data for the above mentioned machine learning implementations in R.

3.2.4.1 Challenges of a small biological data set

The presented data set consists of contextual subsets (e.g. by specific transects or sampled by a given method) which are likely to contain samples more similar to each other than to those of other subsets. To ensure that the model’s decision making was based on the presence of TNT rather than to a particular cruise or experiment, we developed guidelines to assess which samples were appropriate for ML analyses. First, the technical replicates were averaged. Then, if for a given subset of samples all of the following questions could be answered with yes, samples had to be removed from the subset to prevent potential spurious relationships between the presence of TNT and the prediction accuracy:

For all samples from the same cruise (incl. biological replicates) → do they originate from the same experiment? → and the same area? → and do the sediment sampling positions have horizontal distance of less than 20 m → and do they only contain one class (TNT present or TNT absent) OR is there a strong imbalance (e.g. 20 x TNT absent, 1 x TNT present)?

Following this guideline, led to a removal of 17 of the original 167 sediments (Supplementary Material 3.2).

3.2.4.2 Machine learning workflow

The remaining 150 samples were split into a training-validation set (in short: training set) consisting of 112 samples (75 %) and a holdout test set of 38 samples (25 %). This procedure was repeated to yield six different, random and reproducible splits of training and test sets.

In supervised learning, the training set for a model contains the independent variables and the corresponding continuous or discrete response variable. The measured TNT concentrations were categorized as response classes “absent” for concentrations below the detection limit (0.01 $ng \cdot g^{-1}$ or 0.044 $pmol \cdot g^{-1}$ wet sediment) and “present”.

Settings automatically derived from the learning process are called parameters, such as the weights between ANN nodes. Hyperparameters, instead, are model settings chosen before training has started. Random forests are controlled via two main hyperparameters: the number of trees per forest and the number of variables “mtry” to consider for sample separation at each tree node. The default value for mtry for classification tasks is the square root of the total number of independent variables. As this default value might not be optimal for sparse data such as ASV abundance tables, a factor multiplying this number of variables was used instead and will be referred to as “mtry factor” (Hastie et al., 2009).

RF models were trained on various combinations of hyperparameter values and input data to estimate the best performance on the holdout set. This process is called a grid search and combinations were compared using the out-of-bag validation error. A confusion matrix was generated to calculate performance metrics. Balanced accuracy was used as score. It corrects for imbalanced response variables and allowed comparisons across training set splits, which displayed class ratios of 43 - 48 % “TNT present” (Brodersen et al., 2010). The validation results of the six data splits were averaged to select the best performing hyperparameter values and input sets. When predicting the holdout set, the model was trained on the full training-validation set. The holdout predictions for the various input data sets took place after all hyperparameter values were determined. This is required to prevent data leakage.

3.2.4.3 TNT presence prediction based on Random forest grid search

Data sets designed as model input were threefold: a) community data: describing data deriving from 16S rRNA gene amplicon sequencing; b) sediment data: sediment parameters derived from geochemical and sedimentological analyses; c) combined, a combination of both aforementioned input sets.

The grid search with community data was performed as follows: All combinations of relative abundance thresholds, the number of trees and the mtry factor were investigated. ASV had to be more abundant than a given threshold in at least one sample. If so, the ASV remains without change, otherwise it was filtered out. Thresholds were: 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6, 0.8 and 1 %. Each of the resulting input sets was provided to models consisting of 100, 500, 1000, 5000, 10000 and 20000 trees along with mtry factors ranging from 1 to 13 by 2. For each combination 50 models were trained and validated.

Subsequently, the filtered relative ASV abundances were accumulated by taxonomic ranks genus through phylum to train 200 models with the previously identified hyperparameter values of 10000 trees, an mtry factor of 5 and a threshold of 0.08 %.

The sediment data contained 41 independent variables including reactive element contents, sum parameters such as total nitrogen, and the grain size distribution. Hundred models were

trained with 1000, 5000, 10000 trees and mtry factors 1, 3, and 5. For combined input data it was found sufficient to apply the same hyperparameters as were applied to the community data.

Validation and holdout scores were tested separately for significant differences between input data sets. Equal means were tested with unequal variance and one-way analysis of variance. The results of the analysis of variance were further subjected to the Tukey multiple comparisons of means with 95 % family-wise confidence level to identify the pairwise significances.

3.2.4.4 Selection of most important variables

The most important variables for classification were retrieved from models trained with community, sediment and combined data. Importance for community data (0.08 % threshold, genus rank) and combined data was calculated utilizing the corrected Gini impurity (Nembrini et al., 2018), followed by *p* value estimation after Janitza et al. (2018). A 100 models using 10000 trees and an mtry factor of 5 were trained and the results averaged. Variable importance and associated *p* value for sediment data required the permutation-based approach by Altmann et al. (2010). A 1000 permutations with mtry factor 1 and 10000 trees were applied. The analysis involved elements Zr, which likely was not soluble by HCl extraction as well as Ca and Sn, where the measurement by ICP-MS was later identified as unreliable. The elements were still included in the training data, but were not reported as important and removed for other analysis such as the Spearman rank correlation.

The variables were ordered by average importance over all splits. The number of variables for further analyses were selected based on decreasing decline in importance, meaning if the variables became more similarly important to each other, the cutoff was set. Thus, 25 genera were selected with Janitza importance > 0.25, $p < 0.01$ and 9 sediment parameters with Altmann importance > 0.001 and $p < 0.05$. The most important 50 combined variables (equal to Janitza importance > 0.15 and $p < 0.01$) were compared to the 25 community and 9 sediment variables.

3.2.4.5 Random forest's proximity matrix for PCA ordination and correlation

Ordination methods are useful to explore multivariate data sets such as microbial community compositions. The proximity matrix generated by random forests keeps count of samples which end up in the same terminal node of a decision tree and, therefore, is a measure of (dis-) similarity. It can be used with unsupervised classification: a synthetic data set is added to the original data set. This consists of shuffled columns of the actual data, thus breaking all relationships between variables. The model (10000 trees, mtry factor 1) tries to distinguish between permuted and original data and thereby identifies correlations and clusters in the

actual data set. For supervised classification, the actual classes were used and no synthetic data set was required.

Principal component analysis (PCA) was performed based on the proximity matrix for the most important 25 genera. To identify microbial community shaping influences for the unsupervised classification, the sediment parameters were correlated with the PCA ordination. The function *envfit()* from R package vegan v. 2.5-6 (Oksanen et al., 2019) with 9999 permutations was used to achieve this. Correlating parameters with $p < 0.001$ and $R^2 > 0.3$ were displayed. The PCA ordination was performed for sediment data as described above, except the *envfit()* step.

Complementary, Spearman's rank-order correlations between sediment variables were investigated. The results were hierarchically clustered and variables with $p < 0.01$ were marked significant.

3.2.4.6 Assessing robustness of classification with random forest and artificial neural nets

The classification consistency was examined to increase the understanding of the predictions. All 150 samples were used as training and validation set for 1000 models (10000 trees, mtry factor 1). Mean prediction errors < 0.5 % or > 99.5% accuracy were rounded to 0 and 100 %, respectively.

Artificial neural networks (ANN) were additionally deployed to measure classification robustness across algorithms. The input data for ANNs required additional steps including the one-hot encoding of categorical variables and scaling of the independent variables: the mean of each variable was subtracted, and it was divided by the standard deviation. This yielded values centered around 0 with a standard deviation of 1. ANN grid searches were performed complementary to what is described for random forest above. Results suggested that 50 nodes in the first hidden layer and 40 nodes in the second hidden layer were appropriate values, along a mini-batch training size of 4. No regularization was applied. The optimizer function Adaptive Moment Estimation outperformed Root Mean Square Propagation. Binary cross entropy was set as loss function, with accuracy as metric. Learning took a maximum of 100 epochs, stopped by an early callback if the validation loss did not decrease for 2 ongoing epochs. The node within the hidden layers were rectified linear unit-activated whereas the output nodes' activation function was sigmoid. Further hyperparameters and settings were default values of the keras R package.

Performance assessment was achieved by splitting the training data into three different, non-overlapping equally proportioned subsets. Two partitions were used for training and the remaining one for validation. These three subsets were composed differently for each of the

conducted 333 runs. This 333 times repeated 3-fold cross validation yielded a total of 999 predictions.

3.2.5 Data availability

Code, scripts and files are available under GitHub (https://github.com/RJ333/). The R package phyloseq2ML is deposited at https://github.com/RJ333/phyloseq2ML. Sequences were deposited in the NCBI database under BioProject ID PRJNA632711 and SRA accessions SAMN14917999 - SAMN14918370. Geochemical data is included in Supplementary Material 3.2.

3.3 Results

3.3.1 TNT contamination of Kolberger Heide sediments

To provide an overview of the contamination levels and distribution at Kolberger Heide munitions dumpsite, information from all 167 sediments was taken into account. A selection of 150 sediments was then used specifically for ML. Out of 167 original sediments, 148 contained MC: TNT (detected in 70 sediments), 2-ADNT (135), 4-ADNT (144), 2,4-DANT (70), 2,6-DANT (55). None of the other MC (2,4-DNT, 2,6-DNT, DNB, TNB, HMX, RDX, Tetryl) were detected in more than 8 sediments (Supplementary Material 3.2).

TNT was determined to be present at levels less than 25 pmol·g^{-1} wet sediment in 65 samples. Notably, the highest values of 587, 690 and 3485 pmol·g^{-1} were found in three sediments retrieved from a detonation site, where exposed munition chunks were spread over the sea floor.

The heavy metals mercury and lead were used as proxies for primary explosive compounds in conventional ammunition, which potentially could be present at the dumpsite; chemical warfare agents can contain arsenic. Mercury contents ranged in Kolberger Heide sediments from 3.7 to 4503 µg Hg·kg^{-1} dry sediment, with a median of 21 µg Hg·kg^{-1} and 15 samples exceeding 450 µg Hg·kg^{-1}. The maximal content of 4503 µg Hg·kg^{-1} was found during a line transect, where samples were taken every 20 m. The neighboring samples to the maximal value contained 8 and 12 µg Hg·kg^{-1}, demarcating a precise area of elevated Hg presence. Arsenic appeared on level between 0.4 and 4.8 mg kg^{-1} with a median of 0.8 mg·kg^{-1} and lead ranged from 1 to 75 mg·kg^{-1} with a median of 2 mg·kg^{-1}.

The microbial community composition of the sediments was investigated for measurable effects caused by the TNT. A total of 279 16S rRNA gene libraries were generated from the 167 sediments; 259 libraries from 150 sediments were selected to be appropriate for ML purposes. The selected libraries had a mean size of 82219 reads, with the 95 % confidence intervals being 78115 and 86322 reads (Supplementary Material 3.3). Averaging ultimately yielded 150 community tables comprising 66230 ASV, 1703 genera and 78 phyla available for

machine learning. The 150 samples selected for Machine Learning contained TNT in 68 cases (45.3 %). Due to the skewed distribution a binary classification approach was adopted.

3.3.2 Community data predicted TNT presence more accurate than sediment data

A selection of eight input data sets was compared for their potential to predict the presence of TNT. The achieved validation and prediction scores were averaged over the six training/test splits (Figure 3.1). Full sediment contained 41 independent variables and Full community (0.08 % relative abundance threshold) included 542 genera. The mtry factor 5 allowed for 115 genera being considered at each node. The 0.08 % threshold yielded the second highest mean balanced accuracy among the examined threshold values, and showed a more distinct classification distribution (Supplementary Material 3.4), therefore it was applied to all community data sets presented here.

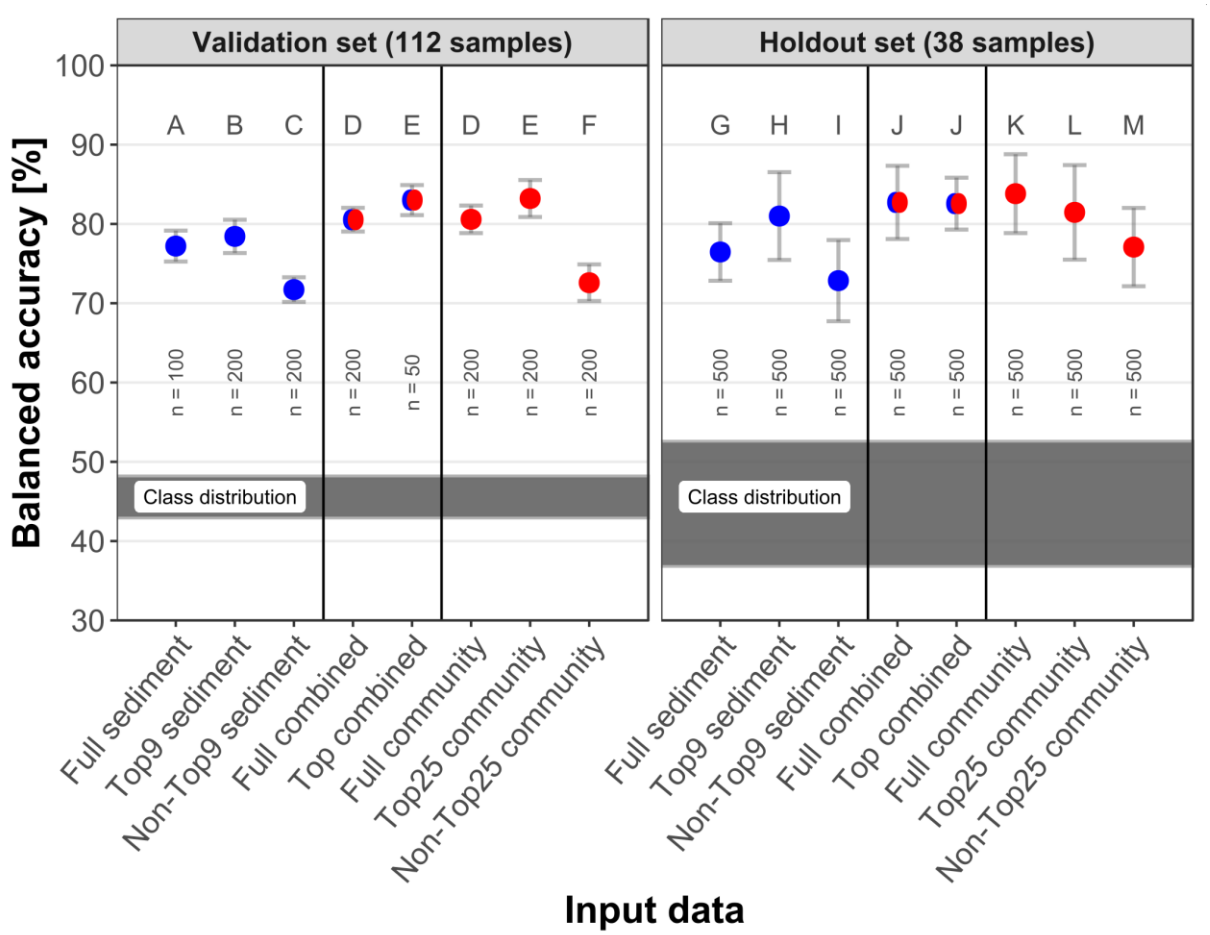


Figure 3.1: Correct TNT classifications per input data in the validation and hold out test set. Red indicates community data, blue symbolizes sediment data and red-blue combined variables. Of each data type, either all variables were utilized by the model ("Full"), or only the best variables based on variable importance ("Top") or all variables except Top ("Non-Top"). Classification performance is displayed as mean and standard deviation of balanced accuracy, the classification results of the six different data set splits were averaged. The validation values are out-of-bag estimates. The letters indicate which groups were significantly (adjusted $p < 0.005$) different to all other groups within the data set. The shaded area indicates the distribution of samples containing TNT among the six data set splits. n indicates the number of models calculated.

With reference to the validation set, which was used to optimize the classification, selections of either the most important 25 genera or 9 sediment parameters yielded more accurate

classifications than by using all variables; the lowest scores were achieved by putting the remaining non-important variables to use. In this order, the mean balanced accuracy for sediment data decreased from 78.4 over 77.2 to 71.7 % and for the community data from 83.2 over 80.6 to 72.6 %. Using the most important variables from both data sets combined also improved the classification from 80.5 to 83.0 %. The Top25 community represents 4.6 % of the genera and increased the balanced accuracy, whereas the other 517 genera significantly reduced it. For each variable selection (Full, Top, Non-Top), the community data performed better than the corresponding sediment data. The combined input data achieved classifications similar to community data alone.

TNT was present in 44 to 48 % of the samples in the six training data sets. The holdout set contained fewer samples; consequentially one sample's classification represented > 2.5 % accuracy. This led to more widespread class ratios from 36 to 52 % and a higher standard deviation. Best predictions reached 83.8 % with Full community and 82.7 and 82.6 % with Full combined and Top combined, respectively. Predictions on the holdout set were slightly better than the corresponding validation scores, excluding Top combined and Top genus. The largest difference between validation and holdout scores was an increase of 4 % for "Non-Top25 community". Validation and holdout scores met the same range from 70 to 85 %, but showed no further relation.

All balanced accuracy means in the validation set were significantly different from each other (adjusted $p < 0.005$) except "Full community" to "Full combined" (D) and "Top25 community" to "Top combined" (E) in the validation set. This extended to all groups in the holdout set but "Full combined" to "Top combined" (J).

The hierarchical structure of the taxonomic annotation allowed investigating the influence of pooling the relative abundance by taxonomic ranks (Figure 3.2) to identify the best compromise between the number of taxa and the information contained in inter-taxa abundance variability. The highest mean balanced accuracy was achieved by ASV (82.9 %) and decreased towards the broader order rank (74.9 %). Training with relative abundance per class (78.8 %) and phylum (76.9 %), however, still resulted in acceptable predictions, yielding more accurate classifications than on order rank. The genus rank (80.6 %) was chosen for further analyses, because it contained fewer and more interpretable variables.

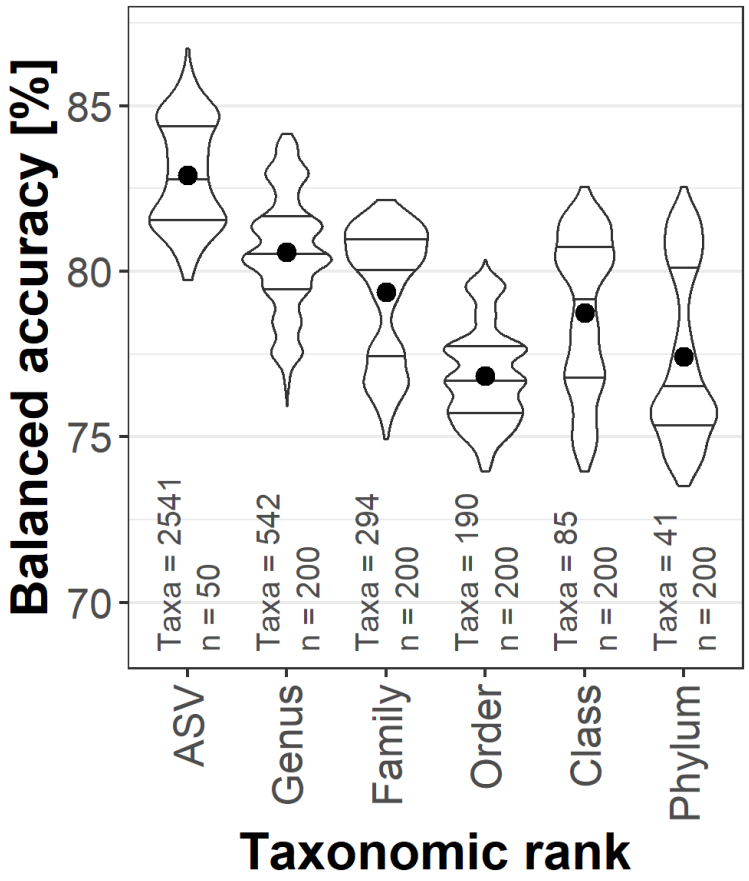


Figure 3.2: Violin plots with median and interquartile range of correct TNT classifications of the validation set. The relative abundances were agglomerated on the taxonomic ranks. The dot represents the mean balanced accuracy; the classification results of the six different data set splits were averaged. n indicates the number of models calculated, Taxa represents the number of variables for each rank.

The distribution of information among samples was then assessed by comparing the validation scores for the six sample compositions. The results showed that Full community was more accurate for each set (Figure 3.3). Between the splits, a range > 5 % in the scores for Full sediment (75.1 - 80.5 % mean balanced accuracy) and Full community (77.9 - 83.2 %) was observed. Shifts in balanced accuracy between splits were not consistent for sediment and community data. For example, comparing Data split 1 and 2 the Full sediment classification performance dropped whereas the Full community balanced accuracy was maintained. These findings signal that the available sediment parameters and taxa abundances did not supply equivalent information.

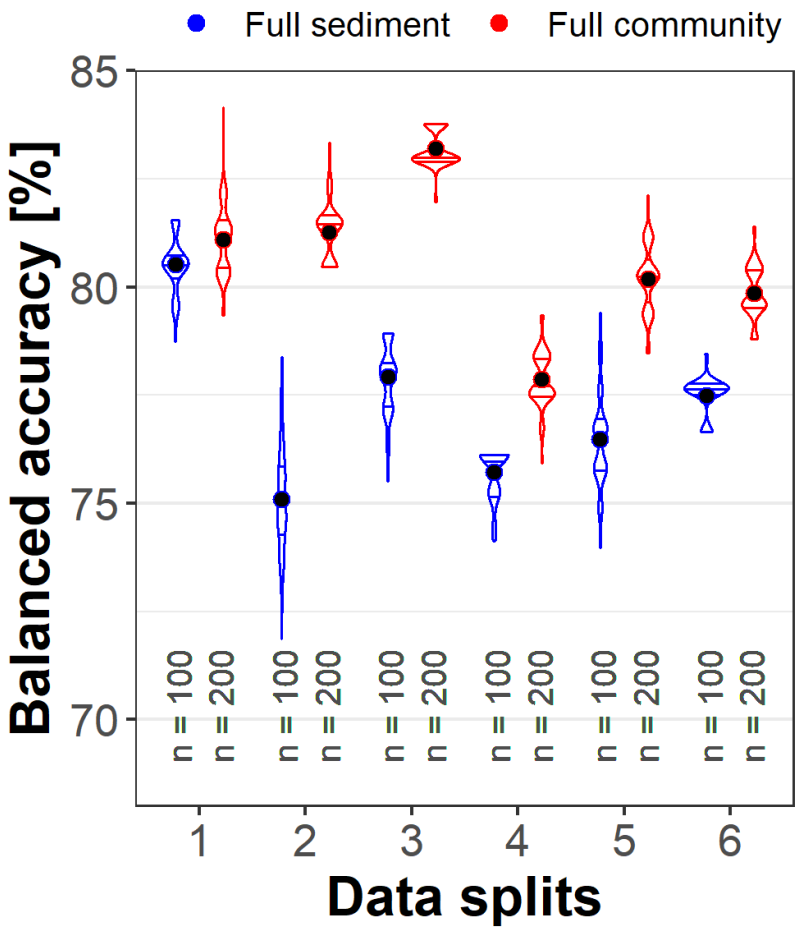


Figure 3.3: Violin plots with median and interquartile range of correct TNT classifications for six different validation sets. Full community (red) always performed better than Full sediment (blue) and their performances changed independent of each other towards the different validation set compositions. The dot represents the mean balanced accuracy; n indicates the number of models calculated.

3.3.3 Grain size distribution as the major driver of community composition

After successful classifications were achieved using community information, TNT was investigated with regard to its potential as important driver of the microbial community composition; as such influence would facilitate the process of prediction. PCA ordination of the Top25 community was performed using the sample proximity obtained by an unsupervised random forest classification. PC 1 explained 56.1 % variation. Along PC1, the grain size fractions < 125 µm were separated from those > 250 µm (Figure 3.4 A).

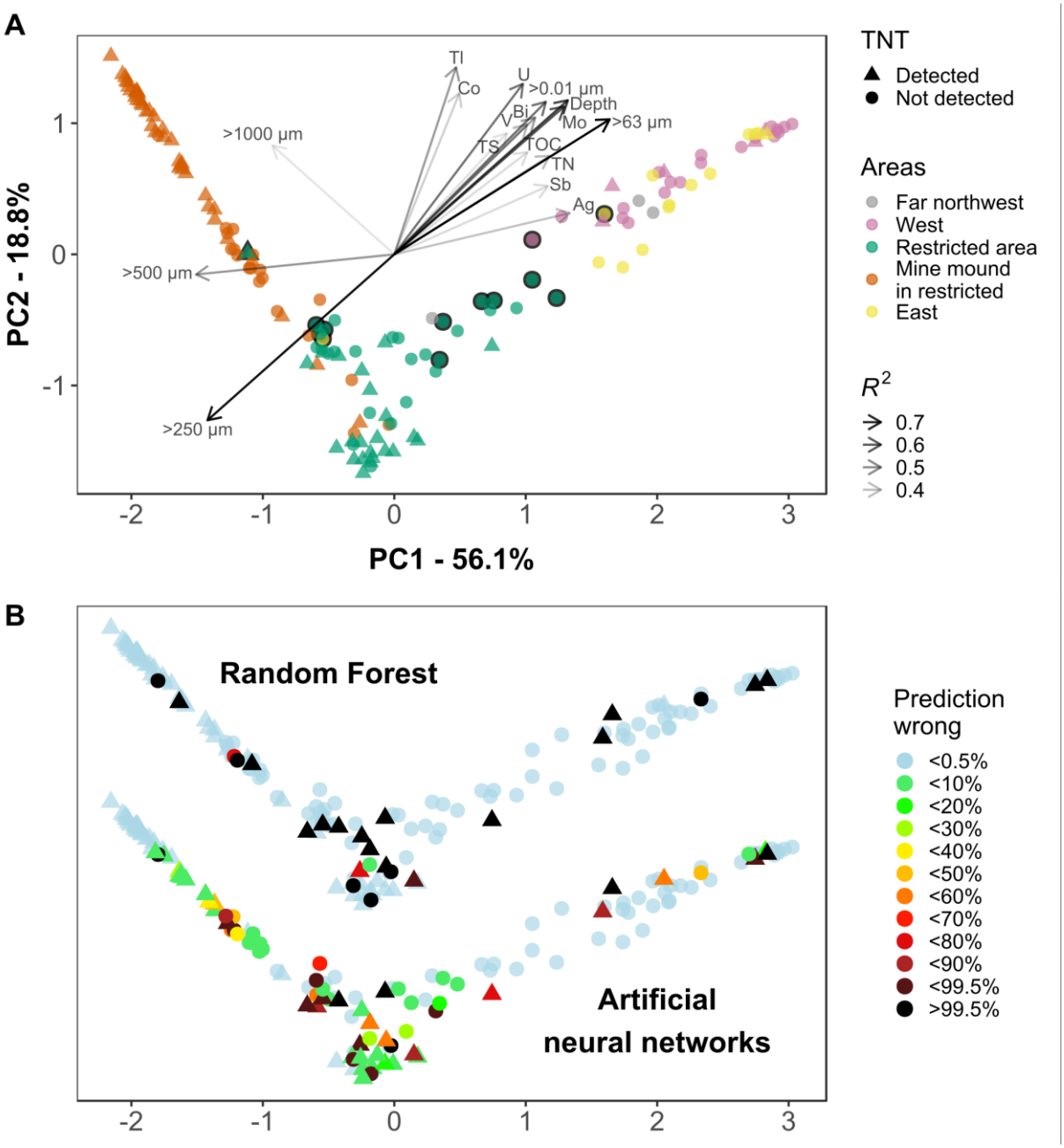


Figure 3.4: PCA ordination based on the abundance of the most important 25 genera. Dissimilarity between samples was calculated using the proximity matrix of an unsupervised random forest. A) The microbial communities were colored by sample area and shaped to indicate the presence of TNT. The length and shade of correlating sediment parameters ($p < 0.001$, $R2 > 0.3$) represents the goodness of fit. The black outline marks East (yellow) and West (purple) samples which were not MUC samples. Similarly, the outline marks Restricted Area samples that were not part of a transect. B) Using the same ordination as in A, the fraction of misclassifications per 1000 (RF, top) and 999 (ANN, bottom) predictions is displayed for each sample. Light blue colored samples were always correctly predicted, black displays consistently misclassified samples. Please note: the y-axis (PC2 in A) was stretched to accommodate the results from both methods.

The latter spread along PC2, which explained 18.8 % variation. The former fractions co-correlated with further sediment parameters; some of those were important variables for random forest when using Full sediment (vanadium, cobalt, total nitrogen). Significant correlations with MC were not found. The highest accordance among MC with the community composition ordination was shown by 2,6-DNT with R^2 of 0.033 and p 0.07. TNT (R^2: 0.014, p:

0.38) was detected across all clusters, but predominantly present in mine mound samples. Only a few core samples contained TNT.

The multicorer samples comprised smaller sized particles than most surface sediments. They were sliced at 2 cm, from the sediment-water interface to 22 cm depth (Figure 3.4 A, West and East areas, no black outline) and formed a prominent cluster, with communities driven by the grain size distribution and presumably the redox potential declining with depth. The region did not play a role for clustering, as cores were collected kilometers east and west of the mine mound, which itself is centrally located in the restricted area (Supplementary Material 3.1).

The samples from the mine mound area (a cluster of about 70 mines) were mostly taken within a defined distance of 0 - 5 m to a mine. Although this is a part within the restricted area, the communities mostly grouped together. Several transects with sampling intervals of 20 m were conducted across the restricted area, surrounding the mine mound (Figure 3.4 A, Restricted Area, no black outline). The corresponding communities formed a distinct cluster, too. Three more samples with no detected TNT were collected multiple kilometers away towards northwest.

An ordination based on only the sediment parameters including the MC was generated to compare with the microbial community ordination. Again, no separation based on TNT presence was observed (Supplementary Material 3.5). Furthermore, the mine mound and the overall restricted area sediments clustered alongside, with eastern samples placed in proximity. In this ordination the MUC samples to the east and west were clearly separated, with west and far northwest samples forming a remote cluster.

The seasonal conditions during sampling should be mentioned, as they might have influenced the community composition more strongly than the sediment parameters. The restricted area was sampled mostly manually in June and September 2017 at the sediment surface by divers; three more sediments were obtained using Van Veen grab samplers. The mine mound samplings by divers took place in December 2016 and November and December 2017, which could explain the division between mine mound and restricted area microbial communities. The cores were collected on one day in October 2017.

Random forest was able to predict TNT using only sediment parameters, although no driving influence by MC were detected in the ordinations. Therefore, Spearman rank-order correlation was performed to investigate which variables significantly ($p < 0.01$) correlated with TNT. A cluster of MC consisting of TNT and its metabolites 2-ADNT, 4-ADNT, 2,4-DANT and 2,6-DANT was identified, which also showed a loose positive correlation with RDX (Supplementary Material 3.6). Another cluster consisted of DNB, HMX, TNB, 2,4-DNT and 2,6-DNT. The latter

two MC are co-contaminants of TNT. However, the MC were not part of the random forest input data set. Furthermore, some weaker correlations with TNT were identified.

The results confirmed that community compositions were primarily controlled by factors other than the presence of TNT; therefore, supervised classification was applied to still extract such a potential impact. Both community and sediment data-based ordination demonstrated as well, that the distribution of TNT containing samples was appropriate to utilize machine learning.

3.3.4 Community information important in combined data sets

Foregoing results indicated that a potential impact of TNT was masked by stronger drivers. Therefore, it was essential to investigate the variables that enabled RF predictions. Potential microbial fingerprints (in case of community data) indicative for the presence of TNT were examined. The variable importances, extended by maximal relative abundances and taxonomic lineage of the genera are provided in Supplementary Material 3.7.

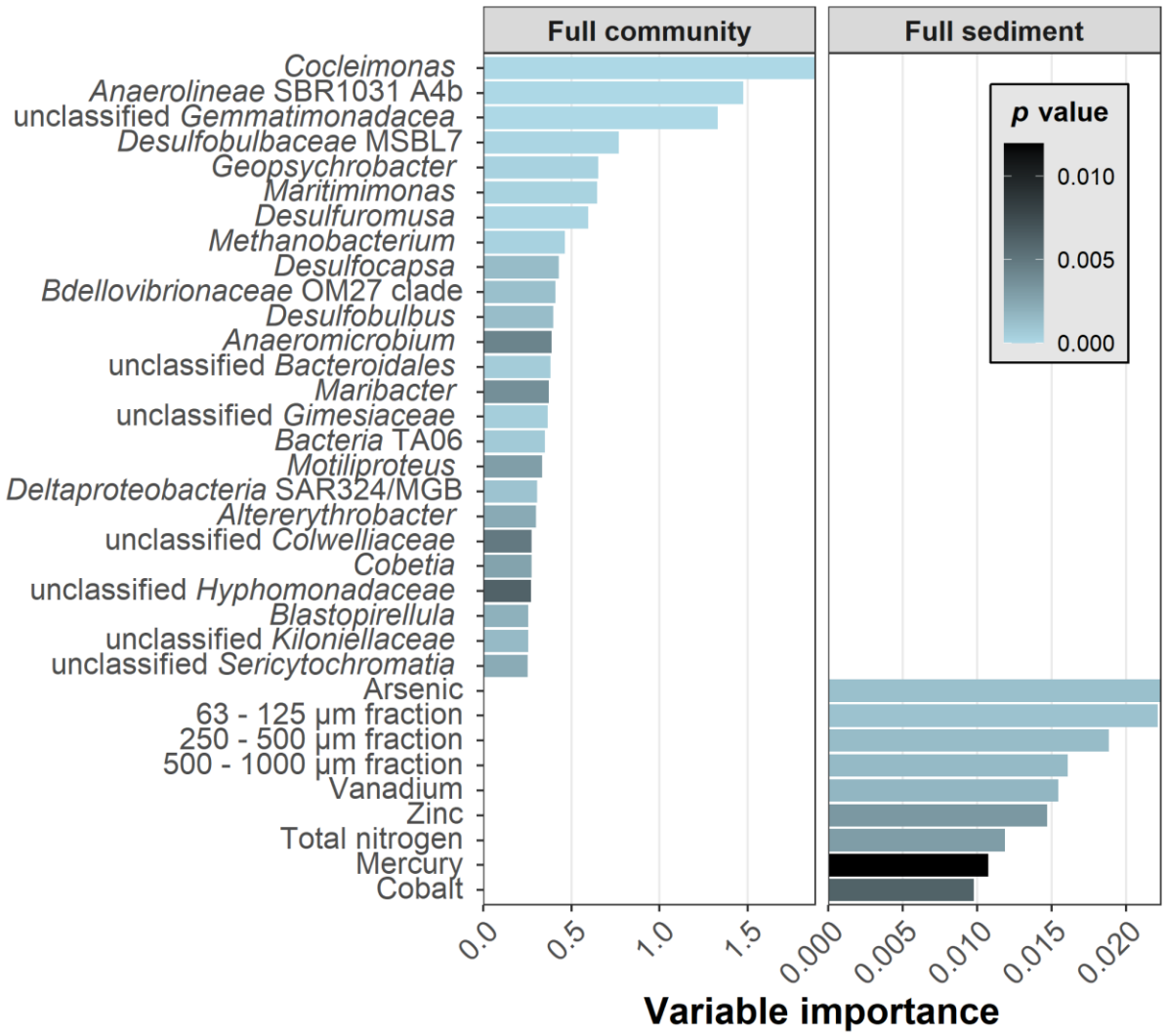


Figure 3.5: The variable importance and *p* values for the classification of TNT presence. Twenty-five genera of the Full community and 9 sediment parameters of Full sediment were selected. The most detailed taxonomic annotation was provided in case none was available at genus rank. Importance and *p* values were generated after Altmann (Full sediment) and Janitza (Full community) for six data split and subsequently averaged.

The most supportive genera (Figure 3.5) were *Cocleimonas (*1.65 % maximal relative abundance), the unclassified *Anaerolineae* SBR1031 A4b (0.11 %) and an unclassified *Gemmatimonadaceae* (0.38 %). Relative abundances of the Top25 genera ranged from 5.65 % for the unclassified Cyanobacterium *Sericytochromatia* to 0.09 % for the unclassified Planctomycete *Gimesiaceae.* The important sediment variables contained grain size fractions, elemental contents, and total nitrogen as a sum parameter for various nitrogen compounds. Among these, arsenic and the 63 – 125 µm fraction were most important. This grain size fraction correlated with sum parameters of sulfur and carbon and element contents of e.g. molybdenum and uranium in direction of the MUC samples.

The 50 most important Full combined variables were then compared against the foregoing top Full community and Full sediment variables. Interestingly, out of 50 variables only 6 were sediment parameters (arsenic (#9), zinc (#21), 63 – 125 µm fraction (#35), vanadium (#40), mercury (#45), cobalt (#48)), all of them were part of the Top9 sediment. The achieved classification score of Full combined was as accurate as by Full community input (Figure 3.1). The 44 genera included all of the Top25 community genera. Further genera were related to them on family or order level, for example *Flavobacteriaceae, Clostridiales, Sphingomonadaceae* and *Desulfobulbaceae*. Overall, recovered variables in the combined data set were as important as in individual data sets. Sediment importance ranking concurred, although they were calculated using two different methods for Full community and Full combined.

3.3.5 Processing of all samples depends on a combination of important variables

To understand the model's approach to classify the samples and to validate a potential indicative fingerprint, the reasons for the determination of important variables had to be identified. By analyzing their relative abundance it became clear that 23 of 25 important genera were in average more abundant in surface than core samples, the opposite was true for the clostridium *Anaeromicrobium* and TA06 (Supplementary Material 3.8, I, Y).

Although the abundance of the most important *Cocleimonas* could be very low in samples regardless of class, it mostly occurred in samples with TNT. Second most important *Anaerolineae* SBR1031 A4b proved to be more abundant overall in samples with TNT. Clade TA06, however, was found in as few as 12 samples, and was abundant in very similar sediments of both classes (Supplementary Material 3.8, H, P, Y). The presence of some genera was linked to grain sizes: *Cobetia* was present in medium to finer sediments, *Colwelliaceae* on the contrary in coarser samples (Supplementary Material 3.8, C, G). This goes along with the finding that in a combined data set the grain size information was not as important anymore. But other important genera such as the up to 4.1 % abundant *Maribacter*, *Maritimonas* (3.5 %) and *Blastopirellula* (4.6 %) were present in 131 to 142 of 150 samples

(Supplementary Material 3.8, B, D, E). In a similar fashion, the concentrations of sediment parameters were displayed in Supplementary Material 3.9.

3.3.6 RF predictions were consistent, with transect samples being most challenging

With achieved classification scores for the presence of TNT well above 80% the inner works of the model for the important variables became understandable, but additional information on misclassified samples was required. By recording the mean of 1000 predictions, it was possible to identify consistently and/or incorrectly classified samples (Figure 3.4 B).

Random Forest had cumulatively 24 of 150 samples misclassified (84 % accuracy), including 5 of 35 core samples and 6 of 58 sediments near the mine mound. These predictions were robust; a classification was either wrong or correct, taken 0.5 % tolerance into account. Only four samples showed varying classifications, being incorrectly classified 1.3, 71, 79 and 93 % of the time.

A PCA ordination based on a TNT classifying model showed the attempt to cluster by class: clusters in top right and bottom center were predominantly TNT-present and in the top left mostly TNT-absent (Supplementary Material 3.10). The center region displayed communities of both classes intermingled. Samples of all areas were observed there, but those from the restricted area were most present with both classes. It is likely that the samples in the center region were more often misclassified. Finding two separate clusters for TNT-present samples indicated that two distinct groups of important variables contained in the model were required to achieve classifications of those samples.

The restricted area achieved the highest misclassification rate. Within a total of 51 sediments for this region, all 13 misclassifications could be attributed to 41 samples collected by four transects (Supplementary Material 3.10, Restricted area, no black outline, see also Figure 3.4 A). The 200 m long transects, each consisting of 9 to 11 sampling points, covered different sections of the restricted area.

In general, the less abundant class in a given region is prone to misclassification; however, minority class samples were also predicted correctly. The inconsistently classified samples can be imagined close to the decision boundaries between predominantly “present” and “absent” groups (Figure 3.4 B, Random Forest).

The robustness test utilizing an ANN gathered 70 wrongly predicted samples in 999 classifications. Sixty-four of those were not robust. More specifically, 30 samples were misclassified less than 10 % of the time and another 11 samples were almost more frequently than 99.5 % misclassified. Furthermore, all samples incorrectly classified by random forest were misclassified by the ANN, too. Regarding the higher prediction variance of the ANN it should be noted that RF is an ensemble classifier (see Discussion).

3.3.7 TNT metabolites containing samples more likely to be classified false positive

The presence of ADNTs or DANTs in sediments indicates that TNT had been present. It was hypothesized that such former TNT-containing sediments might harbor community compositions which “look like” TNT was still present after its dissipation due to resilience. In consequence such samples should be predicted falsely positive. A “clean” sample on the other hand contains neither TNT nor its metabolites, indicating that it was not contaminated with TNT for a longer time.

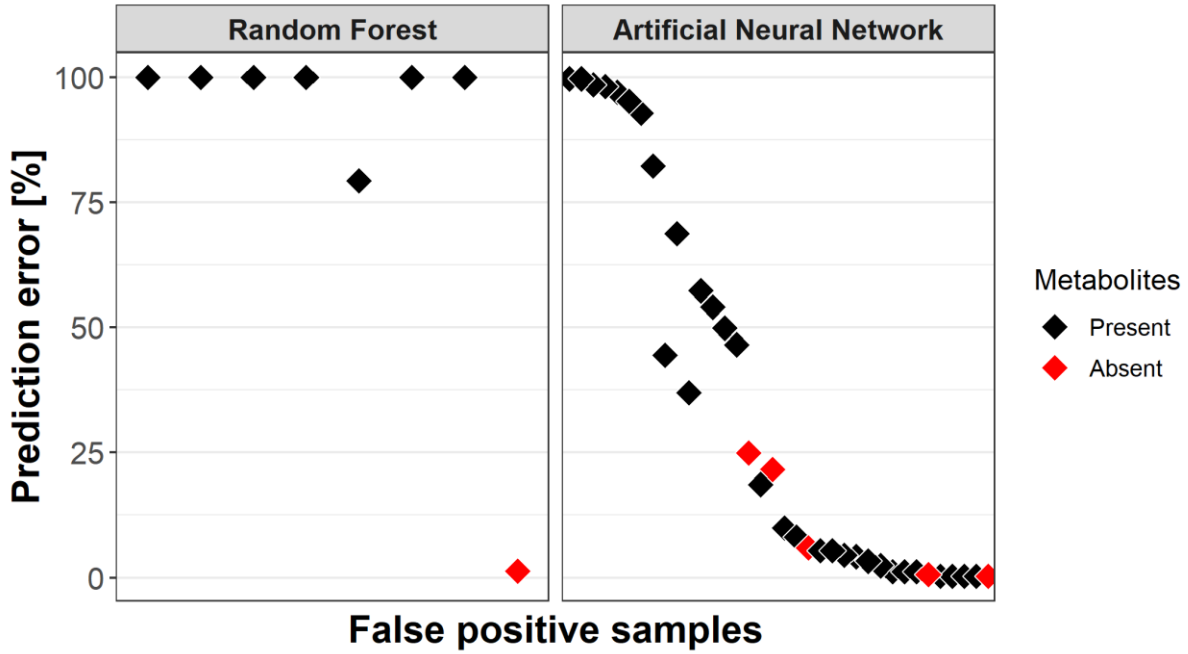


Figure 3.6: Misclassification rates of samples which were predicted “TNT present” but did not contain the explosive (False positive prediction). Red indicates whether a false positive samples contained TNT metabolites, i.e. ADNTs and DANTs. Samples containing metabolites were more likely to be misclassified as false positives.

The RF models predicted eight false positives; two of them were not consistently misclassified (Figure 3.6). Interestingly, seven of the false positives actually contained TNT metabolites and the one “clean” sample was only 1.3 % times incorrectly classified. The ANN predicted 36 false positives, 5 of those without metabolites. Their prediction errors ranged from 0.3 – 25 % with an average of 10.6 %, compared to a mean prediction error of 38.3 % for the remaining false positives with metabolites. Furthermore, prediction rates for false positives did not correlate with the individual or summed concentration of TNT metabolites.

It was additionally verified whether a higher TNT concentration goes together with a stronger impact on the community composition, thereby decreasing the probability of a false negative prediction. However, the RF predictions contained only two suitable false negative samples. For ANN a higher TNT concentration did not lead to better prediction rates of the sample.

3.4 Discussion

In this study, microbial communities were used to predict the presence of TNT in sediments (at pmol·g^{-1} levels) in and around a munitions dumpsite in the German Baltic Sea with about 84 % balanced accuracy. Genera and sediment parameters being most important to reach this value, and the samples that were a challenge to the models, could be identified. Moreover, many TNT false-positive samples had traces of TNT metabolites, indicating that microbial community compositions may conserve information of former TNT presence for a longer period.

3.4.1 Model-relevant genera were related to TNT-degrading taxa

A selection of 9 sediment parameters or 25 bacterial genera predicted TNT as well (holdout set) or even better (validation) compared to using all variables. This is a result similar to the results of Thompson et al. (2019), who conducted a study to predict concentrations of dissolved organic carbon using most effectively a subset of the microbial community compositions of a plant litter decomposition experiment. One reason for such improved performances could be a lower likelihood of overfitting.

The subset was identified by the variable importance metric, which indicates correlation with the response variable. A potential causation between TNT presence and identified important genera is attributable to TNT as a source for biomass generation, energy supply or toxic stress (George et al., 2009; Gallagher et al., 2010). The bacterial enzymatic degradation of TNT is mediated by nitroreductases. Nitroreductases and other common enzyme families have been reported as responsible for the reduction of nitro groups (Esteve-Núñez et al., 2001), which are among the first steps of microbial TNT transformation. Such enzymes are widely distributed among microorganisms, rendering microbial TNT metabolization possible in marine sediments (Roldán et al., 2008). In fact, TNT degradation products as ADNTs and DANTs were present in Kolberger Heide sediments. The ability to degrade TNT was specifically proven for more than 20 different genera, ranging from anaerobic members of the family *Clostridiaceae* to aerobic members of the family *Pseudomonadaceae* (Esteve-Núñez et al., 2001). Relatives of these organisms are important for the models of our study; for instance, the Top25 and Top50 members *Anaeromicrobium* and *Clostridiaceae* sensu stricto 13, respectively, are phylogenetic members of the *Clostridiaceae*. Top25 *Altererythrobacter* is also phylogenetically related to TNT-degrading *Sphingomonas sanguinis* (Habineza et al., 2017). However, deriving bacterial activities from phylogenetic relations has to be handled carefully as phylogeny can be an unreliable indicator of bacterial ecology. Thus, it is also possible that the obligate anaerobic *Anaeromicrobium* acted as redox indicator for reduced conditions in the investigated sediments.

It was furthermore shown that multiple genera were required to separate classes in all samples, because some important taxa, such as clade TA06, were only detected in 12 samples. Consequently, their contribution to classification was limited. However, these genera were likely important, because they allowed classification of otherwise similar samples. In this regard, other variables could not replace this information.

The prediction of TNT was still successful using the available sediment information alone. We assume that, in this case, many samples were separated first and foremost by the grain size distribution, as the finer multicorer samples contained many TNT-free sediments, compared to the coarser mine mound samples, consisting of many TNT-contaminated sediments. The other parameters further on separated within those groups specifically. In order to supplement microbial community variables one might intuitively assume that at least grain size and, where appropriate, redox conditions should be measured as major proxies to inform the model. However, the combined usage of community compositions and sediment parameters did not lead to predictions more accurate than by using the community data on its own. It turned out that the second most important Full sediment variable (63 – 125 µm grain size fraction) was only the 35th most important Full combined variable and the other grain size fractions were not included in the Top50. These findings show that taxa abundances can replace the grain size information because it is reflected by the community data.

More information would be required to conclusively determine the reason why samples from the mine mound area, which is located in the center of the restricted area, formed a distinct cluster in the unsupervised PCA ordination (Fig. 4 A). This was noticeable, as the transect samples formed another distinct cluster, though the transects geographically encircled the mine mound. We suggest the sampling of the mine mound in a different season than the conduct of the transects as a main reason for varying assemblages (Meyer-Reil, 1983), but the proximity to mines as factor cannot be ruled out. Such an influence, however, was not displayed by the measured sediment variables (Supplementary Material 3.5), where sediments from the mine mound and the restricted area clustered more similarly.

3.4.2 The microbial fingerprint requires further data to become indicative

A meaningful indicative microbial fingerprint is equivalent with the abundances of important variables per response class, if they are causally related. Yet the clade TA06 was detected in 12 of 150 samples, which increases the likelihood of being only coincidentally useful; in other words, the sample size is too small to know whether overfitting occurred (Dietterich, 1995). Thus, there is a need to reduce the potential of spurious relationships. To receive a reliable, generalizable and informative fingerprint we propose to: a) maximize the sample to variable ratio by using a minimum number of taxa while still reaching acceptable predictions, e.g. using backward elimination (Guyon et al., 2002); b) add samples of further targeted sites and

conditions, which cover all response classes, and c) perform regression instead of classification as long as the concentration of the response variable is appropriately distributed and covered. Regression yields a more informative relation between response and community composition and avoids arbitrary limits between response classes.

In our study, the 150 samples were split into six different training and test sets. The test set is usually the ultimate benchmark for the predictive potential of the model, but it was likely that not all samples in our data set were equally different from each other. Therefore, the hyperparameters as well as the important variables were based on averaged results from the six sample set compositions. This approach can be seen as extra layer of repeated cross validation and helps to maximize the generality of the fingerprint and the chosen settings. It also resulted in more reliable prediction accuracies, as for an individual sample split mean balanced accuracies > 90 % were achieved. Important is that by this approach a training sample of one split is also a test sample of another split. This results ultimately in information leakage, although in a rather indirect way (Kaufman et al., 2011). We argue that this approach is justifiable for our small data set, where the detection of a generalized TNT-indicative microbial fingerprint as proof of principle was the priority. But in larger data sets, or to compare different prediction methods, regular approaches with a fixed hold out test set should be applied. It should also be remembered that if such a model would be actually deployed, the data to be predicted, e.g. from the next sampling campaign, would not yet exist.

With regard to an indicative fingerprint, we conclude that the presented data set probably contains essential parts of it, but is not yet suited to distinguish accidentally valuable from truly influenced variables. However, we conclude that the first steps were successfully taken to determine a microbial fingerprint indicative of TNT contamination in Kolberger Heide.

3.4.3 An indicative microbial community fingerprint may differ between habitats

Given the existence of such a fingerprint, part of its value is to use it for other areas of interest. In this regard, the usage of microbial community compositions has both advantages and drawbacks. Advantageous is that the features were assigned at least a partial taxonomy; thus, are interpretable and relatable to literature or cultivation dependent complimentary investigations. Yet, using taxa infers using a proxy, depending on many influences such as nutrients, salinity, redox, pH, temperature (Lindh and Pinhassi, 2018) or as described in this study, grain size.

In order to create meaningful fingerprints, communities likely need to originate from a somewhat similar habitat under specific conditions. But, importantly, our models still could predict using data from various habitats - as from deeper multicorer and surface sediment samples - albeit the variable importance would be a mixture of habitat fingerprints and therefore less interpretable. Additionally, the important taxa might not occur everywhere. To address this

issue, higher tax ranks can be used, which are more likely to be found in various areas. Ghannam et al. (2020), for instance, used phyla to differentiate geographic locations on global scale. In our spatially restricted samples the phylum rank also achieved 76.9 % mean balanced accuracy, which is still well above coincidence. But the context of the response variable should be considered, as a higher taxonomic rank is reasonable to cover taxa globally. However, in a previous study we detected distinct reactions to the herbicide glyphosate at OTU-level for *Pseudomonas,* which were not distinguishable anymore on genus level (Janßen et al., 2019a). An alternative is to combine important variables from all taxonomic ranks and train with those.

Furthermore, it is conceivable to target functions (genes or transcripts) directly by shotgun sequencing instead of using taxa as proxy. Alneberg et al. (2020) demonstrated that functional genes from metagenome assembled genomes predicted salinity and depth in Baltic Sea waters.

3.4.4 Misclassified samples define further sampling campaigns

Two mechanistically different ML algorithms were able to predict the presence of TNT in Kolberger Heide sediments using 25 genera. The samples misclassified by RF were also misclassified by the ANN, indicating that the data were insufficient in that case, independent of the algorithm in use. The more consistent predictions of RF stem in part from it being an ensemble classifier (Breiman, 2001a). Thus, all the individual predictions of the tree models are not published, as they are for ANNs, but reduced to a single prediction based on a majority vote. As ANNs do not have this leveling mechanism by default, more variance in cumulated classifications was observed.

It seems reasonable to explore the microbial community composition by proximity matrix-based ordinations, using the same distance metric that is used for the supervised classification. It allows correlating environmental variables, the addition of context data and provides an understanding on the data set dynamics. Combined with the classification robustness it becomes a powerful approach to determine model limitations as well as their overcoming (e.g. more transect samples, Supplementary Material 3.10). It can be compared to the supervised ordination, which indicates the separation by TNT presence or absence and confirmed that many of the samples consistently misclassified were not well separated. For more insights, decision boundaries can be added (for one model at a time [Hastie et al., 2009]).

3.4.5 Resilience of TNT presence as a tool to detect historical contaminations

In addition to investigating whether the composition of microbial communities can indicate TNT-contaminated sediments, it was of interest to us whether these indications could be maintained for a longer period of time, even if the sediment only contained TNT metabolites or was already TNT-free again. In this case, samples would be characterized as being false-positive. Indeed, based on our approach it became apparent that especially samples

containing no metabolites at all had a lower chance of a false positive prediction. Unfortunately, the sample size did not allow a meaningful test of significance yet. The possible implications are relevant though, as shown by Smith et al. (2015), who successfully classified microbial communities affected by the Deepwater Horizon oil spill. Their random forest models classified samples falsely positive, which were once contaminated, yet subsequently the hydrocarbon concentrations had returned to background levels.

To investigate such a phenomenon based on ecological resilience (Shade et al., 2012) at Kolberger Heide, it should be considered whether TNT and its metabolites result in similar variable importance and fingerprints due to their structural similarity as nitroaromatic compounds. In such a case, a test of true resilience after a TNT contamination - and therefore the time span to detect such – would require to work with once contaminated samples then free of TNT and its metabolites. It should also be ensured that the metabolites were not formed e.g. in the water column and subsequently adsorbed to the sediment.

3.4.6 Importance of microbiological surveys as a key component in environmental monitoring

The Kolberger Heide munitions dumpsite is a stressor to blue mussels (*Mytilus edulis,* Strehse et al., 2017; Appel et al., 2018) and dab (*Limanda limanda,* Koske et al., 2020); our study verified the presence of explosives and their transformation products in sediments as well. Furthermore, mines at Kolberger Heide have been proposed as point sources of mercury due to, e.g., mercury(II) fulminate fuses (Bełdowski et al., 2019). However, despite spottily occurring concentrations up to 4503 µg $Hg \cdot kg^{-1}$ dry sediment, no correlation with the distance to mines was detected (Supplementary Material 3.11). Additionally, most mines on-site are registered as discarded munition material (Kampmeier et al., 2020). In comparison to unexploded ordnance, those were not fused and therefore should not contain mercury(II) fulminate.

TNT was found strongly correlated with DANTs and ADNTs, though (Supplementary Material 3.6). The presence of TNT metabolites proves that Kolberger Heide also represents a disturbance towards the microbial community, as it reacted to the explosives. But it is not clear yet to which extent the community is affected. A potential impact of TNT was surpassed by the main driving grain size distribution and correlating factors (Figure 3.4 A), which is expected for such low levels of TNT. Wikström et al. (2000) reported small amounts of degradation and increased microbial growth following the addition of TNT to lake microcosms. However, they did not find a permanent alteration of microbial communities based on random amplified polymorphic DNA analysis. In a study evaluating the toxicity of Harz soil extracts containing TNT, the *Allivibrio fischeri* luminescence test (EN ISO 11348) reported a long-term EC_{20} of 60 – 90 $ng \cdot g^{-1}$ or 264 $pmol \cdot g^{-1}$ - 396 $pmol \cdot g^{-1}$ (assuming 1 mL = 1 g [Frische, 2002]). Such

concentrations were met in the Kolberger Heide in exceptional cases, e.g. at the detonation site. A summary of various studies investigating a disturbing or toxic impact on soil microorganisms can be found in the article of Kuperman et al. (2009), although effects were only observed at soil TNT content 10^3 to 10^6-fold higher than measured in the current study.

The information of a potential MC impact could have been recorded by microbial communities. Such information could be utilized in cases were direct measurements are problematic to realize: it was reported that TNT is hard to detect just in centimeters distance from containments because it slowly dissolves but is rapidly transformed or bound to sediment (Porter et al., 2011; Gledhill et al., 2019). In fact, TNT can be bio-transformed in minutes (Elovitz and Weber, 1999). Therefore, measured TNT concentrations may not fully capture the impact towards the environment and the microbial community specifically. Furthermore, it should be kept in mind that many more sediments contained MC other than TNT; the impact on the environment has to be considered for all MC in terms of combined effects and quantity, especially with the background of continuously corroding of metal housings. There is even an urgent demand to merely identify the actual MC composition of the dumped ammunitions (Beck et al., 2019). The release of MC might also be intermittent (“sudden release”), which emphasizes the advantages of a resilient indicative fingerprint.

We suggest that microbial community data should be included with monitoring strategies and could potentially act as an information repository to complement the snapshot which is generated by standard monitoring methods. In return, monitoring provides a standardized solution to retrieve more and even specifically required samples to overcome the most severe hindrance for ML: limited sample size. With sufficient data, supervised machine learning could identify impacts of contaminants without being main community drivers. Depending on available context information, the sequenced community data can be utilized to train for other variables, e.g., the prediction of heavy metal contents, or to classify communities based on their distance from mines and investigate the corresponding fingerprint.

3.5 Conclusion

This study demonstrated successfully the prediction of TNT presence in Kolberger Heide sediments using microbial community information, and highlighted regions of the munitions dumpsite where further samples should be collected. A possible TNT indicative fingerprint on genus rank was identified as successful proof of principle. Finally, a potential for TNT-dissipation resilient community compositions was observed.

The importance of environmental monitoring including the implementation of the aforementioned approach was laid out, harnessing its predictive potential. In this regard, resilient microbial communities would allow to fill gaps between sporadic samplings; thus, to

identify contamination events not measurable at all times. As surplus, each monitoring event would generate more training data for more accurate predictions. This may ultimately lead to a more fundamental monitoring of marine ecosystems; based on highly-resolved biological variables and potentially automatable or autonomously operable.

Bibliography

Ahtiainen, H., Artell, J., Elmgren, R., Hasselström, L., and Håkansson, C. (2014). Baltic Sea nutrient reductions – What should we aim for? *J. Environ. Manage.* 145, 9–23. doi:10.1016/j.jenvman.2014.05.016.

Allaire, J. J., and Chollet, F. (2020). keras: R interface to “Keras.” Available at: https://cran.r-project.org/package=keras.

Allaire, J. J., and Tang, Y. (2020). tensorflow: R interface to “TensorFlow.” Available at: https://cran.r-project.org/package=tensorflow.

Allison, S. D., and Martiny, J. B. H. (2008). Resistance, resilience, and redundancy in microbial communities. *Proc. Natl. Acad. Sci.* 105, 11512–11519. doi:10.1073/pnas.0801925105.

Alneberg, J., Bennke, C., Beier, S., Bunse, C., Quince, C., Ininbergs, K., et al. (2020). Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Commun. Biol.* 3, 1–10. doi:10.1038/s42003-020-0856-x.

Alneberg, J., Bjarnason, B. S., Bruijn, I. De, Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1114. doi:10.1038/nmeth.3103.

Alneberg, J., Sundh, J., Bennke, C., Beier, S., Lundin, D., Hugerth, L. W., et al. (2018). BARM and BalticMicrobeDB, a reference metagenome and interface to meta-omic data for the Baltic Sea. *Sci. Data* 5, 180146. doi:10.1038/sdata.2018.146.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics* 26, 1340–1347. doi:10.1093/bioinformatics/btq134.

Andersen, J. H., Carstensen, J., Conley, D. J., Dromph, K., Fleming-Lehtinen, V., Gustafsson, B. G., et al. (2017). Long-term temporal and spatial trends in eutrophication status of the Baltic Sea. *Biol. Rev.* 92, 135–149. doi:10.1111/brv.12221.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12, 878. doi:10.15252/msb.20156651.

Appel, D., Strehse, J. S., Martin, H. J., and Maser, E. (2018). Bioaccumulation of 2,4,6-trinitrotoluene (TNT) and its metabolites leaking from corroded munition in transplanted blue mussels (M. edulis). *Mar. Pollut. Bull.* 135, 1072–1078. doi:10.1016/j.marpolbul.2018.08.028.

Backer, H., Leppänen, J. M., Brusendorff, A. C., Forsius, K., Stankiewicz, M., Mehtonen, J., et al. (2010). HELCOM Baltic Sea Action Plan - A regional programme of measures for the marine environment based on the Ecosystem Approach. *Mar. Pollut. Bull.* 60, 642–649. doi:10.1016/j.marpolbul.2009.11.016.

Baho, D. L., Peter, H., and Tranvik, L. J. (2012). Resistance and resilience of microbial communities - Temporal and spatial insurance against perturbations. *Environ. Microbiol.* 14, 2283–2292. doi:10.1111/j.1462-2920.2012.02754.x.

Bateman, A., Martin, M. J., O’Donovan, C., Magrane, M., Alpi, E., Antunes, R., et al. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169.

doi:10.1093/nar/gkw1099.

Battaglin, W. A., Meyer, M. T., Kuivila, K. M., and Dietze, J. E. (2014). Glyphosate and its degradation product AMPA occur frequently and widely in U.S. soils, surface water, groundwater, and precipitation. *JAWRA J. Am. Water Resour. Assoc.* 50, 275–290. doi:10.1111/jawr.12159.

Beck, A. J., van der Lee, E. M., Eggert, A., Stamer, B., Gledhill, M., Schlosser, C., et al. (2019). In situ measurements of explosive compound dissolution fluxes from exposed munition material in the Baltic Sea. *Environ. Sci. Technol.* 53, 5652–5660. doi:10.1021/acs.est.8b06974.

Bełdowski, J., Klusek, Z., Szubska, M., Turja, R., Bulczak, A. I., Rak, D., et al. (2016a). Chemical Munitions Search & Assessment-An evaluation of the dumped munitions problem in the Baltic Sea. *Deep. Res. Part II Top. Stud. Oceanogr.* 128, 85–95. doi:10.1016/j.dsr2.2015.01.017.

Bełdowski, J., Szubska, M., Emelyanov, E., Garnaga, G., Drzewińska, A., Bełdowska, M., et al. (2016b). Arsenic concentrations in Baltic Sea sediments close to chemical munitions dumpsites. *Deep. Res. Part II Top. Stud. Oceanogr.* 128, 114–122. doi:10.1016/j.dsr2.2015.03.001.

Bełdowski, J., Szubska, M., Siedlewicz, G., Korejwo, E., Grabowski, M., Bełdowska, M., et al. (2019). Sea-dumped ammunition as a possible source of mercury to the Baltic Sea sediments. *Sci. Total Environ.* 674, 363–373. doi:10.1016/j.scitotenv.2019.04.058.

Bennke, C. M., Pollehne, F., Müller, A., Hansen, R., Kreikemeyer, B., and Labrenz, M. (2018). The distribution of phytoplankton in the Baltic Sea assessed by a prokaryotic 16S rRNA gene primer system. *J. Plankton Res.* 40, 244–254. doi:10.1093/plankt/fby008.

Bergen, B., Naumann, M., Herlemann, D. P. R., Gräwe, U., Labrenz, M., and Jürgens, K. (2018). Impact of a major inflow event on the composition and distribution of bacterioplankton communities in the Baltic Sea. *Front. Mar. Sci.* 5, 1–14. doi:10.3389/fmars.2018.00383.

Bergström, L., Börjesson, E., and Stenström, J. (2011). Laboratory and lysimeter studies of glyphosate and aminomethylphosphonic acid in a sand and a clay soil. *J. Environ. Qual.* 40, 98–108. doi:10.2134/jeq2010.0179.

Bernard, S., and Papineau, D. (2014). Graphitic carbons and biosignatures. *Elements* 10, 435–440. doi:10.2113/gselements.10.6.435.

Bernstein, A., and Ronen, Z. (2011). “Biodegradation of the explosives TNT, RDX and HMX,” in *Microbial Degradation of Xenobiotics*, ed. S. N. Singh, 135–176. doi:10.1007/978-3-642-23789-8_5.

Bodor, A., Bounedjoum, N., Vincze, G. E., Erdeiné Kis, Á., Laczi, K., Bende, G., et al. (2020). Challenges of unculturable bacteria: environmental perspectives. *Rev. Environ. Sci. Biotechnol.* 19, 1–22. doi:10.1007/s11157-020-09522-4.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170.

Böttcher, C., Knobloch, T., Rühl, N.-P., Sternheim, J., Wichert, U., and Wöhler, J. (2011). Munitionsbelastung der deutschen Meeresgewässer – Bestandsaufnahme und Empfehlungen. Available at: www.munition-im-meer.de.

Bourdès, V., Bonnevay, S., Lisboa, P., Defrance, R., Pérol, D., Chabaud, S., et al. (2010). Comparison of artificial neural network with logistic regression as classification models for variable selection for prediction of breast cancer patient outcomes. *Adv. Artif. Neural Syst.* 2010, 1–11. doi:10.1155/2010/309841.

Brannon, J. M., Price, C. B., Yost, S. L., Hayes, C., and Porter, B. (2005). Comparison of environmental fate and transport process descriptors of explosives in saline and freshwater systems. *Mar. Pollut. Bull.* 50, 247–251. doi:10.1016/j.marpolbul.2004.10.008.

Braun, S., Morono, Y., Littmann, S., Kuypers, M., Aslan, H., Dong, M., et al. (2016). Size and carbon content of sub-seafloor microbial cells at Landsort Deep, Baltic Sea. *Front. Microbiol.* 7, 1–13. doi:10.3389/fmicb.2016.01375.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi:10.1038/nbt.3519.

Breiman, L. (2001a). Random Forests. *Mach. Learn.* 45, 5–32. doi:https://doi.org/10.1023/A:1010933404324.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16, 199–231. doi:10.1214/ss/1009213726.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. in *2010 20th International Conference on Pattern Recognition* (IEEE), 3121–3124. doi:10.1109/ICPR.2010.764.

Broman, E., Sachpazidou, V., Pinhassi, J., and Dopson, M. (2017). Oxygenation of hypoxic coastal Baltic Sea sediments impacts on chemistry, microbial community composition, and metabolism. *Front. Microbiol.* 8, 1–15. doi:10.3389/fmicb.2017.02453.

Bruns, A., Cypionka, H., and Overmann, J. (2002). Cyclic AMP and acyl homoserine lactones increase the cultivation efficiency of heterotrophic bacteria from the central Baltic Sea. *Appl. Environ. Microbiol.* 68, 3978–3987. doi:10.1128/AEM.68.8.3978-3987.2002.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176.

Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. *Front. Neurosci.* 11, 1–23. doi:10.3389/fnins.2017.00543.

Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus machine learning. *Nat. Methods* 15, 233–234. doi:10.1038/nmeth.4642.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi:10.1038/nmeth.3869.

Cao, Y., Geddes, T. A., Yang, J. Y. H., and Yang, P. (2020). Ensemble deep learning in bioinformatics. *Nat. Mach. Intell.* doi:10.1038/s42256-020-0217-y.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci.* 108, 4516–4522. doi:10.1073/pnas.1000080107.

Carles, L., Gardon, H., Joseph, L., Sanchís, J., Farré, M., and Artigas, J. (2019). Meta-analysis of glyphosate contamination in surface waters and dissipation by biofilms. *Environ. Int.* 124, 284–293. doi:10.1016/j.envint.2018.12.064.

Caspi, R., Billington, R., Keseler, I. M., Kothari, A., Krummenacker, M., Midford, P. E., et al. (2020). The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.* 48, D445–D453. doi:10.1093/nar/gkz862.

Charvet, S., Riemann, L., Alneberg, J., Andersson, A. F., von Borries, J., Fischer, U., et al. (2019). AFISsys - An autonomous instrument for the preservation of brackish water samples for microbial metatranscriptome analysis. *Water Res.* 149, 351–361. doi:10.1016/j.watres.2018.11.017.

Chollet, F., and Allaire, J. J. (2018). *Deep Learning with R*. Manning Publications USA Available at: https://www.manning.com/books/deep-learning-with-r.

Christner, B. C., Mosley-Thompson, E., Thompson, L. G., and Reeve, J. N. (2003). Bacterial recovery from ancient glacial ice. *Environ. Microbiol.* 5, 433–436. doi:10.1046/j.1462-2920.2003.00422.x.

Clarke, A. (2014). The thermal limits to life on Earth. *Int. J. Astrobiol.* 13, 141–154. doi:10.1017/S1473550413000438.

Clarke, S. C. (2005). Pyrosequencing: nucleotide sequencing technology with bacterial genotyping applications. *Expert Rev. Mol. Diagn.* 5, 947–953. doi:10.1586/14737159.5.6.947.

Conley, D. J., Björck, S., Bonsdorff, E., Carstensen, J., Destouni, G., Gustafsson, B. G., et al. (2009). Hypoxia-related processes in the Baltic Sea. *Environ. Sci. Technol.* 43, 3412–3420. doi:10.1021/es802762a.

Costerton, J. W., Lewandowski, Z., Caldwell, D. E., Korber, D. R., and Lappin-Scott, H. M. (1995). Microbial biofilms. *Annu. Rev. Microbiol.* 49, 711–745. doi:10.1146/annurev.mi.49.100195.003431.

Dalrymple, G. B. (2001). The age of the Earth in the twentieth century: a problem (mostly) solved. *Geol. Soc. London, Spec. Publ.* 190, 205–221. doi:10.1144/GSL.SP.2001.190.01.14.

Darrell, T., Kloft, M., Pontil, M., Rätsch, G., Rodner, E., License, G. R., et al. (2015). Machine learning with interdependent and non-identically distributed data. doi:10.4230/DagRep.5.4.18.

Davey, M. E., and O'Toole, G. A. (2000). Microbial biofilms: from ecology to molecular genetics. *Microbiol. Mol. Biol. Rev.* 64, 847–867. doi:10.1128/MMBR.64.4.847-867.2000.

de Wit, C. A., Bossi, R., Dietz, R., Dreyer, A., Faxneld, S., Garbus, S. E., et al. (2020). Organohalogen compounds of emerging concern in Baltic Sea biota: Levels, biomagnification potential and comparisons with legacy contaminants. *Environ. Int.* 144, 106037. doi:10.1016/j.envint.2020.106037.

Dellwig, O., Schnetger, B., Meyer, D., Pollehne, F., Häusler, K., and Arz, H. W. (2018). Impact of the major baltic inflow in 2014 on manganese cycling in the Gotland Deep (Baltic Sea). *Front. Mar. Sci.* 5, 1–20. doi:10.3389/fmars.2018.00248.

Dellwig, O., Wegwerth, A., Schnetger, B., Schulz, H., and Arz, H. W. (2019). Dissimilar

behaviors of the geochemical twins W and Mo in hypoxic-euxinic marine basins. *Earth-Science Rev.* 193, 1–23. doi:10.1016/j.earscirev.2019.03.017.

Di Giulio, M. (2003). The universal ancestor and the ancestor of Bacteria were hyperthermophiles. *J. Mol. Evol.* 57, 721–730. doi:10.1007/s00239-003-2522-6.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi:10.1038/nbt.3820.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Comput. Surv.* 27, 326–327. doi:10.1145/212094.212114.

Dodd, M. S., Papineau, D., Grenne, T., Slack, J. F., Rittner, M., Pirajno, F., et al. (2017). Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature* 543, 60–64. doi:10.1038/nature21377.

Dopson, M., Baker-Austin, C., Hind, A., Bowman, J. P., and Bond, P. L. (2004). Characterization of Ferroplasma isolates and Ferroplasma acidarmanus sp. nov., extreme ccidophiles from acid mine drainage and industrial bioleaching Environments. *Appl. Environ. Microbiol.* 70, 2079–2088. doi:10.1128/AEM.70.4.2079-2088.2004.

Douglas, A. E. (2015). Multiorganismal insects: Diversity and function of resident microorganisms. *Annu. Rev. Entomol.* 60, 17–34. doi:10.1146/annurev-ento-010814-020822.

Duke, S. O., and Powles, S. B. (2008). Glyphosate: a once-in-a-century herbicide. *Pest Manag. Sci.* 64, 319–325. doi:10.1002/ps.1518.

Dundar, M., Krishnapuram, B., Bi, J., and Rao, R. B. (2007). Learning classifiers when the training data is not IID. *IJCAI Int. Jt. Conf. Artif. Intell.*, 756–761.

Edlund, A. (2007). Microbial Diversity in Baltic Sea Sediments. *Dr. Thesis*.

Elovitz, M. S., and Weber, E. J. (1999). Sediment-mediated reduction of 2,4,6-trinitrotoluene and fate of the resulting aromatic (poly)amines. *Environ. Sci. Technol.* 33, 2617–2625. doi:10.1021/es980980b.

Esteve-Núñez, A., Caballero, A., and Ramos, J. L. (2001). Biological degradation of 2,4,6-trinitrotoluene. *Microbiol. Mol. Biol. Rev.* 65, 335–352. doi:10.1128/MMBR.65.3.335.

Fahy, A., Lethbridge, G., Earle, R., Ball, A. S., Timmis, K. N., and McGenity, T. J. (2005). Effects of long-term benzene pollution on bacterial diversity and community structure in groundwater. *Environ. Microbiol.* 7, 1192–1199. doi:10.1111/j.1462-2920.2005.00799.x.

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science (80-.).* 320, 1034–1039. doi:10.1126/science.1153213.

Fayyad, U., Piatetsky-Shapiro, and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Mag.* 17, 37–54.

Fei-Fei, L., Deng, J., and Li, K. (2010). ImageNet: Constructing a large-scale image database. *J. Vis.* 9, 1037–1037. doi:10.1167/9.8.1037.

Fernandes, G., Aparicio, V. C., Bastos, M. C., Gerónimo, E. De, Labanowski, J., Damian, P. O., et al. (2019). Indiscriminate use of glyphosate impregnates river epilithic biofilms in

southern Brazil. *Sci. Total Environ.* 651, 1377–1387. doi:10.1016/j.scitotenv.2018.09.292.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 15, 3133–3181. doi:10.1080/13216597.1999.9751892.

Fioravanti, D., Giarratano, Y., Maggio, V., Agostinelli, C., Chierici, M., Jurman, G., et al. (2018). Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics* 19, 1–13. doi:10.1186/s12859-018-2033-5.

Fischer, S. G., and Lerman, L. S. (1980). Separation of random fragments of DNA according to properties of their sequences. *Proc. Natl. Acad. Sci.* 77, 4420–4424. doi:10.1073/pnas.77.8.4420.

Flemming, H.-C., and Wuertz, S. (2019). Bacteria and archaea on Earth and their abundance in biofilms. *Nat. Rev. Microbiol.* 17, 247–260. doi:10.1038/s41579-019-0158-9.

Frische, T. (2002). Screening for soil toxicity and mutagenicity using luminescent bacteria—a case study of the explosive 2,4,6-trinitrotoluene (TNT). *Ecotoxicol. Environ. Saf.* 51, 133–144. doi:10.1006/eesa.2001.2124.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi:10.1093/bioinformatics/bts565.

Gallagher, E. M., Young, L. Y., McGuinness, L. M., and Kerkhof, L. J. (2010). Detection of 2,4,6-trinitrotoluene-utilizing anaerobic bacteria by 15N and 13C incorporation. *Appl. Environ. Microbiol.* 76, 1695–1698. doi:10.1128/AEM.02274-09.

Garnaga, G., Wyse, E., Azemard, S., Stankevičius, A., and de Mora, S. (2006). Arsenic in sediments from the southeastern Baltic Sea. *Environ. Pollut.* 144, 855–861. doi:10.1016/j.envpol.2006.02.013.

George, I. F., Liles, M. R., Hartmann, M., Ludwig, W., Goodman, R. M., and Agathos, S. N. (2009). Changes in soil Acidobacteria communities after 2,4,6-trinitrotoluene contamination. *FEMS Microbiol. Lett.* 296, 159–166. doi:10.1111/j.1574-6968.2009.01632.x.

Gerhard, W. A., and Gunsch, C. K. (2019). Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environ. Int.* 124, 312–319. doi:10.1016/j.envint.2018.12.038.

Ghannam, R. B., Schaerer, L. G., Butler, T. M., and Techtmann, S. M. (2020). Biogeographic patterns in members of globally distributed and dominant taxa found in port microbial communities. *mSphere* 5. doi:10.1128/mSphere.00481-19.

Glasby, T. M., and Underwood, A. J. (1996). Sampling to differentiate between pulse and press perturbations. *Environ. Monit. Assess.* 42, 241–252. doi:10.1007/BF00414371.

Glasl, B., Bourne, D. G., Frade, P. R., Thomas, T., Schaffelke, B., and Webster, N. S. (2019). Microbial indicators of environmental perturbations in coral reef ecosystems. *Microbiome* 7, 1–13. doi:10.1186/s40168-019-0705-7.

Gledhill, M., Beck, A. J., Stamer, B., Schlosser, C., and Achterberg, E. P. (2019). Quantification of munition compounds in the marine environment by solid phase extraction – ultra high performance liquid chromatography with detection by electrospray

ionisation – mass spectrometry. *Talanta* 200, 366–372. doi:10.1016/j.talanta.2019.03.050.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* 8, 1–6. doi:10.3389/fmicb.2017.02224.

Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. in *JMLR W&CP: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, 249–256.

Grasshoff, K., Kremling, K., and Ehrhardt, M. eds. (1999). *Methods of Seawater Analysis*. Third. Wiley doi:10.1002/9783527613984.

Green, M. R., and Sambrook, J. (2012). *Molecular cloning: a laboratory manual*. Fourth Edi. Cold Spring Harbor Laboratory Press.

Greinert, J. (2019). UDEMM - Practical guide for environmental monitoring of conventional munitions in the seas. *Berichte aus dem GEOMAR Helmholtz-Zentrum für Ozeanforsch. Kiel* 54. doi:10.3289/GEOMAR_REP_NS_54_2019.

Grote, J., Labrenz, M., Pfeiffer, B., Jost, G., and Jürgens, K. (2007). Quantitative distributions of Epsilonproteobacteria and a Sulfurimonas subgroup in pelagic redoxclines of the central Baltic Sea. *Appl. Environ. Microbiol.* 73, 7155–7161. doi:10.1128/AEM.00466-07.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422. doi:10.1023/A:1012487302797.

Habineza, A., Zhai, J., Mai, T., Mmereki, D., and Ntakirutimana, T. (2017). Biodegradation of 2,4,6-trinitrotoluene (TNT) in contaminated soil and microbial remediation options for treatment. *Period. Polytech. Chem. Eng.* 61, 171–187. doi:10.3311/PPch.9251.

Hall, M., National, H., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., et al. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explor. Newsl.* 11, 10–18. doi:10.1145/1656274.1656278.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Second Edi. New York, NY: Springer New York doi:10.1007/b94608.

Häusler, K., Dellwig, O., Schnetger, B., Feldens, P., Leipe, T., Moros, M., et al. (2018). Massive Mn carbonate formation in the Landsort Deep (Baltic Sea): Hydrographic conditions, temporal succession, and Mn budget calculations. *Mar. Geol.* 395, 260–270. doi:10.1016/j.margeo.2017.10.010.

He, Z., Zhang, P., Wu, L., Rocha, A. M., Tu, Q., Shi, Z., et al. (2018). Microbial functional gene diversity predicts groundwater contamination and ecosystem functioning. *MBio* 9, 1–15. doi:10.1128/mBio.02435-17.

Heinänen, A. (1991). Bacterial numbers, biomass and productivity in the Baltic Sea: a cruise study. *Mar. Ecol. Prog. Ser.* 70, 283–290. doi:10.3354/meps070283.

HELCOM (2018). State of the Baltic Sea - Second HELCOM holistic assessment 2011-2016. *Balt. Sea Environ. Proc.* 155, 1–155. Available at: www.helcom.fi/baltic-sea-trends/holistic-assessments/state-of-the-baltic-sea-2018/reports-and-materials/.

Herlemann, D. P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., and Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* 5, 1571–1579. doi:10.1038/ismej.2011.41.

Hove-Jensen, B., Zechel, D. L., and Jochimsen, B. (2014). Utilization of glyphosate as phosphate source: Biochemistry and genetics of bacterial carbon-phosphorus lyase. *Microbiol. Mol. Biol. Rev.* 78, 176–197. doi:10.1128/MMBR.00040-13.

Huang, J., Su, Z., and Xu, Y. (2005). The evolution of microbial phosphonate degradative pathways. *J. Mol. Evol.* 61, 682–690. doi:10.1007/s00239-004-0349-4.

Huang, X.-L., and Zhang, J.-Z. (2011). Phosphorus sorption on marine carbonate sediment: Phosphonate as model organic compounds. *Chemosphere* 85, 1227–1232. doi:10.1016/j.chemosphere.2011.07.016.

Inagaki, F., Hinrichs, K. U., Kubo, Y., Bowles, M. W., Heuer, V. B., Hong, W. L., et al. (2015). Exploring deep microbial life in coal-bearing sediment down to 2.5 km below the ocean floor. *Science (80-.).* 349, 420–424. doi:10.1126/science.aaa6882.

Ininbergs, K., Bergman, B., Larsson, J., and Ekman, M. (2015). Microbial metagenomics in the Baltic Sea: Recent advancements and prospects for environmental monitoring. *Ambio* 44, 439–450. doi:10.1007/s13280-015-0663-7.

Jacob, G. S., Garbow, J. R., Hallas, L. E., Kimack, N. M., Kishore, G. M., and Schaefer, J. (1988). Metabolism of glyphosate in Pseudomonas sp. strain LBr. *Appl. Environ. Microbiol.* 54, 2953–2958.

Janitza, S., Celik, E., and Boulesteix, A. L. (2018). A computationally fast variable importance test for random forests for high-dimensional data. *Adv. Data Anal. Classif.* 12, 885–915. doi:10.1007/s11634-016-0276-4.

Janßen, R., Beck, A. J., Werner, J., Dellwig, O., Alneberg, J., Kreikemeyer, B., et al. (2020). Machine learning predicts the presence of 2,4,6-trinitrotoluene in sediments of a Baltic Sea munitions dumpsite using microbial community compositions. *Front. Microbiol.*

Janßen, R., Skeff, W., Werner, J., Wirth, M. A., Kreikemeyer, B., Schulz-Bull, D., et al. (2019a). A glyphosate pulse to brackish long-term microcosms has a greater impact on the microbial diversity and abundance of planktonic than of biofilm assemblages. *Front. Mar. Sci.* 6, 1–17. doi:10.3389/fmars.2019.00758.

Janßen, R., Zabel, J., von Lukas, U., and Labrenz, M. (2019b). An artificial neural network and Random Forest identify glyphosate-impacted brackish communities based on 16S rRNA amplicon MiSeq read counts. *Mar. Pollut. Bull.* 149, 110530. doi:10.1016/j.marpolbul.2019.110530.

Kampmeier, M., van der Lee, E. M., Wichert, U., and Greinert, J. (2020). Exploration of the munition dumpsite Kolberger Heide in Kiel Bay, Germany: Example for a standardised hydroacoustic and optic monitoring approach. *Cont. Shelf Res.* 198, 104108. doi:10.1016/j.csr.2020.104108.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi:10.1093/molbev/mst010.

Kaufman, S., Rosset, S., and Perlich, C. (2011). Leakage in data mining: Formulation, detection, and avoidance. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 556–563. doi:10.1145/2020408.2020496.

Khadra, M., Planas, D., Girard, C., and Amyot, M. (2018). Age matters: Submersion period shapes community composition of lake biofilms under glyphosate stress. *FACETS* 3, 934–951. doi:10.1139/facets-2018-0019.

Klátyik, S., Takács, E., Mörtl, M., Földi, A., Trábert, Z., Ács, É., et al. (2017). Dissipation of the herbicide active ingredient glyphosate in natural water samples in the presence of biofilms. *Int. J. Environ. Anal. Chem.* 97, 901–921. doi:10.1080/03067319.2017.1373770.

Knights, D., Costello, E. K., and Knight, R. (2011). Supervised classification of human microbiota. *FEMS Microbiol. Rev.* 35, 343–359. doi:10.1111/j.1574-6976.2010.00251.x.

Koch, A. L. (2001). Oligotrophs versus copiotrophs. *BioEssays* 23, 657–661. doi:10.1002/bies.1091.

Koske, D., Straumer, K., Goldenstein, N. I., Hanel, R., Lang, T., and Kammann, U. (2020). First evidence of explosives and their degradation products in dab (Limanda limanda L.) from a munition dumpsite in the Baltic Sea. *Mar. Pollut. Bull.* 155, 111131. doi:10.1016/j.marpolbul.2020.111131.

Kuperman, R. G., Simini, M., Siciliano, S. D., and Gong, P. (2009). “Effects of energetic materials on soil organisms,” in *Ecotoxicology of Explosives*, eds. G. I. Sunahara, G. Lotufo, R. G. Kuperman, and J. Hawari (CRC Press), 35–76. doi:10.1201/9781420004342.

Kwiatkowska, M., Jarosiewicz, P., Michałowicz, J., Koter-Michalak, M., Huras, B., and Bukowska, B. (2016). The impact of glyphosate, its metabolites and impurities on viability, ATP level and morphological changes in human peripheral blood mononuclear cells. *PLoS One* 11, e0156946. doi:10.1371/journal.pone.0156946.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* 10, 1–8. doi:10.1038/s41467-019-08987-4.

Larsen, P. E., Field, D., and Gilbert, J. A. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nat. Methods* 9, 621–625. doi:10.1038/nmeth.1975.

Leamon, J. H., and Rothberg, J. M. (2009). “DNA sequencing and genomics,” in *Encyclopedia of Microbiology* (Elsevier), 148–161. doi:10.1016/B978-012373944-5.00024-9.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi:10.1038/nature14539.

Leipe, T., Moros, M., Kotilainen, A., Vallius, H., Kabel, K., Endler, M., et al. (2013). Mercury in Baltic Sea sediments - Natural background and anthropogenic impact. *Geochemistry* 73, 249–259. doi:10.1016/j.chemer.2013.06.005.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., et al. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102, 3–11. doi:10.1016/j.ymeth.2016.02.020.

Li, L., Rakitsch, B., and Borgwardt, K. (2011). ccSVM: correcting support vector machines for confounding factors in biological data classification. *Bioinformatics* 27, i342–i348. doi:10.1093/bioinformatics/btr204.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News* 2, 18–22. Available at: https://cran.r-project.org/doc/Rnews/.

Lidbury, I. D. E. A., Murphy, A. R. J., Scanlan, D. J., Bending, G. D., Jones, A. M. E., Moore, J. D., et al. (2016). Comparative genomic, proteomic and exoproteomic analyses of three Pseudomonas strains reveals novel insights into the phosphorus scavenging capabilities of soil bacteria. *Environ. Microbiol.* 18, 3535–3549. doi:10.1111/1462-2920.13390.

Lin, C., Jain, S., Kim, H., and Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Res.* 45, e156–e156. doi:10.1093/nar/gkx681.

Lindh, M. V., and Pinhassi, J. (2018). Sensitivity of bacterioplankton to environmental disturbance: A review of Baltic Sea field studies and experiments. *Front. Mar. Sci.* 5, 1–17. doi:10.3389/fmars.2018.00361.

Lindh, M. V., Sjöstedt, J., Andersson, A. F., Baltar, F., Hugerth, L. W., Lundin, D., et al. (2015). Disentangling seasonal bacterioplankton population dynamics by high-frequency sampling. *Environ. Microbiol.* 17, 2459–2476. doi:10.1111/1462-2920.12720.

Lipok, J., Owsiak, T., Młynarz, P., Forlani, G., and Kafarski, P. (2007). Phosphorus NMR as a tool to study mineralization of organophosphonates—The ability of Spirulina spp. to degrade glyphosate. *Enzyme Microb. Technol.* 41, 286–291. doi:10.1016/j.enzmictec.2007.02.004.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8, 833–835. doi:10.1038/nmeth.1681.

López-Rodas, V., Flores-Moya, A., Maneiro, E., Perdigones, N., Marva, F., García, M. E., et al. (2007). Resistance to glyphosate in the cyanobacterium Microcystis aeruginosa as result of pre-selective mutations. *Evol. Ecol.* 21, 535–547. doi:10.1007/s10682-006-9134-8.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8.

Lozano, V. L., Vinocur, A., Sabio y García, C. A., Allende, L., Cristos, D. S., Rojas, D., et al. (2018). Effects of glyphosate and 2,4-D mixture on freshwater phytoplankton and periphyton communities: a microcosms approach. *Ecotoxicol. Environ. Saf.* 148, 1010–1019. doi:10.1016/j.ecoenv.2017.12.006.

Lu, G.-H., Zhu, Y.-L., Kong, L.-R., Cheng, J., Tang, C.-Y., Hua, X.-M., et al. (2017). Impact of a glyphosate-tolerant soybean line on the Rhizobacteria, revealed by Illumina MiSeq. *J. Microbiol. Biotechnol.* 27, 561–572. doi:10.4014/jmb.1609.09008.

Makarova, K. S., Aravind, L., Wolf, Y. I., Tatusov, R. L., Minton, K. W., Koonin, E. V., et al. (2001). Genome of the extremely radiation-resistant bacterium Deinococcus radiodurans viewed from the perspective of comparative genomics. *Microbiol. Mol. Biol. Rev.* 65, 44–79. doi:10.1128/MMBR.65.1.44-79.2001.

Marshall, K. C. (2013). “Planktonic versus sessile life of prokaryotes,” in *The Prokaryotes* (Berlin, Heidelberg: Springer Berlin Heidelberg), 191–201. doi:10.1007/978-3-642-30123-0_49.

Martin, C. W. (2020). Impact of sea-dumped ammunition on microbial communities in Baltic Sea sediments analysed through shotgun metagenomics.

Martinez, A., Tyson, G. W., and Delong, E. F. (2010). Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ. Microbiol.* 12, 222–238. doi:10.1111/j.1462-2920.2009.02062.x.

Martínez, A., Ventouras, L. A., Wilson, S. T., Karl, D. M., and DeLong, E. F. (2013). Metatranscriptomic and functional metagenomic analysis of methylphosphonate utilization by marine bacteria. *Front. Microbiol.* 4, 1–18. doi:10.3389/fmicb.2013.00340.

Maser, E., and Strehse, J. S. (2020). “Don’t Blast”: blast-in-place (BiP) operations of dumped World War munitions in the oceans significantly increase hazards to the environment and the human seafood consumer. *Arch. Toxicol.* 94, 1941–1953. doi:10.1007/s00204-020-02743-0.

McGrath, J. W., Ternan, N. G., and Quinn, J. P. (1997). Utilization of organophosphonates by environmental microorganisms. *Lett. Appl. Microbiol.* 24, 69–73. doi:10.1046/j.1472-765X.1997.00350.x.

McLaren, M. (2020). speedyseq: Faster implementations of common phyloseq functions. Available at: https://github.com/mikemc/speedyseq.

McMurdie, P. J., and Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8, e61217. doi:10.1371/journal.pone.0061217.

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10. doi:10.1371/journal.pcbi.1003531.

Menze, B. H., Kelm, B. M., Splitthoff, N., and Hamprecht, F. A. (2011). “On oblique random forests,” in *Machine Learning and Knowledge Discovery in Databases*, 453–469.

Mercurio, P., Flores, F., Mueller, J. F., Carter, S., and Negri, A. P. (2014). Glyphosate persistence in seawater. *Mar. Pollut. Bull.* 85, 385–390. doi:10.1016/j.marpolbul.2014.01.021.

Meredith, H. R., Andreani, V., Ma, H. R., Lopatkin, A. J., Lee, A. J., Anderson, D. J., et al. (2018). Applying ecological resistance and resilience to dissect bacterial antibiotic responses. *Sci. Adv.* 4, eaau1873. doi:10.1126/sciadv.aau1873.

Meyer-Reil, L.-A. (1983). Benthic response to sedimentation events during autumn to spring at a shallow water station in the Western Kiel Bight. *Mar. Biol.* 77, 247–256. doi:10.1007/BF00395813.

Meyer-Reil, L. A. (1994). Microbial life in sedimentary biofilms - The challenge to microbial ecologists. *Mar. Ecol. Prog. Ser.* 112, 303–311. doi:10.3354/meps112303.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Moitinho-Silva, L., Steinert, G., Nielsen, S., Hardoim, C. C. P., Wu, Y. C., McCormack, G. P., et al. (2017). Predicting the HMA-LMA status in marine sponges by machine learning. *Front. Microbiol.* 8, 1–14. doi:10.3389/fmicb.2017.00752.

Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digit. Signal Process. A Rev. J.* 73, 1–15.

doi:10.1016/j.dsp.2017.10.011.

Morono, Y., Ito, M., Hoshino, T., Terada, T., Hori, T., Ikehara, M., et al. (2020). Aerobic microbial life persists in oxic marine sediment as old as 101.5 million years. *Nat. Commun.* 11. doi:10.1038/s41467-020-17330-1.

Myers, J. P., Antoniou, M. N., Blumberg, B., Carroll, L., Colborn, T., Everett, L. G., et al. (2016). Concerns over use of glyphosate-based herbicides and risks associated with exposures: a consensus statement. *Environ. Heal.* 15, 19. doi:10.1186/s12940-016-0117-0.

Nembrini, S., König, I. R., and Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics* 34, 3711–3718. doi:10.1093/bioinformatics/bty373.

Newman, M. M., Lorenz, N., Hoilett, N., Lee, N. R., Dick, R. P., Liles, M. R., et al. (2016). Changes in rhizosphere bacterial gene expression following glyphosate treatment. *Sci. Total Environ.* 553, 32–41. doi:10.1016/j.scitotenv.2016.02.078.

Nguyen, N. G., Tran, V. A., Ngo, D. L., Phan, D., Lumbanraja, F. R., Faisal, M. R., et al. (2016). DNA sequence classification by convolutional neural network. *J. Biomed. Sci. Eng.* 09, 280–286. doi:10.4236/jbise.2016.95021.

Ni, J., Yan, Q., and Yu, Y. (2013). How much metagenomic sequencing is enough to achieve a given goal? *Sci. Rep.* 3, 1968. doi:10.1038/srep01968.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834. doi:10.1101/gr.213959.116.

Økland, R. H. (2007). Wise use of statistical tools in ecological field studies. *Folia Geobot.* 42, 123–140. doi:10.2307/41245506.

Oksanen, J. (2015). Multivariate analysis of ecological communities in R: vegan tutorial. Available at: https://linkinghub.elsevier.com/retrieve/pii/0169534788901243.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2019). vegan: Community Ecology Package. Available at: https://cran.r-project.org/package=vegan.

Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K., and Sekiya, T. (1989). Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc. Natl. Acad. Sci.* 86, 2766–2770. doi:10.1073/pnas.86.8.2766.

Orwin, K. H., and Wardle, D. A. (2004). New indices for quantifying the resistance and resilience of soil biota to exogenous disturbances. *Soil Biol. Biochem.* 36, 1907–1912. doi:10.1016/j.soilbio.2004.04.036.

Paliy, O., and Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Mol. Ecol.* 25, 1032–1057. doi:10.1111/mec.13536.

Paluszynska, A., and Biecek, P. (2017). randomForestExplainer: Explaining and visualizing random forests in terms of variable importance. Available at: https://cran.r-project.org/package=randomForestExplainer.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996–1004. doi:10.1038/nbt.4229.

Patterson, J., and Gibson, A. (2017). *Deep learning*. , eds. M. Loukides and T. McGovern O'Reilly Media.

Pedotti, M., Rosini, E., Molla, G., Moschetti, T., Savino, C., Vallone, B., et al. (2009). Glyphosate resistance by engineering the flavoenzyme glycine oxidase. *J. Biol. Chem.* 284, 36415–36423. doi:10.1074/jbc.M109.051631.

Pernthaler, J. (2017). Competition and niche separation of pelagic bacteria in freshwater habitats. *Environ. Microbiol.* 19, 2133–2150. doi:10.1111/1462-2920.13742.

Pfeiffer, F. (2009). Bericht über die in-situ-Begleituntersuchungen zur Munitionssprengung in der Ostsee vom 18.2.2009. Available at: https://www.schleswig-holstein.de/DE/Fachinhalte/M/meeresschutz/Downloads/Bericht_Begleituntersuchung_2009.pdf.

Porter, J. W., Barton, J. V., and Torres, C. (2011). "Ecological, radiological, and toxicological effects of naval bombardment on the coral reefs of Isla de Vieques, Puerto Rico," in *Warfare ecology: A new synthesis for peace and security*, eds. G. E. Machlis, T. Hanson, Z. Špirić, and J. E. McKendry, 65–121. doi:10.1007/978-94-007-1214-0_8.

Porter, K. G., and Feig, Y. S. (1980). The use of DAPI for identifying and counting aquatic microflora. *Limnol. Ocean.* 25, 943–948. doi:10.4319/lo.1980.25.5.0943.

Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019). Application of machine learning in microbiology. *Front. Microbiol.* 10, 1–10. doi:10.3389/fmicb.2019.00827.

Qu, X., Ren, Z., Zhang, H., Zhang, M., Zhang, Y., Liu, X., et al. (2017). Influences of anthropogenic land use on microbial community structure and functional potentials of stream benthic biofilms. *Sci. Rep.* 7, 15117. doi:10.1038/s41598-017-15624-x.

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria Available at: https://www.r-project.org/.

R Core Team, Team, R. C., and others (2017). R: A language and environment for statistical computing. 3. doi:ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rantajärvi, E., Flinkman, J., Ruokanen, L., Hällfors, S., Stipa, T., Suominen, T., et al. (2003). Alg@line in 2003: 10 years of innovative plankton monitoring and research an operational information service in the Baltic Sea.

Ratcliff, A. W., Busse, M. D., and Shestak, C. J. (2006). Changes in microbial community structure following herbicide (glyphosate) additions to forest soils. *Appl. Soil Ecol.* 34, 114–124. doi:10.1016/j.apsoil.2006.03.002.

Reese, A. T., Savage, A., Youngsteadt, E., McGuire, K. L., Koling, A., Watkins, O., et al. (2016). Urban stress is associated with variation in microbial species composition—but not richness—in Manhattan. *ISME J.* 10, 751–760. doi:10.1038/ismej.2015.152.

Rheinheimer, G. (1998). Pollution in the Baltic Sea. *Naturwissenschaften* 85, 318–329. doi:10.1007/s001140050508.

Rieck, A., Herlemann, D. P. R., Jürgens, K., and Grossart, H.-P. (2015). Particle-associated differ from free-living bacteria in surface waters of the Baltic Sea. *Front. Microbiol.* 6. doi:10.3389/fmicb.2015.01297.

Rodríguez, J. (2020). Bacterial communities in polluted Baltic Sea environments in a changing climate.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4, e2584. doi:10.7717/peerj.2584.

Roldán, M. D., Pérez-Reinado, E., Castillo, F., and Moreno-Vivián, C. (2008). Reduction of polynitroaromatic compounds: the bacterial nitroreductases. *FEMS Microbiol. Rev.* 32, 474–500. doi:10.1111/j.1574-6976.2008.00107.x.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408. doi:10.1037/h0042519.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi:10.1038/s42256-019-0048-x.

Ryan, F. J. (2019). Application of machine learning techniques for creating urban microbial fingerprints. *Biol. Direct* 14, 13. doi:10.1186/s13062-019-0245-x.

Rykiel Jr., E. J. (1985). Towards a definition of ecological disturbance. *Aust. J. Ecol.* 10, 361–365. Available at: http://www.buyteknet.info/fileshare/data/analisis_lect/towards_a_definition_of_ecological_disturbance_122.pdf.

Samhita, L., and Gross, H. J. (2013). The “Clever Hans Phenomenon” revisited. *Commun. Integr. Biol.* 6, e27122. doi:10.4161/cib.27122.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi:10.1128/AEM.01541-09.

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks* 61, 85–117.

Schulz, H. N., and Jørgensen, B. B. (2001). Big bacteria. *Annu. Rev. Microbiol.* 55, 105–137. doi:10.1146/annurev.micro.55.1.105.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi:10.1093/bioinformatics/btu153.

Shade, A., Peter, H., Allison, S. D., Baho, D. L., Berga, M., Bürgmann, H., et al. (2012). Fundamentals of microbial community resistance and resilience. *Front. Microbiol.* 3, 1–19. doi:10.3389/fmicb.2012.00417.

Shirani, M., Afzali, K. N., Jahan, S., Strezov, V., and Soleimani-Sardo, M. (2020). Pollution and contamination assessment of heavy metals in the sediments of Jazmurian playa in southeast Iran. *Sci. Rep.* 10, 4775. doi:10.1038/s41598-020-61838-x.

Skeff, W., Neumann, C., and Schulz-Bull, D. E. (2015). Glyphosate and AMPA in the estuaries of the Baltic Sea method optimization and field study. *Mar. Pollut. Bull.* 100, 577–585. doi:10.1016/j.marpolbul.2015.08.015.

Skeff, W., Recknagel, C., and Schulz-Bull, D. E. (2016). The influence of salt matrices on the reversed-phase liquid chromatography behavior and electrospray ionization tandem mass spectrometry detection of glyphosate, glufosinate, aminomethylphosphonic acid and 2-aminoethylphosphonic acid in water. *J. Chromatogr. A* 1475, 64–73. doi:10.1016/j.chroma.2016.11.007.

Smith, M. B., Rocha, A. M., Smillie, C. S., Olesen, S. W., Paradis, C., Wu, L., et al. (2015). Natural bacterial communities serve as quantitative geochemical biosensors. *MBio* 6, 1–13. doi:10.1128/mBio.00326-15.

Snoeijs-Leijonmalm, P., and Andrén, E. (2017). "Why is the Baltic Sea so special to live in?," in *Biological Oceanography of the Baltic Sea*, eds. P. Snoeijs-Leijonmalm, H. Schubert, and T. Radziejewska (Dordrecht: Springer Netherlands), 23–84. doi:10.1007/978-94-007-0668-2_2.

Soneson, C., Gerster, S., and Delorenzi, M. (2014). Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS One* 9, e100335. doi:10.1371/journal.pone.0100335.

Spain, J. C. (1995). Biodegradation of nitroaromatic compounds. *Annu. Rev. Microbiol.* 49, 523–555. doi:10.1146/annurev.micro.49.1.523.

Sprinkhuizen-Kuyper, I. G., and Boers, E. J. W. (1996). The error surface of the simplest XOR network has only global minima. *Neural Comput.* 8, 1301–1320. doi:10.1162/neco.1996.8.6.1301.

Stachowski-Haberkorn, S., Becker, B., Marie, D., Haberkorn, H., Coroller, L., and de la Broise, D. (2008). Impact of Roundup on the marine microbial community, as shown by an in situ microcosm experiment. *Aquat. Toxicol.* 89, 232–241. doi:10.1016/j.aquatox.2008.07.004.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Steel, E. A., Kennedy, M. C., Cunningham, P. G., and Stanovick, J. S. (2013). Applied statistics in ecology: common pitfalls and simple solutions. *Ecosphere* 4, art115. doi:10.1890/ES13-00160.1.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B* 36, 111–133. doi:10.1111/j.2517-6161.1974.tb00994.x.

Strehse, J. S., Appel, D., Geist, C., Martin, H. J., and Maser, E. (2017). Biomonitoring of 2,4,6-trinitrotoluene and degradation products in the marine environment with transplanted blue mussels (M. edulis). *Toxicology* 390, 117–123. doi:10.1016/j.tox.2017.09.004.

Sviridov, A. V., Shushkova, T. V, Ermakova, I. T., Ivanova, E. V, Epiktetov, D. O., and Leontievsky, a a (2015). Microbial degradation of glyphosate herbicides (Review). *Appl. Biochem. Microbiol.* 51, 188–195. doi:10.1134/S0003683815020209.

Sviridov, A. V, Shushkova, T. V, Zelenkova, N. F., Vinokurova, N. G., Morgunov, I. G., Ermakova, I. T., et al. (2012). Distribution of glyphosate and methylphosphonate catabolism systems in soil bacteria Ochrobactrum anthropi and Achromobacter sp. *Appl. Microbiol. Biotechnol.* 93, 787–796. doi:10.1007/s00253-011-3485-y.

Sweitzer, J., Langaas, S., and Folke, C. (1996). Land cover and population density in the Baltic Sea drainage basin: A GIS database. *Ambio* 25, 191–198. doi:10.2307/4314452.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2014). Intriguing properties of neural networks. in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 1–10.

Takahashi, S., Tomita, J., Nishioka, K., Hisada, T., and Nishijima, M. (2014). Development of

a prokaryotic universal primer for simultaneous analysis of bacteria and archaea using next-generation sequencing. *PLoS One* 9, e105592. doi:10.1371/journal.pone.0105592.

Tan, M., and Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *36th Int. Conf. Mach. Learn. ICML 2019* 2019-June, 10691–10700.

Thompson, J., Johansen, R., Dunbar, J., and Munsky, B. (2019). Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. *PLoS One* 14, 1–16. doi:10.1371/journal.pone.0215502.

Thureborn, P., Franzetti, A., Lundin, D., and Sjöling, S. (2016). Reconstructing ecosystem functions of the active microbial community of the Baltic Sea oxygen depleted sediments. *PeerJ* 4, e1593. doi:10.7717/peerj.1593.

Thureborn, P., Lundin, D., Plathan, J., Poole, A. M., Sjöberg, B.-M., and Sjöling, S. (2013). A metagenomics transect into the deepest point of the Baltic Sea reveals clear stratification of microbial functional capacities. *PLoS One* 8, e74983. doi:10.1371/journal.pone.0074983.

Tlili, A., Corcoll, N., Bonet, B., Morin, S., Montuelle, B., Bérard, A., et al. (2011). In situ spatio-temporal changes in pollution-induced community tolerance to zinc in autotrophic and heterotrophic biofilm communities. *Ecotoxicology* 20, 1823–1839. doi:10.1007/s10646-011-0721-2.

Topçuoğlu, B. D., Lesniak, N. A., Ruffin, M. T., Wiens, J., and Schloss, P. D. (2020). A framework for effective application of machine learning to microbiome-based classification problems. *MBio* 11. doi:10.1128/mBio.00434-20.

Uritskiy, G. V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6, 158. doi:10.1186/s40168-018-0541-1.

Van Bruggen, A. H. C., He, M. M., Shin, K., Mai, V., Jeong, K. C., Finckh, M. R., et al. (2018). Environmental and health effects of the herbicide glyphosate. *Sci. Total Environ.* 616–617, 255–268. doi:10.1016/j.scitotenv.2017.10.309.

Vera, M. S., Lagomarsino, L., Sylvester, M., Pérez, G. L., Rodríguez, P., Mugni, H., et al. (2010). New evidences of Roundup® (glyphosate formulation) impact on the periphyton community and the water quality of freshwater ecosystems. *Ecotoxicology* 19, 710–721. doi:10.1007/s10646-009-0446-7.

Vieira-Silva, S., and Rocha, E. P. C. (2010). The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* 6, e1000808. doi:10.1371/journal.pgen.1000808.

Villmoare, B., Kimbel, W. H., Seyoum, C., Campisano, C. J., DiMaggio, E. N., Rowan, J., et al. (2015). Early Homo at 2.8 Ma from Ledi-Geraru, Afar, Ethiopia. *Science (80-.).* 347, 1352–1355. doi:10.1126/science.aaa1343.

Wang, C., Lin, X., Li, L., Lin, L. X., and Lin, S. (2017). Glyphosate shapes a dinoflagellate-associated bacterial community while supporting algal growth as sole phosphorus source. *Front. Microbiol.* 8. doi:10.3389/fmicb.2017.02530.

Wang, C., Lin, X., Li, L., and Lin, S. (2016a). Differential growth responses of marine phytoplankton to herbicide glyphosate. *PLoS One* 11, 1–20. doi:10.1371/journal.pone.0151633.

Wang, S., Seiwert, B., Kästner, M., Miltner, A., Schäffer, A., Reemtsma, T., et al. (2016b). (Bio)degradation of glyphosate in water-sediment microcosms - A stable isotope co-labeling approach. *Water Res.* 99, 91–100. doi:10.1016/j.watres.2016.04.041.

Weaver, M. A., Krutz, L. J., Zablotowicz, R. M., and Reddy, K. N. (2007). Effects of glyphosate on soil microbial communities and its mineralization in a Mississippi soil. *Pest Manag. Sci.* 63, 388–393. doi:10.1002/ps.1351.

Weinbauer, M. G., Fritz, I., Wenderoth, D. F., and Höfle, M. G. (2002). Simultaneous extraction from bacterioplankton of total RNA and DNA suitable for quantitative structure and function analyses. *Appl. Environ. Microbiol.* 68, 1082–1087. doi:10.1128/AEM.68.3.1082-1087.2002.

White, A. K., and Metcalf, W. W. (2004). Two C-P lyase operons in Pseudomonas stutzeri and their roles in the oxidation of phosphonates, phosphite, and hypophosphite. *J. Bacteriol.* 186, 4730–4739. doi:10.1128/JB.186.14.4730.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wikström, P., Andersson, A. C., Nygren, Y., Sjöström, J., and Forsman, M. (2000). Influence of TNT transformation on microbial community structure in four different lake microcosms. *J. Appl. Microbiol.* 89, 302–308. doi:10.1046/j.1365-2672.2000.01111.x.

Wilkins, D., Leung, M. H. Y., and Lee, P. K. H. (2017). Microbiota fingerprints lose individually identifying features over time. *Microbiome* 5, 1. doi:10.1186/s40168-016-0209-7.

Wirth, M. A., Schulz-Bull, D. E., and Kanwischer, M. (2021). The challenge of detecting the herbicide glyphosate and its metabolite AMPA in seawater - method development and application in the Baltic Sea. *Chemosphere*.

Wirth, M. A., Sievers, M., Habedank, F., Kragl, U., Schulz-Bull, D. E., and Kanwischer, M. (2019). Electrodialysis as a sample processing tool for bulk organic matter and target pollutant analysis of seawater. *Mar. Chem.* 217, 103719. doi:10.1016/j.marchem.2019.103719.

Wood, K. (2019). Microbial ecology: Complex bacterial communities reduce selection for antibiotic resistance. *Curr. Biol.* 29, R1143–R1145. doi:10.1016/j.cub.2019.09.017.

Wright, M. N., and Ziegler, A. (2017). ranger: A fast implementation of Random Forests for high dimensional data in C++ and R. *J. Stat. Softw.* 77. doi:10.18637/jss.v077.i01.

Wu, W., May, R. J., Maier, H. R., and Dandy, G. C. (2013). A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. *Water Resour. Res.* 49, 7598–7614. doi:10.1002/2012WR012713.

Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A., and Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2, 26. doi:10.1186/2049-2618-2-26.

Ye, Y., and Doak, T. G. (2011). “A parsimony approach to biological pathway reconstruction/inference for metagenomes,” in *Handbook of Molecular Microbial Ecology I* (Hoboken, NJ, USA: John Wiley & Sons, Inc.), 453–460. doi:10.1002/9781118010518.ch52.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The

SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Res.* 42, D643–D648. doi:10.1093/nar/gkt1209.

Yu, H., Samuels, D. C., Zhao, Y., and Guo, Y. (2019). Architectures and accuracy of artificial neural network for disease classification from omics data. *BMC Genomics* 20, 167. doi:10.1186/s12864-019-5546-z.

Zaborska, A. (2014). Anthropogenic lead concentrations and sources in Baltic Sea sediments based on lead isotopic composition. *Mar. Pollut. Bull.* 85, 99–113. doi:10.1016/j.marpolbul.2014.06.013.

Zhao, H., Tao, K., Zhu, J., Liu, S., Gao, H., and Zhou, X. (2015). Bioremediation potential of glyphosate-degrading *Pseudomonas* spp. strains isolated from contaminated soil. *J. Gen. Appl. Microbiol.* 61, 165–170. doi:10.2323/jgam.61.165.

Zhong, G., Ling, X., and Wang, L. (2019). From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures. *WIREs Data Min. Knowl. Discov.* 9, 1–14. doi:10.1002/widm.1255.

Ziesemer, K. A., Mann, A. E., Sankaranarayanan, K., Schroeder, H., Ozga, A. T., Brandt, B. W., et al. (2015). Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Sci. Rep.* 5, 16498. doi:10.1038/srep16498.

Zinke, L. A., Glombitza, C., Bird, J. T., Røy, H., Jørgensen, B. B., Lloyd, K. G., et al. (2018). Microbial organic matter degradation potential in Baltic Sea sediments is influenced by depositional conditions and in situ geochemistry. *Appl. Environ. Microbiol.* 85, 1–18. doi:10.1128/AEM.02164-18.

List of figures

List of tables

List of abbreviations

p,µ,n,m,c,k,M	Pico, micro, nano, milli, centi, kilo, mega
^{2}, ^{3}	Squared, cubic

16S rRNA	16S ribosomal ribonucleic acid
2-ADNT	2-amino-4,6-dinitrolouene
2,4-DANT	2,4-diamino-6-nitrotoluene
2,6-DANT	2,6-diamino-4-nitrotoluene
2,4-DNT	2,4-dinitrotoluene
2,6-DNT	2,6-dinitrotoluene
4-ADNT	4-amino-2,6-dinitrolouene
ABW	Artificial brackish water
ADAM	Adaptive Moment Estimation
AMPA	Aminomethylphosphonic acid
ANN	Artificial neural network
ASV	Amplicon sequence variant
B	Byte
Bagging	Bootstrap aggregating
BARM	Baltic Sea Reference Metagenome
BLAST	Basic Local Alignment Search Tool
BLUEPRINT	Biological lenses using gene prints
BSAP	Baltic Sea Action Plan
C	Carbon
°C	Degree Celsius
CCA	Canonical correspondence analysis
cDNA	Complimentary DNA
CNN	Convolutional neural network
CPU	Central processing unit
CV	Cross validation
d	Days
DAPI	4',6-diamidino-2- phenylindole
de.NBI	German Network for Bioinformatics Infrastructure
DIP	Dissolved inorganic phosphorus
DMA	Direct mercury analyzer
DNA	Deoxyribonucleic acid
DNB	1,3-dinitrobenzene
DOC	Dissolved organic carbon
DON	Dissolved organic nitrogen
ESI	Electrospray ionization
EtOH	Ethanol
EU	European Union
FMOC	Fluorenylmethoxycarbonyl group
g	Gram
x g	Times gravity
GF/F	Glass microfibre filters, grade GF/F
GPU	Graphics processing unit

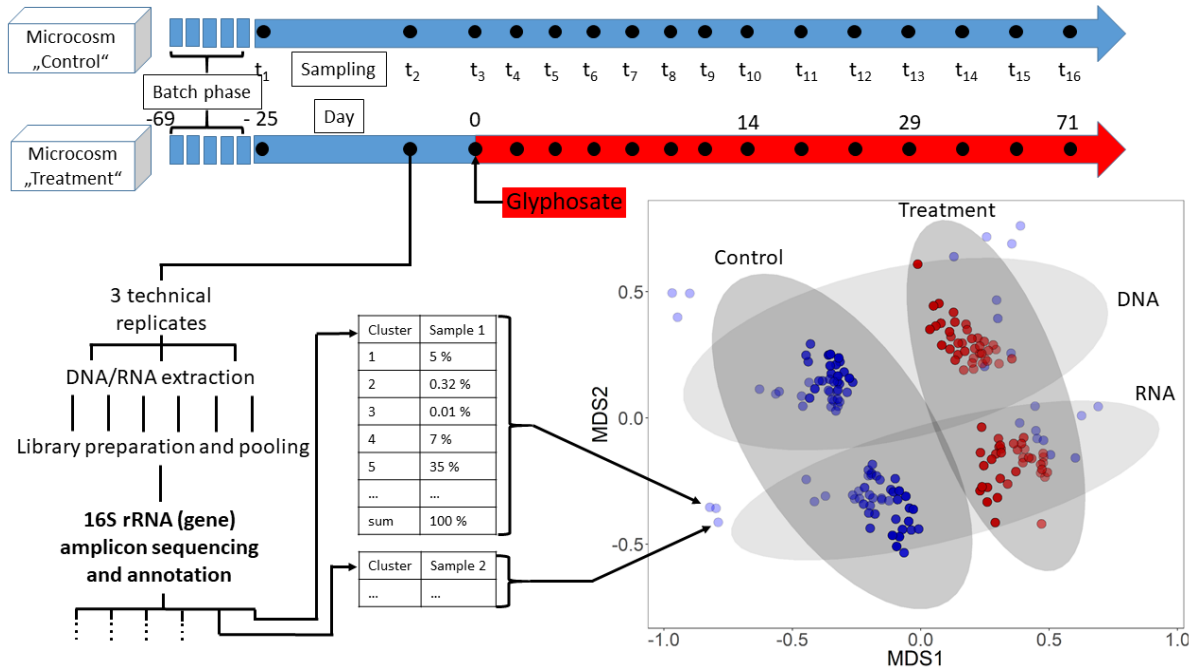
h	Hours
HCl	Hydrochloric acid
HELCOM	Helsinki commission
HMX	Octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine
HPLC	High performance liquid chromatography
ICP-MS	Inductively coupled plasma mass spectrometry
ICP-OES	Inductively coupled plasma optical emission spectrometry
i.i.d.	Independent and identically distributed
L	Liter
LOD	Limit of detection
m	Meter
max.	Maximum
MBSS	Mecklenburg Bay Sediment Standard
MC	Munition compounds
MilliQ	Ultrapure water
min	Minutes
M	Mole per litre
ML	Machine learning
mol	Mole
MS	Mass spetrometry
mtry	Number of variables to randomly select from to split a decision tree node
N	Nitrogen
n	Number, e.g of samples ("n > p") or repetitions
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
nMDS	Non-metric Multidimensional Scaling
nt	Nucleotides
OOB	Out-of-bag
OTU	Operational taxonomic unit
p	Number of variables (e.g. in "n>p"), not to confuse with *p* value
P	Phosphorus
PCA	Principal component analysis
PCoA	Principal coordinate analysis
PCR	Polymerase chain reaction
PERMANOVA	Permutational multivariate analysis of variance
pH	Potential of hydrogen, acidity of aqueous solutions
POC	Particulate organic carbon
PON	Particulate organic nitrogen
ppb	Parts per billion
ppm	Parts per million
RDA	Redundancy analysis
RF	Random Forest
rcf	Relative centrifugal force
RDX	1,3,5-trinitroperhydro-1,3,5-triazine
rmsprop	Root Mean Square Propagation
RNAseq	Total RNA sequencing

rpm	Revolutions per minute
S	Svedberg (e.g. in "16S") / Sulfur
sp.	Species
spp.	Species pluralis
SRA	Short read archive
T	Temperature
TC	Total carbon
TIC	Total inorganic carbon
TN	Total nitrogen
TNB	1,3,5-trinitrobenzene
TNT	2,4,6-trinitritoluene
TOC	Total organic carbon
TS	Total sulfur
UHPLC	Ultrahigh performance liquid chromatography
UniProtKB	UniProt Knowledgebase
UXO	Unexploded ordnance

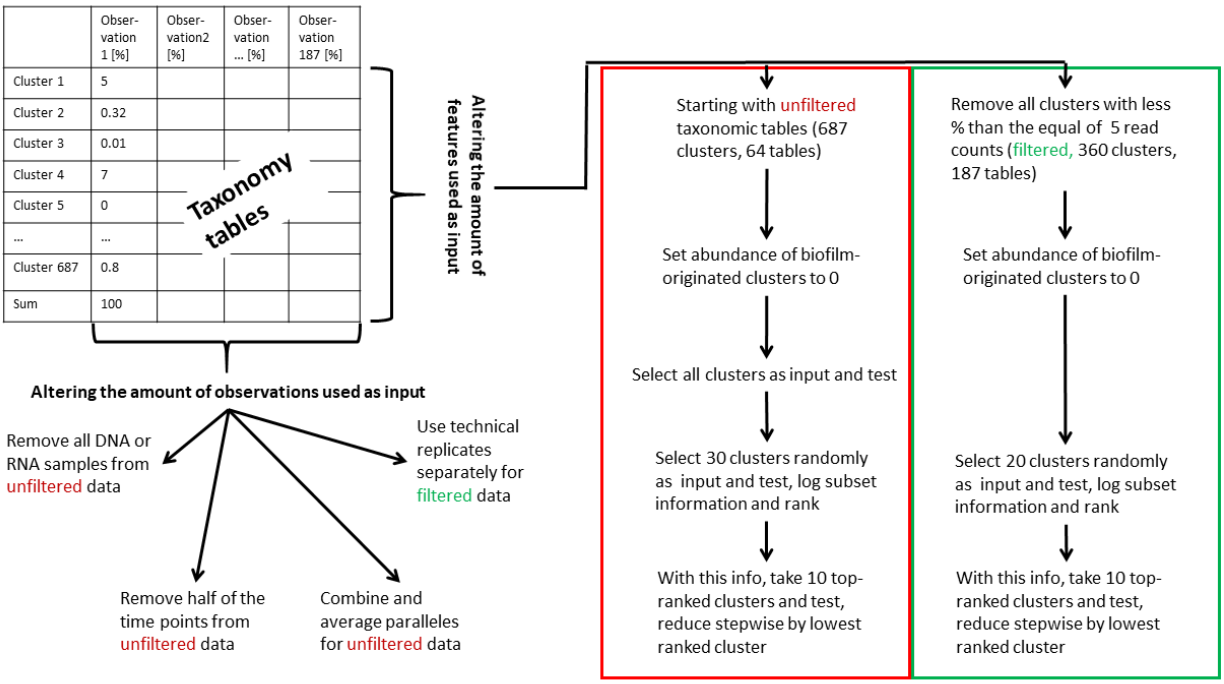
Supplementary materials

Chapter I

Supplementary Figures

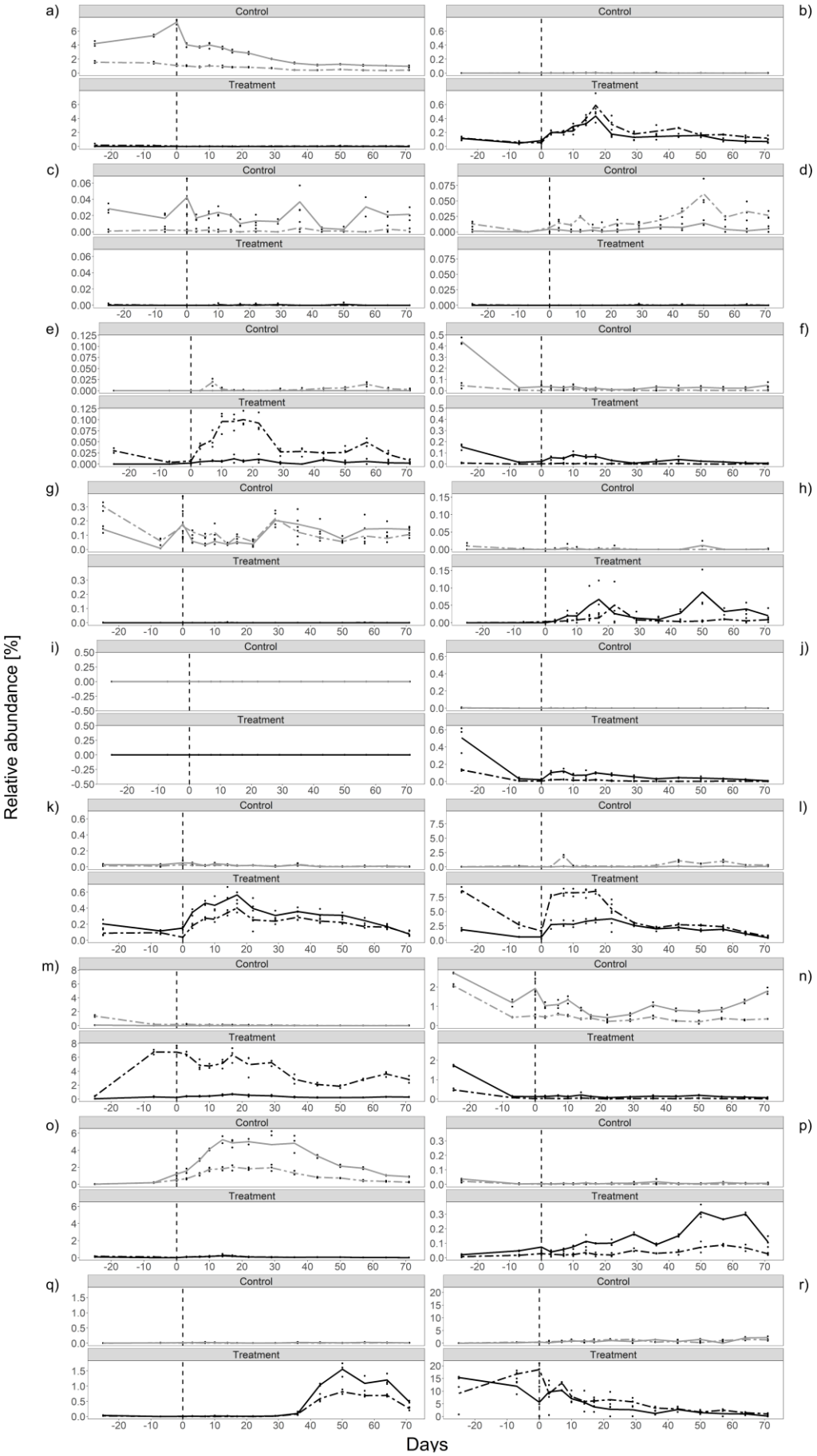


Supplementary Material 1.1: A timeline of the laboratory work flow followed by wet lab downstream processing, MiSeq sequencing and bioinformatic analysis. The taxonomic annotation was performed by the SILVAngs pipeline and the NMDS ordination plot was generated using the metaMDS function from R package vegan based on Bray Curtis dissimilarity. Red labeled samples experienced contact with glyphosate. The ascending alpha gradient indicates passing time.



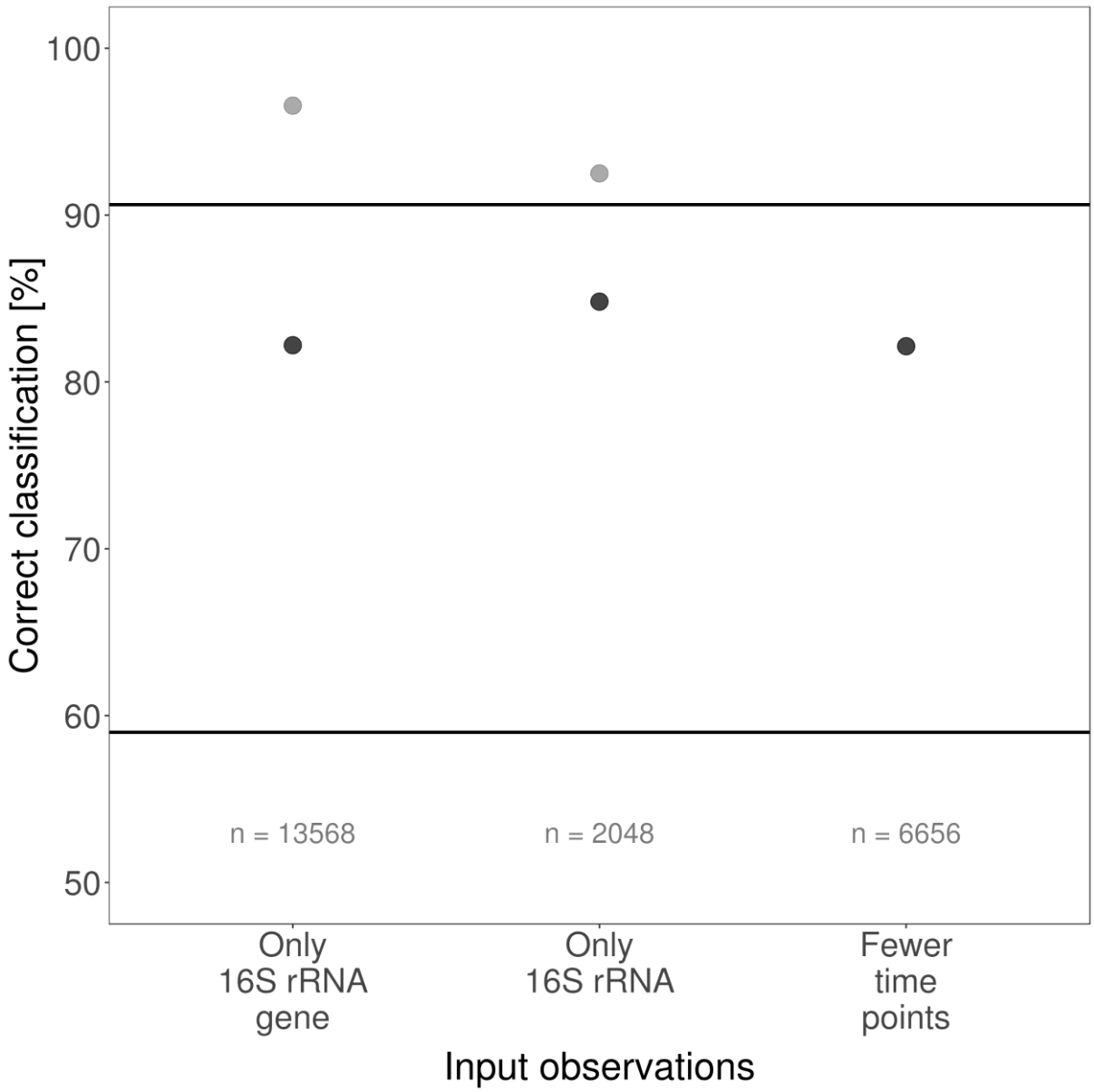
	Obser-vation 1 [%]	Obser-vation2 [%]	Obser-vation ... [%]	Obser-vation 187 [%]
Cluster 1	5			
Cluster 2	0.32			
Cluster 3	0.01			
Cluster 4	7			
Cluster 5	0			
...	...			
Cluster 687	0.8			
Sum	100			

Supplementary Material 1.2: A flow chart displaying the various approaches to detect the limits of reasonable classification by the ANN by reducing the amount of features and observations. Steps on unfiltered data are marked in red, filtered in green.


a)
b)
c)
d)
e)
f)
g)
h)
i)
j)
k)
l)
m)
n)
o)
p)
q)
r)
Control
Treatment
Relative abundance [%]
Days

Supplementary Material 1.3: All top 10 ranked clusters (Table 1.3) from the filtered and unfiltered data plus *Limnohabitans* spp. were displayed based on their relative abundance with 16S rRNA gene- and 16S rRNA- derived data for both microcosms. The technical replicates are shown as dots, the mean as line. The 16S rRNA gene is shown as continuous and 16S rRNA as broken line. The black vertical line demarks the addition of glyphosate. Due to abundance differences in orders of magnitude, the y scale is adjusted for each plot:

a) *Massilia* spp.; b) *Parvibaculum* spp.; c) *Dokdonella* spp.; d) *Reyranella* spp.; e) B38/*Gammaproteobacteria*; f) *Loktanella* spp.; g) *Caulobacter* spp.; h) *Aminobacter* spp.; i) *Nesiotobacter* spp.; j) *Idiomarina* spp.; k) *Hyphomonas* spp.; l) *Gallaecimonas* spp.; m) *Thalassobaculum* spp.; n) *Sphingopyxis* spp.; o) *Rhizobium* spp.; p) *Brevundimonas* spp.; q) *Sphingomonas* spp.; r) *Limnohabitans* spp.



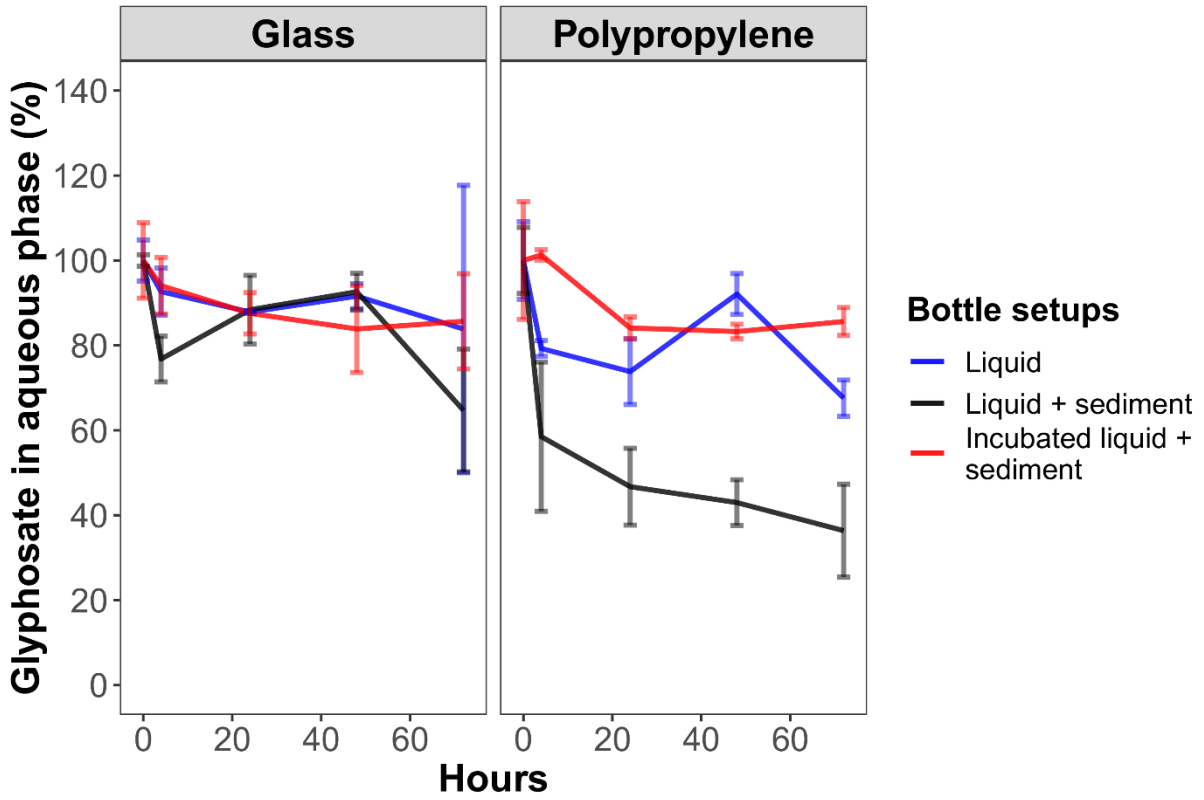
Supplementary Material 1.4: Classification rates after removal of observations. *n* is the number of classifications performed with the respective setup by the ANN. The horizontal bar at 59% displays the classification achievable by pure guessing, the upper bar marks the threshold for a classification which both separates the microcosms and before and after glyphosate addition. None of the ANN setups was able to reach the upper threshold, whereas RF-based classification was successful using solely 16S rRNA or 16S rRNA gene samples.

Supplementary Tables

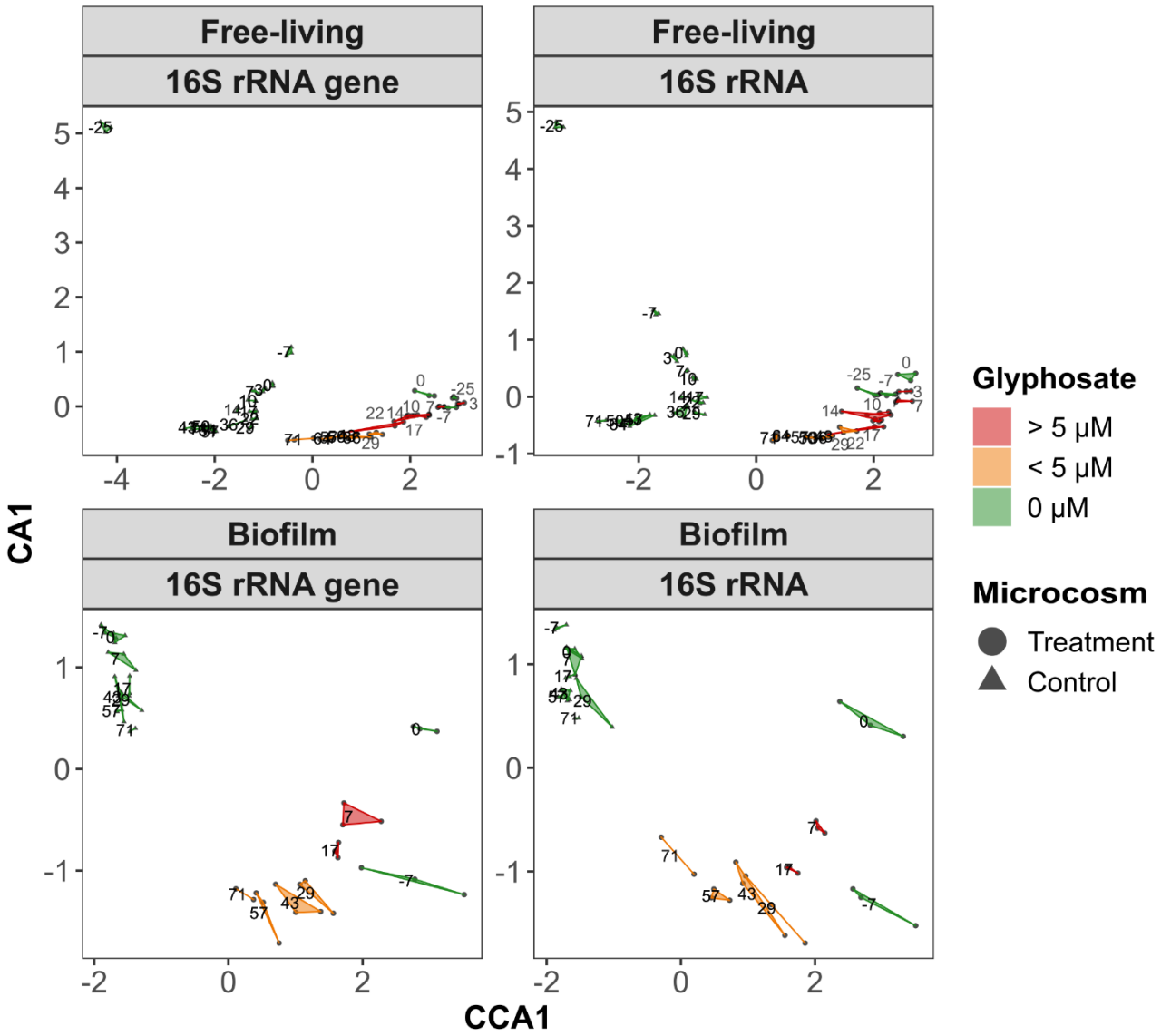
Supplementary Material 1.5 can be found as digital appendix and in the final publication as Supplementary Table 1.

Chapter II

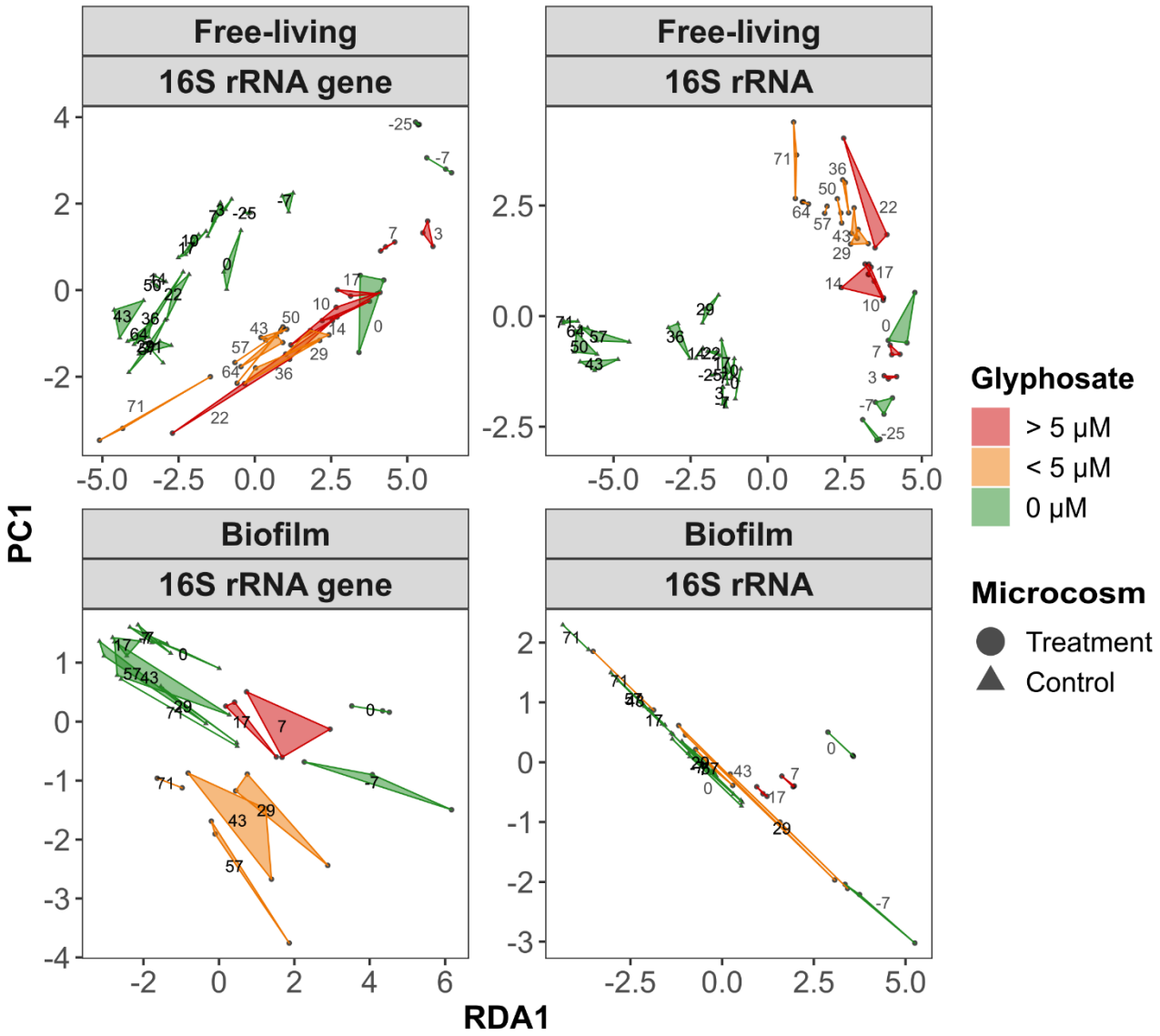
Supplementary Figures



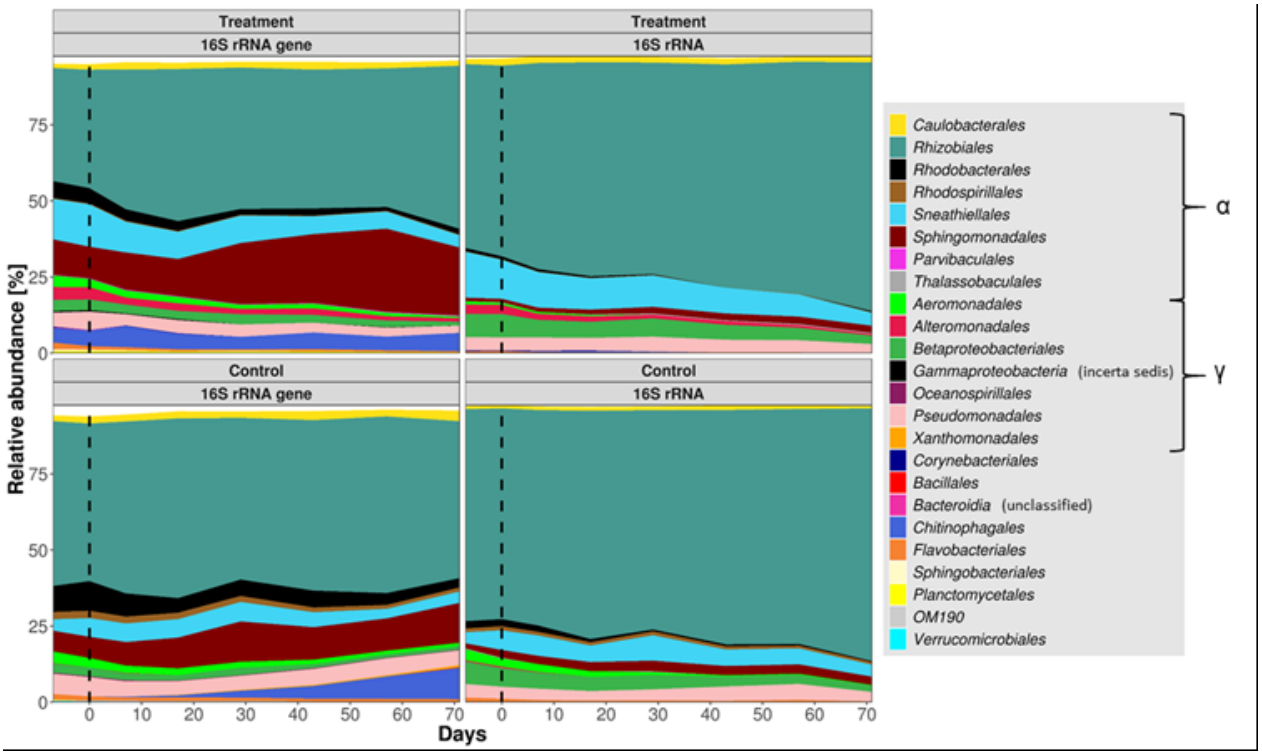
Supplementary Material 2.1: Glyphosate adsorption test. A possible adsorption effect of glyphosate to various surfaces was described previously (Bergström et al., 2011; Huang and Zhang, 2011). Adsorption contributes to the decrease and acts as glyphosate reservoir in the experiment. As glyphosate was measured in the water column, adsorption to biofilm or surfaces would be not distinguishable from dissipation. To assess the behavior of glyphosate in presence of biofilms in a microcosm, the following conditions were set up: a) 500 mL ABW (blue); b) 500 mL ABW and 250 g quartz sand (black) and c) 500 mL ABW and 250 g quartz sand and ½ inoculum filter (red), respectively, each prepared in 1 L glass bottles and in 1 L polypropylene bottles. The inoculation for c) took place for 5 days before the filter were removed. Bottles a) and b) were set up after the inoculation step for c). Glyphosate (Dr. Ehrenstorfer GmbH, Augsburg, Germany) was added to all bottles at the same time to a final concentration of 0.296 µM. The bottles were thoroughly mixed and the first sample (t_0, 800 µL) was taken in triplicate from each bottle. The bottles were further incubated at room temperature, stirred at 100 rpm and samples were taken in triplicate after 4 h, 24 h, 48 h and 72 h. The samples were stored at -20°C until measurement. The figure shows the results of the glyphosate adsorption test in glass and polypropylene bottles. The biota incubated bottles displayed smallest loss and fluctuation in glyphosate concentration. The glyphosate concentration at the end of the microcosm experiment was 1.01 µM and we suggest that in this range no degradation appears in a nutrient-rich environment.



Supplementary Material 2.4 A: Canonical correspondence analysis. Free-living communities are partially over-clustering, but a clear separation between treatment and control and largely as well for the different glyphosate concentrations was achieved. Biofilm samples are better separated compared to the NMDS ordination (Figure 3). Treatment communities' direction of succession changes from day -7 to 0 compared to day 0 to the samples treated with glyphosate (interpreting day 0 as a return point). This change of direction can be assumed in the free-living communities from day -25 to day 0 compared to the following samples.



Supplementary Material 2.4 B: Redundancy analysis. In the free-living communities, the “turning point” described in Supplementary Material 2.4 A can be observed again. 16S rRNA gene samples overlap partially, control and treatment samples are clearly separated. Except shortly after glyphosate addition, the change along the axes converges with the temporal gradient or the glyphosate concentration decrease. The 16S rRNA gene samples are well separated, but show a relative proximity between control and treatment communities. The results for Biofilm 16S rRNA communities do not help to explain the impact of glyphosate.



Supplementary Material 2.5: Relative biofilm community composition in the treatment and control microcosms based on 16S rRNA gene and 16S rRNA abundance. Taxa were cumulated on order level, sorted by class. α = *Alphaproteobacteria*, γ = *Gammaproteobacteria*. All orders > 0.15 % relative abundance are displayed. Glyphosate addition is marked by a vertical dashed line. Notice the dominance of *Rhizobiales* and the overall stability of the communities, especially based on the 16S rRNA gene.

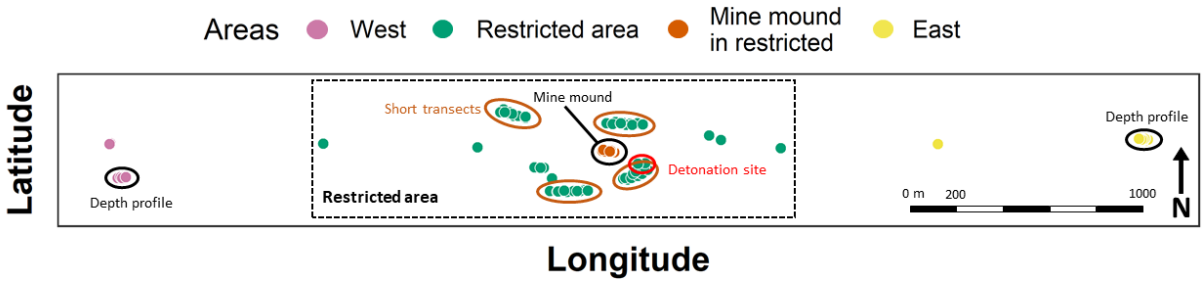
Supplementary Material 2.6 can be found as digital appendix and in the final publication as Supplementary Material 6.

Supplementary Tables

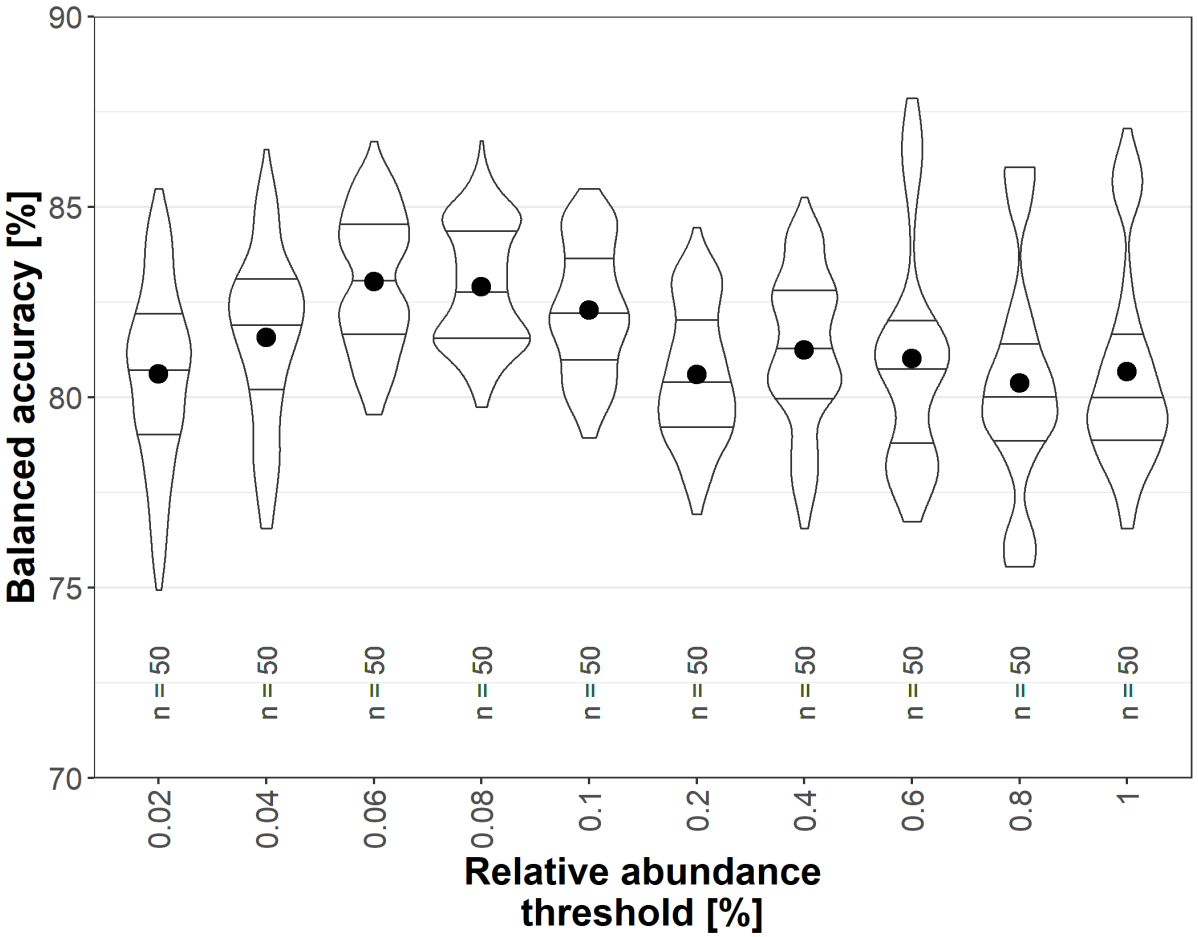
Supplementary Material 2.2, 2.3, 2.7 and 2.8 can be found as digital appendix and in the final publication as Supplementary Material 2, 3, 7 and 8.

Chapter III

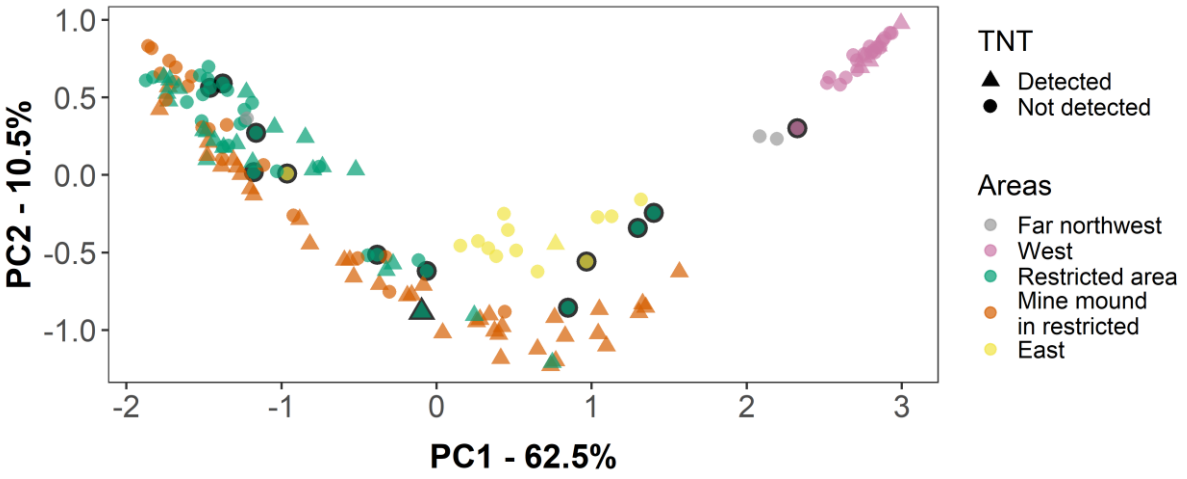
Supplementary Figures



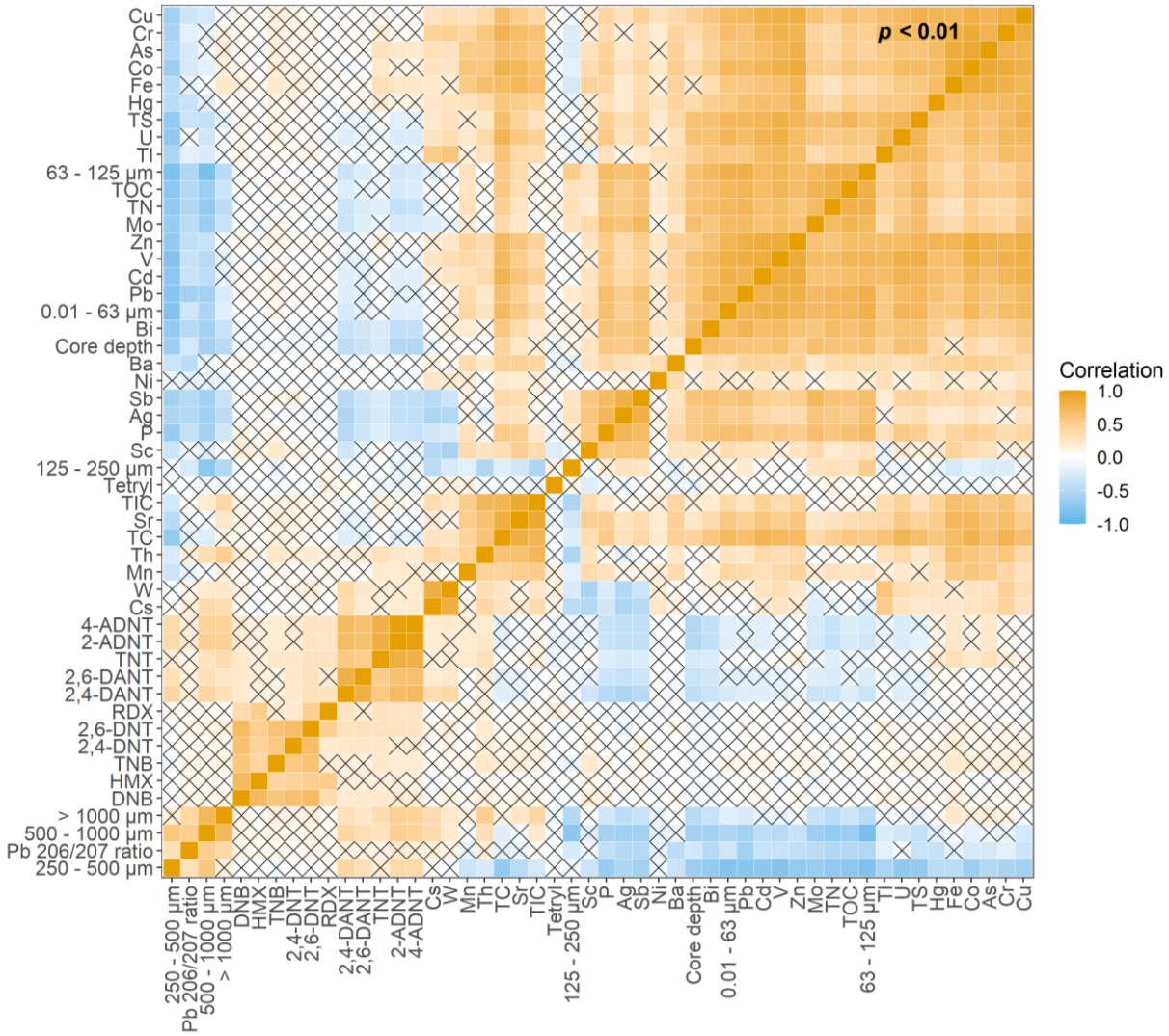
Supplementary Material 3.1: A map of the sampling sites at Kolberger Heide munitions dumpsite, located in the Baltic Sea near the city of Kiel, Germany. The restricted area is demarked by a dashed box. The multicorer sampling took place at the sites names “Depth profile”. Sampling sites featured in the study within the restricted area were the short transects of 200 m total length, with samplings every 20 m around the mine mound. The mine mound was subject to several sampling campaigns, including the sampling in defined distances to an individual mine. Craters caused by detonation of munition are located at the “Detonation site”, MC concentrations were there about 1000 times higher than in average. For more details the reader is referred to Kampmeier et al., (2020).



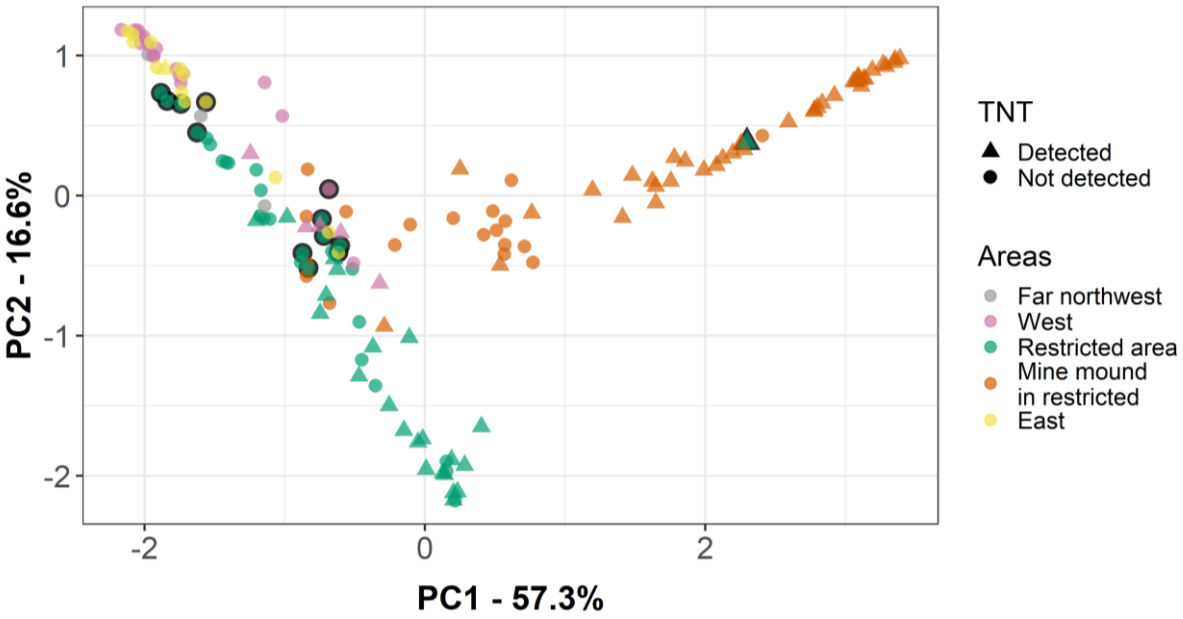
Supplementary Material 3.4: Violin plots of correct TNT classifications using different thresholds on relative abundance per ASV for the validation set. The dot represents the mean balanced accuracy, averaged over six different data set splits. n indicates the number of models calculated. The random forest models consisted of 10000 trees with an mtry factor of 5. The mean balanced accuracy ranged from 80.4 – 83.0 %. 0.08 % was chosen as the distribution became more distinct compared to the slightly better performing 0.06 % threshold.



Supplementary Material 3.5: PCA ordination for sediment data. The proximity matrix was generated by an unsupervised random forest classifying Full sediment data. In comparison to the PCA ordination based on the Top25 community, the core samples (West and East without black outline) were well separated herein. Furthermore, samples from the mine mound and the overall restricted area are more similar based on sediment parameters. PC1 explained 62.5 % variation, which likely correlated mostly with grain size fractions, the coarser directed to the left and the finer towards the right. It is shown that samples with and without TNT were well intermixed, which might be a reason for the lower classification scores achieved by sediment data.



Supplementary Material 3.6: Spearman rank correlation of sediment parameters. Correlations of $p > 0.01$ were signed as insignificant with an X. This analysis was performed to investigate which sediment parameters could be useful to predict TNT due to correlations. Positive correlations were found between TNT and its metabolites such as ADNTs and DANTs as well as iron, thallium, manganese, arsenic and cobalt. Furthermore, the grain size fractions 500 – 1000 µm and > 1000 µm, distinctive of the predominantly TNT-present mine mound samples, were identified. TNT was also negatively correlated with total nitrogen, antimony, silver, phosphorus, bismuth and sediment depth, hinting at the mostly TNT-absent MUC samples. Arsenic, cobalt, total nitrogen and grain sizes were important variables for the random forest model. In further leading investigations, lead, the lead isotope ratio $^{206}Pb/^{207}Pb$ and mercury (proposed to leak from UXO) were not found to correlate with any MC except for a weak negative tie between 2,4-DANT and lead. Two groups of MC were discernable: TNT and its metabolites and secondly, DNTs, TNB, HMX and DNB. RDX was loosely connected to both groups and Tetryl showed no correlation to any other MC. These results suggested that predicting TNT using other sediment parameter than its metabolites' concentrations would turn out challenging.



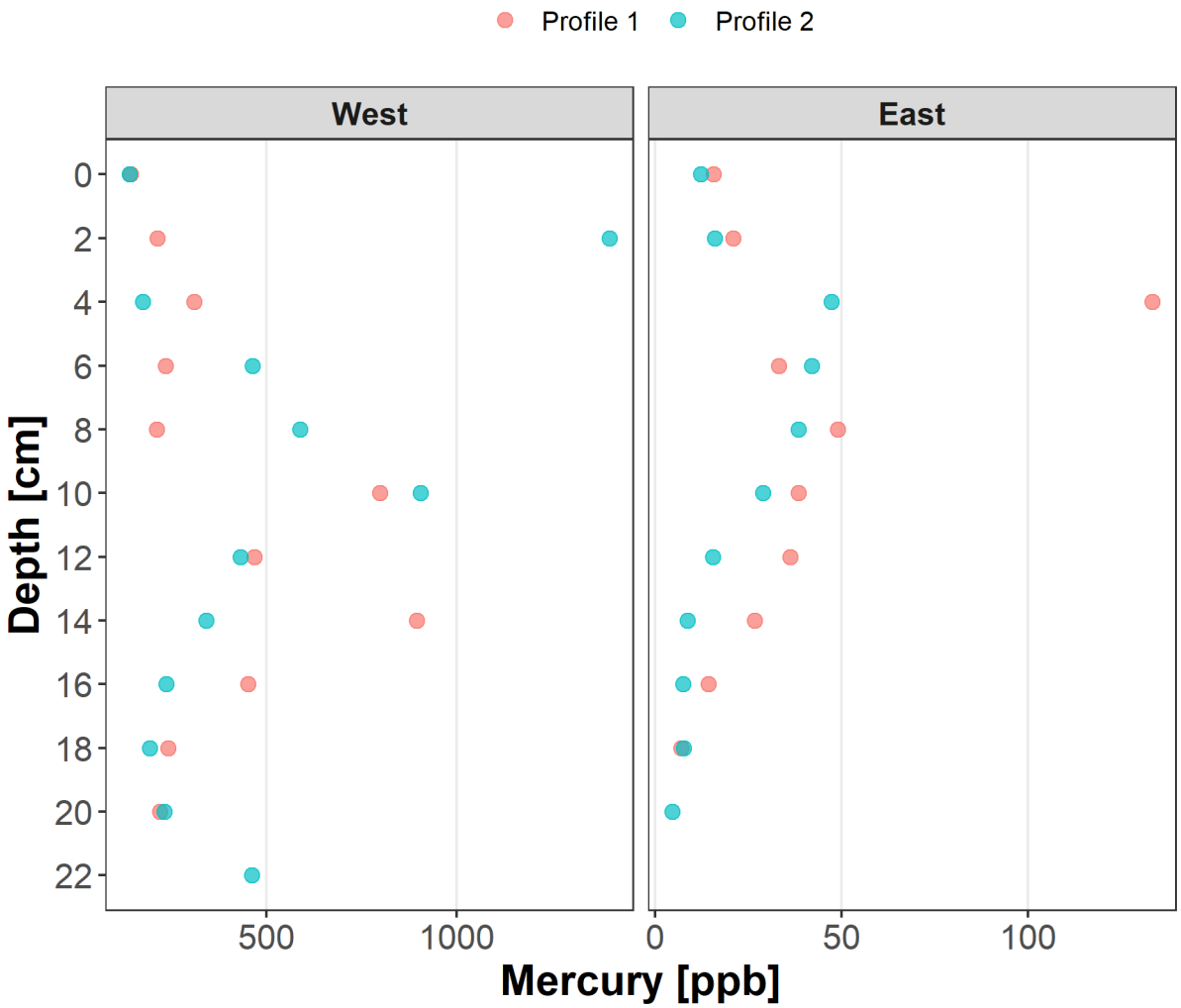
Supplementary Material 3.10: PCA ordination based on the abundance of the most important 25 genera. Dissimilarity calculated using the proximity matrix of TNT-classifying supervised random forest. The microbial communities were colored by sample area and shaped indicating the presence of TNT. The East (yellow) and West (purple) samples with a black outline were not MUC samples. The restricted area samples with a black outline were not part of a transect. Sediments containing TNT could be separated to the top right and to the center bottom, absent samples were located in the top left. The central area contained sediments with and without TNT.

Supplementary Figure 3.11: This document contains further information on the presence of certain heavy metals within all collected sediments and more specifically along depth profiles and in defined distances to mines.

Results and discussion:

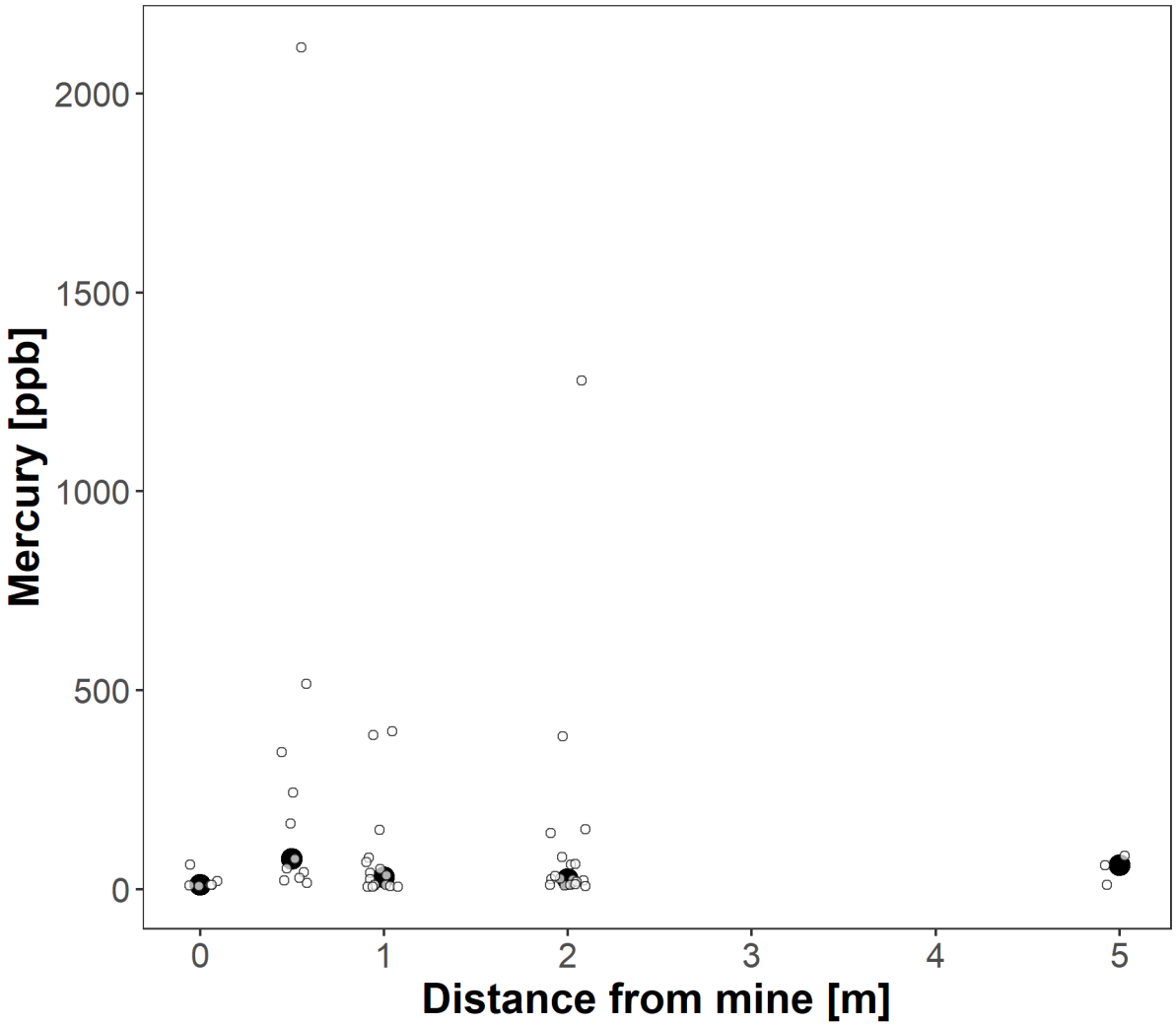
In 167 sediments, Hg ranged from 3.7 to 4503.4 µg $Hg \cdot kg^{-1}$ dry sediment, with a median of 20.5 µg and 15 samples exceeding 450 µg. The maximal concentration was found during a line transect, where the neighboring samples in 20 m distance contained 8 and 12 µg. Arsenic was detected from 0.4 to 4.8 ppm with a median of 0.8 and lead ranged from 1 to 75 ppm with a median of 2.

UXO have been proposed as point sources of heavy metals, especially mercury and lead. They were installed as highly toxic primary explosives mercury(II) fulminate, lead azide and lead styphnate. The mercury background in the Baltic Sea was estimated at 20 to 50 µg $Hg \cdot kg^{-1}$ dry sediment (Leipe et al., 2013). They also mentioned 250 $\mu g \cdot kg^{-1}$ as highest surface value in the northern Baltic sea and 450 $\mu g \cdot kg^{-1}$ several cm deeper of. A similar trend was shown within the western MUC cores, although concentrations at 10 cm depth reached up to 900 µg $Hg \cdot kg^{-1}$.



Mercury concentrations along the depth profiles taken with a multicorer east and west of the Kolberger Heide. Please not the different scales on the x axis.

Bełdowski et al. (2019) detected a maximal concentration of 322.2 µg $Hg \cdot kg^{-1}$ mercury in 8 Kolberger Heide top layer sediments with high variance between sediments. In our study spottily occurring high values of up to 4503 µg $Hg \cdot kg^{-1}$ were detected, too. Within 2 m of a mine the highest values were measured at 0.5 m distance (mean 329 µg $Hg \cdot kg^{-1}$, median 75.6 µg $Hg \cdot kg^{-1}$) distance. However, the other sediments within 2 m radius of the same mine did not show such elevated levels, potentially because the mines at Kolberger Heide are classified as discarded munition material. In comparison to unexploded ordnance, those were not fused and therefore should not contain mercury(II) fulminate.



Mercury concentrations in sediments sampled in a maximal distance of 5 m around mines. The black filled dot is the median concentration. Samples from 0.5 to 2 m distance originated from a cardinal direction wise sampling around 3 distinct mines. Sediments of 0 and 5 m stem from a linear distance sampling. All sampling took place in the mine mound area.

It has yet to be determined why rare samples demonstrate such high concentrations. There was no significant correlation over all sediment samples for Hg with TNT, though both substances would likely be transported differently if originating from the same mine. Lead concentrations fitted within expected Baltic Sea sediment background (Zaborska, 2014). The important variable arsenic caught our attention, as it also is a compound of chemical warfare agents. However, its median concentration did not exceed e.g. the average southeastern Baltic Sea background of 3.4 ppm (Garnaga et al., 2006). and chemical warfare agents were to our knowledge not disposed of in the Kolberger Heide (Böttcher et al., 2011; Bełdowski et al., 2016b).

Supplementary Material 3.8 and 3.9 can be found as digital appendix and in the final publication as Supplementary Figures 8 and 9.

Supplementary Tables

Supplementary Material 3.2, 3.3 and 3.7 can be found as digital appendix and in the final publication as Supplementary Table 2, 3 and 7.

Digital appendix

The appendix includes the Supplementary Material, that was left out for the printed thesis due to formatting restrictions

- Chapter I: Supplementary Table 1 (here Supplementary Material 1.5)
- Chapter II: Supplementary Tables 2, 3, 7, 8 (here Supplementary Material 2.2, 2.3, 2.7, 2.8), Supplementary Figure 6 (here Supplementary Material 2.6)
- Chapter III: Supplementary Figures 8, 9 (here Supplementary Material 3.8, 3.9), Supplementary Tables 2, 3, 7 (here Supplementary Material 3.2, 3.3, 3.7)

Furthermore, the tables used as input for analysis and plotting, including the taxa abundance tables, taxonomy tables and further meta data as well as the machine learning results are provided.

This data will be uploaded to

https://owncloud.io-warnemuende.de/index.php/s/0nvEnzEbiFtrC5c

The password is “my_thesis”

The code for the Chapters and the R package can be found in the according repos at https://github.com/RJ333/

The code for Chapter III is still under development, as the manuscript is only submitted: https://github.com/RJ333/Kolberger_Heide_manuscript

Acknowledgements

Diese Doktorarbeit ist das Resultat einer langen Zeit, die ich mit Unterbrechungen am IOW verbracht habe. Sie beinhaltet den Einfluss vieler Menschen, und ohne diese würde es die Arbeit vermutlich gar nicht geben. Trotzdem werde ich Leute hier vergessen haben zu nennen, das tut mir leid. Es gibt keine Reihenfolge bei der Nennung.

This thesis is the final consequence of a very long time I've spent intermittently at the IOW. The influence of many people is somehow included in this work and the thesis would likely not exist without these experiences. I first want to apologize to those I forgot here, I have the feeling this list is not complete (and I know it is already quite long). The listing is in no particular order. The originally german parts were translated because it's fun.

Es sind nun fast 9 Jahre vergangen, seitdem ich an der Uni Rostock in 2011 meinen Masterstudiengang Mikrobiologie begonnen hatte und 2012 zum ersten Mal als HiWi für Matthias und Joachim Kuss gearbeitet habe. Es ging um Quecksilber und Cyanobakterien in der Ostsee. Mir war nicht wirklich klar, warum man das untersuchen würde, also fragte ich als Biotechnologe von einer FH naiv: „und was wollt ihr dann damit anfangen?". Das war mein erster Kontakt mit Grundlagenforschung und Ökologie. Verschiedene HiWis, Anstellungen (u.a. bei Regine und Stephan Hüttmann, wofür ich sehr dankbar bin) und Forschungsprojekte später verstehe ich nun etwas von der Ostsee und warum Bakterien in der Umwelt doch eine gewisse Relevanz besitzen. Ich habe (bzw. musste) mir ein paar Kenntnisse im Bereich Bioinformatik und Statistik angeeignet und lernte vor allem darüber, wie Wissenschaft funktioniert (und wie nicht) und wie ich funktioniere (und wie nicht). Die Doktorandenzeit mit Dir war eine ungewöhnliche Zeit, mit vielen Wendungen und viel Verantwortung und Selbstständigkeit, aber dadurch auch sehr viel Flexibilität: Machine Learning oder TNT stand z.B. ursprünglich auf keinem Plan. Ich fange jetzt erst an zu erkennen, dass Du zu Beginn anscheinend deutlich mehr Vertrauen in meine Fähigkeiten attest als ich. Jahre später dann dachte ich mir manchmal, dass Du mir nichts mehr beibringen kannst, vor allem, wo die Bioinformatik eine große Rolle eingenommen hatte. Tja, wie falsch ich damit lag. Bis zur Abgabe der Doktorarbeit und vermutlich darüber hinaus habe ich direkt oder indirekt von Dir gelernt. Es hat mir immer geholfen, dass Du deine Kritik genau formulieren konntest, so dass ich sie nachvollziehen konnte (wenn ich sei eingesehen habe). Die Doktorarbeit war ein langer und ziemlich anstrengender Weg, aber es war wohl auch der Weg und die Zeit, die ich benötigt habe. Das war nur möglich durch deine Betreuung und deinen persönlichen Einsatz und dafür möchte ich mich ganz herzlich bedanken. Vielleicht konntest Du auch was von mir lernen, z.B., dass die einzig angemessene Beinbekleidung für jemanden, der den ganzen Tag vor dem Rechner sitzt, eine Jogginghose ist (oder eine Badehose).

Autotranslated: Almost 9 years have passed since I started my master's degree in Microbiology at the University of Rostock in 2011 and worked as a HiWi for Matthias and Joachim Kuss for the first time in 2012. It was about mercury and cyanobacteria in the Baltic Sea. I didn't really know why they would investigate this, so as a biotechnologist from a university of applied sciences I asked naively: "and what do you want to do with it? That was my first contact with basic research and ecology. Various hiWis, jobs (e.g. with Regine and Stephan Hüttmann, for which I am very grateful) and research projects later I now understand something about the Baltic Sea and why bacteria in the environment have a certain relevance after all. I acquired (or had to acquire) some knowledge in bioinformatics and statistics and learned mainly about how science works (and how not) and how I work (and how not). The PhD time with you was an unusual time, with many twists and turns and a lot of responsibility and independence, but also a lot of flexibility: Machine Learning or TNT, for example, was not originally on any plan. I'm only now beginning to realize that at the

beginning you seemed to have much more confidence in my abilities than I did. Years later, I sometimes thought that you couldn't teach me anything anymore, especially when bioinformatics had played a major role. Well, how wrong I was about that. Until I handed in my doctoral thesis and probably beyond I learned directly or indirectly from you. It always helped me that you were able to formulate your criticism precisely so that I could understand it (if I accepted it). The doctoral thesis was a long and rather exhausting way, but it was probably also the way and the time I needed. This was only possible because of your supervision and personal commitment and I would like to thank you very much for that. Maybe you could also learn something from me, e.g. that the only appropriate legwear for someone who sits in front of the computer all day long is a pair of jogging pants (or swimming trunks).

I would like to sincerely thank the further reviewers of this thesis, Rudolf Amann, Alexander Probst and Stephen Techtmann, the latter two I could meet in person at the ISME in Leipzig. At that time I would not have thought that you would end up as my reviewers, thank you for that!

Bei Heide möchte ich mich für deine Unterstützung meines etwas kurvenreichen Weges am IOW bedanken. Außerdem haben mir deine pragmatischen und direkten Hinweise viel geholfen.

Autotranslated: I would like to thank Heide for your support of my somewhat winding path at the IOW. Your pragmatic and direct advice also helped me a lot.

Wenn es Probleme mit Verwaltungssachen, Terminen, Finanzierung, etc. gab (oder, wenn ich es mal wieder nicht geschafft habe, einen Bestellschein auszufüllen), bin ich zu Solveig gegangen. Danach gab es kein Problem mehr. Danke! Vielen Dank auch an Siggi, Swen und Bernd, die meinen Rechner unzählige Male begutachtet haben und sich bis zum Ende nicht sicher waren, was der eigentlich für ein Problem hat. Und auf jeden Fall vielen Dank an die Werkstatt Crew, mit denen ich unsere eigene Magnetaufreinigungsplatte entworfen habe (besser und ca. 1/10 so teuer wie das gekaufte Produkt), und die immer für verrückte Ideen (Hutbau inklusive) aufgeschlossen waren.

Autotranslated: Whenever there were problems with administrative matters, appointments, financing, etc. (or when I once again failed to fill out an order form), I went to Solveig. After that there was no problem anymore. Thanks! Many thanks also to Siggi, Swen and Bernd, who checked my computer countless times and were not sure until the end what the problem was. And in any case many thanks to the workshop crew, with whom I designed our own magnetic cleaning plate (better and about 1/10 as expensive as the purchased product), and who were always open for crazy ideas (hat making included).

Meine ersten IOW Laborerfahrungen (ebenfalls 2012) waren beaufsichtigt von Christian M. Von deiner PCR-Einführungen habe ich die gesamte Laborzeit profitiert, meine PCRs haben fast immer funktioniert! Außerdem warst Du genauso wie Christian B. und auch Christian S. immer gute Gesellschaft fürs Mittagessen. Als es mit der Doktorarbeit losging war plötzlich Christin L. da. Ohne Dich wäre ein Großteil meiner Laborarbeit gar nicht, oder deutlich schlechter, oder deutlich langsamer gewesen. Außerdem hast Du mir Arbeit abgenommen, wenn es wirklich dringend (oder echt langweilig) war. Kultivierung ist spannend, aber auch ziemlich mühselig ;-) Ich möchte mich bei allen TAs und Methodenspezialisten bedanken, die mir Molekularbio (Heike) oder im Schnellverfahren geologische und sedimentologische Analysen gezeigt und erklärt haben (Anne, Ines, Olaf, Sascha, Mischa, Peter) oder für mich gemessen haben (Steph, Annett, Ronny, Jenny, Wael, Marisa, Rainer, …). Ich werde versuchen, die Vakuumpumpen in Zukunft nicht mehr mit Sand zu befüllen.

Autotranslated: My first IOW lab experiences (also in 2012) were supervised by Christian M. I benefited from your PCR introductions the whole lab time, my PCRs almost always worked! Furthermore you were always good company for lunch, just like Christian B. and Christian S. When the doctoral thesis started, suddenly Christin L. was there. Without you, most of my laboratory work would not have worked at all, or would have been much worse, or much slower. You also took work off my hands when it was really urgent (or really boring). Cultivation is exciting, but also quite tedious ;-) I would like to thank all the TA and method specialists who showed and explained molecular bio (Heike) or in a fast procedure geological and sedimentological analyses (Anne, Ines, Olaf, Sascha, Mischa, Peter) or measured for me (Steph, Annett, Ronny, Jenny, Wael, Marisa, Rainer, ...). I will try not to fill the vacuum pumps with sand in the future.

Bei Christin B. möchte ich mich bedanken, dass Du mich an Blueprint beteiligt und mich mit nach Schweden genommen hast, ebenso für die ganze Organisation.

Autotranslated: I would like to thank Christin B. for involving me in Blueprint and taking me to Sweden, as well as for the whole organization.

I'm very happy and lucky that I joined the Environmental Microbiology group. I enjoyed all our group activities (including the working group meetings, where I could make most effectively fun about microplastics :-)). Thank you all for your support, comments, ideas and just for being there. Also thanks to Juliana for always trying to get the people together!

Going twice to Anders' group at SciLifeLab to learn about bioinformatic processing (the first time) and analyze the glyphosate metagenome with Johannes (the second time) was the beginning of my encounters with bioinformatics and programming in general. I remember asking (and probably confusing) Anders in advance whether I need to install Biolinux on my laptop, because I thought that would show good preparation. I later realized that the work does not happen on the laptop ;-) I'm very grateful that I could come to Sweden effectively as appendix to Christin's Blueprint analysis. Anders continued to provide extremely valuable ideas to my work (and also, in my memory, he was able to take whole ideas of mine apart with a single question) as part of my thesis committee, while also being just a very nice person to be around. You even took us to the Schären islands (I guess *skärgård)* to renovate historical boats. It was an exciting time to live in Stockholm and Uppsala, but the best part was to get to know Johannes A., Luisa and Yue, who all did not even try to hide from the many questions I had. It was great to be a guest to Yue's defense (and the party afterwards). Johannes spent a lot of time with my data and teaching me, Christin and Emma (who was great to learn together with, although she did not enjoy Subway food as much as I did) about bioinformatics and how to work resource-efficiently on UPPMAX ("oh yes, you just threw a whole fat node on this simple task…"). He also taught me a very important rule: "When the computer is working, it means I'm working. Let's have Fika". This visit restarted my interest in programming after very disappointing trials in school. I was happy to meet new people (e.g. Jürg) and former colleagues (Carlo) during lunch or after hours. I also appreciate greatly that Malin (although I never met her) and Anke let us/me stayed in their apartments for my both visits.

I would like to thank Mercè for many more things then showing me around in Gamla Uppsala, but most influential were the introduction into R ("I don't use R Studio", "merging is the best") and learning about Catalan culture (despite our **Castellano**-German tandem learning) in Germany and Catalunya. ;-) I'm looking forward to our next Soziedad Alkoholika concert and maybe once I'll dare to participate in Carneval de Vic!

Der wohl prägendste Teil meiner IOW Zeit war das Büroleben in 301, nachdem mit Katis Einzug drei Doktoranden ihr Leid, ihre Sorgen, ihre Frustrationen (vor allem PC-bezogen)

und ihre Einkäufe beim Bäcker und Edeka (es gab auch positive Momente) teilen konnten, während sich das Büro in einen Dschungel und zentrale Anlaufstelle entwickelte. Es sind viel zu viele Momente, um sie hier aufzuzählen, aber unsere gemeinsamen (oft auch hitzigen) Diskussionen sind mir wichtig. Kati, Du bist in ganz vielen Fällen meine erste Ansprechpartnerin (und Korrekturleserin) gewesen, professionell und privat. Lars, Du hast mir gezeigt, dass man für seine Prinzipien einstehen sollte und hat so manche Lücke in meiner Argumentation offenbart, ihr beide habt mich zu einem besseren Wissenschaftler gemacht hat. Zu dem IOW-Gefühl stark beigetragen haben weiterhin Franzi, Philipp und Jan, z.T. durch wissenschaftliche Diskussionen, aber vor allem durch das nach dem Klopfen folgende: „Habt ihr Essen dabei?“, „Jemand Lust auf nen Kaffee?“ bzw. „Geht wer schwimmen?“. Ich fand es motivierend zu sehen, wie ihr alle Experten auf euren Gebieten geworden seid.

Autotranslated: Probably the most formative part of my IOW time was the office life in 301, after Kati's arrival three PhD students shared their suffering, worries, frustrations (mainly PC related) and shopping at the bakery and Edeka (there were also positive moments), while the office developed into a jungle and central contact point. There are far too many moments to list here, but our joint (often heated) discussions are important to me. Kati, you have been my first point of contact (and proofreader) in many cases, professionally and privately. Lars, you have shown me that one should stand up for one's principles and have revealed many a gap in my argumentation, you both have made me a better scientist. Franzi, Philipp and Jan also contributed to the IOW feeling, partly through scientific discussions, but especially through the following after the knocking: "Do you have any food with you?", "Anyone in the mood for a coffee?" and "Does anyone want to go swimming?” I found it motivating to see how you all have become experts in your fields.

I want to thank Brittan for advice on my work and especially for helping me with my English in manuscripts. Furthermore, I really appreciate your support for my US holiday trip and hope that we can have more interesting discussion about US and German cultural stuff. The same goes for Alex, who hopefully enjoyed the Jehacktet concert we went to years ago. Thank you very much for your English check on this thesis and good luck with your proposal!

Sophie, you were my boss during my second IOW HiWi. I want to thank you for teaching me about RNA-grade extractions and being prepared before starting. I want to thank you but way more for your help, suggestions and kindness. I was very happy and thankful to meet you again in the US and wish you all the best! Hope to see you again!

An die Munitionsgruppe um Claus Böttcher, Jens Sternheim, Aaron Beck, Edmund Maser, Jennifer Strehse, Daniel Appel, Mareike Kampmeier und Munitect: Danke für die Möglichkeit, Vorträge über meine Arbeit zu halten und dass ihr mir diese tollen und spannenden Proben zur Verfügung gestellt habt und mich in den UDEMM-Zirkel mit aufgenommen habt. Es war eine Bereicherung, mit Leuten aus verschiedenen Disziplinen und Sichtweisen (Wissenschaft, Politik, Behörde) zu kooperieren, die so kompetent sind und sich gegenseitig unterstützen. Ein besonderer Dank gilt Claus dafür, dass Du alles getan hat um die Fäden zusammenzuführen, Aufmerksamkeit zu generieren und finanzielle Mittel aufzutreiben sowie Aaron, der Du mir viele, viele Fragen beantworten musstest und beim Thesiskomitee mitgewirkt hast.

Autotranslated: To the ammunition group around Claus Böttcher, Jens Sternheim, Aaron Beck, Edmund Maser, Jennifer Strehse, Daniel Appel, Mareike Kampmeier and Munitect: Thank you for the opportunity to give lectures about my work and for making these great and exciting samples available to me and for including me in the UDEMM circle. It was an enrichment to cooperate with people from different disciplines and points of view (science, politics, authorities) who are so competent and support each other. A special thanks to Claus

for doing everything to bring the threads together, to generate attention and to raise funds, and to Aaron for answering many, many questions and for helping with the thesis committee.

Thomas und Jakob ehemals vom Fraunhofer IGD: Es war spannend mit Euch zusammenzuarbeiten und Unterhaltungen über eine mir völlig unbekannte Disziplin zu führen, genauso wie der Versuch, euch die Welt der Molekularbiologie näherzubringen. Es ist schade, dass die Kooperation nicht weiterging, da ich nun etwas mehr Schnittstellenkompetenz zum Machine Learning einbringen könnte ;-) Es war ein sehr glücklicher Zufall für mich, da ML schließlich zum beherrschenden Thema meiner Arbeit geworden ist und mich die Herangehensweise der Modelle so an die Prüfungsphasen der Unizeit erinnert: Fragen und Antworten in Altklausuren anschauen und den Zusammenhang dazwischen rausfinden.

Autotranslated: Thomas and Jakob formerly from Fraunhofer IGD: It was exciting to work with you and to have conversations about a discipline completely unknown to me, as well as the attempt to introduce you to the world of molecular biology. It's a pity that the cooperation didn't continue, because now I could bring a little more interface competence to machine learning ;-) It was a very lucky coincidence for me, because ML has finally become the dominating topic of my work and the approach of the models reminds me so much of the examination phases at university: looking at questions and answers in old exams and finding out the connection between them.

Bei Sebastian Jordan möchte ich mich für die Chance und das Vertrauen bedanken, seine Methanoxidierer 16S auszuwerten, es hat sowohl mit Dir als auch mit den Daten Spaß gemacht zu arbeiten. Sag Bescheid, wenn Du mal wieder umziehst ;-)

Autotranslated: I would like to thank Sebastian Jordan for the chance and the confidence to evaluate his methane oxidizers’ 16S, it was fun to work with you as well as with the data. Let me know when you move again ;-)

Bei Christian S. möchte ich für seinen Rat, seine Anteilnahme und sein Interesse gerade zu Beginn meiner Zeit am IOW bedanken (gleiches gilt für Falk und Sonja). Ich konnte immer zu Dir kommen und Du hast mir direkt geholfen oder mich entsprechend weitergeleitet. Außerdem hast Du mich ab und an dran erinnert, dass Freizeit auch einen gewissen Stellenwert besitzt und dass man nicht zu streng mit sich sein sollte. Ich weiß noch, wie Du mir nach einem fehlgeschlagenen Versuch gesagt hast: „Und weißt du, was beim nächsten Mal passiert? Es wird wieder schiefgehen! Und dann wird es wieder schiefgehen! Bis es irgendwann klappt, so ist das halt.“

Autotranslated: I would like to thank Christian S. for his advice, his sympathy and his interest especially at the beginning of my time at the IOW (the same goes for Falk and Sonja). I could always come to you and you helped me directly or forwarded me accordingly. You also reminded me from time to time that leisure time is also very important and that one should not be too strict with oneself. I still remember how you told me after a failed attempt: "And do you know what happens next time? It will fail again! And then it will go wrong again! Until one day it will work, that's the way it is."

Ich möchte mich ganz herzlich bei Jerry, Patrick, Kerstin und Moritz bedanken, dass ihr euch meine Sorgen angehört habt, für Ablenkung gesorgt habt und mich in Rostock besucht habt. Irgendwann wohne ich auch wieder näher bei euch. Wenn alles gut geht, sitzen wir dieses Jahr wohl mit 2 Doktoren unterm Weihnachtsbaum ;-) Der gleiche Dank geht an alle ehemaligen aus Krefeld und Umgebung, die den weiten Weg bis nach Rostock auf sich genommen haben. Die Berge rufen schon.

Autotranslated: I would like to thank Jerry, Patrick, Kerstin and Moritz for listening to my concerns, providing distraction and visiting me in Rostock. Sometime I will live closer to you again. If all goes well, we will probably sit under the Christmas tree with 2 PhDs this year ;-) The same thanks goes to all the former ones from Krefeld and surroundings who took the long way to Rostock. The mountains are already calling.

Im Namen meiner geistigen Gesundheit und der Abwechslung möchte ich mich bei Torsten, Ols und Robert (Radio Lohro Metaltörn), Brit und der Housedance Gruppe (Tanzland) sowie den Doppelkopfleuten bedanken, die mir sehr geholfen haben, den Kopf frei zu kriegen und mich für andere Sachen zu begeistern.

Autotranslated: In the name of my mental health and distraction from work I would like to thank Torsten, Ols and Robert (Radio Lohro Metaltörn), Brit and the Housedance Group (Tanzland) as well as the Doppelkopf people who helped me a lot to clear my head and to get enthusiastic about other things.

Bei Johannes W. möchte ich mich bedanken, dass Du meine Programmierfähigkeiten so stark vorangetrieben hast. Ohne deinen Antrieb gäbe es jetzt kein R package und ich hätte keine Ahnung von code review, git, merge requests, unit tests oder continuous integration oder warum Java vom Teufel persönlich erdacht wurde. Alles Gute in Tübingen und grüß Max!

Autotranslated: I would like to thank Johannes W. for pushing my programming skills so hard. Without your drive there would be no R package now and I would have no idea about code review, git, merge requests, unit tests or continuous integration or why Java was invented by the devil himself. All the best in Tübingen and greet Max!

Bei Carla Martin möchte ich mich dafür bedanken, dass Du die TNT Metagenome in Angriff genommen hast. Das war vor allem für eine Bachelorarbeit eine ziemliche anspruchsvolle Tätigkeit, die Du gut gemeistert hast. Ich hoffe, dass es Dir auch Spaß gemacht hat.

Autotranslated: I would like to thank Carla Martin for tackling the TNT metagenomes. Especially for a bachelor thesis this was quite a challenging task, which you mastered well. I hope that you enjoyed it as well.

I would like to thank an enormous number of unknown or anonymous people that provided online resources on programming, statistics and machine learning and even answered to my specific questions, in particular Stack Overflow and the Statquest youtube channel.

Bei Fritzi möchte ich bedanken, dass sie mir die Korngrößenbestimmung am neuen Gerät gezeigt hat und meine Arbeit Korrektur gelesen hat. Vielmehr bin ich aber froh, dass du in „meine“ WG gezogen bist, einen starken Einfluss auf die letzten zwei Jahre hattest und es auch während der stressigen Schreibphase noch mit mir ausgehalten hast. Auch Martin ist hier gedankt, der sich immer fragt, wie jemand so lange am Rechner sitzen kann. Aber dafür hast Du mir ja die entsprechenden Home office Übungen gezeigt ;-)

Autotranslated: I would also like to thank Fritzi for showing me how to determine grain size on the new instrument and for proofreading my work. But I am much more happy that you moved into "my" flat share, had a strong influence on the last two years and that you still put up with me during the stressful writing phase. I would also like to thank Martin, who always wonders how someone can sit at the computer for so long. But you showed me the corresponding home office exercises ;-)

Ich möchte mich bei Katrin, die nun auch ihren Doktor hat, bedanken für all deine Unterstützung, Zuwendung und das Teilen des Doktorandenseins.

Autotranslated: I would like to thank Katrin, who now also has her doctorate, for all your support, attention and sharing of being a doctoral student.

Hannah: für alles, was Du verändert hast.

Autotranslated ☺: Hannah: for everything that you have changed.

Declaration of authenticity

Ich versichere hiermit an Eides statt, dass ich die vorliegende Arbeit selbstständig angefertigt und ohne fremde Hilfe verfasst habe, keine außer den von mir angegebenen Hilfsmitteln und Quellen dazu verwendet habe und die den benutzten Werken inhaltlich und wörtlich entnommenen Stellen als solche kenntlich gemacht habe.

Rostock, den 18.09.2020