Diss. ETH No. 28093

MALDI-TOF mass spectrometry based clinical antimicrobial resistance prediction using machine learning

A dissertation submitted to attain the degree of

Doctor of Sciences of ETH Zurich
(Dr. sc. ETH Zurich)

presented by

Caroline Viktoria Weis
M. Sc. ETH, Biotechnology

born on 23 April 1991
citizen of Germany

accepted on the recommendation of

Prof. Dr. Karsten Borgwardt
Prof. Dr. Dr. Adrian Egli
Prof. Dr. Jean-Philippe Vert

2022

Abstract

Antimicrobial resistance has emerged as one of the most severe infectious disease threats in the 21st century, with the World Health Organisation declaring it one of the ten major global public health challenges facing humanity. With ample usage of antimicrobial drugs both in farming and healthcare along a rise in the occurrence of new resistances threatening human lives, early characterisation of an infection in a patient along with targeted administration of antimicrobial drugs is of utmost importance. This has led to the credo of *antimicrobial stewardship*, which works towards the goal of quantifying and improving the usage of antimicrobials, targeting both prescription by physicians as well as intake by patients. These steps are critical for effective treatment of infections, reducing harms resulting from unnecessary antimicrobial admission, and combating the development of new resistances. Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) mass spectrometry (MS) is the most widely-used technology for rapid microbial species characterisation in clinics. The extensive information contained in MALDI-TOF mass spectra has the potential to provide knowledge beyond pathogen identification and to predict antimicrobial resistance by harnessing the power of machine learning. A systematic literature review at the beginning of this thesis concludes that efforts towards large-scale machine learning for resistance prediction on MALDI-TOF mass spectra are rare, stymied by the lack of machine learning model development tailored to MALDI-TOF MS and an absence of large benchmark databases.

This thesis presents several advancements towards the goal of clinically-applied antimicrobial resistance prediction based solely on rapidly-available MALDI-TOF mass profiles. At the outset, we introduce *DRIAMS*, a newly curated dataset of unprecedented size that combines more than 300,000 MALDI-TOF MS mass profiles with more than 750,000 antimicrobial resistance labels, allowing for large-scale machine learning analysis of MALDI-TOF MS based phenotype prediction. We establish a predictive performance baseline employing several widely-used machine learning models—logistic regression, light gradient boosting machines and multi-layer perceptrons—on the following prediction tasks: ceftriaxone resistance prediction in *E. coli* and *K. pneumoniae* and oxacillin resistance prediction in *S. aureus*. Auxiliary analyses indicate that recent samples and resistance prediction stratified by species lead to the most favourable performance results. The models reach high predictive performances with an AUROC of 0.74 for ceftriaxone resistance prediction both in *E. coli* and *K. pneumoniae* and 0.80 for oxacillin resistance in *S. aureus*.

After establishing the rich potential of MALDI-TOF mass spectra, we present GP–PIKE, the first machine learning model tailored to phenotype prediction from MALDI-TOF MS profiles, using a novel kernel—PIKE—combined with a Gaussian Process classifier. The kernel is designed to exploit the properties of MALDI-TOF profiles and to take a reduced data representation of MALDI-TOF mass spectra as input. GP–PIKE outperforms the baseline methods, logistic regression and a Gaussian Process classifier with an RBF kernel, by a large margin. Furthermore, a behavioural analysis of GP–PIKE's maximum class probability indicates its usefulness as a well-calibrated confidence estimate. The results obtained with GP–PIKE suffer from a large standard deviation between train–test splits. We conjecture that this variation stems from some underlying phylogenetic structure that was previously not considered in the stratification. In this thesis, we introduce a stratification procedure enhanced by hierarchical clustering, aiming to infer phylogenetic relatedness from MALDI-TOF mass spectra and to enforce a similar distribution of the inferred structure between train and test. While we do not observe a decrease in standard deviation, the results indicate improved prediction results through our hierarchical stratification procedure.

Further, we evaluate the transferability of predictors trained at a specific source site to profiles collected at specific other medical institutions. Our results suggest that all models require retraining on samples native to the prediction site; however, large MALDI-TOF MS datasets collected at other sites can improve predictive performance further. The low transferability likely stems from distribution shifts between datasets from different collection sites. To mitigate these distribution shifts, we introduce a new method based on adversarial representation learning. The presented approach is able to balance out separations between distributions, however, it has not been shown to be able to improve the predictive performance at the target site. Lastly, we outline the steps necessary for the full development of a clinically applicable predictor and the potential of antimicrobial phenotype prediction based on other data types.

By providing the first large-scale, publicly available database for MALDI-TOF MS based clinical antimicrobial resistance prediction, demonstrating the potential of established machine learning models as well as developing a new kernel tailored to this data type, this thesis constitutes a major step towards MALDI-TOF MS based clinical antimicrobial resistance prediction and more generally leveraging digital approaches for antimicrobial stewardship.

Zusammenfassung

Resistenz gegen antimikrobielle Substanzen hat sich zu einer der größten Gefahren durch Infektionskrankheiten des 21. Jahrhunderts entwickelt. Die Weltgesundheitsorganisation zählt diese Resistenzen zu den zehn größten Bedrohungen für die öffentliche Gesundheit. Da antimikrobielle Medikamente umfangreich eingesetzt werden (vom landwirtschaftlichen Sektor bis zum Gesundheitswesen) und gleichzeitig immer neue lebensbedrohlichen Resistenzen auftreten, ist eine frühzeitige Charakterisierung der Infektion und eine gezielte Verabreichung von antimikrobiellen Medikamenten von höchster Wichtigkeit.

Diese Problematik hat zum Credo des *Antimicrobial Stewardship* geführt. Sie zielt darauf ab, den Einsatz von antimikrobiellen Substanzen zu quantifizieren und zu verbessern, sowohl die Verschreibung durch Ärzte als auch die Einnahme durch Patienten. Dies ist entscheidend für die effektive Behandlung von Infektionen, Vermeidung von Schäden durch unnötige Antibiotikaeinnahme, und Bekämpfung von sich neu entwickelden Resistenzen. Die Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) Massenspektrometrie (MS) ist die meistverwendete Methode zur Speziesidentifikation in der Klinik. Die MALDI-TOF-Massenspektren bilden die Zellen so umfassend ab, dass sie neben der Erregeridentifikation das Potential haben weiter Informationen zu liefern, zum Beispiel die Vorhersage von Antibiotikaresistenzen durch Methoden des maschinellen Lernens (Machine Learning). Eine systematische Übersichtsarbeit zu Beginn der dieser Arbeit kommt zu dem Schluss, dass die Entwicklung von Resistenzvorhersagemodellen anhand von MALDI-TOF Massenspektren durch fehlende auf MALDI-TOF MS zugeschnittene Machine-Learning-Modellen oder großen öffentlich zugänglichen Datensätzen eingeschränkt wird.

Diese Doktorarbeit stellt mehrere Weiterentwicklungen in Richtung einer klinisch anwendbaren Vorhersage von Antibiotikaresistenzen durch MALDI-TOF Massenprofilen vor. Zu Beginn steht *DRIAMS*, ein neu kuratierter Datensatz von noch nie dagewesener Größe. Dieser verknüpft mehr als 300000 MALDI-TOF MS Massenprofile mit über 750000 antimikrobiellen Resistenzen, die eine große Analyse der MALDI-TOF MS-basierten Phänotypvorhersage ermöglichen. Wir analysieren *DRIAMS* mit drei weit verbreiteten Lernmodellen—logistische Regression, LightGBM und das sogenannte multilayer perceptrons—im Hinblick auf drei Vorhersageaufgaben: die Vorhersage von Resistenz gegen Ceftriaxon bei *E. coli* und *K. pneumoniae* und der Oxacillin Resistenzvorhersage bei *S. aureus*. Zusätzliche Analysen zeigen, dass zeitnahe Daten und Resistenzvorhersage auf Speziesebene zu den besten Vorhersagen führen. Die Modelle führen zu starken Vorhersagen mit einem AUROC Wert von 0,74 für die Vorhersage von Ceftriaxon-Resistenz bei *E. coli* und *K. pneumoniae* und 0,80 für die Oxacillin-Resistenz bei *S. aureus*.

Als nächstes stellen wir GP–PIKE vor, das erste Machine-Learning-Modell welches auf die Vorhersage von Phänotypen aus MALDI-TOF MS Profilen zugeschnitten ist und einen neuartigen Kernel—PIKE—mit einem Gaußschen Prozess-Klassifikator kombiniert. Der Kernel ist konzipiert um die Eigenschaften von MALDI-TOF Massenprofilen auszunutzen und eine reduzierte Darstellung von MALDI-TOF Massenspektren als Eingabe zu verwenden. Vorhersagen durch GP–PIKE übertreffen die Vergleichsmethoden, logistische Regression und ein Gaußschen Prozess-Klassifikator mit einem RBF-Kernel, massgeblich. Außerdem wurde das Verhalten der maximalen Klassenwahrscheinlichkeit von GP–PIKE analysiert und auf seine Nützlichkeit als gut kalibrierter Schätzer der Glaubwürdigkeit von Vorhersagen hin bewertet. Die mit GP–PIKE erzielten Resultate leiden unter einer großen Standardabweichung zwischen den verschiedenen Trainings-Test-Splits. Wir vermuten, dass diese Varianz von einer zugrundeliegenden phylogenetischen Struktur herrührt, die in bisherigen Stratifizierungen nicht berücksichtigt wurde. Wir führen ein durch hierarchisches Clustering verbessertes Stratifikationsverfahren ein, dass darauf abzielt phylogenetische Verwandtschaft aus MALDI-TOF Massenspektren abzuleiten und eine ähnliche Verteilung der resultierenden Struktur zwischen Trainings und Testdaten zu erzwingen. Obwohl wir anhand der Resultate keine Verkleinerung der Standardabweichung beobachten können, deuten die Ergebnisse dennoch auf verbesserte Vorhersage durch unser hierarchisches Stratifikationsverfahren hin. Darüber hinaus bewerten wir die Übertragbarkeit der Vorhersagegüte von Klassifikatoren, die auf Daten einer medizinischen Einrichtung trainiert und evaluiert wurden, auf MALDI-TOF Massenprofile einer anderen Einrichtung. Unsere Ergebnisse deuten darauf hin, dass alle Modelle mit Proben aus dem gleichen Labor neu trainiert werden müssen; ein großer MALDI-TOF MS Datensatz, der an anderen Standorten gesammelt wurde, kann die Vorhersageleistung jedoch weiter verbessern. Die geringe Übertragbarkeit ist wahrscheinlich auf unterschiedliche Verteilungen innerhalb der Datensätze von verschiedenen Orten zurückzuführen. Wir stellen einen neuen Ansatz vor, der auf dem “feindlichem” (adversarial) Lernen neuer Datenrepräsentationen basiert, um diese Verschiebungen in der Verteilung zu mildern. Der vorgestellte Ansatz ist in der Lage eine Diskrepanz in der Datenverteilungen auszugleichen, hat sich jedoch nicht als fähig erwiesen die Vorhersageleistung am Zielort zu verbessern.

Am Ende der Doktorarbeit skizzieren wir die Schritte, die für die vollständige Entwicklung eines klinisch anwendbaren Klassifikators erforderlich sind, sowie das Potenzial der Vorhersage antimikrobieller Phänotypen auf der Grundlage anderer Arten von Daten.

Durch die Bereitstellung des ersten großangelegten, öffentlich zugänglichen Datensatzes für MALDI-TOF MS-basierte Vorhersage antimikrobieller Resistenzen in der Klinik, die Demonstration des Potenzials etablierter Machine-Learning-Modelle und die Entwicklung eines neuen, auf diesen Datentyp zugeschnittenen Kernels, stellt diese Arbeit einen wichtigen Schritt in Richtung MALDI-TOF MS-basierter Vorhersage von Antibiotikaresistenzen und im Allgemeinen der Nutzung digitaler Ansätze für Antimicrobial Stewardship, also den Umgang mit antimikrobiellen Mitteln, dar.

Acknowledgements

First and foremost, I would like to express my deep gratitude to my doctoral advisor Prof. Dr. Karsten Borgwardt for giving me the opportunity to pursue my doctoral studies in his group, and for his constant support, supervision, and guidance. I am particularly grateful to Karsten for the freedom he gave me in my research, his sage advice, and for being a reliable support system through these years. I truly appreciate the environment I could conduct my doctoral studies in. I would like to thank Prof. Dr. med. et Dr. phil. Adrian Egli and Prof. Dr. Jean-Philippe Vert very much for being part of my doctoral committee, and Prof. Dr. Barbara Treutlein for chairing the doctoral examination.

Next, I wish to express my sincerest gratitude all my co-authors, collaborators, students and discussion partners who contributed enormously to all the work described in this thesis. I am particularly thankful to Dr. Bastian Rieck for being a champion supervisor, helping me improve my programming and writing skills, and being the coding magician who could answer any question that googling could not solve. On top of that, I want to thank Bastian for providing the LaTeX template for this thesis. A heartfelt thank you goes to Dr. Catherine Jutzeler for being an expert mentor and researcher, lighting up every coffee break, and for encouraging me to climb mountains higher than I believed I could conquer—figuratively and literally—and cheering me on throughout the way. I want to thank my main co-authors throughout many projects: Aline Cuénod for patiently answering all my questions regarding the clinical microbiology practices, Max Horn for being a true help in all matters concerning code implementation and machine learning, and Dr. Felipe Llinares-López for answering questions and guiding me throughout the beginning stages of this thesis. I would also like to give extra thanks to Leslie O'Bray, Katharina Heinrich, Benjamin Hepp, Bastian and Catherine for proofreading this thesis.

I would like to thank the entire team at the University Hospital Basel-Stadt and the University of Basel, who were my direct collaborators throughout the majority of my doctoral studies. In addition to Adrian and Aline, I would like to give special thanks to PD Dr. med. Michael Osthoff for taking the time to give his perspective as a clinical specialist on infectious diseases and Dr. Helena Seth-Smith for her feedback on microbiology matters and proofreading manuscripts. I appreciate the support and their feedback on manuscripts of all my collaborators from Canton Hospital Basel-Land, Canton Hospital Aarau and Viollier. I would also like to extend my gratitude to the entire laboratory staff at these medical institutions who—even though seldom through direct interaction with me—provided the foundation for all the research in this thesis.

The time in my doctoral studies would have not been as great without the current and past members of the MLCB Lab: Dr. Michael Adamer, Christian Bock, Dr. Dean Boden-

ham, Lucie Bourguignon, Dr. Sarah Brüningk, Dr. Dexiong Chen, Bowen Fan, Dr. Lukas Folkman, Elisabetta Ghisu, Udo Gieraths, Dr. Dominik Grimm, Thomas Gumbsch, Anja Gumpinger, Dr. Xiao He, Dr. Katharina Heinrich, Dr. Felix Hensel, Max Horn, Dr. Catherine Jutzeler, Dr. Juliane Klatt, Tim Kucera, Dr. Felipe Llinares-López, Michael Moor, Giulia Muzio, Leslie O'Bray, Dr. Laetitia Papaxanthos, Dr. Bastian Rieck, Dr. Damián Roqueiro, Dr. Matteo Togninalli, and Dr. Daisuke Yoneoka. You were an absolute joy to work with, and I truly appreciate what I learned from each one of you. A special thank you goes to Katharina, a true generalist who is always able to provide feedback, be it mathematics, manuscript proofreading, organizational or administrative matters. I want to acknowledge all the people adjacent to MLCB who provided support and joy throughout my doctoral studies, including fellow PhD students at D-BSSE and friends from my time in the D-BSSE Master program together with whom I embarked on the PhD journey. Finally, I would like to thank Cindy Malnasi for all her help and administrative assistance throughout my doctorate, and all D-BSSE IT and Facility Management team members for helping with all the minor problems throughout the way.

Last but not least, I am very grateful for all my friends and family supporting me on this journey. In particular, I want to thank my mother who taught me to take pride in my interest in science and mathematics already as a young child and my brother Christoph who encouraged me to pursue a degree at ETH. And finally, I would like to thank Benjamin for being my rock and keeping my head level throughout these years.

Contents

I Defining the current state and obstacles of rapid antimicrobial resistance prediction using MALDI-TOF MS based machine learning techniques 1

1 Introduction to MALDI-TOF MS based antimicrobial resistance prediction 3
1.1 The thread of antimicrobial resistance to global health 3
1.2 MALDI-TOF mass spectrometry 4
1.3 MALDI-TOF MS based phenotype prediction 6
1.4 Contributions of this thesis 8

2 Systematic review of machine learning for microbial identification and antimicrobial resistance prediction on MALDI-TOF mass spectra 13
2.1 Preferred reporting items for systematic reviews and meta-analysis . . . 13
2.2 Summary of literature 16
2.3 Identified shortcomings and discussion 19

II Large-scale full-spectrum MALDI-TOF MS based clinical antimicrobial resistance prediction 25

3 *DRIAMS*: A database for MALDI-TOF MS based resistance prediction 27
3.1 Clinical routine data 29
3.2 Preprocessing and quality control 30
3.3 Confounding factors in real-world clinical data 32
3.4 Summary of the *DRIAMS* datasets 34

4 Large-scale species-specific antimicrobial resistance prediction 41
4.1 Species-stratified antimicrobial resistance prediction 41
4.2 Calibrated classifiers for interpretable prediction scores 56
4.3 Interpreting mass peak contributions through Shapley values 58
4.4 Retrospective clinical case study 61
4.5 Summary and discussion 63

III Improving the predictive performance and transferability of MALDI-TOF MS based resistance prediction through kernel methods and representation learning 67

5 Kernel-based microbial phenotype prediction from MALDI-TOF mass spectra 69
- 5.1 Topology-based peak detection 70
- 5.2 GP-PIKE: A kernel method designed for MALDI-TOF mass spectra 72
- 5.3 Evaluation of topological and kernel methods for MALDI-TOF mass-peak based resistance prediction 78
- 5.4 Confidence analysis of predictions with GP–PIKE 82
- 5.5 Summary and discussion 85

6 Mitigating the effect of phylogenetic variance on MALDI-TOF MS based phenotype prediction through hierarchical stratification 87
- 6.1 Hierarchical clustering 88
- 6.2 Data stratification based on inferred hierarchical structure 92
- 6.3 Summary and discussion 96

7 Learning domain independent MALDI-TOF mass spectrum representations 101
- 7.1 Leveraging large-scale multi-site data for improved local prediction . . . 102
- 7.2 Learning site-independent data representations 105
- 7.3 Summary and discussion 113

IV Outlook on the path to clinical MALDI-TOF MS based antimicrobial resistance prediction 115

8 Necessary steps to reach clinical applicability 117

V Appendix 123

Data availability 125

Software availability 127

List of Figures 129

List of Tables 135

Acronyms 139

Glossary 141

Bibliography 143

Part I

Defining the current state and obstacles of rapid antimicrobial resistance prediction using MALDI-TOF MS based machine learning techniques

1

Introduction to MALDI-TOF MS based antimicrobial resistance prediction

1.1 The thread of antimicrobial resistance to global health

Antimicrobial resistance (AMR) has been recognised as a major epidemiological thread for decades [131], and poses a continuously growing threat to public health [42]. Antimicrobial resistance is the resistance to antimicrobial drugs presented by infectious agents—e.g. , bacteria, viruses, fungi and parasites—which can be inherent or acquired by the inappropriate use of medicines [42]. For the last two decades, the World Health Organization (WHO) has lead the global response to the risk posed by AMR and declared it one of the top 10 global public health threats to humanity in 2019 [42]. In their 2021 report on 'Global antimicrobial resistance and use surveillance system', they identify two factors to be crucial indicators for the severity of the threat, namely (i) prevalence of bloodstream infections (bacteremia), and (ii) trends in antibiotic consumption.

In 2017, the WHO published a list of *priority pathogens* defining the 12 bacterial families which pose the greatest threat to human health [41]. The list of families receiving the rating *critical* includes the *Enterobacteriaceae* group, which includes both *Escherichia coli* (*E. coli*) and the *Klebsiella* genus. Infections with these strains cause severe (and often fatal) infections such as pneumonia and infections in the bloodstream. The second tier, rated *high*, includes *Staphylococcus aureus* (*S. aureus*), the leading cause for bacteremia and infective endocarditis (infections of the endocardial surface of the heart) among a number of other infections [116].

The terms *antibiotic* and *antifungal drug* refer to drugs treating infectious caused by either bacteria or fungi respectively, while the overarching term for both types of drugs is *antimicrobial drugs*.

Antimicrobial stewardship. Antimicrobial stewardship refers to an approach with the goal to quantify and improve the usage of antibiotics, both prescription by physicians and intake by patients. Improving these aspects is critical for effective treatment of infections, reducing harms resulting from unnecessary antimicrobial admission, and combat development of resistances.

1.2 MALDI-TOF mass spectrometry

In recent decades, Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) mass spectrometry (MS) has revolutionized clinical diagnostics by facilitating rapid, reliable and cost-effective microbial detection [23] [78] [71]. The technique produces MALDI-TOF mass spectra, which will be the main data type analysed in this thesis. In the following, a brief introduction into the technique and properties of the MALDI-TOF mass spectra is given.

Mass spectrometry for proteomics. Mass spectrometry is a technology that characterizes molecules by ionizing individual particles and determining their mass-to-charge ratio (m/z–ratio). The measurement output is a mass spectrum, depicting the ion signal as a function of the m/z–ratio and thereby providing information about the possible content of a probe. MALDI—the technique to obtain charged particles containing intact biomolecules—was developed to expand the applicability of mass spectrometry to large molecules such as biopolymers including proteins. The MALDI matrix solution keeps the large molecules mostly intact through desorption, with most ions only receiving a single charge, which makes the inference of the original protein relatively easy. The high throughput and speed associated with complete automation has made MALDI-TOF MS the preferred technique for large-scale proteomics [109].

MALDI-TOF MS. MALDI-TOF MS is an analytical technique which decomposes and ionises a biological sample into charged molecules with a laser and determines the m/z–ratio of the ions [89]. Before a MALDI-TOF mass spectrometry measurement can be performed, the microbial sample must be cultured for around 24 h overnight to enrich enough bacterial cell amount for measurement. For one spectrum, material from a single culture is then transferred to a MALDI-TOF target plate, where the matrix solution is added, allowing for the larger molecules to stay stable while the laser fragments and ionises the sample. In most cases, the lasers used for MALDI are within the ultraviolet (UV) range, e.g. nitrogen lasers, but infrared lasers are used also [89]. The laser beam irradiates the isolate on the target plate, causing desorption and ionisation of the particles which are then accelerated by an electric field generator. The intensity and m/z–ratio of molecules are measured by the time-of-flight analyser. Assuming equal charge, light molecules reach the detector earlier than heavy molecules [48].

Despite the fact that fragmentation is an inherently stochastic process, the output spectrum over the m/z–ratio of the particles is known to be highly characteristic for different microbial species. Each recording typically contains several ten thousand measurement points in a range of 2 kDa to 20 kDa. The results are summarised in a so-called MALDI-TOF mass spectrum or MALDI-TOF mass profile. An example MALDI-TOF mass spectrum is depicted in Figure 1.1. MALDI-TOF mass profiles provide an overview of the microbial composition and therefore lay a foundation for predicting bacterial phenotype, such as species or antimicrobial resistance properties [129]. The current application of MALDI-TOF MS lies in rapid species identification. The two main

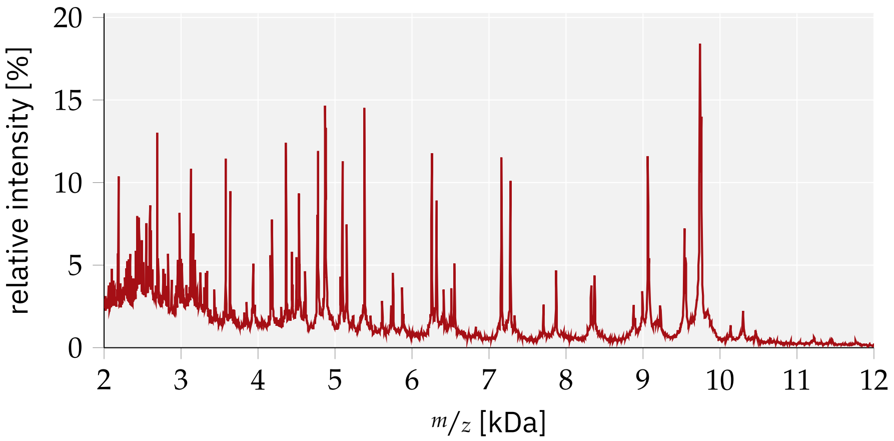


Figure 1.1: **Exemplary MALDI-TOF MS spectrum** depicting an *E.coli* sample. The spectrum is in its raw form as it was extracted from the MALDI-TOF MS instrument, i.e., no spectral preprocessing has been performed. The MALDI-TOF MS baseline signal is clearly observable up until 11.5 kDa. The mass spectrum was trimmed to 2 kDa to 12 kDa for illustration purposes.

MALDI-TOF MS instrument manufacturers, *Biomérieux* and *Bruker Daltonics* [9, 12], provide the customer with a full analysis pipeline, providing both the instrument as well as the software performing species identification. While the specific methodology and algorithms are kept private by the manufacturing companies, the species are determined through a similarity comparison between the spectrum of interest and a reference database employing statistical methods.

The fact that the x–axis depicts the m/z–ratio of a measured molecule leads to a noteworthy property of mass spectra: particles corresponding to a specific molecule can show up at several positions at the mass spectrum, depending on the number of charges. Although most particles will receive a single charge during ionisation, ions with *multiple* charges can occur and thus confound the signal in the resulting mass spectrum [65, 89]. For instances, a molecule with mass 6,000 Da receiving a double charge appears at the same m/z–ratio as a molecule weighting 3,000 Da with a single charge [65]. Considering this property of MALDI-TOF mass spectra during phenotype prediction has the potential to improve classification performance, which will be discussed in Section 5 for model development. The likelihood of multiple charges decreases with the number of charges received, and the most likely one being a single-charge. Therefore that a MALDI-TOF mass peak corresponds to a particle with exactly the mass of its m/z–ratio. In order to reduce noise and to accentuate peaks, a single output MALDI-TOF MS measurement was constructed through repeated measurements merged into one MALDI-TOF mass spectrum. Additionally, the output profiles all include a baseline signal produced by the matrix solution and slightly varying measured intensities which contributes to an increased noise level. The reduction of these noise signals will be addressed by preprocessing described in Chapter 3.

1.3 MALDI-TOF MS based phenotype prediction

Phenotype prediction from MALDI-TOF mass profiles is an active and expanding field of research. Despite MALDI-TOF MS being widely-employed for species identification, several research fields aim to exploit further signals displayed in MALDI-TOF mass spectra for a more fine-grained characterisation of microbial isolates. In the following, we briefly discuss a number of current research topics.

Conventionally, the analysis of MALDI-TOF mass spectra relies on a small number of attributes, such as peak height and area under the peak, that have been empirically linked to microbial species. While this is a valid approach and works fairly well at species level, there is a wealth of information contained in these spectra that remains unused. To fully exploit the information contained in MALDI-TOF mass spectra, researchers have been implementing machine learning algorithms in their efforts to refine species identification. This information has proven useful for identification and differentiation of species, particularly those that are phylogenetically proximal, as well as sublineages within species [13, 39, 69]. Moreover, it has been recently recognized that information contained in MALDI-TOF mass spectra can also aid antibiotic resistance profiling [13, 39]. However, while species prediction provides reliable identification applicable in the clinic, machine learning methods predicting antimicrobial resistance face several challenges preventing swift progress, such as few data sources and lack of proper validation. This thesis commences with a comprehensive assessment of the current state-of-the-art of MALDI-TOF MS based phenotype prediction in Chapter 2, focusing on both species and resistance prediction.

Subspecies characterisation. The furthest advanced application of phenotype prediction on MALDI-TOF mass profiles is the subspecies identification from bacterial specimen [20, 112, 129]. A number of microbial phylogenetic lineages are known to cause serious infections. For these lineages in particular, a fast and high-throughput identification method is needed. Quick, robust and cheap subspecies identification method are essential for infectious disease control. Applying and developing machine learning methods to MALDI-TOF mass spectra is generally cheaper than current subspecies identification methods such as multi-locus sequence typing (MLST) [20, 112]. Prior research in subspecies discrimination includes typing of *Mycoplasma pneumoniae* [133], discrimination between contagious and environmental strains of *Streptococcus uberis* [32] and strain typing of *Staphylococcus haemolyticus* [20]. Additionally, MALDI-TOF MS has been shown useful for rapid and cheap identification of clonal complexes, e.g., methicillin-resistant *Staphylococcus aureus* (MRSA), vancomycin-intermediately resistant *Staphylococcus aureus* (VISA) and heterogeneous VISA (hVISA) [15, 138]. Furthermore, research into single-cell MALDI-aerosol TOF MS has shown potential to reduce the initial 24 h culture step, which is necessary to acquire sufficient biomass for the MALDI-TOF MS analysis [81].

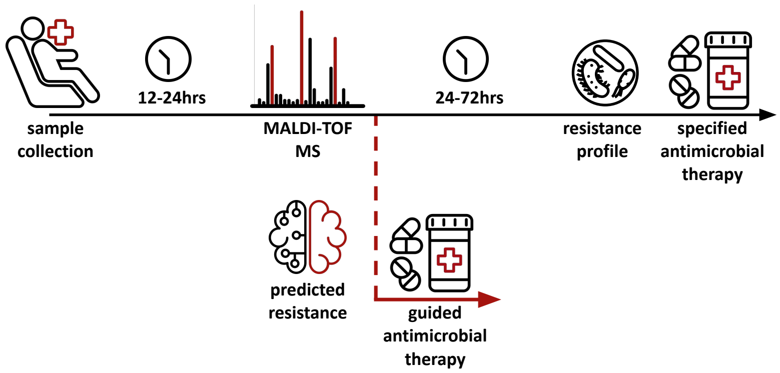


Figure 1.2: **The current usage of MALDI-TOF MS in the clinic and the potential inclusion of a resistance predictor.** The upper timeline (following the black arrow) depicts the antimicrobial species and resistance workflow as it is currently implemented in many clinics, including the culture growth waiting times between steps and the point of MALDI-TOF MS measurement. The lower part (following from black to dashed and red arrow) depicts the vision of a machine learning driven prediction approach running in parallel to the established diagnostics path. Antimicrobial therapy guided by a MALDI-TOF MS based predictor could lead to an informed treatment decision 24 h to 72 h earlier than in the current workflow focused solely on diagnostics. All icons are listed in the Noun Project [1].

Antimicrobial resistance prediction. In recent years, the field of MALDI-TOF MS based machine learning has shifted towards the prediction of antimicrobial resistance [129]. These resistance predictors hold the potential to reduce the time required to determine effective antimicrobial treatment by 24 h to 72 h and optimise the use of broad-spectrum antibiotics [126]. Antimicrobial resistance prediction based on MALDI-TOF mass spectra has been shown to be effective for several antimicrobial–species scenarios, including carbapenem resistance in *Klebsiella pneumoniae* [52], intermediate resistance to vancomycin in *Staphylococcus aureus* [120] and carbapenem resistance in *Bacteroides fragilis* [49]. A full systematic evaluation of the literature on MALDI-TOF MS based phenotype prediction is part of Chapter 2.

Extending the current clinical workflow through guided treatment. A depiction of the current antimicrobial susceptibility testing workflow can be found in Figure 1.2. The upper timeline illustrates the current clinical workflow, from collecting a microbial sample of an infected patient to the obtaining the resistance profile from antimicrobial resistance testing and making a treatment decision. The timeframes indicate the duration of each step, i.e. 12 h to 24 h needed for the culture phase before MALDI-TOF MS and 24 h to 72 h growth phase before the resistance is determined. The lower part depicts the vision driving this thesis—a MALDI-TOF MS based predictor guiding early antimicrobial treatment decisions. Such a predictor would be inserted into the pipeline at the time of obtaining the MALDI-TOF mass spectrum, providing a machine learning based treatment recommendation 24 h to 72 h before obtaining the phenotype testing based resistance profile. In a clinical setting, the rapid and reliable identification of potential pathogens is of utmost importance for a timely initiation of appropriate antimicrobial treatment. In this thesis, we envision a reliable predictor for antimicrobial resistance based on MALDI-TOF MS aimed at guiding the early treatment decisions. We hypothesise that such a tailored antimicrobial treatment would result in improved patient outcomes and a decrease of unnecessary use of broad-band antimicrobials.

1.4 Contributions of this thesis

The aim of this dissertation is to improve antimicrobial resistance prediction based on MALDI-TOF MS data, moving the field towards a clinically-applicable treatment guidance model. Previous studies have recognized the potential of advancing MALDI-TOF mass spectrum based phenotype prediction through machine learning, specifically for antimicrobial resistance prediction.

I Defining the current state and obstacles of rapid antimicrobial resistance prediction using MALDI-TOF MS based machine learning techniques

In the second chapter of Part I, as a first step, we establish the current state-of-the-art with a systematic literature review, which is based on the author's systematic review

> C. Weis, C. R. Jutzeler, and K. Borgwardt. "Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review".
> *Clinical Microbiology and Infection* 26:10, 2020, pp. 1310–1317.
> doi: `https://doi.org/10.1016/j.cmi.2020.03.014`

This work is the first study accumulating all publications on MALDI-TOF MS based phenotype prediction and providing a structured assessment of the factors limiting the field. The comprehensive assessment of the current literature identifies several shortcomings, specifically (i) the lack of studies analysing large datasets, (ii) only few efforts exist to develop machine learning models tailored for the datatype of MALDI-TOF mass spectra, and (iii) a large number of studies omit validating their findings on an external dataset. This chapter provides the foundation for the entire thesis, with the remaining parts addressing different aspects of the identified shortcomings.

II Large-scale full-spectrum MALDI-TOF MS based clinical antimicrobial resistance prediction

Starting off the second part of this thesis, Chapter 3 introduces *DRIAMS*, the dataset curated by the author, which marks the largest database collected for MALDI-TOF MS based phenotype prediction to date. Further we describe the curation and properties of this database, made publicly available, to facilitate further research and progress in the field (V Software Availability). Chapter 4 applies established machine learning models to resistance prediction based on *DRIAMS*, setting up a predictive performance baseline and demonstrating the benefits of the large dataset. The models are comprised of logistic regression, LightGBM and multi-layer perceptrons (MLP), which could reach predictive performance baselines of 0.74 AUROC for ceftriaxone resistance in *E. coli* (LightGBM), 0.74 AUROC for ceftriaxone resistance in *K. pneumoniae* (MLP) and 0.80 AUROC for oxacillin resistance in *S. aureus* (LightGBM). In addition to building a well-established machine learning pipeline, we (i) build a species-stratified single-antimicrobial prediction scenario as a simple but effective machine learning set-up, (ii) apply probability calibration using Platt scores to obtain outputs that can be readily interpreted by clinical staff and used for sample rejection, (iii) conduct a biological literature validation by calculating Shapley values and interpreting the highest-contributing features with resistance-associated peaks described in the MALDI-TOF MS literature, and (iv) assess the potential impact a antimicrobial resistance predictor on MALDI-TOF mass spectra through a retrospective clinical case study on 63 patients. Both Chapter 3 and Chapter 4 are based on the study published as a preprint

C. Weis, A. Cuénod, B. Rieck, F. Llinares-López, O. Dubuis, S. Graf, C. Lang, M. Oberle, K. K. Soegaard, M. Osthoff, M. Brackmann, K. Borgwardt, and A. Egli. "Direct Antimicrobial Resistance Prediction from clinical MALDI-TOF mass spectra using Machine Learning". *accepted in Nature Medicine*, 2021. doi: `https://doi.org/10.1101/2020.07.30.228411`

III Improving the predictive performance and transferability of MALDI-TOF MS based resistance prediction through kernel methods and representation learning

In the next and final part, this thesis introduces several machine learning concepts tailored specifically to the data type of MALDI-TOF mass spectra. The two chapters 5 and 6 employ a reduced but more accurate representation of the MALDI-TOF mass spectra for prediction, namely only a set of signal peaks determined for each spectrum. Chapter 5 introduces the first kernel specifically tailored to MALDI-TOF mass spectra, the Peak Information Kernel (PIKE), and is based on the author's publication

C. Weis, M. Horn, B. Rieck, A. Cuénod, A. Egli, and K. Borgwardt. "Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra". *OUP Bioinformatics* 36, 2020, pp. i30–i38.
doi: `https://doi.org/10.1093/bioinformatics/btaa429`

This kernel demonstrates superior prediction performance in comparison to logistic regression and the established RBF kernel. Further, we demonstrate its property to provide class probabilities that work as reliable and easily interpretable confidence estimates, which are imperative to clinical applications of machine learning models. This work is the first study to address reliability estimation in MALDI-TOF based phenotype prediction.

Chapter 6 explores whether prior knowledge specific to MALDI-TOF mass spectra training datasets—namely that the mass spectra depict microbial samples that are part of strains related to each other—can be utilised to create improved stratification splits for training. This analysis was motivated by the large standard deviation observed between the data splits in the GP–PIKE experiments. To this end, hierarchical clustering is employed to facilitate an informed train–test split and its benefit to the classification performance is assessed. This chapter is based on unpublished study

C. Weis, B. Rieck, S. Balzer, A. Cuénod, A. Egli, and K. Borgwardt. "Improved MALDI-TOF MS based antimicrobial resistance prediction through hierarchical stratification". *Unpublished*, 2020

While the results indicate that the approach does not decrease the standard deviation between data splits, they report higher classification performance through the introduced hierarchical stratification procedure, e.g. improving piperacillin-tazobactam resistance prediction in *K. pneumoniae* from 0.36 to 0.41 AUPRC.

Chapter 7 is dedicated to the transferability of antimicrobial resistance prediction from one site to another. In a first step, the transferability between all sites in *DRIAMS* is assessed. The results suggest that the prediction models require regular retraining on spectra native to the prediction site. Further we show that large MALDI-TOF MS datasets from other medical institutions can increase the predictive performance. We conjecture that the low transferability is most likely caused by distribution shifts between different datasets. A new approach is introduced that leverages complex machine learning methods inspired by adversarial deep learning frameworks to mitigate domain shifts between sites. We explore an adversarial learning framework to learn site independent representations in Chapter 7. The assessment of cross-site transferability is again based on the aforementioned preprint [126], while the adversarial learning framework in this chapter is based on the unpublished study

C. Weis, M. Horn, B. Rieck, A. Cuénod, A. Egli, and K. Borgwardt. "Domain adaptation for transferable antimicrobial resistance prediction from MALDI-TOF mass spectra". *Unpublished*, 2021

We conclude the thesis with an summary and outlook in Chapter 8. Here we sketch a roadmap summarising all tasks that need to be tackled in order to develop an antimicrobial resistance predictor fit for clinical deployment, formed on the collected experience from all studies performed in this thesis. We propose several research directions with potential to improve antimicrobial resistance prediction from MALDI-TOF mass spectra. Further, we briefly discuss another type of antimicrobial resistance prediction rapidly developing in recent years: applying machine learning models to genomic data from bacteria to infer phenotypes such as resistance properties.

Specific contributions to each publication. With regard to Weis *et al.* [126], C.W. designed and implemented all machine learning experiments in collaboration with co-author B.R., contributed to the implementation of the *DRIAMS-A* and *DRIAMS-B* curation and preprocessing pipeline together with co-author A.C., solely performed the implementation of the curation and preprocessing pipeline of *DRIAMS-C* and *DRIAMS-D*, designed twelve out of sixteen display items, and wrote major parts of the manuscript.

In Weis *et al.* [129], C.W. contributed to the data acquisition, data analysis, quality analysis, interpretation and writing the manuscript. All contributions are in equal parts to the co-author C.J..

Both in Weis *et al.* [128] and Weis *et al.* [127], C.W. contributed to the data preprocessing. C.W. contributed to the design and implementation of all machine learning experiments and writing the manuscript in equal proportion to co-authors B.R. and M.H..

C.W. designed and contributed to the implementation of all machine learning experiments in Weis *et al.* [130]. Further, C.W. contributed to the data preprocessing and drafting the manuscript.

2

Systematic review of machine learning for microbial identification and antimicrobial resistance prediction on MALDI-TOF mass spectra

C. Weis, C. R. Jutzeler, and K. Borgwardt. “Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review”. *Clinical Microbiology and Infection* 26:10, 2020, pp. 1310–1317.
doi: `https://doi.org/10.1016/j.cmi.2020.03.014`

Before diving into the novel research work and contributions of this thesis, we establish the current state-of-the-art, compile the current literature, and identify any shortcomings in the field. To ensure completeness and in order to adhere to the current standards of literature reviews, we conduct a *systematic review* with the aim to analyse and evaluate all studies that employ machine learning for phenotype prediction based on MALDI-TOF mass spectra. Each study is assessed with respect to two objectives: (i) compile information on each study regarding the species investigated, machine learning algorithms employed and model performance, and (ii) assess the reproducibility, robustness, generalisability and clinical significance of the presented machine learning models.

2.1 Preferred reporting items for systematic reviews and meta-analysis

This study was conducted in accordance to the state-of-the-art recommendations for systematic literature reviews. To that effect, we follow the guidelines provided by the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement [69] and registered the study in the International prospective register of systematic reviews (PROSPERO) at `CRD42020165579`.

Search methods for identification of studies. A thorough search of original research articles was performed using the scholarly search platforms PubMed/Medline, Scopus and Web of Science. The search string was constructed to include publications analysing MALDI-TOF mass spectra using machine learning:

```
(‘machine learning’ OR ‘classification algorithm’ OR
   ‘support vector machine’ OR ‘random forest’ OR
   ‘logistic regression’ OR ‘neural network’) AND
                        ‘maldi-tof’
```

Additionally, a manual search was performed by reviewing references and review articles. All studies published in the time range of the platforms respective inception dates to the 31st of January 2020 were included.

Selection of studies. The author of this thesis carried out the initial screening of retrieved articles and applied the inclusion and exclusion criteria (as listed in the next paragraph). Then a co-author on the study independently reviewed all studies to satisfied the inclusion criteria. In case of disagreements, a consensus decision was made through a common discussion.

Inclusion and exclusion criteria. To be included in the review, a study is required to meet the following criteria: (i) presentation of an original research article, (ii) application of machine learning methods to MALDI-TOF mass spectra for microbial species and antimicrobial susceptibility identification, (iii) provide information on the machine learning algorithms, and (iv) provide information on the studied species. We exclude (i) studies that do not analyse antimicrobial phenotypes, e.g. MALDI-TOF mass spectra analysis of single proteins and peptides, cancer or genomics, (ii) paediatric studies, (iii) case studies, and (iv) review articles.

Data extraction and synthesis. We extract the following information from all studies (main text and supplemental material if available): (a) publication characteristics (first author’s last name, publication time), (b) study objectives (species discrimination, identification or antimicrobial susceptibility testing), (c) cohort selection (genera, sample size), (d) technical information on MALDI-TOF instrument used, (e) model selection (algorithm, platforms, software and software and packages, model parameters), (f) reported model performance (the metrics reported for model evaluation, mean and measure of variance), and (g) regularization methods ensuring generalization and external validation strategies.

Quality assessment of machine learning studies. We choose a list of criteria to judge the quality of the included studies, based on a previous review [90] and the relevance to our study objective: (a) unmet needs addressed in study, (b) reproducibility (feature engineering, hyperparameters, software and hardware), (c) robustness (valid methods to address model overfitting, stability of results), (d) generalisability (validation on external data), and (e) clinical significance (interpretation of predictors, suggested clinical use). These criteria were assessed for presence or absence (‘yes’ or ‘no’) and summarised in Table 2.3.

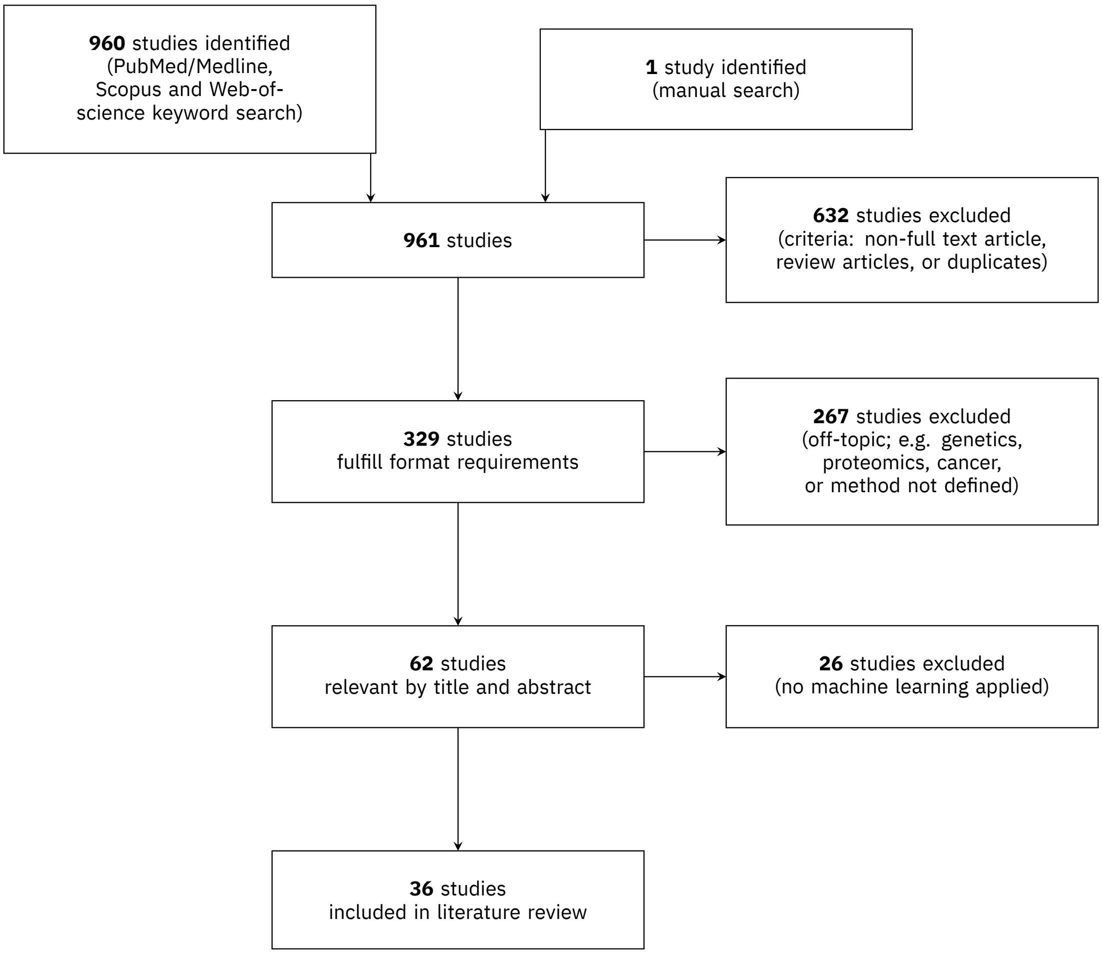

960 studies identified
(PubMed/Medline,
Scopus and Web-of-science keyword search)
1 study identified
(manual search)
961 studies
632 studies excluded
(criteria: non-full text article,
review articles, or duplicates)
329 studies
fulfill format requirements
267 studies excluded
(off-topic; e.g. genetics,
proteomics, cancer,
or method not defined)
62 studies
relevant by title and abstract
26 studies excluded
(no machine learning applied)
36 studies
included in literature review

Figure 2.1: **PRISMA flowchart** illustrating the complete literature search pipeline. In total, 960 studies were found using the search phrase and one additional publication was identified by manual search of references. 632 studies are excluded based on their article type or being duplicates in the search results. We exclude 267 studies as their research topic was not among the topic of antimicrobial phenotype prediction. Reading through the articles revealed that 26 studies do not apply machine learning methods. The remaining 36 studies are included in the review. Figure adapted from Weis *et al.* [129].

machine learning algorithm	abbreviation	n	%
support vector machine	SVM	18	50.0
genetic algorithm	GA	15	41.7
artificial / supervised neural network	ANN / SNN	13	36.1
quick classifier	QC	11	30.5
random forest	RF	9	5.0
clustering / hierachical cluster analysis	HC	8	22.2
k-nearest neighbors	kNN	5	13.9
decision tree	DT	4	11.1
logistic regression	LR	3	8.3
Aristotle classifier	-	1	2.8
linear discrimant analysis	LDA	1	2.8
naïve Bayes	-	1	2.8

Table 2.1: **Machine learning algorithms applied in the 36 reviewed studies.** Multiple algorithms can be applied in a single study.

Results. The full literature search—including screening studies, eligibility assessment and articles reviewed—is depicted in Figure 2.1. Out of 36 published studies chosen for assessment, 27 analysed microbial species identification [16, 17, 21, 25, 27, 33, 35, 51, 56, 59, 61, 62, 63, 74, 75, 85, 96, 100, 102, 103, 113, 121, 122, 123, 134, 139, 140] and nine analysed antibiotic resistance prediction [3, 4, 26, 50, 53, 67, 110, 115, 120]. A total of 924 studies are excluded, as they do not meet all inclusion criteria.

2.2 Summary of literature

Tables 2.4 and 2.5 give an overview over all the 27 studies analysing microbial species identification and nine studies investigating antibiotic resistance prediction, respectively.

Bacterial species and antimicrobial drugs. The bacterial genera primarily investigated are *Staphylococcus* ($n = 14$), *Streptococcus* ($n = 6$), *Escherichia* ($n = 4$), and *Klebsiella* ($n = 3$). Among the papers focusing on antimicrobial resistance prediction, vancomycin ($n = 3$) and carbapenems ($n = 1$) are the most widely-used broad-spectrum antibiotics studied. Further, the narrow-spectrum antibiotic methicillin ($n = 3$) and the antifungal drug fluconazole ($n = 1$) were also studied in the literature. Additionally, we observe a high variance in the number of samples included in each study, ranging from less than 50 [33, 63, 96, 115] to 787 isolates [122].

Machine learning models. Table 2.1 provides an overview of the wide range of machine learning algorithms that were used in the literature, with the most commonly applied model being support vector machines (SVM, $n = 18$), genetic algorithms (GA, $n = 15$), artificial/supervised neural networks (ANN, $n = 13$) and quick classifiers (QC,

n = 11). As these algorithms will not be revisited later on in this thesis, a brief description is given in Table 2.2. We observe a substantial overlap between studies in terms of applied machine learning algorithms, namely that the classifiers GA, SVM, SNN, and QC are applied most frequently, and are frequently applied together in the same study. This trend is caused by most analyses being performed on manufacturer-provided software, such as `flexAnalysis` and `ClinProTools` from Bruker Daltonics, employing preprogrammed machine learning algorithms. `ClinProTools` provides the four most frequently employed algorithms, namely GA, SVM, SNN, and QC [17], causing the described behaviour. Other algorithms applied in the literature are clustering/hierarchical cluster analysis (UHCA), random forests (RF), decision trees (DT), k-nearest neighbors (kNN), multiple logistic regression (MLR), naïve Bayes and Aristotle classifiers.

Software. The most frequently employed software to perform mass spectra analysis was the `ClinProTools` Software by the MALDI-TOF MS instrument manufacturer Brucker Daltonics (n = 17). Notice that when selecting the 'SVM' or 'genetic algorithm' option in `ClinProTools`, the software only employs the respective algorithms for peak selection, while classification itself is performed using a kNN algorithm based on the selected peaks. The remaining studies employed `R` or `R Studio` (n = 9), `MATLAB` (n = 7), `Python` (n = 1), `MALDI Biotools 3.0` (n = 1), `Statistics Program for Social Sciences` (n = 1), `Mathematica` (n = 1) or a combination thereof to perform their analyses.

Model generalisation and external validation. All studies reviewed applied one form of cross-validation to avoid their models overfitting to the training data; either 5-fold, 10-fold or leave-one-out. Validation on out-of-distribution data usually requires external MALDI-TOF data and was only performed in four studies: (1) Wang *et al.* [122] collected their main dataset from the biobank of a teaching hospital in northern Taiwan. They obtained an independent external validation dataset from a bacterial biobank and two teaching hospitals in middle and southern Taiwan. (2) The study by Esener *et al.* [33] aims at discriminating contagious from environmental strains of *Streptococcus uberis* in dairy herds. Data was collected from 29 farms; the data of 19 farms was used for the main analysis and data from the remaining ten farms were held-out for external validation. (3) Fangous *et al.* [35] collected 40 *Mycobacterium abscessus* isolates across France for the main part of the analysis. The subsequent external validation was conducted on another 40 *M. abscessus* isolates, obtained from the French National Reference Centre for Mycobacteria and Resistance of Mycobacteria to Antituberculosis. (4) Rodrigues *et al.* [96] aimed at precisely identifying different species within the *Klebsiella* group, basing their analysis on 46 strains collected from different sources around the globe (e.g. human, environment, water, plant). For validation, 49 isolates belonging to *K. pneumoniae* phylogroups derived from 49 faecal samples of humans in Madagascar were analysed.

machine learning algorithm	description
genetic algorithm (GA)	Genetic algorithms aim to select a combination of peaks that separate the classes in a way that maximizes the variance between classes. The steps in the algorithm are inspired by biological processes, such as mutations, crossover and selection, by evolving a collection of candidate solutions towards a better performing solution.
artificial neural network (ANN)	Artificial neural networks were developed drawing inspiration from the structure of mammalian brain neural networks. The network consists of several stacked layers of relatively simple mathematical units, which combine the input information of the previous layer's neurons and direct the output to neurons in the next layer. Models labelled 'supervised neural networks' or 'back propagation neural networks' are instances of ANNs.
support vector machine (SVM)	Support vector machines are a supervised learning algorithm that finds the best separating maximum margin hyperplane between the classes in a higher dimensional representation of the instances. During optimization the hyperplane maximizing the gap between the plane and the instances is determined. The data is mapped into a higher-dimensional space using a kernel function, e.g. the radial basis function kernel or the polynomial kernel.
quick classifier (QC)	The Quick Classifier calculates the average area of each peak together and provides a p-value per class. During classification, the peak areas are sorted by the univariate sorting algorithm and an average over all peaks is calculated which indicates class membership.
k-nearest neighbor (kNN)	The k-nearest neighbor classification algorithm bases its classification of unseen instances on the similarity between the instance and each training data point. The assigned class is chosen to be the majority class of the closest k training data point classes. A frequently used similarity measure is the Euclidean distance.

Table 2.2: **Description of the most frequent machine learning algorithms** employed in the current literature [129].

2.3 Identified shortcomings and discussion

In this section, we conduct a systematic quality assessment of the machine learning analyses we find in the literature and discuss several shortcomings, while aiming for the development of a clinically applicable phenotype prediction tool to be incorporated in the routine diagnostics.

Quality of reviewed studies. The results obtained through the quality assessment are included in Table 2.3. The resulting quality scores range from poor (<60%) to very good (100%). Only one study fulfilled all quality criteria and obtained a quality score of 100%. Four quality requirements were met by more than 97% of the studies (35 out of 36), namely (i) highlighting the limits in current non-machine learning approaches in the introduction (ii) providing information on hardware and software used in the study, (iii) employing valid methods to avoid model overfitting (i.e. cross-validation), and (iv) providing information on the clinical relevance. However, only 11% of the studies validated their machine learning models on an external dataset.

External model evaluation. Robustness, reliability and validity are critical evaluation factors when developing a machine learning framework intended for clinical application. The quality assessment (Tables 2.4 and 2.5) reveals that the majority of the reviewed studies assess the robustness of their models using k-fold cross-validation and report standard deviation, confidence intervals, or other stability metrics. However, 39% of the reviewed studies do not provide the hyperparameters obtained through cross-validation, impeding replication and comparison with other studies.

Generalization of models' results is essential to allow for comparisons and application sharing between hospitals. MALDI-TOF mass spectra measured on different MALDI-TOF MS instruments and at different locations are known to suffer from batch effects [79, 126], likely stemming from differences in laboratory routine or machine settings. As a result, models trained on MALDI-TOF mass spectra collected at one hospital display decreased predictive performance on out-of-hospital (i.e. out-of-distribution) data. This topic will be discussed in more detail in Chapter 7. The generalization capabilities of a model are best assessed through so-called external validation in which the model is presented with unseen, out-of-distribution data. With only 11% of the reviewed studies including an external validation, we want to highlight this aspect as a massive shortcoming in assessing the validity of reported predictive performances and the potential of MALDI-TOF MS based phenotype prediction in general. We speculate that the cost and large-scale effort required to curate datasets from multiple sites poses too large a challenge and proves infeasible for most studies.

Interpretability of predictors. Being able to understand the decision-making process and functionality of a machine learning predictor is crucial to building tools that are meant to be applied in clinical care. Confidence in the model is built by identifying the MALDI-TOF MS peaks that contribute most of the information to the prediction. Practically all reviewed studies provide an investigation into the predictive peaks. For instances, Esener *et al.* [33] cross-references the mass of identified proteins with the NCBI protein database to identify corresponding proteins. This analysis uncovered that peak predictors in their study correspond to bacteriocins and ribosomal proteins. Another study [110] compared predictive peaks reported by their model with known fragments of the methicillin resistance-causing penicillin-binding protein (PBP) in methicillin-resistant *Staphylococcus aureus* (MRSA). Generally however, biological interpretation of feature peaks is limited by the lack of prior knowledge as most resistance mechanisms have not been previously analysed through MALDI-TOF MS.

MALDI-TOF MS tailored machine learning. The development of machine learning models specifically designed for input of the datatype MALDI-TOF mass spectra is still very much in its beginning stages, demonstrated by the lack of any MALDI-TOF MS tailored algorithms listed in 2.4 and 2.5. The systematic literature review reveals three limitations that prevent the advancement of machine learning techniques: (i) small sample size, (ii) lack of external validation, and (iii) poor reproducibility. The sample sizes employed in the reviewed studies were noticeably low for a machine learning application, with numbers ranging between dozens to hundreds of isolates, which then had to be further split for training and testing. Even the largest study reviewed, including 787 isolates, can hardly reflect the microbial diversity that is present in a population and would occur in the clinical routine. Small samples sizes, which cannot cover and represent the entire data distribution, lead to machine learning models with low generalisability and large false discovery rates [14]. The severity of this problem is further increased by the second limitation, as external validation is crucial to reliably judge the generalisability of a trained model. With the lack of external validation in many reviewed studies, it is difficult to assess whether the reported predictive performances and significance of specific MALDI-TOF mass peaks can be reproduced for new data. These two limitations could be addressed through combining available MALDI-TOF MS datasets and making them publicly available for other researchers, to be used as external validation datasets. Many of the currently available public MALDI-TOF datasets can be found in the ‘MassIVE’ repository [84], a community database for mass spectrometry data mainly focusing on proteomics. Alternatively, a large MALDI-TOF dataset is provided by the Robert Koch Institute, including 6264 MALDI-TOF mass spectra of highly pathogenic microbes [60]. However, (i) both data sources do not provide the AMR profiles corresponding to the MALDI-TOF mass spectra, and (ii) many studies do not make their datasets public after publication. In addition to posing a drawback to increase the number of publicly available datasets, withholding the dataset renders any attempt to reproduce the study results impossible. The systematic review reveals that only nine studies [3, 21, 26, 27, 53, 100, 113, 120, 122] provided the MALDI-TOF MS (and AMR information if applicable) and only four studies [3, 26, 27, 100] made the machine learning code available after publication. Additional information facilitating reproducibility is contained in the model hyperparameters selected during training. We find that most studies also do not provide details on the optimized hyperparameters (Table 2.3).

Exacerbating the problem, most analyses are performed on software provided by the manufacturer, such as `flexAnalysis` and `ClinProTools` from Bruker Daltonics [16, 17, 33, 35, 50, 51, 56, 74, 75, 84, 110, 113, 122, 123, 134, 139]. Therefore the employed analysis workflows are not accessible, obstructing any external attempts to review and potentially improve the state-of-the-art pipelines. In conclusion, to unleash and assess the full potential of MALDI-TOF MS based phenotpye prediction, a joint effort between clinicians and machine learning reseachers is needed, where code and clinical data are shared and made publicly available to foster study reproducibility and foster the development of new and advanced machine learning algorithms.

Limitations of this systematic review. To conclude the systematic review, we address the limitations hampering the review process. The literature search was restricted to articles listed in the scholar databases PubMed/Medline, Scopus and Web of Science (complemented by a publication discovered through reference search). Considering that none of these platforms have a strong focus on machine learning publications and the pace of research in the field, it is likely that the list of reviewed papers has grown significantly since conducting the systematic review. The literature search also excluded preprints, reports and conference proceedings, which commonly precede journal publication in the research field of machine learning. The lack of any studies reporting poor or mediocre predictive performance likely indicates a strong publication bias, caused by preferential submission and publication of positive research results.

In summary, the application of machine learning to microbial phenotype prediction is still in its infancy. The main limitations to swift progress are (i) lack of large datasets, (ii) lack of external validation and information on generalization, and (iii) poor reproducibility and code sharing, which drives the development of new task-specific machine learning algorithms.

	Limits in current non-machine learning approaches	Feature engineering	Hardware and software	Hyper-parameters	Valid methods for overfitting	Stability of results	External data validation	Interpretation of predictors	Suggested clinical use	Total
[3]	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	78.0%
[4]	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	78.0%
[16]	Yes	No	Yes	No	Yes	No	No	Yes	Yes	56.0%
[17]	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	78.0%
[21]	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	78.0%
[25]	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	67.0%
[26]	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	89.0%
[27]	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	89.0%
[33]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	100.0%
[35]	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	89.0%
[50]	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	78.0%
[51]	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	67.0%
[53]	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	78.0%
[56]	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	67.0%
[59]	Yes	No	Yes	No	Yes	No	No	Yes	Yes	56.0%
[61]	Yes	Yes	Yes	No	Yes	No	No	No	No	44.0%
[62]	Yes	No	Yes	No	Yes	Yes	No	Yes	Yes	67.0%
[63]	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	78.0%
[67]	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	78.0%
[74]	Yes	No	Yes	No	Yes	No	No	Yes	Yes	56.0%
[75]	Yes	No	Yes	No	Yes	Yes	No	Yes	Yes	67.0%
[85]	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	67.0%
[96]	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	78.0%
[100]	Yes	No	Yes	No	Yes	No	No	Yes	Yes	56.0%
[103]	Yes	No	Yes	No	Yes	No	No	Yes	Yes	56.0%
[102]	Yes	No	Yes	No	Yes	No	No	Yes	Yes	56.0%
[110]	Yes	No	Yes	No	Yes	No	No	Yes	Yes	56.0%
[113]	Yes	No	Yes	No	Yes	Yes	No	Yes	Yes	67.0%
[115]	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	89.0%
[123]	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	78.0%
[120]	Yes	No	Yes	No	Yes	Yes	No	Yes	Yes	67.0%
[121]	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	78.0%
[122]	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	89.0%
[134]	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	78.0%
[140]	No	No	Yes	Yes	Yes	No	No	No	Yes	44.0%
[139]	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	78.0%
Grand total	97.00%	39.00%	100%	53.00%	100%	56.00%	11.00%	89.00%	97%	

Table 2.3: **Quality assessment of reviewed studies** All studies analysed in the systematic review are assessed with respect to nine quality criteria, which are described in more detail in Section 2.1.

	Species of interest	Number of isolates	Availability code / data	Machine learning algorithms	Analysis platforms	Generalization
[16]	Staphylococcus aureus	95	no / no	GA, QC, SNN	`ClinProTools`	cross-validation
[17]	Streptococcus	75	no / no	GA, QC, SNN, SVM	`ClinProTools`	cross-validation
[21]	Staphylococcus haemolyticus	254	no / ask	DT, MLR, RF, SVM	`R`	5-fold cross-validation
[25]	Fructobacillus, Lactococcus, Leuconostoc	59	ask / no	HC, kNN, RF, SVM	`R`	10-fold cross-validation
[27]	20 unspecified bacterial strains	571	yes / yes	Aristotle Classifier	`RStudio`	5-fold cross-validation
[33]	Streptococcus uberis	19	no / no	GA, QC, SNN	`ClinProTools`	10-fold cross-validation
[35]	Mycobacterium abscessus	92	no / no	DT	`ClinProTools`	external validation
[51]	Staphylococcus, Streptococcus, Escherichia coli, Klebsiella pneumoniae, Salmonella, Pseudomonas	132	no / no	GA, HC, QC, SVM	`ClinProTools`	cross-validation
[56]	Escherichia coli and Shigella	138	ask / no	GA	`ClinProTools`, `MATLAB`	cross-validation
[59]	Bacillus	423	no / no	ANN, HC	`MATLAB`	5-fold cross-validation
[61]	Mycobacterium abscessus, Mycobacterium fortuitum	594	no/ no	SVM	IDSys LT (ASTA)	cross-validation
[62]	Extended spectrum beta-lactamases-producing gram-negative bacteria	69	no / no	GA, QC, SNN	`ClinProTools`	cross-validation
[63]	Escherichia coli and Shigella	23	no / no	BPNN	`MATLAB`	5-fold cross-validation
[74]	Streptococcus pneumoniae	574	no / no	GA, QC, SNN	`ClinProTools`	cross-validation
[75]	Enterococcus faecium	132	no / no	GA, HC, QC, SNN	`ClinProTools`	cross-validation
[85]	Staphylococcus aureus	52	no / no	GA, QC, SNN	`ClinProTools`	cross-validation
[96]	*Klebsiella*	46	no / no	SVM	`FlexAnalysis`	cross-validation
[100]	Escherichia coli, Streptococcus pyogenes, Bacillus anthracis, Streptococcus pneumoniae	540	yes / yes	RF	Microsoft `VBA`	bootstrapping
[103]	Enterobacteriaceae and Pseudomonas aeruginosa	76	no / no	HC, SVM	`MATLAB`	10-fold cross-validation
[102]	Clostridium botulinum and Clostridium sporogenes	123	no / no	HC, SVM	`MATLAB`	10-fold cross-validation
[113]	Leptospira	101	no / ask	GA, QC, SNN	`ClinProTools`	cross-validation
[121]	Methicillin-resistant Staphylococcus aureus	125	no / no	DT, kNN, SVM	`MATLAB`, Python, R, SPSS	nested cross-validation
[123]	Staphylococcus aureus	306	no / no	GA-kNN, QC, SNN	`ClinProTools`	5-fold cross-validation
[122]	B Streptococcus serotypes	787	no / ask	RF, SVM, UHCA	`ClinProTools`	5-fold cross-validation
[134]	Mycoplasma pneumoniae	68	no / no	GA	`ClinProTools`	cross-validation
[140]	Bacillus and Staphylococcus	69	no / no	RBF-ANNs	`MATLAB`	l.o.o. cross-validation
[139]	Staphylococcus aureus	290	no / no	Clustering, GA	`ClinProTools`	cross-validation

Table 2.4: **Summary of reviewed studies addressing species prediction.** Abbreviations: Algorithm abbreviations follow Table 2.1. 'ask' indicates that the study authors claim to give out code or data upon request. l.o.o. for 'leave-one-out'. The depicted columns address the limitations discussed in Section 2.3. A comprehensive version of this table can be found in Weis et al. [129].

	Species of interest	Number of isolates	Availability code / data	Machine learning algorithms	Analysis platforms	Generalization
[3]	Staphylococcus aureus	171	yes / yes	RF	Brucker (varSelRF), R	l.o.o. cross-validation
[4]	Staphylococcus aureus	727	no / no	GA, SVM	not reported	cross-validation
[26]	Candida albicans	33	yes / yes	LDA, LR, RF	R	cross-validation
[50]	Bacteroides fragilis	424	no / no	GA, QC, SNN, SVM	`ClinProTools`	cross-validation
[53]	Klebsiella pneumoniae	95	no/ yes	kNN, LR, naïve Bayes, RF, SVM	Mathematica	l.o.o. cross-validation
[67]	Staphylococcus aureus	80	no / no	SVM	R	10-fold cross-validation
[110]	Staphylococcus aureus	160	no / no	SVM	Biotools 3.0	10-fold cross-validation
[115]	Staphylococcus aureus	20	no / no	RF, SVM-RFE	R	l.o.o. cross-validation
[120]	Staphylococcus aureus	125	no / ask	kNN-DT, RF, SVM	R	5-fold cross-validation

Table 2.5: **Summary of reviewed studies addressing antimicrobial resistance prediction.** Abbreviations: Algorithm abbreviations follow Table 2.1. 'ask' indicates that the study authors claim to give out code or data upon request. l.o.o. for 'leave-one-out'. The depicted columns address the limitations discussed in Section 2.3. A comprehensive version of this table can be found in Weis et al. [129].

Part II

Large-scale full-spectrum MALDI-TOF MS based clinical antimicrobial resistance prediction

3

DRIAMS: Database of ResIstance against Antimicrobials with MALDI-TOF Mass Spectrometry

C. Weis, A. Cuénod, B. Rieck, F. Llinares-López, O. Dubuis, S. Graf, C. Lang, M. Oberle, K. K. Soegaard, M. Osthoff, M. Brackmann, K. Borgwardt, and A. Egli. "Direct Antimicrobial Resistance Prediction from clinical MALDI-TOF mass spectra using Machine Learning". *accepted in Nature Medicine*, 2021. doi: `https://doi.org/10.1101/2020.07.30.228411`

The systematic review in Chapter 2 reveals that the major limitation to advancement and novel algorithm development for MALDI-TOF MS based phenotype prediction is the lack of large-scale MALDI-TOF MS datasets, both for model development and external validation. Consequentially, we embark on the journey towards a clinically applicable antimicrobial resistance classifier by collecting a clinical routine MALDI-TOF MS dataset suitable for phenotype prediction such as species prediction and antimicrobial resistance prediction. This dataset constitutes the largest database of its type at the time of writing this thesis, and includes MALDI-TOF mass spectra, species information, antimicrobial resistance profiles and clinical meta-data spanning four different diagnostic laboratories in Switzerland. We term this dataset *DRIAMS* (Database of ResIstance against Antimicrobials with MALDI-TOF Mass Spectrometry) and make it available to the public to advance the entire field of MALDI-TOF MS based phenotype prediction. The raw dataset comprises a total of 303,195 mass spectra and 768,300 antimicrobial resistance phenotypes, representing 803 different species of bacterial and fungal pathogens. The dataset can be accessed here[1]. This chapter will be devoted to the collection and content of *DRIAMS*, starting off with a description of the collection sites in Section 3.1, followed by a detailed walk-through of the preprocessing and quality control steps taken when building the database in Section 3.2. Section 3.3 includes a short discussion on the confounding factors in the dataset and we conclude with a summary of all contents in *DRIAMS* in Section 3.4. Figure 3.1 depicts all steps to obtain the raw and preprocessed *DRIAMS* datasets giving an overview of all steps described in chapters 3.1 and 3.2.

---

[1] doi:10.5061/dryad.bzkh1899q

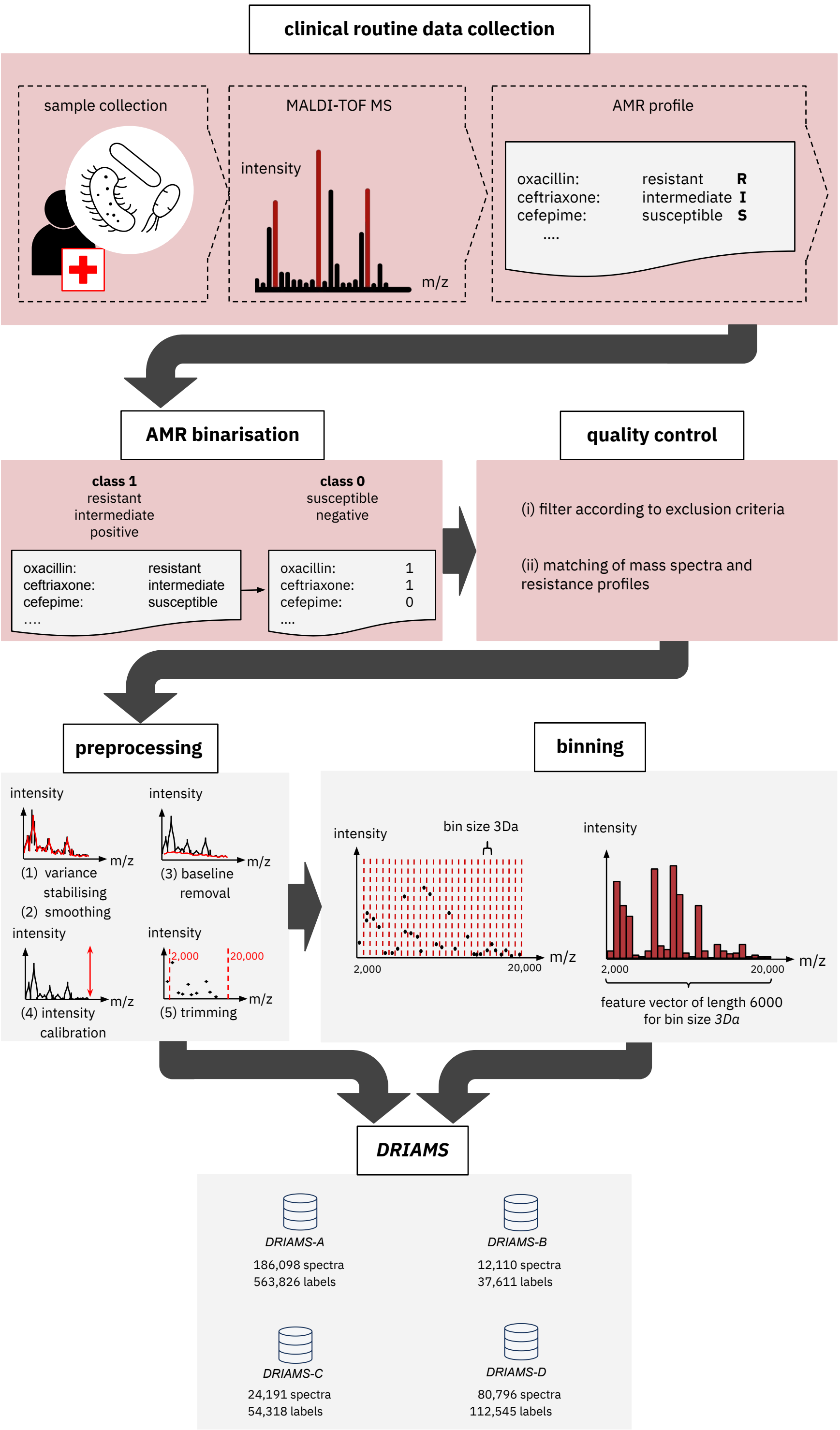


Figure 3.1: ***DRIAMS* data collection and curation**. The upper row indicates the steps taken to collect and filter samples, consisting of MALDI-TOF mass spectra and AMR profiles. The lower row is a depiction of the creation of both the preprocessed and binned MALDI-TOF mass spectra, which (along with the raw spectra and AMR profiles) comprise all four subdatasets of *DRIAMS*. Figure adapted from Weis *et al.* [126].

3.1 Clinical routine data

The data contained in *DRIAMS* was collected during daily clinical routine treatment at ISO/IEC 17025 accredited diagnostic routine laboratories and extracted from hospital records for the purpose of this thesis. All medical laboratories are located in Switzerland. For easy reference, we assign the four subdatasets constituting *DRIAMS* their own labels, *DRIAMS-A* to *DRIAMS-D*. The four diagnostic laboratories sites in this study are (1) University Hospital Basel-Stadt (providing *DRIAMS-A*), (2) Canton Hospital Basel-Land (providing *DRIAMS-B*), (3) Canton Hospital Aarau (providing *DRIAMS-C*), and (4) laboratory service provider Viollier (providing *DRIAMS-D*). For each site a collection timeframe was chosen and all extracted data (passing quality control, see Section 3.2) from that time frame was included in *DRIAMS*. The collection time frames for each site are (1) *DRIAMS-A*: 34 months (11/2015–08/2018), (2) *DRIAMS-B*: 6 months (01/2018–06/2018), (3) *DRIAMS-C*: 8 months (01/2018–08/2018), and (4) *DRIAMS-D*: 6 months (01/2018–06/2018).

MALDI-TOF mass spectra measurements. All MALDI-TOF mass spectra in *DRIAMS* were measured using the Microflex Biotyper System by Bruker Daltonics (Bremen, Germany), a widely-employed MALDI-TOF MS system in microbiological clinical routine diagnostics both in Europe [44, 94] as well as North America [28]. Both hospitals in Basel-Land and Aarau use the Microflex Biotyper LT/SH System, while Viollier uses the Microflex smart LS System. The diagnostic laboratory at the University Hospital Basel-Stadt uses both systems in parallel. These two systems employ a different type of laser gas but use the same reference spectra database for species identification, indicating a close similarity between the produced MALDI-TOF mass spectra, therefore we included spectra from both Microflex Biotyper systems.

Along with the MALDI-TOF mass spectra themselves, the species assigned to each spectrum was extracted from the hospital records. The species were identified at the time of measurement using the Microflex Biotyper Database (MBT 7854 MSP Library, BDAL V8.0.0.0_7311-7854) provided by the flexControl Software (Bruker Daltonics flexControl v.3.4).

Antimicrobial susceptibility testing profiles. The antimicrobial resistance profiles were routinely acquired at the same four microbiological laboratories and during the same time frames as the MALDI-TOF MS data. The phenotypic resistance in bacterial strains was determined by either (i) microdilution assays (VITEK® 2, BioMérieux, Marcy-l'Étoile, France), (ii) minimal inhibitory concentration (MIC) stripe tests (Liofilchem, Roseto degli Abruzzi, Italy), or (iii) disc diffusion tests (ThermoFisher Scientific, Waltham, USA). The resistance of yeast isolates was determined through Sensititre Yeast One (Thermofisher). All breakpoint measurements were categorized to be either susceptible, intermediate, or resistant following CLSI (2015 M45; 2017 M60) and recommendations by the European Committee on Antimicrobial Susceptibility Testing (EUCAST) [34] at the time of measurement (EUCAST v6-v8). Collecting a dataset by accessing clinical records retrospectively allows for the aggregation of an unprecedented amount of data, as MALDI-TOF MS has been employed by many clinics along with antimicrobial susceptibility testing for years. However, real-world clinical routine data is impaired by the constantly fluctuating measurement environments it was collected in and the missing information that cannot be retrieved retrospectively. These factors can cause both the over- or underestimation of predictive performance, e.g. through confounding factors linked to the phenotype of interest, or through measurement noise hindering the detection of predictive signals. A number of confounding influences will be discussed in detail later in Section 3.3.

3.2 Preprocessing and quality control

This section is devoted to the steps required to convert the raw data extracted from the hospital file storage systems into a dataset ready for machine learning analysis. Curating the new MALDI-TOF MS database comes with a number of challenges. Due to the nature of real-world clinical routine data collection, the raw data contains failed instrument calibration and other measurements not suitable for inclusion into the database. Additionally, we face a major challenge when combining two types of measurements performed on the same microbial sample, since MALDI-TOF MS measurements and antimicrobial susceptibility testing were not intended to be matched in the clinical routine. Therefore, the records do not always archive all information needed to match measurements unambiguously. As we combine data from four different diagnostics site, requirements are different for all four raw input datasets and the data aggregation process has to be adjusted accordingly. Lastly, we discuss the conversion into the final data representation using in the machine learning analysis.

Mass spectra exclusion criteria. We excluded mass spectra files that were either empty (due to e.g. faulty execution or an empty sample plate) or measured during the calibration process of the MALDI-TOF MS machine. During this calibration process an *E. coli* probe modified with chemicals producing specific mass peaks is repeatedly measured while the instrument parameters are adjusted.

Matching two data sources. A dataset suitable for MALDI-TOF MS based antimicrobial resistance prediction must consist of entries representing microbial isolates, with both the mass spectrum and the antimicrobial resistance label known for a large number of instances. In order to construct such a dataset from clinical records, the MALDI-TOF MS and resistance profile measurements belonging to the same microbial isolate have to be matched. We face the same challenge at all medical sites: the mass spectra and antimicrobial susceptibility testing of the same isolate are carried out in separate procedures in the clinical routine, and as both measurements do not need to be matched for clinical diagnostics, are not recorded in a way that allows one to identify the resistance profile record and mass spectrum file stemming from the same isolate.

For ease of communication, here we term the report documenting antimicrobial resistance profiles as the ‘laboratory report’. The species of the isolate is determined through MALDI-TOF MS species identification and is also added to the laboratory report. By only transferring the species identified through MALDI-TOF MS, the laboratory report entry is decoupled from the mass spectrum file. Each laboratory report entry is identified by a code, which we refer to as the ‘sample ID’, linking the entry to a patient or a unique sample taken from a patient. If multiple probes have been taken from the same patient, multiple entries with the same sample ID may occur. The spectra measured by the Bruker Microflex systems are labelled with an ambiguous code corresponding to the non-unique sample ID in the laboratory report, linking MALDI-TOF MS measurements to patients. We construct a new code by combining the sample ID and the determined genus of a sample to identify the MALDI-TOF mass spectrum corresponding to each laboratory report entry. The new code allows for unambiguous identification in most cases, as the primary reason for repeated sample IDs is the presence of several genera samples in the patient sample, leading to several measurements. Samples for which the new sample ID-genus code was not unique are omitted.

Antimicrobial nomenclature. The antimicrobial drugs are labelled by their German name in the *DRIAMS* ID files and are anglicized during data read-in by our machine learning pipeline. Additionally, inconsistencies are unified between sites as well, including spelling variants and different names for the same drugs, such as cotrimoxazol and trimethoprim/sulfamethoxazole. *DRIAMS-A* has an additional suffix nomenclature structure: (i) *high level* indicates the higher of two dosages listed for gentamycin (intravenous administration) in EUCAST, with the standard dose being 5 mg/kg and the *high level* dose 7 mg/kg, (ii) *meningitis*, *pneumoniae*, *endocarditis*, and *uncomplicated UTI* (urinary tract infection) indicate an infection-specific breakpoint for the respective infection in EUCAST, (iii) *screen* in cefoxitin indicates that this test is used as a MRSA screen in the clinical routine diagnostic at University Hospital Basel, (iv) *GRD* stands for ‘glycopeptide resistance detection’ used at University Hospital Basel in very rare cases to detect glycopeptide intermediate *S. aureus*, and (v) *1mg_l* indicates rifampicin concentration in liquid culture, typically for *Mycobacterium tuberculosis*. If this suffix is entered for other species, it is a mistake made when entering information into the laboratory information system.

Spectral representation. The MALDI-TOF mass spectra are extracted from the Bruker Flex machine in the Bruker Flex data format. The following standard MALDI-TOF mass spectra preprocessing protocol is applied: (1) intensity transformation with a square-root method to stabilize the variance, (2) Savitzky-Golay smoothing with half-window-size 10, (3) baseline estimation is removed in 20 iterations using the SNIP algorithm, (4) intensity calibration using the total-ion-current (TIC), and (5) spectra trimming to values in 2 kDa to 20 kDa. All steps were implemented in `R` using the package `MaldiQuant` [43] v1.19, with detailed parameter values given in the code. Raw MALDI-TOF mass spectra constitute a list of a number of tuples, each containing a m/z–ratio and the respective intensity. The number of measured tuples varies between mass spectra. Most machine learning models process input data in the form of a fixed-length feature vector. Therefore, we convert the raw MALDI-TOF mass spectra into vectors of fixed length by partitioning the measured intensities ranging from 2 kDa to 20 kDa into a feature grid of disjoint, equal-sized bins, and summing up all intensities falling into the same bin. Initial exploratory experiments indicate that a bin size of 3 Da provides a suitable feature vector representing the mass spectrum, allowing for separation and distribution of mass peaks with similar m/z–ratio values, while being large enough not to cause redundancies between features. Therefore, we obtain a vector of dimensionality 6,000. In the public version of *DRIAMS*, the raw, preprocessed as well as the binned spectra representations are included.

Antimicrobial resistance phenotype binarisation. Throughout this thesis we represent the prediction as a binary classification problem. Therefore, the values reported in the antimicrobial resistance profiles were assigned to two classes. The profiles are already represented by EUCAST and CLSI [22, 34] based resistance categories. For antimicrobial resistances that are reported by RSI values, the positive class was assigned when the category indicated a resistant (R) or intermediate (I) sample and the negative class was assigned to susceptible (S) samples. The choice for grouping samples with the intermediate category to the resistant sample class was made deliberately in accordance to their ultimate effect in patient treatment: both resistant and intermediate samples prevent the prescription of an antibiotic.

3.3 Confounding factors in real-world clinical data

In this section we discuss several aspects that affect the population covered in *DRIAMS* and potentially confound the signal for antimicrobial resistance.

Patient clientèle. The mark-up of patients and respective infections treated at each medical location is reflected in the microbial population captured by each dataset in *DRIAMS*. Samples included in *DRIAMS-A* (University Hospital Basel-Stadt) originate primarily from patients residing in the city of Basel and its surroundings seeking out- or inpatient treatment. The patient population in *DRIAMS-B* mostly stem from towns surrounding the city of Basel, while patients from the entirety of the Swiss canton Aargau are included in *DRIAMS-C*. *DRIAMS-D* differs substantially from the other medical institutions in the sense that its collection site Viollier is a service provider performing species identification from samples collected in medical practices and hospitals originating from all over Switzerland. The implications include a shift in infections depicted—milder infections will be overrepresented in *DRIAMS-D* due to the medical practice data, while highly complex infections will appear at a lower ratio.

dataset	total	hospital hygiene	blood	deep tissue	genital	respiratory	stool	urine	varia
E. coli	$n = 4961$	659	1190	1073	24	364	5	1473	173
(ceftriaxone)	$\% = 100$	13.3	24	21.6	0.5	7.3	0.1	29.7	3.5
K. pneumoniae	$n = 2860$	229	273	204	5	268	15	1790	76
(ceftriaxone)	$\% = 100$	8	9.6	7.1	0.2	9.4	0.5	62.6	2.7
S. aureus	$n = 3790$	379	708	1356	34	517	0	187	609
(oxacillin)	$\% = 100$	10	18.7	35.8	0.9	13.6	0	4.9	16.1

Table 3.1: **Distribution of samples over workstations** for three species–antibiotic datasets from *DRIAMS-A*.

Clinical research. As *DRIAMS-A* was collected at a hospital with a university affiliation, laboratory equipment is used for research experiments in parallel to the clinical routine measurements. The spectra were not specifically labelled when originating from non-routine experiments and the *DRIAMS* dataset was not filtered for them. As we cannot identify these spectra, we can only speculate in regards to the implications: strains that are subject of local research projects might be overrepresented, or only therefore included, and therefore bias the dataset.

Workstations. *DRIAMS-A* clinical routine isolates are analysed at one of nine workstations categorised by isolation material or procedure: (i) urine isolates, (ii) blood culture isolates, (iii) stool isolates, (iv) genital tract isolates, (v) isolates with a polymerase chain reaction (PCR)-based test, (vi) respiratory tract isolates, (vii) isolates from deep (usually sterile) material, (viii) isolates from a hospital hygiene department, and (ix) remaining isolates. Specifically, the isolates collected by the hospital hygiene department create substantial confounding. The purpose of these measurements is to prevent within-hospital (nosocomial) transmissions of multidrug-resistant pathogens, by testing samples collected from within the hospital (e.g. door handles and other surfaces). As the objective is to detect resistant pathogens, the collected isolates are cultured on selective growth media (i.e. growth media containing antimicrobial drugs) that only allow for the growth of resistant strains. The growth medium affects the microbial proteome, and is therefore reflected in its MALDI-TOF mass spectrum [104] and could confound antimicrobial resistance prediction if a mass peak indicating a selective growth medium is used as a signal for antimicrobial resistance. The individual sample sizes per workstation are listed in Table 3.1.

Patient cases. The *DRIAMS-A* laboratory report includes information on patient case affiliation. A clinical case defines a single hospital stay, i.e. the duration between hospital entry and exit of a specific patient. Repeated hospital stays of the same patient are identified as separate patient cases. MALDI-TOF mass spectra affiliated with the same patient case number likely stem from the same infection, i.e. the same microbial strains with identical resistance profiles. Clinical cases should be considered during machine learning experimental design to avoid information leakage and confounded results. For the three remaining subdatasets, *DRIAMS-B*, *DRIAMS-C* and *DRIAMS-D*, no information on patient cases is provided.

3.4 Summary of the *DRIAMS* datasets

Tables 3.2 to 3.6 provide statistics for each antimicrobial drug contained in the datasets *DRIAMS-A* to *DRIAMS-D* respectively. For each drug, the number of available spectra n, the positive (resistant/intermediate) class ratio and the three most frequent species with their number of spectra are listed. These tables allow one to judge the vast possibilities of species-antimicrobial scenario combinations that can be studied with our curated and publicly available dataset *DRIAMS*.

antimicrobial drug	n	% class1	most frequent species
ciprofloxacin	30543	24.4	*Escherichia coli* (4911), *Staphylococcus epidermidis* (4679), *Staphylococcus aureus* (3757)
meropenem	29531	17.4	*Escherichia coli* (4928), *Staphylococcus epidermidis* (4261), *Staphylococcus aureus* (3643)
imipenem	29391	23.4	*Escherichia coli* (4923), *Staphylococcus epidermidis* (4261), *Staphylococcus aureus* (3640)
cefepime	28476	22.9	*Escherichia coli* (4890), *Staphylococcus epidermidis* (4261), *Staphylococcus aureus* (3640)
piperacillin-tazobactam	28398	23.1	*Escherichia coli* (4799), *Staphylococcus epidermidis* (4261), *Staphylococcus aureus* (3640)
ampicillin-amoxicillin	26871	81.7	*Escherichia coli* (4866), *Staphylococcus epidermidis* (4371), *Staphylococcus aureus* (3556)
cotrimoxazole	26640	18.3	*Escherichia coli* (4888), *Staphylococcus epidermidis* (4545), *Staphylococcus aureus* (3741)
ceftriaxone	26545	27.5	*Escherichia coli* (4961), *Staphylococcus epidermidis* (4261), *Staphylococcus aureus* (3640)
amoxicillin-clavulanic acid	25228	39.3	*Escherichia coli* (4826), *Staphylococcus epidermidis* (4262), *Staphylococcus aureus* (3640)
levofloxacin	20784	19.1	*Escherichia coli* (4858), *Klebsiella pneumoniae* (2830), *Pseudomonas aeruginosa* (2356)
colistin	18333	15.5	*Escherichia coli* (4930), *Pseudomonas aeruginosa* (3234), *Klebsiella pneumoniae* (2854)
tobramycin	18190	9.3	*Escherichia coli* (4876), *Pseudomonas aeruginosa* (3231), *Klebsiella pneumoniae* (2846)
ceftazidime	17392	14.1	*Escherichia coli* (4822), *Klebsiella pneumoniae* (2832), *Pseudomonas aeruginosa* (2459)
amikacin	17222	5.7	*Escherichia coli* (4858), *Klebsiella pneumoniae* (2830), *Pseudomonas aeruginosa* (2372)
vancomycin	15076	1.2	*Staphylococcus epidermidis* (4777), *Staphylococcus aureus* (3791), *Enterococcus faecium* (1183)
ertapenem	14753	2.0	*Escherichia coli* (4983), *Klebsiella pneumoniae* (2859), *Enterobacter cloacae* (1249)
penicillin	13406	73.7	*Staphylococcus epidermidis* (4371), *Staphylococcus aureus* (3553), *Staphylococcus hominis* (709)
linezolid	12288	0.1	*Staphylococcus epidermidis* (4407), *Staphylococcus aureus* (3639), *Enterococcus faecium* (1124)
tigecycline	12280	0.4	*Staphylococcus epidermidis* (4365), *Staphylococcus aureus* (3640), *Enterococcus faecium* (1128)
clindamycin	11612	31.3	*Staphylococcus epidermidis* (4192), *Staphylococcus aureus* (3575), *Staphylococcus hominis* (685)
daptomycin	11384	0.9	*Staphylococcus epidermidis* (4761), *Staphylococcus aureus* (3780), *Staphylococcus hominis* (715)
erythromycin	11079	40.9	*Staphylococcus epidermidis* (4232), *Staphylococcus aureus* (3598), *Staphylococcus hominis* (685)
oxacillin	10985	42.2	*Staphylococcus epidermidis* (4664), *Staphylococcus aureus* (3790), *Staphylococcus hominis* (685)
rifampicin	10966	4.9	*Staphylococcus epidermidis* (4758), *Staphylococcus aureus* (3774), *Staphylococcus hominis* (717)
fusidic acid	10637	32.1	*Staphylococcus epidermidis* (4589), *Staphylococcus aureus* (3766), *Staphylococcus hominis* (692)
gentamicin	10579	21.8	*Staphylococcus epidermidis* (4221), *Staphylococcus aureus* (3629), *Staphylococcus hominis* (683)
cefuroxime	10578	42.3	*Staphylococcus epidermidis* (4261), *Staphylococcus aureus* (3640), *Staphylococcus hominis* (668)
cefazolin	10036	42.1	*Staphylococcus epidermidis* (4261), *Staphylococcus aureus* (3640), *Staphylococcus hominis* (668)
tetracycline	9918	31.1	*Staphylococcus epidermidis* (4132), *Staphylococcus aureus* (3609), *Staphylococcus hominis* (670)
teicoplanin	7691	2.9	*Staphylococcus aureus* (3629), *Enterococcus faecium* (1124), *Enterococcus faecalis* (757)
cefpodoxime	6720	34.8	*Escherichia coli* (2075), *Klebsiella pneumoniae* (1826), *Proteus mirabilis* (470)
fosfomycin-trometamol	6129	21.6	*Klebsiella pneumoniae* (1809), *Escherichia coli* (1499), *Proteus mirabilis* (465)
norfloxacin	6105	14.3	*Klebsiella pneumoniae* (1814), *Escherichia coli* (1488), *Proteus mirabilis* (461)
mupirocin	3815	0.7	*Staphylococcus aureus* (3633), *Staphylococcus epidermidis* (84), *MIX!Staphylococcus aureus* (31)
nitrofurantoin	2108	19.5	*Escherichia coli* (1498), *Enterococcus faecium* (464), *Enterococcus faecalis* (54)
aztreonam	856	70.6	*Pseudomonas aeruginosa* (763), *Pseudomonas stutzeri* (22), *Escherichia coli* (13)
caspofungin	686	5.4	*Candida albicans* (292), *Candida glabrata* (171), *Candida parapsilosis* (65)
gentamicin_high_level	686	15.5	*Enterococcus faecium* (273), *Enterococcus faecalis* (152), *Streptococcus oralis* (43)
5-fluorocytosine	680	3.7	*Candida albicans* (292), *Candida glabrata* (174), *Candida parapsilosis* (65)
micafungin	680	15.1	*Candida albicans* (290), *Candida glabrata* (174), *Candida parapsilosis* (65)

Table 3.2: ***DRIAMS-A* summary (Part 1)** n refers to the number of MALDI-TOF mass spectra with a phenotype for the respective drug. '% class' indicates the percentage of these mass spectra having a positive class, meaning either intermediate or resistant phenotype. Antimicrobial nomenclature and suffixes are described in detail in Section 3.2.

antimicrobial drug	n	% class1	most frequent species
anidulafungin	675	28.3	*Candida albicans* (287), *Candida glabrata* (173), *Candida parapsilosis* (65)
fluconazole	675	39.3	*Candida albicans* (283), *Candida glabrata* (174), *Candida parapsilosis* (65)
itraconazole	668	37.7	*Candida albicans* (283), *Candida glabrata* (172), *Candida parapsilosis* (65)
amoxicillin	629	21.6	*Enterococcus faecium* (91), *Haemophilus influenzae* (56), *Enterococcus faecalis* (52)
amphotericin b	616	0.0	*Candida albicans* (290), *Candida glabrata* (176), *Candida parapsilosis* (65)
voriconazole	502	5.8	*Candida albicans* (286), *Candida parapsilosis* (65), *Candida tropicalis* (43)
posaconazole	412	12.4	*Candida albicans* (282), *Candida parapsilosis* (65), *Candida tropicalis* (43)
moxifloxacin	411	22.1	*Finegoldia magna* (23), *Bacteroides fragilis* (19), *Parvimonas micra* (15)
penicillin_with_endokarditis	330	53.9	*Streptococcus oralis* (65), *Streptococcus parasanguinis* (56), *Streptococcus mitis* (39)
penicillin_without_endokarditis	325	43.4	*Streptococcus oralis* (65), *Streptococcus parasanguinis* (56), *Streptococcus mitis* (39)
clarithromycin	313	18.8	*Streptococcus pneumoniae* (255), *Streptococcus agalactiae* (7), *Streptococcus dysgalactiae* (7)
penicillin_with_pneumonia	293	4.1	*Streptococcus pneumoniae* (256), *Streptococcus pseudopneumoniae* (7), *Streptococcus anginosus* (7)
penicillin_with_other_infections	292	17.5	*Streptococcus pneumoniae* (256), *Streptococcus pseudopneumoniae* (7), *Streptococcus anginosus* (7)
penicillin_with_meningitis	291	19.9	*Streptococcus pneumoniae* (255), *Streptococcus pseudopneumoniae* (7), *Streptococcus anginosus* (7)
metronidazole	247	5.7	*Bacteroides fragilis* (37), *Clostridium difficile* (18), *Clostridium perfringens* (13)
quinolones	211	5.7	*Haemophilus influenzae* (55), *Staphylococcus aureus* (19), *Staphylococcus epidermidis* (17)
meropenem_with_meningitis	210	13.3	*Haemophilus influenzae* (70), *Streptococcus pneumoniae* (56), *Haemophilus parainfluenzae* (40)
chloramphenicol	203	13.8	*Haemophilus influenzae* (55), *Staphylococcus aureus* (19), *Staphylococcus epidermidis* (17)
meropenem_without_meningitis	142	3.5	*Haemophilus influenzae* (72), *Haemophilus parainfluenzae* (40), *MIX!Haemophilus parainfluenzae* (11)
aminoglycosides	122	16.4	*Staphylococcus aureus* (19), *Staphylococcus epidermidis* (15), *Corynebacterium macginleyi* (14)
doxycycline	105	11.4	*Staphylococcus epidermidis* (37), *Finegoldia magna* (6), *Campylobacter jejuni* (5)
azithromycin	95	12.6	*Neisseria gonorrhoeae* (67), *MIX!Neisseria gonorrhoeae* (6), *Campylobacter fetus* (5)
fosfomycin	81	58.0	*Pseudomonas aeruginosa* (37), *Staphylococcus epidermidis* (14), *Klebsiella pneumoniae* (7)
amoxicillin-clavulanic acid_ uncomplicated_hwi	80	30.0	*Escherichia coli* (47), *Citrobacter freundii* (14), *Proteus mirabilis* (9)
minocycline	74	29.7	*Staphylococcus epidermidis* (25), *Burkholderia multivorans* (10), *Stenotrophomonas maltophilia* (7)
cefixime	74	8.1	*Neisseria gonorrhoeae* (63), *MIX!Neisseria gonorrhoeae* (6), *Neisseria meningitidis* (3)
meropenem_with_pneumonia	70	0.0	*Streptococcus pneumoniae* (56), *Streptococcus pseudopneumoniae* (3), *Streptococcus anginosus* (3)
cefoxitin_screen	52	9.6	*Staphylococcus aureus* (23), *Staphylococcus epidermidis* (13), *Staphylococcus pseudintermedius* (8)
vancomycin_grd	12	0.0	*Staphylococcus aureus* (12)
teicoplanin_grd	12	16.7	*Staphylococcus aureus* (12)
rifampicin_1mg-l	10	20.0	*Peptoniphilus harei* (3), *Corynebacterium jeikeium* (2), *Staphylococcus aureus* (2)

Table 3.3: ***DRIAMS-A* summary (Part 2)** n refers to the number of MALDI-TOF mass spectra with a phenotype for the respective drug. '% class1' indicates the percentage of these mass spectra having a positive class, meaning either intermediate or resistant phenotype. Antimicrobial nomenclature and suffixes are described in detail in Section 3.2.

antimicrobial drug	n	% class1	most frequent species
ciprofloxacin	2019	16.8	*Staphylococcus aureus* (348), *Staphylococcus epidermidis* (220), *Escherichia coli* (213)
amoxicillin-clavulanic acid	1990	26.9	*Staphylococcus aureus* (346), *Staphylococcus epidermidis* (220), *Escherichia coli* (213)
cotrimoxazol	1749	11.2	*Staphylococcus aureus* (345), *Staphylococcus epidermidis* (218), *Escherichia coli* (213)
cefepime	1689	15.9	*Staphylococcus aureus* (346), *Staphylococcus epidermidis* (220), *Escherichia coli* (213)
gentamicin	1547	8.6	*Staphylococcus aureus* (348), *Staphylococcus epidermidis* (220), *Escherichia coli* (213)
fosfomycin	1535	19.1	*Staphylococcus aureus* (346), *Staphylococcus epidermidis* (218), *Escherichia coli* (213)
ampicillin	1348	53.0	*Escherichia coli* (213), *Enterococcus faecalis* (171), *Klebsiella pneumoniae* (152)
vancomycin	1236	1.3	*Staphylococcus aureus* (346), *Staphylococcus epidermidis* (218), *Enterococcus faecalis* (171)
imipenem	1215	7.6	*Escherichia coli* (213), *Enterococcus faecalis* (170), *Klebsiella pneumoniae* (152)
levofloxacin	1204	16.4	*Staphylococcus aureus* (348), *Staphylococcus epidermidis* (220), *Enterococcus faecalis* (172)
piperacillin-tazobactam	1175	16.0	*Escherichia coli* (213), *Enterococcus faecalis* (171), *Klebsiella pneumoniae* (151)
linezolid	1142	0.0	*Staphylococcus aureus* (346), *Staphylococcus epidermidis* (218), *Enterococcus faecalis* (172)
ceftriaxone	1085	10.6	*Escherichia coli* (213), *Klebsiella pneumoniae* (152), *Proteus mirabilis* (85)
benzylpenicillin	1075	57.5	*Staphylococcus aureus* (348), *Staphylococcus epidermidis* (220), *Streptococcus agalactiae* (79)
clindamycin	1033	20.4	*Staphylococcus aureus* (348), *Staphylococcus epidermidis* (216), *Streptococcus agalactiae* (79)
nitrofurantoin	1015	52.6	*Escherichia coli* (213), *Enterococcus faecalis* (171), *Klebsiella pneumoniae* (152)
tigecycline	957	0.2	*Staphylococcus aureus* (345), *Staphylococcus epidermidis* (217), *Enterococcus faecalis* (171)
erythromycin	948	24.7	*Staphylococcus aureus* (347), *Staphylococcus epidermidis* (216), *Streptococcus agalactiae* (79)
ceftazidime	940	11.3	*Escherichia coli* (213), *Klebsiella pneumoniae* (152), *Pseudomonas aeruginosa* (138)
amikacin	926	2.8	*Escherichia coli* (205), *Klebsiella pneumoniae* (147), *Pseudomonas aeruginosa* (135)
meropenem	916	1.2	*Escherichia coli* (205), *Klebsiella pneumoniae* (147), *Pseudomonas aeruginosa* (129)
tetracycline	877	14.8	*Staphylococcus aureus* (343), *Staphylococcus epidermidis* (172), *Streptococcus agalactiae* (79)
clindamycin_induced	877	11.6	*Staphylococcus aureus* (348), *Staphylococcus epidermidis* (220), *Streptococcus agalactiae* (77)
ertapenem	805	2.5	*Escherichia coli* (213), *Klebsiella pneumoniae* (152), *Proteus mirabilis* (85)
cefuroxime	803	37.9	*Escherichia coli* (213), *Klebsiella pneumoniae* (152), *Proteus mirabilis* (84)
rifampicin	802	1.6	*Staphylococcus aureus* (348), *Staphylococcus epidermidis* (220), *Aerococcus urinae* (44)
cefoxitin	800	29.2	*Escherichia coli* (213), *Klebsiella pneumoniae* (152), *Proteus mirabilis* (85)
norfloxacin	800	17.9	*Escherichia coli* (213), *Klebsiella pneumoniae* (152), *Proteus mirabilis* (85)
teicoplanin	746	1.3	*Staphylococcus aureus* (346), *Enterococcus faecalis* (171), *Enterococcus faecium* (41)
cefoxitin_screen	739	25.2	*Staphylococcus aureus* (348), *Staphylococcus epidermidis* (220), *Staphylococcus hominis* (38)
fusidic acid	736	23.9	*Staphylococcus aureus* (346), *Staphylococcus epidermidis* (218), *Staphylococcus hominis* (38)
mupirocin	736	1.4	*Staphylococcus aureus* (346), *Staphylococcus epidermidis* (218), *Staphylococcus hominis* (38)
oxacillin	731	25.4	*Staphylococcus aureus* (346), *Staphylococcus epidermidis* (220), *Staphylococcus hominis* (35)
daptomycin	719	0.1	*Staphylococcus aureus* (338), *Staphylococcus epidermidis* (214), *Staphylococcus hominis* (35)
gentamicin_high_level	215	16.7	*Enterococcus faecalis* (166), *Enterococcus faecium* (41), *Enterococcus avium* (6)
clarithromycin	209	17.2	*Streptococcus agalactiae* (79), *Streptococcus pyogenes* (49), *Streptococcus dysgalactiae* (27)
strepomycin_high_level	82	31.7	*Enterococcus faecalis* (70), *Enterococcus faecium* (7), *Enterococcus avium* (5)
esbl	57	100.0	*Escherichia coli* (40), *Klebsiella pneumoniae* (16), *Klebsiella oxytoca* (1)
metronidazole	32	9.4	*Finegoldia magna* (9), *Prevotella disiens* (6), *Bacteroides fragilis* (4)
benzylpenicillin_others	27	7.4	*Streptococcus pneumoniae* (23), *Streptococcus pseudopneumoniae* (2), *Streptococcus vestibularis* (1)
benzylpenicillin_with_pneumonia	27	3.7	*Streptococcus pneumoniae* (23), *Streptococcus pseudopneumoniae* (2), *Streptococcus vestibularis* (1)
benzylpenicillin_with_meningitis	27	7.4	*Streptococcus pneumoniae* (23), *Streptococcus pseudopneumoniae* (2), *Streptococcus vestibularis* (1)
mrsa	18	100.0	*Staphylococcus aureus* (18)
minocycline	2	0.0	*Staphylococcus aureus* (1), *Streptococcus dysgalactiae* (1)

Table 3.4: ***DRIAMS-B* summary** n refers to the number of MALDI-TOF mass spectra with a phenotype for the respective drug. '% class1' indicates the percentage of these mass spectra having a positive class, so either intermediate or resistant phenotype. Antimicrobial nomenclature and suffices are described in detail in Section 3.2.

antimicrobial drug	n	% class1	most frequent species
ampicillin	4547	66.5	*Escherichia coli* (884), *Staphylococcus aureus* (738), *Enterococcus faecalis* (530)
amoxicillin-clavulanic acid	4544	30.7	*Escherichia coli* (902), *Staphylococcus aureus* (738), *Enterococcus faecalis* (530)
cotrimoxazole	4075	30.7	*Escherichia coli* (896), *Staphylococcus aureus* (738), *Klebsiella pneumoniae* (366)
gentamicin	3991	13.8	*Escherichia coli* (911), *Staphylococcus aureus* (738), *Klebsiella pneumoniae* (366)
ciprofloxacin	3781	10.5	*Escherichia coli* (889), *Staphylococcus aureus* (738), *Klebsiella pneumoniae* (366)
ceftriaxone	2875	24.9	*Escherichia coli* (916), *Klebsiella pneumoniae* (366), *Pseudomonas aeruginosa* (357)
cefuroxime	2842	37.9	*Escherichia coli* (909), *Klebsiella pneumoniae* (360), *Pseudomonas aeruginosa* (357)
polymyxin b	2791	17.4	*Escherichia coli* (927), *Klebsiella pneumoniae* (366), *Pseudomonas aeruginosa* (357)
ceftazidime	2774	12.8	*Escherichia coli* (915), *Klebsiella pneumoniae* (365), *Pseudomonas aeruginosa* (357)
piperacillin-tazobactam	2494	12.4	*Escherichia coli* (623), *Enterococcus faecalis* (530), *Pseudomonas aeruginosa* (357)
imipenem	1856	2.4	*Escherichia coli* (623), *Pseudomonas aeruginosa* (357), *Klebsiella pneumoniae* (185)
amikacin	1846	1.8	*Escherichia coli* (624), *Pseudomonas aeruginosa* (357), *Klebsiella pneumoniae* (185)
cefepime	1841	14.1	*Escherichia coli* (622), *Pseudomonas aeruginosa* (357), *Klebsiella pneumoniae* (185)
nitrofurantoin	1751	18.6	*Escherichia coli* (471), *Enterococcus faecalis* (362), *Klebsiella pneumoniae* (230)
oxacillin	1561	45.5	*Staphylococcus aureus* (738), *Enterococcus faecalis* (530), *Enterococcus faecium* (93)
norfloxacin	1331	14.4	*Escherichia coli* (446), *Klebsiella pneumoniae* (230), *Proteus mirabilis* (131)
fosfomycin	1255	40.1	*Escherichia coli* (471), *Klebsiella pneumoniae* (230), *Proteus mirabilis* (131)
clindamycin	1223	32.5	*Staphylococcus aureus* (651), *Enterococcus faecalis* (175), *Enterococcus faecium* (93)
penicillin	1136	63.7	*Staphylococcus aureus* (738), *Staphylococcus lugdunensis* (67), *Haemophilus influenzae* (60)
vancomycin	1124	1.4	*Staphylococcus aureus* (651), *Enterococcus faecalis* (174), *Enterococcus faecium* (93)
clarithromycin	1050	20.5	*Staphylococcus aureus* (651), *Staphylococcus lugdunensis* (63), *Haemophilus influenzae* (60)
doxycycline	837	5.3	*Staphylococcus aureus* (650), *Staphylococcus lugdunensis* (63), *Streptococcus pneumoniae* (24)
rifampicin	768	0.5	*Staphylococcus aureus* (648), *Staphylococcus lugdunensis* (63), *Staphylococcus epidermidis* (21)
fusidic acid	745	4.2	*Staphylococcus aureus* (651), *Staphylococcus lugdunensis* (63), *Staphylococcus epidermidis* (21)
mupirocin	675	79.3	*Enterococcus faecalis* (530), *Enterococcus faecium* (93), *Staphylococcus aureus* (25)
linezolid	307	1.0	*Enterococcus faecalis* (175), *Enterococcus faecium* (93), *Enterococcus avium* (8)
metronidazole	105	1.9	*Bacteroides fragilis* (27), *Prevotella bivia* (12), *Peptoniphilus harei* (10)
novobiocin	100	7.0	*Staphylococcus aureus* (89), *Staphylococcus saprophyticus* (7), *Staphylococcus lugdunensis* (4)
moxifloxacin	70	5.7	*Streptococcus pneumoniae* (24), *Granulicatella adiacens* (7), *Actinomyces turicensis* (5)
tetracycline	69	33.3	*Campylobacter jejuni* (52), *Pasteurella multocida* (8), *Enterococcus faecium* (3)
azithromycin	55	3.6	*Campylobacter jejuni* (52), *Neisseria gonorrhoeae* (2), *Streptococcus gallolyticus* (1)
erythromycin	54	5.6	*Campylobacter jejuni* (52), *Kingella kingae* (2)
cefalotin-cefazolin	52	98.1	*Campylobacter jejuni* (52)
meropenem	26	30.8	*Proteus mirabilis* (6), *Klebsiella pneumoniae* (4), *Aerococcus urinae* (4)
amphotericin b	13	0.0	*Candida albicans* (6), *Candida parapsilosis* (3), *Candida glabrata* (2)
voriconazole	13	0.0	*Candida albicans* (6), *Candida parapsilosis* (3), *Candida glabrata* (2)
5-fluorocytosine	13	0.0	*Candida albicans* (6), *Candida parapsilosis* (3), *Candida glabrata* (2)
caspofungin	13	0.0	*Candida albicans* (6), *Candida parapsilosis* (3), *Candida glabrata* (2)
fluconazole	13	0.0	*Candida albicans* (6), *Candida parapsilosis* (3), *Candida glabrata* (2)
colistin	12	41.7	*Pseudomonas aeruginosa* (6), *Klebsiella pneumoniae* (6)
ertapenem	12	0.0	*Escherichia coli* (4), *Citrobacter freundii* (3), *Klebsiella pneumoniae* (2)
cefotaxime	10	0.0	*Pasteurella multocida* (10)
daptomycin	8	0.0	*Staphylococcus aureus* (7), *Staphylococcus epidermidis* (1)
levofloxacin	8	0.0	*Enterococcus faecium* (3), *Parvimonas micra* (2), *Propionibacterium avidum* (1)
meropenem-vaborbactam	6	0.0	*Proteus mirabilis* (6)
teicoplanin	6	0.0	*Enterococcus faecalis* (3), *Enterococcus faecium* (3)
ofloxacin	3	100.0	*Enterococcus faecium* (3)
ceftolozane-tazobactam	2	50.0	*Pseudomonas aeruginosa* (1), *Klebsiella pneumoniae* (1)
tobramycin	1	0.0	*Pseudomonas aeruginosa* (1)

Table 3.5: ***DRIAMS-C* summary** n refers to the number of MALDI-TOF mass spectra with a phenotype for the respective drug. ‘% class1’ indicates the percentage of these mass spectra having a positive class, so either intermediate or resistant phenotype.

antimicrobial drug	n	% class1	most frequent species
fosfomycin	9616	15.2	*Staphylococcus aureus* (2163), *Klebsiella pneumoniae* (2151), *Escherichia coli* (1994)
gentamicin	7008	4.8	*Klebsiella pneumoniae* (2151), *Escherichia coli* (1996), *Proteus mirabilis* (623)
cefepime	6944	4.9	*Klebsiella pneumoniae* (2151), *Escherichia coli* (1996), *Proteus mirabilis* (623)
ceftazidime	6930	8.0	*Klebsiella pneumoniae* (2151), *Escherichia coli* (1995), *Proteus mirabilis* (623)
ampicillin	6877	70.4	*Klebsiella pneumoniae* (2151), *Escherichia coli* (1992), *Proteus mirabilis* (623)
piperacillin-tazobactam	6857	8.4	*Klebsiella pneumoniae* (2148), *Escherichia coli* (1968), *Proteus mirabilis* (610)
imipenem	6815	8.1	*Klebsiella pneumoniae* (2067), *Escherichia coli* (1943), *Proteus mirabilis* (594)
ciprofloxacin	6800	11.6	*Klebsiella pneumoniae* (2102), *Escherichia coli* (1939), *Proteus mirabilis* (621)
ceftriaxone	6618	8.0	*Klebsiella pneumoniae* (2151), *Escherichia coli* (1994), *Proteus mirabilis* (623)
ertapenem	6597	1.8	*Klebsiella pneumoniae* (2151), *Escherichia coli* (1994), *Proteus mirabilis* (623)
amoxicillin-clavulanic acid	6596	23.0	*Klebsiella pneumoniae* (2151), *Escherichia coli* (1994), *Proteus mirabilis* (623)
linezolid	3332	0.1	*Staphylococcus aureus* (2163), *Staphylococcus epidermidis* (371), *Staphylococcus saprophyticus* (200)
tigecycline	3068	0.1	*Staphylococcus aureus* (2106), *Staphylococcus epidermidis* (242), *Staphylococcus saprophyticus* (169)
tetracycline	3046	9.8	*Staphylococcus aureus* (2163), *Staphylococcus epidermidis* (372), *Staphylococcus saprophyticus* (200)
rifampicin	3008	0.7	*Staphylococcus aureus* (2160), *Staphylococcus epidermidis* (372), *Staphylococcus saprophyticus* (191)
erythromycin	2431	28.9	*Staphylococcus aureus* (1735), *Staphylococcus epidermidis* (257), *Staphylococcus saprophyticus* (172)
amikacin	2015	1.7	*Escherichia coli* (726), *Klebsiella pneumoniae* (339), *Pseudomonas aeruginosa* (277)
meropenem	1949	1.5	*Escherichia coli* (725), *Klebsiella pneumoniae* (337), *Pseudomonas aeruginosa* (240)
cotrimoxazole	1280	100.0	*Escherichia coli* (518), *Klebsiella pneumoniae* (233), *Proteus mirabilis* (193)
daptomycin	392	11.2	*Staphylococcus aureus* (273), *Staphylococcus epidermidis* (44), *Staphylococcus haemolyticus* (19)
clindamycin	354	38.4	*Staphylococcus aureus* (108), *Staphylococcus epidermidis* (104), *Staphylococcus lugdunensis* (69)
tobramycin	325	3.7	*Pseudomonas aeruginosa* (257), *Acinetobacter sp* (45), *Acinetobacter baumannii* (11)
colistin	313	6.1	*Pseudomonas aeruginosa* (247), *Acinetobacter sp* (44), *Acinetobacter baumannii* (11)
vancomycin	313	2.6	*Enterococcus faecium* (171), *Enterococcus faecalis* (83), *Enterococcus dispar* (32)
teicoplanin	287	2.8	*Enterococcus faecium* (171), *Enterococcus faecalis* (84), *Enterococcus dispar* (32)
ampicillin-sulbactam	287	50.5	*Enterococcus faecium* (171), *Enterococcus faecalis* (84), *Enterococcus dispar* (32)
piperacillin	252	13.1	*Pseudomonas aeruginosa* (247), *Pseudomonas sp* (4), *Escherichia coli* (1)
ticarcillin-clavulan acid	235	44.3	*Pseudomonas aeruginosa* (230), *Pseudomonas sp* (4), *Escherichia coli* (1)
ticarcillin	229	70.3	*Pseudomonas aeruginosa* (224), *Pseudomonas sp* (4), *Escherichia coli* (1)
aztreonam	228	10.5	*Pseudomonas aeruginosa* (224), *Pseudomonas sp* (3), *Escherichia coli* (1)
levofloxacin	210	22.4	*Pseudomonas aeruginosa* (133), *Acinetobacter sp* (32), *Streptococcus pneumoniae* (27)
moxifloxacin	27	0.0	*Streptococcus pneumoniae* (27)
telithromycin	27	0.0	*Streptococcus pneumoniae* (27)
chloramphenicol	26	3.8	*Streptococcus pneumoniae* (26)
cefotaxime	25	4.0	*Streptococcus pneumoniae* (25)

Table 3.6: ***DRIAMS-D* summary** n refers to the number of MALDI-TOF mass spectra with a phenotype for the respective drug. ‘% class1’ indicates the percentage of these mass spectra having a positive class, so either intermediate or resistant phenotype.

4 Large-scale species-specific antimicrobial resistance prediction

C. Weis, A. Cuénod, B. Rieck, F. Llinares-López, O. Dubuis, S. Graf, C. Lang, M. Oberle, K. K. Soegaard, M. Osthoff, M. Brackmann, K. Borgwardt, and A. Egli. "Direct Antimicrobial Resistance Prediction from clinical MALDI-TOF mass spectra using Machine Learning". *accepted in Nature Medicine*, 2021. doi: `https://doi.org/10.1101/2020.07.30.228411`

As this thesis is the first work to have access to a dataset as comprehensive as *DRIAMS*, no baselines exist establishing (i) the predictive performances achieved by out-of-the-box machine learning algorithms, (ii) whether non-linear and complex models outperform simpler models, or (iii) task-specific properties and requirements for a successful antimicrobial resistance prediction workflow. This section explores different approaches to antimicrobial resistance prediction to provide a foundation for experiments in the field. Initially, a species-specific antimicrobial resistance predictor is built and the behaviour of predictions is analysed when the training samples are out-of-date, calibrated, or are a combination of MALDI-TOF mass spectra of species. We further build an understanding of the decision-making process of the predictors by calculating the Shapley values of feature bins and cross-referencing highly-contributing feature bins with known resistance associated MALDI-TOF mass peaks. All analyses of Chapter 4 are based on *DRIAMS-A*.

4.1 Species-stratified antimicrobial resistance prediction

4.1.1 Machine learning analysis framework

During this chapter, we continuously employ the data representation as provided in *DRIAMS*, with the 6,000-dimensional fixed-length feature vector data representation using a bin length of 3 Da. For more details on data preprocessing and representation in *DRIAMS*, please refer to Section 3.2. In order to establish the baseline, we focus on one collection site representing the largest subdataset in *DRIAMS*— *DRIAMS-A*. The loaded dataset is split into train and test datasets, while keeping a similar antimicrobial class ratio in both. In order to avoid that the machine learning classifier recognises culture medium-specific peaks in the MALDI-TOF mass spectra from the selective media, samples that were collected for the hospital hygiene department from *DRIAMS-A* are excluded from further analysis. For *DRIAMS-A*, this split also ensures that all samples associated with a specific patient case are either part of the training dataset or the testing dataset, but not both. We apply three machine learning models for antimicrobial resistance classification: (i) logistic regression (LR), (ii) gradient-boosted decision trees (LightGBM), and (iii) a multi-layer perceptron (MLP). The three models were selected based on their different model complexity. Throughout this thesis, the three models use the `sklearn` [82] implementation in `Python`. The predictive performances are reported in the commonly used metrics ‘area under the receiver operator characteristic curve’ (AUROC) and ‘area under the precision-recall curve’ (AUPRC) as performance metrics. In the following paragraphs, we explain both the applied machine learning algorithms and the metrics used throughout this thesis (Chapters 4 to 7) in detail.

Logistic regression. Logistic regression is a widely employed classification model $p(y|\boldsymbol{x};\boldsymbol{\theta})$, where $\boldsymbol{x}$ is a fixed-dimensional input vector $\boldsymbol{x} \in \mathbb{R}$, y defines the class label $y \in \{1, ..., C\}$, and $\boldsymbol{\theta}$ refers to the model parameters [72]. Throughout this thesis, labels consist of two classes only i.e. $y \in \{0, 1\}$, hence a ‘binary logistic regression’. The full model of a binary logistic regression is given by

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = Ber\Big(y|\sigma\Big(\boldsymbol{\beta}^T\boldsymbol{x} + b\Big)\Big), \tag{4.1}$$

where Ber is the Bernoulli distribution, σ is the sigmoid function, $\boldsymbol{\beta}$ are the weight parameters, and b is a real-valued bias term. The model parameters consist of $\boldsymbol{\theta} = (\boldsymbol{\beta}, b)$. The sigmoid function is defined as

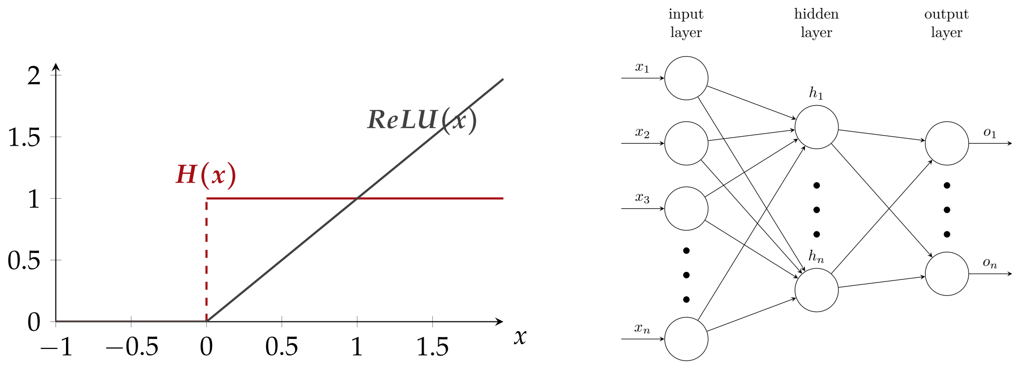
$$\sigma(a) = \frac{1}{1 + e^{-a}}, \tag{4.2}$$

where

$$a = \boldsymbol{\beta}^T\boldsymbol{x} + b \tag{4.3}$$

is called log-odds, logit or pre-activation function [72]. The sigmoid function states the probability that the class label is positive, or $y = 1$, hence it corresponds to $p(y = 1|\boldsymbol{x};\boldsymbol{\theta})$. If the probability of the positive class $p(y = 1|\boldsymbol{x};\boldsymbol{\theta})$ is higher than for the negative class $p(y = 0|\boldsymbol{x};\boldsymbol{\theta})$, then class 1 is predicted, which converts to $a > 0$ or $p(y = 1|\boldsymbol{x};\boldsymbol{\theta}) > 0.5$. The parameters in the logistic regression are estimated through maximum likelihood estimation.

LightGBM. The Light Gradient Boosting Machine (LightGBM) [55] [68] is an optimized implementation of the gradient boosted decision tree, a frequently applied machine learning algorithm, popular due to its efficiency, accuracy, and interpretability. In contrast to previous implementations, LightGBM introduced gradient-based one-side sampling and exclusive feature bundling accelerating the training process and improving memory usage [55]. Gradient boosting employs ensembles of weak prediction models in a boosting framework, i.e. sequential training, with a learner working to improve



(a) activation functions (b) multi-layer perceptron structure

Figure 4.1: **Multi-layer perceptron model** (a) Activation functions used in neural networks, such as a heaviside step function $\boldsymbol{H}(\boldsymbol{x})$ and rectified linear units $\boldsymbol{ReLU}(\boldsymbol{x})$. (b) Structure of multiple layers of node in a multi-layer perceptron model, where all nodes in hidden layers take the output of previous layers as their input.

mistakes of the previous learner in the ensemble. In the case of a gradient boosting decision tree these weak learners consist of decision trees [46].

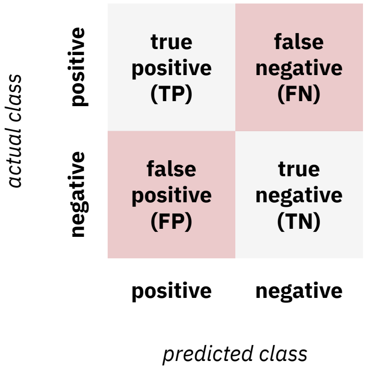
Multi-layer perceptron. A multi-layer perceptron (MLP) model constitutes a simple neural network, consisting of a stack of perceptrons [72]. Perceptrons are simple models inspired by biological neurons, defined by

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\beta}^T\boldsymbol{x} + b),$$

where $\boldsymbol{x}$ is the input feature vector, the model parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}, b)$, and $\boldsymbol{H} \in \{0,1\}$ is the heaviside step function. $\boldsymbol{H}$ acts as a linear threshold activation function as depicted in Figure 4.1a. However, perceptrons are hard to train as the heaviside step function is non-differentiable. Hence, most neural networks employ activation functions, such as sigmoid or, most frequently, rectified linear units (ReLU) $\boldsymbol{ReLU}(x) = max(0,x)$ (Figure 4.1a). Layers of perceptrons are connected in multiple layers, where perceptrons take the output of the previous layer as inputs, as depicted in Figure 4.1b.

4.1.2 Evaluating the binary classification performance

In binary classification problems, the classifier assigns each instance a class, either positive (which corresponds to the resistant/intermediate antimicrobial resistance category) or negative (susceptible antimicrobial resistance category). Regarding the classification outcome, there are four possibilities: (1) an instance that is positive and classified as positive is regarded as a *true positive*, (2) an instance that is positive and classified as negative is regarded as a *false negative*, (3) an instance that is negative and classified as negative is regarded as a *true negative*, and (4) an instance that is negative



(a) confusion matrix

true positive rate = $\frac{TP}{P}$

false positive rate = $\frac{FP}{N}$

precision = $\frac{TP}{TP+FP}$

recall = $\frac{TP}{P}$

accuracy = $\frac{TP+TN}{P+N}$

(b) metrics derived from confusion matrix

Figure 4.2: **Confusion matrix and resulting metrics** (a) two-by-two confusion matrix depicting the four possible outcomes in a binary classification, defined by the actual and predicted class (b) commonly reported machine learning metrics that are derived from the confusion matrix, including true and false positive rate (required for AUROC), precision and recall (AUPRC), and accuracy.

and classified as positive is regarded as a *false positive*. Given a test dataset consisting of a set of instances, the assignment of one of the four cases described above forms a two-by-two matrix, termed a confusion matrix or contingency table. A depiction of the table and a number of metrics can be found in Figure 4.2. The confusion matrix forms the basis of many commonly used machine learning metrics. The subsequent paragraphs introduce and explain these metrics.

Area under the receiver operating characteristic (AUROC). The area under the receiver operating characteristic (AUROC) graphs can be understood as the probability of correctly classifying a pair of instances, i.e. a positive instance and a negative instance. Receiver operating characteristic (ROC) graphs provide a depiction of a classifier and allow one to compare and select models based on their performance [37]. ROC graphs plot the true positive rate (tpr) on the y–axis against the false positive rate (fpr) on the x–axis of a discrete classifier, depicting the trade-off between detection of desired samples (true positives) and cost (false positives) in a single point in the plot. A classifier is considered discrete if it unambiguously predicts a class for each instance, i.e. either the positive or negative class in the binary classification setting. Generally, classifiers are not discrete, but will assign a score between 0 and 1. The assigned class is determined through a decision threshold, which is often assigned to 0.5 by default. Therefore a classifier can be represented in the ROC graph by varying a threshold from $-\infty$ to ∞ and tracing a curve through ROC space. The predictive performance is better if a point lies further to the top left corner ($\mathrm{tpr} = 1$ and $\mathrm{fpr} = 0$), while points on the diagonal represent classifiers with a random performance. These scenarios correspond to AUROC values of 1.0 and 0.5 respectively. Please note that generally, the best predictive performance occurs at a classification threshold other than 0.5.

Area under the precision recall curve (AUPRC). The area under the precision recall curve (AUPRC) quantifies the ability to correctly classify instances from the less frequent of two classes while minimising the false discovery rate. Precision-recall graphs plot the precision on the y–axis against the recall on the x–axis of a discrete classifier. Similarly to ROC curves, the curve is built by varying the decision threshold from $-\infty$ to ∞ and drawing a curve in the graph. The AUPRC value of a perfect classifier is 1.0, while a random classifier will take the class ratio of the minority class.

4.1.3 Building antimicrobial resistance predictors specifically for one species and one antimicrobial drug

The next section devotes itself to constructing the first antimicrobial resistance classifiers on *DRIAMS*, employing out-of-the-box (i.e. not tailored to the data type of MALDI-TOF mass spectra) machine learning models and introduce some analyses to establish the characteristics of MALDI-TOF MS based antimicrobial resistance prediction.

MALDI-TOF mass spectra contain more predictive information than species information alone. First, we reiterate the premise of large-scale MALDI-TOF MS based antimicrobial resistance prediction by demonstrating the superior predictiveness of information contained in MALDI-TOF mass spectra compared to species information alone. A predictive performance based on MALDI-TOF mass spectra higher than based on species information alone is evidence that a MALDI-TOF MS based predictor provides additional information for clinical decision-making, and at an earlier time point than antimicrobial susceptibility testing results. Knowledge of the originating species of a MALDI-TOF mass spectrum is a large indicator towards possible antimicrobial resistances, as resistance prevalences vary widely amongst microbial species. We ensure that MALDI-TOF mass spectra contain more information useful towards antimicrobial resistance prediction than just the species they represent. To this end, we compare the predictive performance of separate logistic regression classifiers trained to predict resistance to 42 antimicrobial drugs in *DRIAMS-A* based on (i) MALDI-TOF mass spectra information alone, and (ii) species information (previously identified through MALDI-TOF MS) alone of each instance. *DRIAMS-A* was selected for this analysis as it is the most comprehensive dataset in terms of the number of MALDI-TOF mass spectra and collection timespan. The resulting AUROC values are depicted in Figure 4.3. For 31 of the 42 antimicrobial drugs investigated, the respective logistic regression classifier had an AUROC value above 0.80, implying accurate resistance predictions based on MALDI-TOF MS. For 22 antimicrobial drugs, the results indicate a statistically significant improvement in prediction performance when predicting from MALDI-TOF mass spectra as compared to using only species information. The results clearly show the superior predictiveness of information captured in MALDI-TOF mass spectra.

Focusing to species-stratified training data yields superior predictions. Previous work indicates that the resistance mechanism against a specific antimicrobial drug dif-

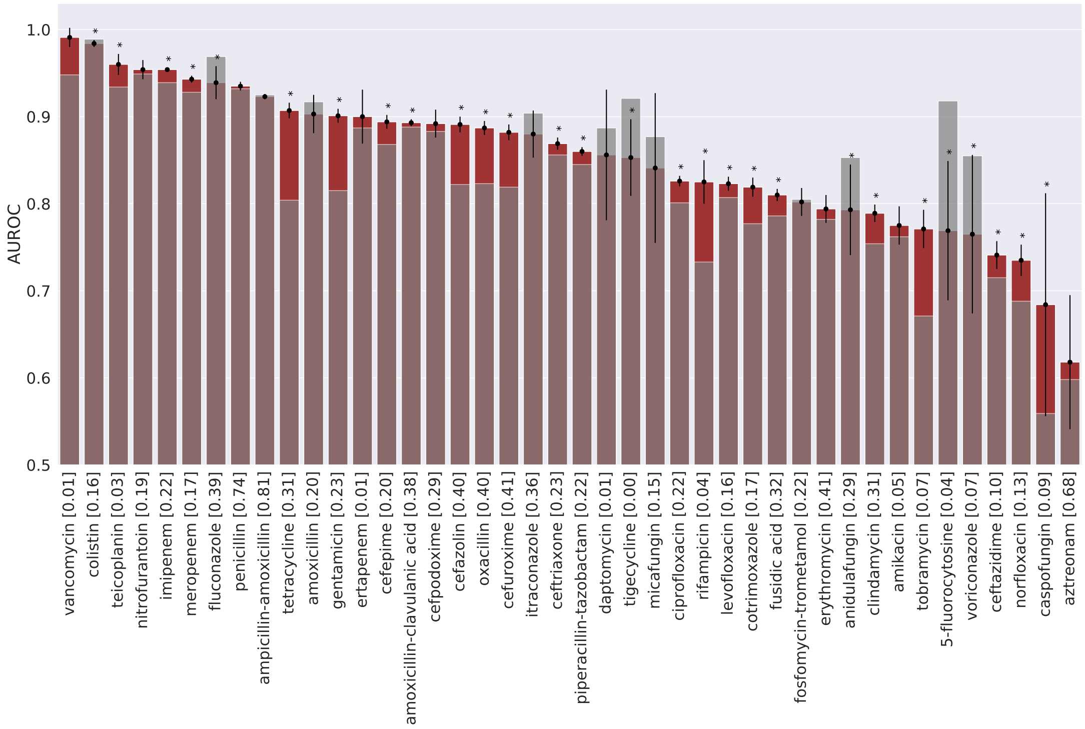


Figure 4.3: **Predictive performance increase through MALDI-TOF MS compared to species information**. For each antimicrobial drug resistance prediction task in *DRIAMS-A*, two logistic regression classifiers are trained: one predicting resistance for MALDI-TOF mass spectra and one based on species information alone. Note that for each drug, the dataset is comprised of all samples with a resistance label available for the respective drug and contains a number of different species. The red bars depict the AUROC when basing prediction on MALDI-TOF MS information, the grey bars indicating the predictive performance based on species information alone. The percentages of positive class (resistant/intermediate) samples in the training data are stated in brackets after the antibiotic name. The AUROC values and error bars depict the mean and standard deviation of ten random train-test-splits, while the asterisks indicate a significance level of less than 5% between the reported metrics of both approaches determined by a non-equal-variance Welch's t-test. Figure adapted from Weis *et al.* [126].

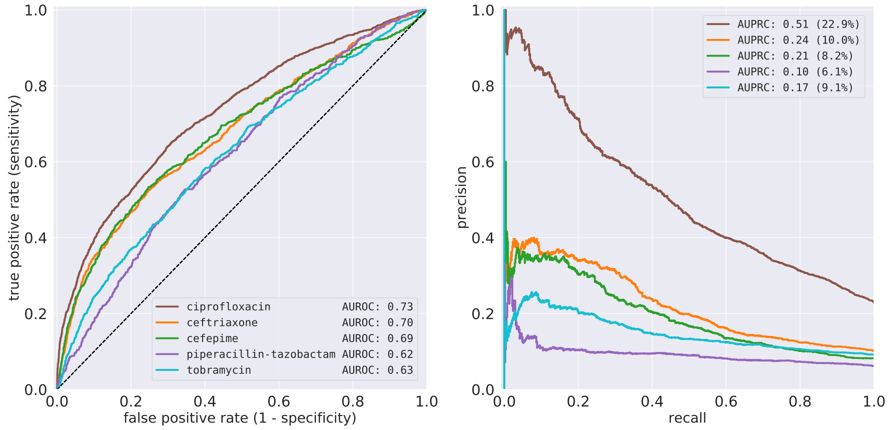
fers between different species of bacteria. Looking at the example of beta-lactam antibiotic resistance: gram-negative bacteria (such as *E. coli* and *K. pneumoniae*) produce beta-lactamses such as *CTX-M*, *TEM*, and *SHV* or carbapenemases [8, 57, 86, 88], while gram-positive bacteria (such as *S. aureus*) produce a penicillinase (*blaZ*) [83] or have an alteration within the penicillin-binding protein (PBP2a) [64]. In the next step, we train classifiers separate for different species. Requiring species information to build the antimicrobial resistance prediction workflow is in line with the development of a clinical application, as (a) the species is identified along with MALDI-TOF MS measurement and is therefore available along with every MALDI-TOF mass spectrum, and (b) antimicrobial drugs considered for treatment, and thus the resistance of interest that shall be investigated, differ for each species. The following analysis focuses on three pathogens — *S. aureus*, *E. coli*, and *K. pneumoniae* — which are all on the WHO list of priority pathogens (see Chapter 1).

Performance results predicting a number of resistances along with their ROC and precision-recall curves can be found in Figures 4.4, 4.5 and 4.6. Table 4.1 provides a direct overview of the predictive performances reached in all three species. The applied machine learning models are the aforementioned (i) logistic regression (LR), (ii) gradient-boosted decision trees (LightGBM), and (iii) a multi-layer perceptron (MLP).

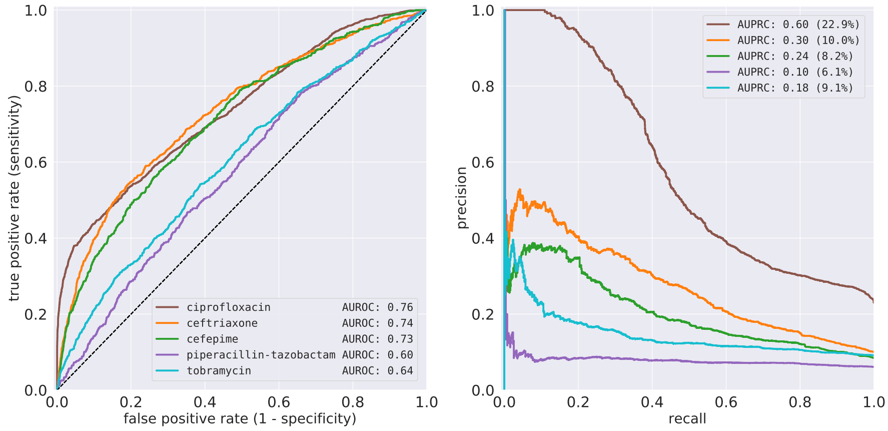
Even with this reduction in studied species, antimicrobial drugs and machine learning models, the combinatorial explosion of cases that can be studied is evident in the number of curves that can analysed. Detailed analysis later on in this thesis requires further selection of antimicrobials and models. To this end, we select an antimicrobial drug of interest for each species studied based on drugs frequently applied in the clinic and good performance (Figures 4.4, 4.5 and 4.6 or Table 4.1). The selected antimicrobials are ceftriaxone resistance in *E. coli* and *K. pneumoniae*—as markers for extended spectrum or other beta-lactamases (ESBL)—and oxacillin in *S. aureus*—as a marker for methicillin-resistant *S. aureus* (MRSA).

The predictors for ceftriaxone resistance in *E. coli* and *K. pneumoniae* reach AUROC values of 0.74 in both species with AUPRC values of 0.30 and 0.33, at a positive (i.e. resistant/intermediate) class ratio of 10.0% and 8.2%, respectively. The oxacillin resistance predictor in *S. aureus* reports an AUROC of 0.80 and AUPRC of 0.49 at a positive class ratio of 10.0%. Oxacillin resistance prediction in *S. aureus* takes a particularly important place in *DRIAMS-A*, as the reported susceptibility of resistance against all beta-lactam antibiotics is inferred from the oxacillin resistance results. The selected drugs represent clinically relevant treatment scenarios, and assessing the potential to predict resistance against each antimicrobial drug in the respective species harbors high impact on patient treatment. Note that the selected species-antibiotic scenario will be frequently abbreviated in this thesis, namely E-CEF for ceftriaxone resistance in *E. coli*, K-CEF for ceftriaxone resistance in *K. pneumoniae*, and S-OXA for oxacillin resistance in *S. aureus*. We further select one machine learning model per scenario to focus the analysis in the rest of this chapter on. Note that for the remainder of Chapter 4, all analyses focus on interpreting and understanding the functionality of these antimicrobial resistance predictors. Hence, fixing the machine learning model does not restrain the development of the predictor. We select the models based on the pre-

(a) *E. coli* logistic regression



(b) *E. coli* LightGBM



(c) *E. coli* MLP

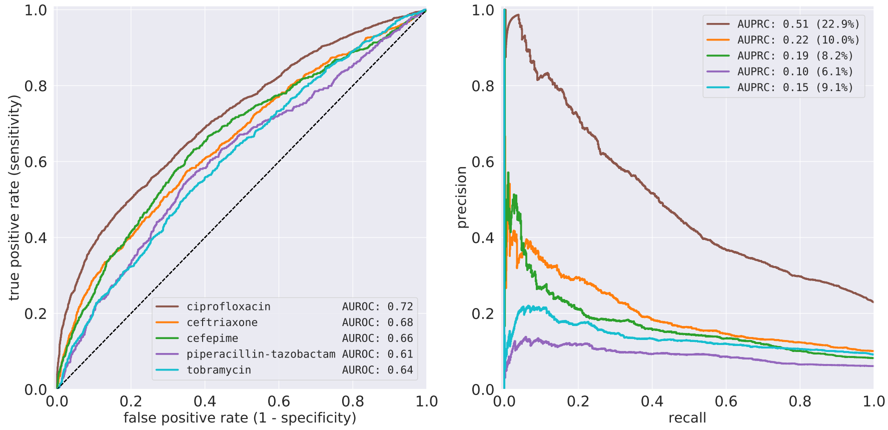
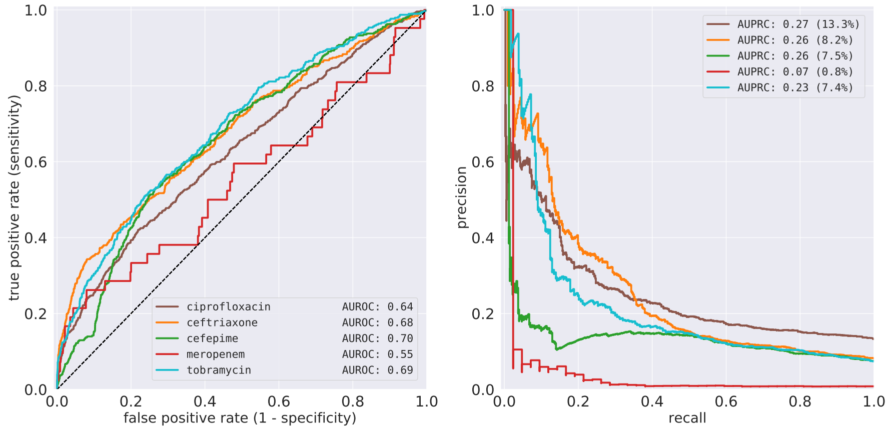
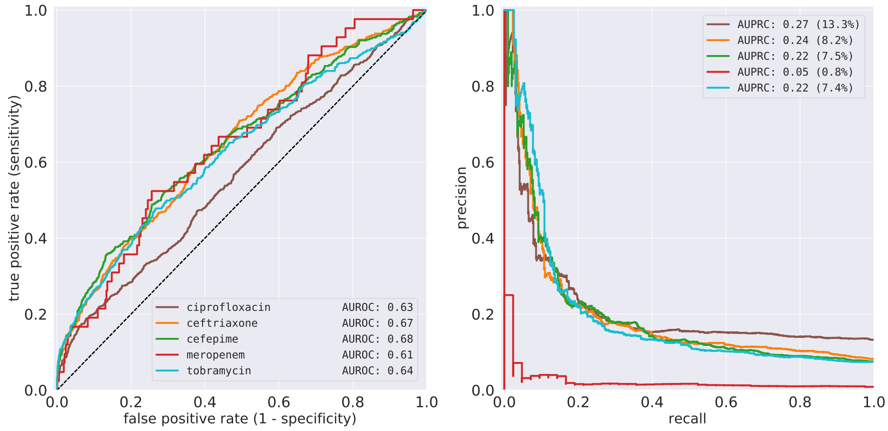


Figure 4.4: **ROC and precision-recall curves** of three machine learning models for resistance in *E. coli* to various different antibiotics. Figure adapted from Weis *et al.* [126].

(a) *K. pneumoniae* logistic regression



(b) *K. pneumoniae* LightGBM



(c) *K. pneumoniae* MLP

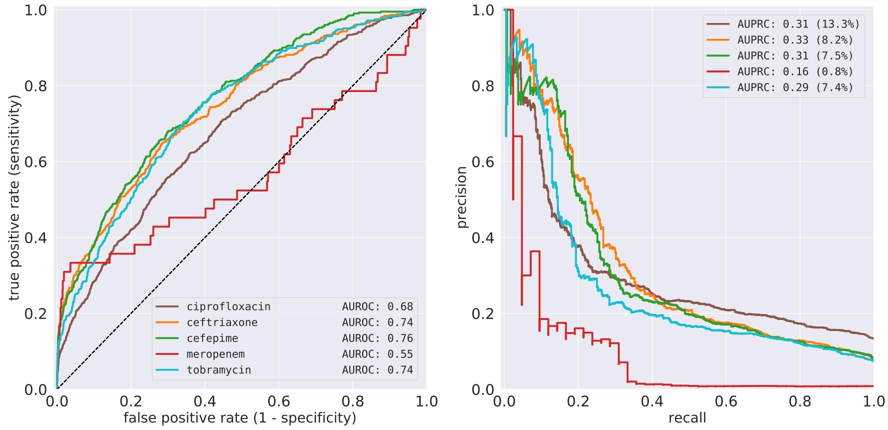
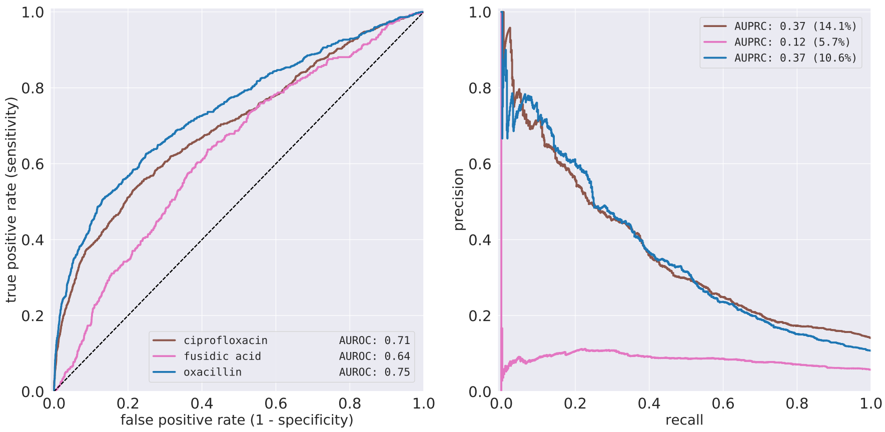
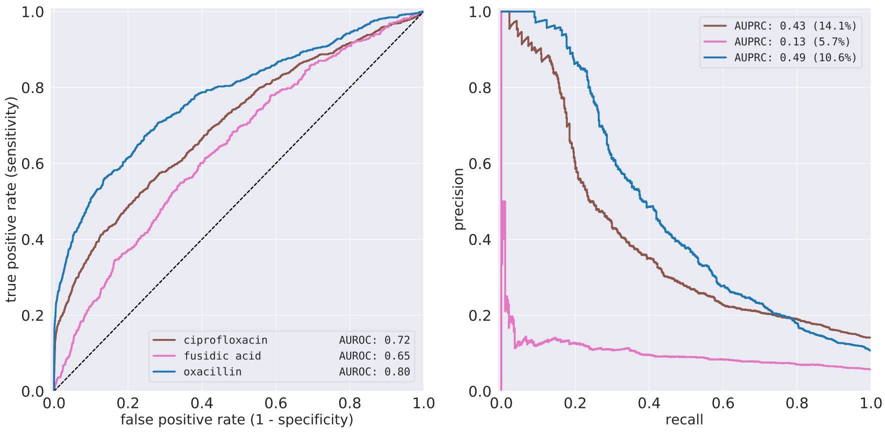


Figure 4.5: **ROC and precision-recall curves** of three machine learning models for resistance in *K. pneumoniae* to various different antibiotics. Figure adapted from Weis *et al.* [126].

(a) *S. aureus* logistic regression



(b) *S. aureus* LightGBM



(c) *S. aureus* MLP

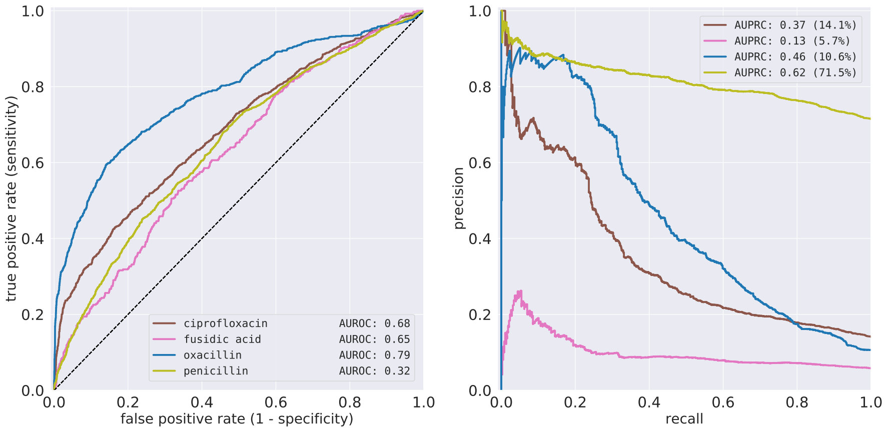


Figure 4.6: **ROC and precision-recall curves** of three machine learning models for resistance in *S. aureus* to various different antibiotics. Figure adapted from Weis *et al.* [126].

species	antibiotic	abbreviation	model	AUROC	AUPRC
E. coli	ceftriaxone	**E-CEF**	logistic regression	0.70	0.24
			LightGBM	**0.74**	0.30
			MLP	0.68	0.22
E. coli	ciprofloxacin	E-CIP	logistic regression	0.73	0.51
			LightGBM	**0.76**	0.60
			MLP	0.72	0.51
E. coli	cefepime	E-PIME	logistic regression	0.69	0.21
			LightGBM	0.73	0.24
			MLP	0.66	0.19
K. pneumoniae	ceftriaxone	**K-CEF**	logistic regression	0.68	0.26
			LightGBM	0.67	0.24
			MLP	**0.74**	0.33
K. pneumoniae	cefepime	K-PIME	logistic regression	0.70	0.26
			LightGBM	0.68	0.22
			MLP	**0.76**	0.31
K. pneumoniae	tobramycin	K-TOB	logistic regression	0.69	0.23
			LightGBM	0.64	0.22
			MLP	**0.74**	0.29
S. aureus	oxacillin	**S-OXA**	logistic regression	0.75	0.37
			LightGBM	**0.80**	0.49
			MLP	**0.79**	0.46
S. aureus	ciprofloxacin	S-CIP	logistic regression	0.71	0.37
			LightGBM	0.72	0.43
			MLP	0.68	0.37
S. aureus	fusidic acid	S-FAC	logistic regression	0.64	0.12
			LightGBM	0.65	0.13
			MLP	0.65	0.13

Table 4.1: **Species-stratified antimicrobial resistance predictors** for several antimicrobial drugs in *DRIAMS-A*. Three machine learning models are applied, with LightGBM performing consistently best for *E. coli* and *S. aureus* and MLP for *K. pneumoniae*. The scenarios selected based on predictive performance and clinical applicability are marked.

dictive performance displayed in Table 4.1. The selected machine learning models are therefore the LightGBM for E-CEF and S-OXA and a MLP for K-CEF.

The AUROC values reported in Table 4.1 indicate a high predictiveness for resistance in the respective scenarios. In the next step, we validate our initial assumption that resistance prediction on a combination of MALDI-TOF mass spectra from different species will not improve predictive performance. Arguments can be made for both an improvement or a decrease in prediction accuracy. As explained in the beginning of the paragraph, different mechanisms cause resistances in different species. Therefore, pooling MALDI-TOF mass spectra across species and predicting antimicrobial resistance through a joint model regardless of input species poses a more complex learning task than merely predicting antimicrobial resistance for one specific species. However, stratifying the training dataset by species reduces the number of instances available for model training and can therefore have a negative impact on generalisation abilities of the trained predictor.

The following study is designed to analyse the trade-off between both arguments: We compare models that are trained on either (i) samples pooled across several species (termed 'ensemble'), or (ii) training sample stratified to only contain one bacterial species (termed 'single'), while subsampling the number of instances. The curves depicting the predictive performance of this ablation study are included in Figure 4.7, both in terms of AUROC and AUPRC. Each point on the curve corresponds to one predictor trained on the number of instances stated on the x–axis, with the rightmost point corresponding to all instances available for the respective scenario. The curves indicate that training on data stratified by species leads to performance improvement in all species when comparing same number of training instances. When comparing the performance at the largest possible number of samples, i.e. the rightmost point of each curve, the improvement persists for *E. coli* and *K. pneumoniae*, while converging to a similar value for *S. aureus*. It should be emphasised that all training instances used for training the predictor resulting in the rightmost point were also included in the pool of training instances used to train the rightmost ensemble predictor. As illustrated in Figure 4.7, despite having access to the same information and more than the rightmost predictor in the single setting, the predictor learning on all instances in the ensemble setting never reaches a higher performance for *E. coli* and *K. pneumoniae*. While the ensemble curves seemingly reach a plateau in the large-sample size scenarios, the species-stratified predictors increase more sharply with the last additions of more training instances, demonstrating the higher complexity of antimicrobial resistance prediction in the ensemble setting and the benefits of large numbers of training instances. Note that this experiment compares the prediction of samples of one specific species that is represented in the training data in high numbers.

It is not clear how a predictor is expected to predict resistance for an instances representing a species and resistance mechanism not included in the training dataset. The ablation study further consolidates the approach of focusing on species-stratified training datasets for MALDI-TOF MS based antimicrobial resistance prediction.

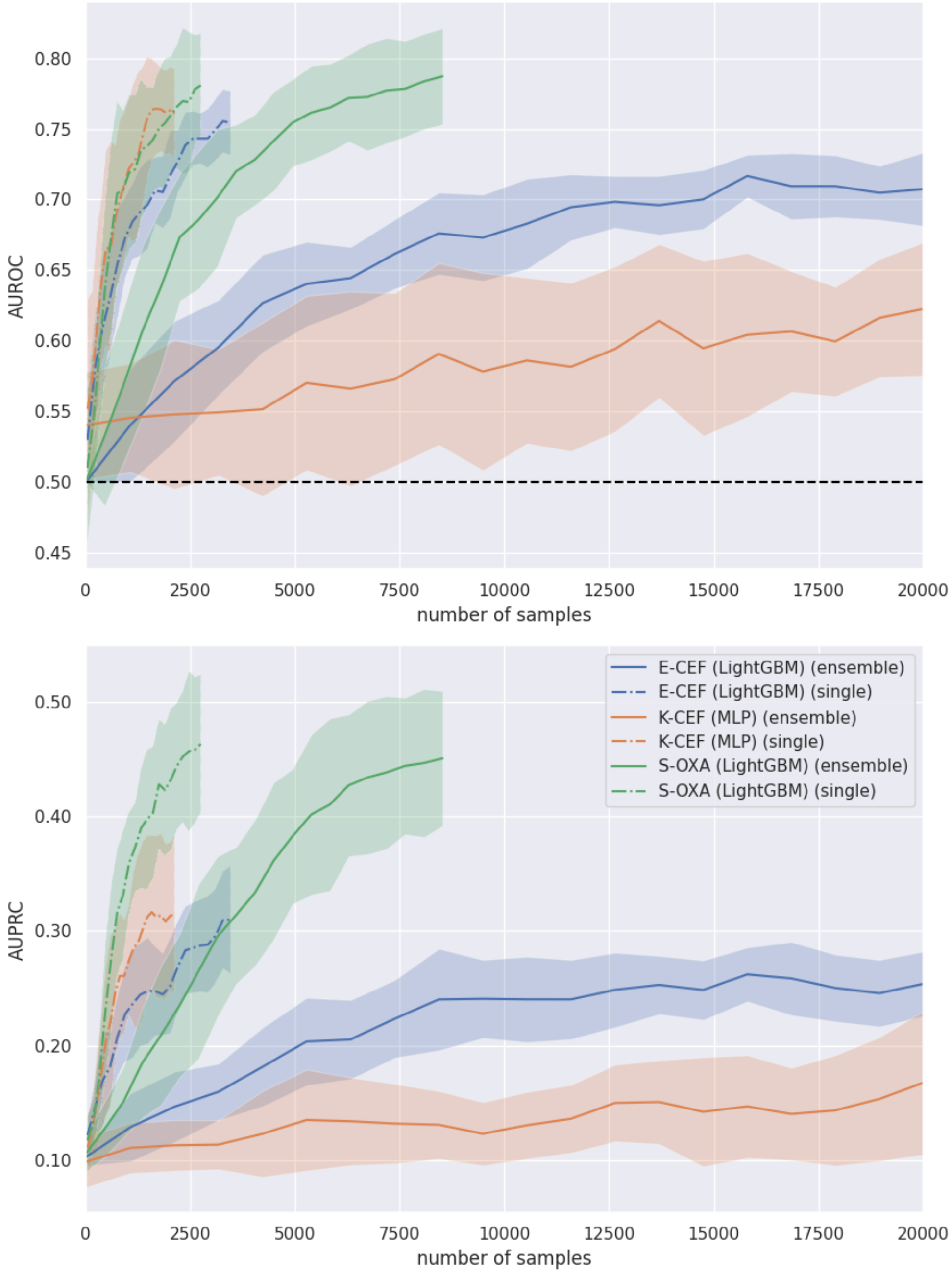
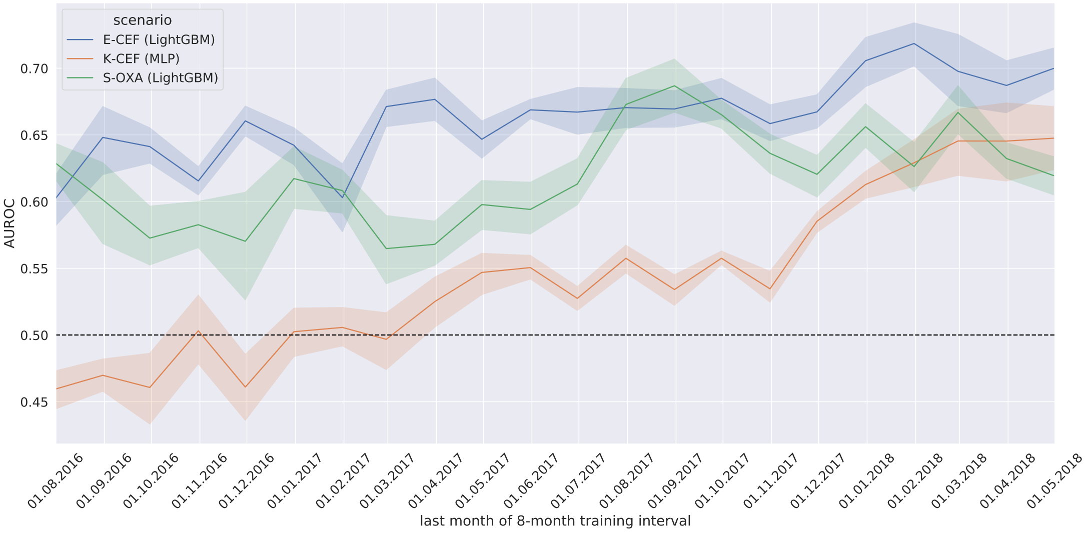


Figure 4.7: **Relationship between sample size and predictive performance** for training data consisting of a single species (dash-dotted line) vs. an ensemble (solid line) of species. The test data only includes MALDI-TOF mass spectra from the target species. The depicted species-antibiotic scenarios and models are selected based on Table 4.1. Curves trained on single-species data increase much more rapidly and often outperform even the full ensemble dataset. Figure adapted from Weis *et al.* [126].

(a) AUROC



(b) AUPRC

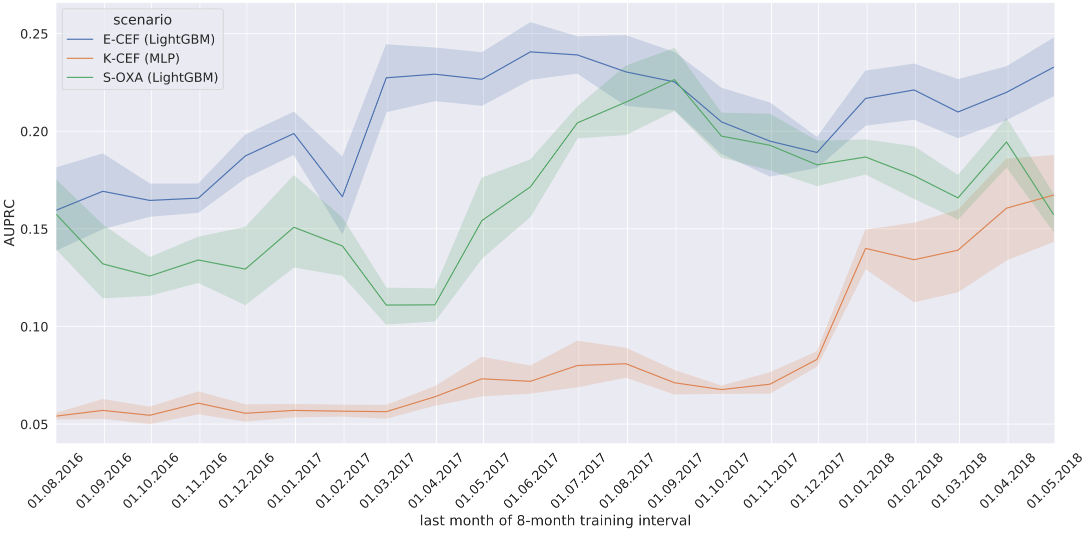


Figure 4.8: **Sliding eight-month training window illustrates advantage of contemporary samples**. Each point depicts the performance of a predictor trained on all samples collected during the eight month window, ending on the date indicated on the x-axis. The test data is comprised of all instances sampled in the four months starting on the rightmost date on the x-axis. The dates are given in the format *dd.mm.yyyy*. The curves indicate a performance decrease with larger distance between the test and training window. All instances stem from *DRIAMS-A*. Scenario abbreviations in the legend follow Table 4.1. Figure adapted from Weis *et al.* [126].

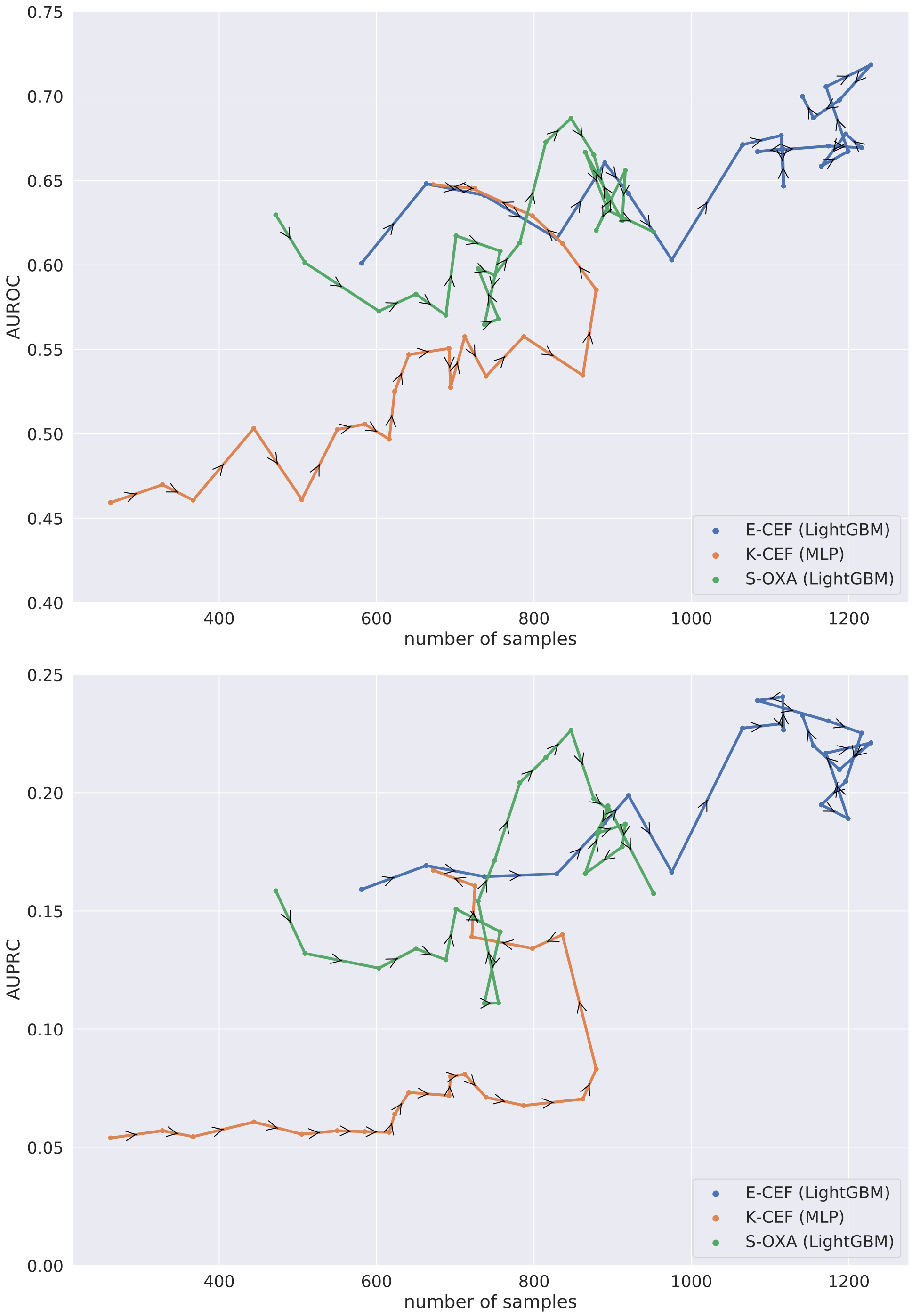


Figure 4.9: **Connection between number of training samples and predictive performance**. Each datapoint corresponds to a eight month training interval and value in Figure 4.8, with arrows indicating the direction of time passing. The progression towards the upper right corner indicates a correlation between increasing sample size and an increase in predictive performance. All instances stem from *DRIAMS-A*. Scenario abbreviations in the legend follow Table 4.1. Figure adapted from Weis *et al.* [126].

Impact of outdated mass spectra on prediction performance. MALDI-TOF mass spectra are subject to various influences that change over time, e.g. changes influencing the instrument (e.g. maintenance, laser replacement or adjustment of internal spectra processing parameters through machine calibration) or changing hospital policies on using MALDI-TOF MS (e.g. increasing the number of hospital divisions that rely on MALDI-TOF MS or acquiring a second instrument). The following experiment is designed to assess the degree to which prediction results differ for outdated MALDI-TOF mass spectra compared to contemporary ones. We studied how training on recently collected instances compares to training on data collected at an earlier date over an equal time range. To this end, we define a fixed test dataset comprised of all instances collected in the latest four months at *DRIAMS-A*. Training is conducted on all instances of an eight month time window sliding over the remaining months in *DRIAMS-A*. As the eight month training window is increasing in temporal distance to the four month test window, the results simulate the effects of using older instances for training. The instances in each training window are oversampled to match the class ratio of the test data. Due to oversampling dynamics and changes in MALDI-TOF MS usage over time, sample sizes can vary between training windows. Curves illustrating the predictive performance for each training window are depicted in Figure 4.8. The results indicate a slight decrease in predictive performance — both in AUROC and AUPRC — with increasing temporal distance between the train and test collection window. The decrease is particularly large for *K. pneumoniae*. Note that while the class ratio is constant for all training datasets, two factors are varying: the temporal distance to the test data and the number of instances. We add a more detailed analysis of the connection between all three factors—AUPRC, number of instances, and sample time—in Figure 4.9, plotting the progression of AUPRC and AUROC against the number of instances for each training window in Figure 4.8. The general trend of each line progressing to the upper right indicates that increasing predictive performance is correlated with recently collected samples and it is also connected to an increased sample size. The increased number of instances collected in more recent months is explained by an increased usage of the MALDI-TOF MS technology at *DRIAMS-A* over time. Nonetheless, upon close inspection of the curves, we see that for *K. pneumoniae*, the number of instances decrease in the most recent training windows, while both performance metrics increase. Overall these results highlight that more recent samples are most advantageous for accurate antimicrobial resistance prediction.

4.2 Calibrated classifiers for interpretable prediction scores

This section addresses a method highly relevant in the development of machine learning predictors aimed at healthcare applications — calibrating probability scores of binary classifiers. In most supervised classification settings, the trained classifier will assign a *prediction score* (or *predicted probability*) between 0.0 and 1.0 to each instance it is presented with. When the class for an instance must be determined, the class is assigned through a threshold t, often $t = 0.5$ by default, with the positive class

scenario	model	AUROC	AUROC (calibrated)	accuracy	accuracy (calibrated)
E-CEF	LR	70.22±3.36	70.22±3.36	81.74±3.95	**89.47**±0.60
	LightGBM	74.02±1.93	74.02±1.93	89.86±0.74	**89.91**±0.72
	MLP	67.52±3.40	67.52±3.40	87.66±1.20	**88.57**±0.94
K-CEF	LR	67.65±4.15	67.65±4.15	89.18±1.51	**92.15**±1.02
	LightGBM	66.96±2.92	66.96±2.92	91.97±1.21	**91.98**±1.32
	MLP	74.07±3.94	74.07±3.94	91.90±1.36	**92.14**±1.20
S-OXA	LR	74.89±4.06	74.89±4.06	82.78±6.93	**89.78**±1.14
	LightGBM	79.86±3.41	79.86±3.41	91.02±1.01	**91.29**±1.20
	MLP	78.72±3.05	78.72±3.05	89.69±1.05	**90.12**±1.06

Table 4.2: **Predictive performance comparison between non-calibrated and calibrated classifiers** reported in AUROC and accuracy. Accuracy values are improved in all cases through calibration. AUROC values are not affected by calibration (neither are AUPRC values — not depicted). Scenario abbreviations follow Table 4.1; logistic regression is abbreviated as LR.

assigned if the score is larger than t and the negative class otherwise (see Section 4.1.2 for more details). While one is intuitively led to interpret these scores as the probability with which an instance belongs to the positive class, this assumption is often not correct [76]. Many machine learning models suffer from biases distorting the relationship between prediction score and the true posterior class probability $P(\text{class}|\text{input})$, stemming e.g. from model assumptions that do not hold in reality [76]. Employing probability calibration can improve prediction accuracy, while also allowing for more precise interpretation of model scores, which is of importance when assessing the uncertainty and possible rejection of a model prediction. We briefly introduce the theoretical aspects of probability calibration, before continuing on to the effects of calibrated prediction scores in this chapter's analysis.

Probability calibration using Platt Scaling. Platt [87] proposed a transformation from support vector machine (SVM) prediction scores to posterior probabilities $P(\text{class}|\text{input})$ by applying a sigmoid function. Let the output of a trained predictor be $f(x)$, then the output is passed through a sigmoid function

$$P(y = 1|f, x) = \frac{1}{1 + exp(Af(x) + B)} \tag{4.4}$$

to obtain calibrated probabilities, with A and B being learnable parameters optimized through maximum likelihood estimation. The parameters are learned on an independent calibration dataset through 5-fold cross-validation. The need for an independent calibration dataset does not constitute a disadvantage as the same dataset can be used for model and calibration parameter optimisation.

Calibration improves classification accuracy. We analyse the influence of probability calibration on the three species-antibiotic scenarios chosen in the beginning of the chapter—E-CEF, K-CEF and S-OXA. A comparison of AUROC and accuracy values before and after calibration is depicted in Table 4.2. Neither AUROC nor AUPRC values are affected by probability calibration (influence on AUPRC not depicted in Table 4.2). This is easily explained, as both metrics are calculated by varying the decision thresholds along all possible thresholds and calculating the performance metrics for each possible threshold. The probability scores are only stretched and contracted along the $[0.0, 1.0]$ axis they all lie on with equal scores being transformed to the same calibrated score, but not changed in order. Applying each possible decision threshold leads to the same metrics as before calibration. The considered thresholds are defined by all probability scores in the test dataset, so in relative terms, there is no change in the dataset as the considered threshold changes along with the calibration. However, the case is different for the metric accuracy, which is based on contingency tables. Here, a single, fixed decision threshold is applied (by default $t = 0.5$) and the contingency table is calculated. Instances change their class assignment if calibration changes the 'side' of t they lie on. As the calibrated scores aim to reflect the true posterior class probability $P(\text{class}|\text{input})$ more accurately, their confusion matrix values improve. This improvement in accuracy is observed in Table 4.2 for all scenarios and models.

4.3 Interpreting mass peak contributions through Shapley values

For many applications, understanding the functionality of a machine learning model is equally important as accurate predictions, particularly in the context of clinical applications. Assessing information on not only *what* resistance label is predicted, but also *why* it is predicted, is essential to (i) gaining (new) biological insights into which signals in MALDI-TOF MS include information on antimicrobial resistance, (ii) ruling out confounding signals in the MALDI-TOF mass spectrum being used for predictions, and (iii) building trust in the predictions with medical professionals using the machine learning-based decision support. Many models directly produce importance measures that allow for direct analysis and interpretation as to which feature is given a certain weight by the model. For instance, linear models such as logistic regression directly weight each feature contribution to obtain an optimal output through parameters β. However, complex models such as neural networks have the ability to model intricate feature interactions. As a result, quantifying how each feature contributes to the model output becomes more difficult. In the context of this thesis, an additional challenge is that multiple prediction scenarios are analysed and model specific interpretation methods may vary between scenarios. In this section we employ a method quantifying the contribution of each feature to the model output that is suitable for any machine learning model — Shapley values. Subsection 4.3.1 gives an overview over the theoretical foundations continuing with results and a biological interpretation of MALDI-TOF mass peak contributions in Subsection 4.3.2.

4.3.1 Shapley values

In game theory, a common problem statement addresses the question of how to distribute gain and cost fairly to several players working in coalition [70]. This question can be transferred to the context of feature contributions in machine learning models by assuming that each feature value within a specific instance is a player in a game and the model output is the payout. Shapley values address how to fairly distribute the payout among the features [107]. Their value quantifies the contribution that each feature brings to the prediction made by the model. The Shapley value for feature j, which is denoted by ϕ_j, is defined as

$$\phi_j(v) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} \left(v(S \cup \{x_j\}) - v(S)\right) \tag{4.5}$$

where v is a value function, and S is a subset of features used in the predictor on a dataset X with p features $\{x_1,\ldots,x_p\}$. It describes the value that feature j contributed to the model outcome for the current sample compared to the overall model outcome for the entire dataset. While other concepts from coalition theory address the same problem (e.g. Banzhaf value [5]), Shapley values are the only method that satisfy all four axioms of a *fair payout*, namely (i) *efficiency*, (ii) *symmetry*, (iii) *null-player property*, and (iv) *additivity* [70]. Consider a dataset X with p features $\{x_1,\ldots,x_p\}$ with the value function v and model output function f: The efficiency axiom demands the the feature contributions over all features must add up to the overall difference between prediction output of x and the average prediction output:

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

The symmetry axiom states that interchangeable players receive equal payoffs, i.e. if for two features j and k

$$v(S \cup \{x_j\}) = v(S \cup \{x_k\}) \qquad \forall\, S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j, x_k\},$$

then

$$\phi_j = \phi_k.$$

If feature j is a null-player in the sense that it does not change the predicted value in any coalition then it should receive a Shapley value of 0, i.e.

$$v(S \cup \{x_j\}) = v(S) \qquad \forall\, S \subseteq \{x_1,\ldots,x_p\},$$

then

$$\phi_j = 0.$$

Lastly, the additivity axiom states that if two models are applied to the same features, the Shapley value of feature j of each individual model should add up to the value it would have received if the models would have been treated as a single combined model. The main disadvantage to employing Shapley values for feature importance interpretation is the high computation time.

4.3.2 Highly contributing feature bins can be associated with known MALDI-TOF mass peaks in literature

The Shapley values are determined for all scenarios and models chosen in Section 4.1.3. Figure 4.12 depicts the average (barplot) and sample-specific (scatter distribution plot) Shapley value for the 30 features receiving the highest average contribution. With the quick drop of average contributions, seen in the barplots on the left of each column, it is evident that three to ten m/z–bins contribute more to the model output than the remaining features. In the distribution plots to the right, it can be observed that the tails of each scatter plot for each feature are coloured with either the highest or lowest feature value. The predictor is using either the presence of a high intensity value (dark pink) or the absence of measured intensity (light blue) for positive class prediction. In the case of *S. aureus* it is particularly the presence of a MALDI-TOF mass peak that indicates a positive class prediction, while for *E. coli* and *K. pneumoniae* both presence and absence of certain mass peaks are indicative. Further, most feature bins included in Figure 4.12 stem from the lower half of the typical m/z–ratio of MALDI-TOF MS, i.e. less than 10,000 Da. This is in line with the properties of MALDI-TOF MS, as more mass particles are measured in the lower m/z–regime and thus more information is conveyed in the corresponding feature bins.

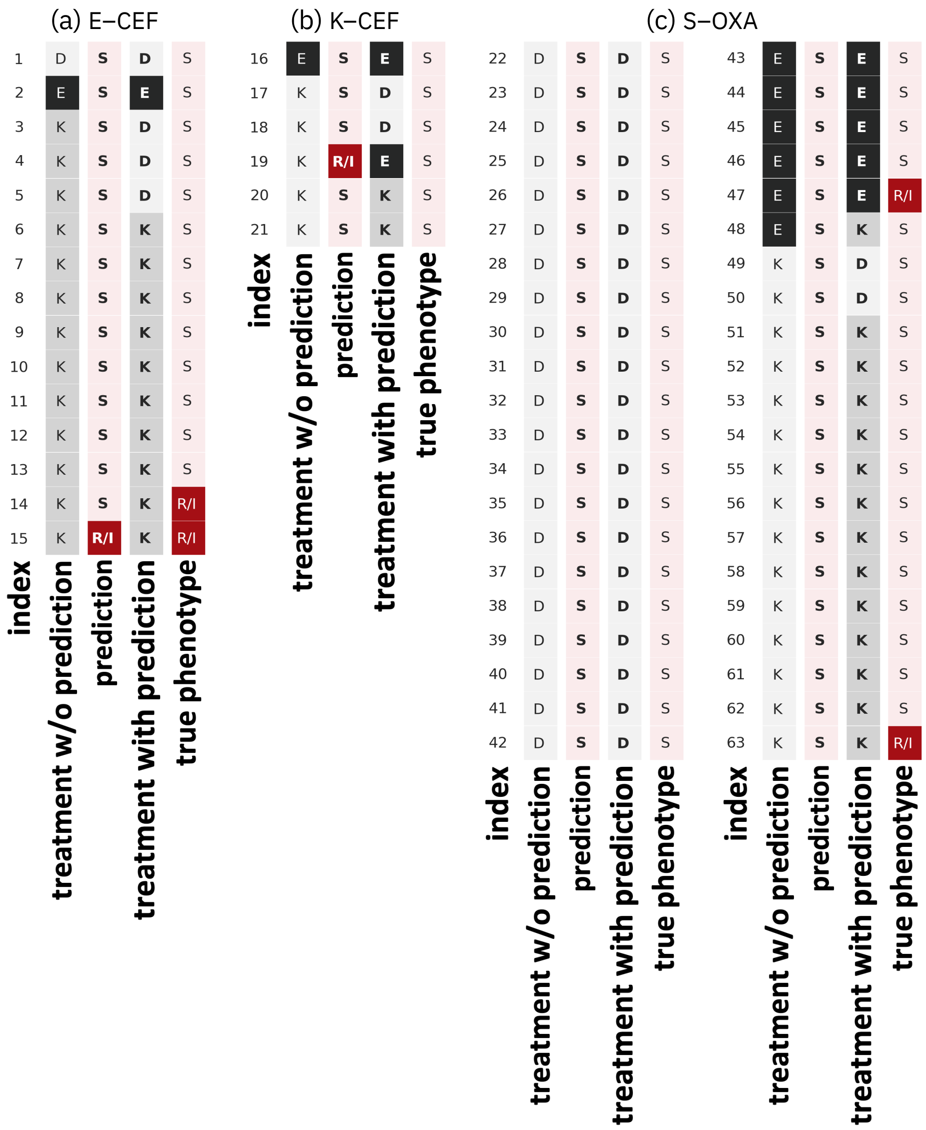
The primary aim of the current analysis is to develop some biological interpretation of the decision-making process of each antimicrobial resistance predictor. We therefore compare the feature bin contributing highly to the model outcome with MALDI-TOF mass peaks that have been shown to be associated with resistance in the literature. Several feature bins that contributed substantially to the *S. aureus* (oxacillin) and *E. coli* (ceftriaxone) classifiers can be annotated with proteins associated to the respective resistance identified in prior studies (Table 4.3). The corresponding feature bins of all MALDI-TOF mass peaks are marked with an asterisk in Figure 4.12. The largest number of mass peaks corresponding to feature bins are identified through studies aiming to identify resistant bacterial strains for oxacillin resistance in *S. aureus* based on MALDI-TOF MS. All prior work in Table 4.3 focused on either differentiation of MSSA and MRSA or differentiation of MRSA sublineages [16, 54, 80, 101, 111, 132, 139]. A number of identified discriminatory peaks correspond to housekeeping genes or other peptides, such as stress hormones or toxins [54]. Nine oxacillin resistance associated MALDI-TOF mass peaks described in the literature can be attributed to the top twenty contributing feature bins. The spread of new multidrug resistant strains in *E. coli* can be attributed to a few clonal lineages, e.g. sequence type (ST) 131 [66]. Three ST131-specific MALDI-TOF mass peaks have been identified in previous studies [58, 73], which can be attributed to five of the top ten most contributing feature bins. This congruence of Shapley values and independent literature references confirm the discriminatory power of highly contributing feature bins and underline their generalisability.

4.4 Retrospective clinical case study

With the aim of developing a clinically-applicable resistance predictor, we have to consider how the classifier would integrate and interact with the overall treatment decision-making progress. The process deciding on an antimicrobial therapy of a patient is in any case highly complex and influenced by several factors, such as the species of the infection, medical history and the current condition of the patient. In this section, we estimate how much influence the presence of a clinical resistance predictor could have on real-world hospital treatment and the resulting benefit.

The optimal way to analyse the effect of a MALDI-TOF MS based resistance predictor would be a prospective clinical study. However, the *DRIAMS-A* based classifier is not suitable for predictions on current patients, as accurate predictions require timely samples (see Section 4.1.3). Therefore, our clinical collaborators performed a retrospective clinical case study: applying a predictor to the most recent data contained in *DRIAMS-A*, analysing the patient records at the time of MALDI-TOF MS, judging whether they would have changed patient treatment and comparing the result to the antimicrobial phenotype determined later on. Our collaborating infectious diseases specialist (referred to as *clinician*) analysed patients with invasive and serious bacterial infections. The patients were in hospital care between May and August 2018; the last four months of *DRIAMS-A* data. From that time period, 63 patient cases with blood culture or deep tissue infections stemming from either *E. coli*, *K. pneumoniae* or *S. aureus* were reviewed by the clinician regarding antibiotic treatment. New classifiers were trained for each of the three common scenarios, on all *DRIAMS-A* data until end of April 2018, and applied to the MALDI-TOF mass spectra corresponding to the patient cases. For each of these 63 cases, our collaborators retrospectively reviewed the medical files to estimate whether a different antibiotic therapy would have been given, had the prediction been available at the time of MALDI-TOF MS. The predicted and true antimicrobial resistances, and the recommended treatments with and without the machine learning depictions, are depicted in Figure 4.10.

For the vast majority of patients—54 out of 63—the presence of the prediction would not have changed the antibiotic treatment recommended at the time of MALDI-TOF MS: For 22 patients, a deescalation strategy towards a narrow-spectrum antibiotic was suggested, while for 25 patients, the given antibiotic regimen would have continued, and for seven patients an escalation of the antibiotic treatment to a broad-spectrum antibiotic would have been ordered. These 54 patients include three cases in which the classifier predicted the susceptible class, while the phenotypic testing concluded an antimicrobial resistance later on. However, none of these false predictions would have caused a treatment less effective than without the algorithm: Two patients had a known MRSA colonisation, which our collaborating clinician prioritised higher than the machine learning prediction. For the third patient, two species—*E. coli* and *K. pneumo-*



(a) E–CEF

index	treatment w/o prediction	prediction	treatment with prediction	true phenotype
1	D	**S**	**D**	S
2	E	**S**	**E**	S
3	K	**S**	**D**	S
4	K	**S**	**D**	S
5	K	**S**	**D**	S
6	K	**S**	**K**	S
7	K	**S**	**K**	S
8	K	**S**	**K**	S
9	K	**S**	**K**	S
10	K	**S**	**K**	S
11	K	**S**	**K**	S
12	K	**S**	**K**	S
13	K	**S**	**K**	S
14	K	**S**	**K**	R/I
15	K	**R/I**	**K**	R/I

(b) K–CEF

index	treatment w/o prediction	prediction	treatment with prediction	true phenotype
16	E	**S**	**E**	S
17	K	**S**	**D**	S
18	K	**S**	**D**	S
19	K	**R/I**	**E**	S
20	K	**S**	**K**	S
21	K	**S**	**K**	S

(c) S–OXA

index	treatment w/o prediction	prediction	treatment with prediction	true phenotype
22	D	**S**	**D**	S
23	D	**S**	**D**	S
24	D	**S**	**D**	S
25	D	**S**	**D**	S
26	D	**S**	**D**	S
27	D	**S**	**D**	S
28	D	**S**	**D**	S
29	D	**S**	**D**	S
30	D	**S**	**D**	S
31	D	**S**	**D**	S
32	D	**S**	**D**	S
33	D	**S**	**D**	S
34	D	**S**	**D**	S
35	D	**S**	**D**	S
36	D	**S**	**D**	S
37	D	**S**	**D**	S
38	D	**S**	**D**	S
39	D	**S**	**D**	S
40	D	**S**	**D**	S
41	D	**S**	**D**	S
42	D	**S**	**D**	S

index	treatment w/o prediction	prediction	treatment with prediction	true phenotype
43	E	**S**	**E**	S
44	E	**S**	**E**	S
45	E	**S**	**E**	S
46	E	**S**	**E**	S
47	E	**S**	**E**	R/I
48	E	**S**	**K**	S
49	K	**S**	**D**	S
50	K	**S**	**D**	S
51	K	**S**	**K**	S
52	K	**S**	**K**	S
53	K	**S**	**K**	S
54	K	**S**	**K**	S
55	K	**S**	**K**	S
56	K	**S**	**K**	S
57	K	**S**	**K**	S
58	K	**S**	**K**	S
59	K	**S**	**K**	S
60	K	**S**	**K**	S
61	K	**S**	**K**	S
62	K	**S**	**K**	S
63	K	**S**	**K**	R/I

Figure 4.10: **Retrospective clinical case study**. The 63 patient cases are grouped by species. Treatment recommendations are displayed in grey tones, with escalation (E) highlighted in black, deescalation (D) in light grey and continuation (K) of the current treatment in grey. The resistance class is indicated by red panels, with the positive class (R/I) is highlighted in dark red. Both the predicted resistance and the treatment decisions employing the prediction are highlighted in bold font. Figure adapted from Weis *et al.* [126].

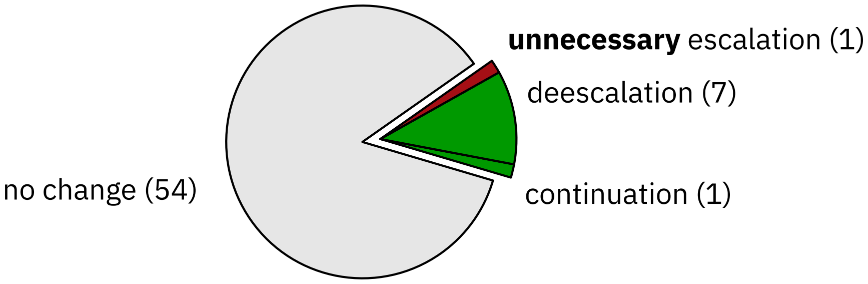


Figure 4.11: **Summary of clinical impact of predictor**. In 54 cases the algorithm would have made no difference to the treatment decision. For nine cases, the treatment recommendation by the specialist is influenced by the predictor: For seven patients the antibiotic therapy was correctly deescalated and in one clinical case the classifier caused a changed in treatment decision from escalation to continuation of therapy. For one patient however, the algorithm lead to an unnecessary escalation of antibiotic therapy. In summary, the retrospective clinical case study indicates that in eight out of nine cases (89%) the resistance predictor would have induced a beneficial treatment decision.

niae—were found in the blood culture samples. The clinician recommended to keep the antibiotic treatment for *E. coli* with or without the machine learning prediction, as no indication for resistance is present in either case, and escalation to a broad-band antibiotic was implemented after obtaining the phenotypic resistance.

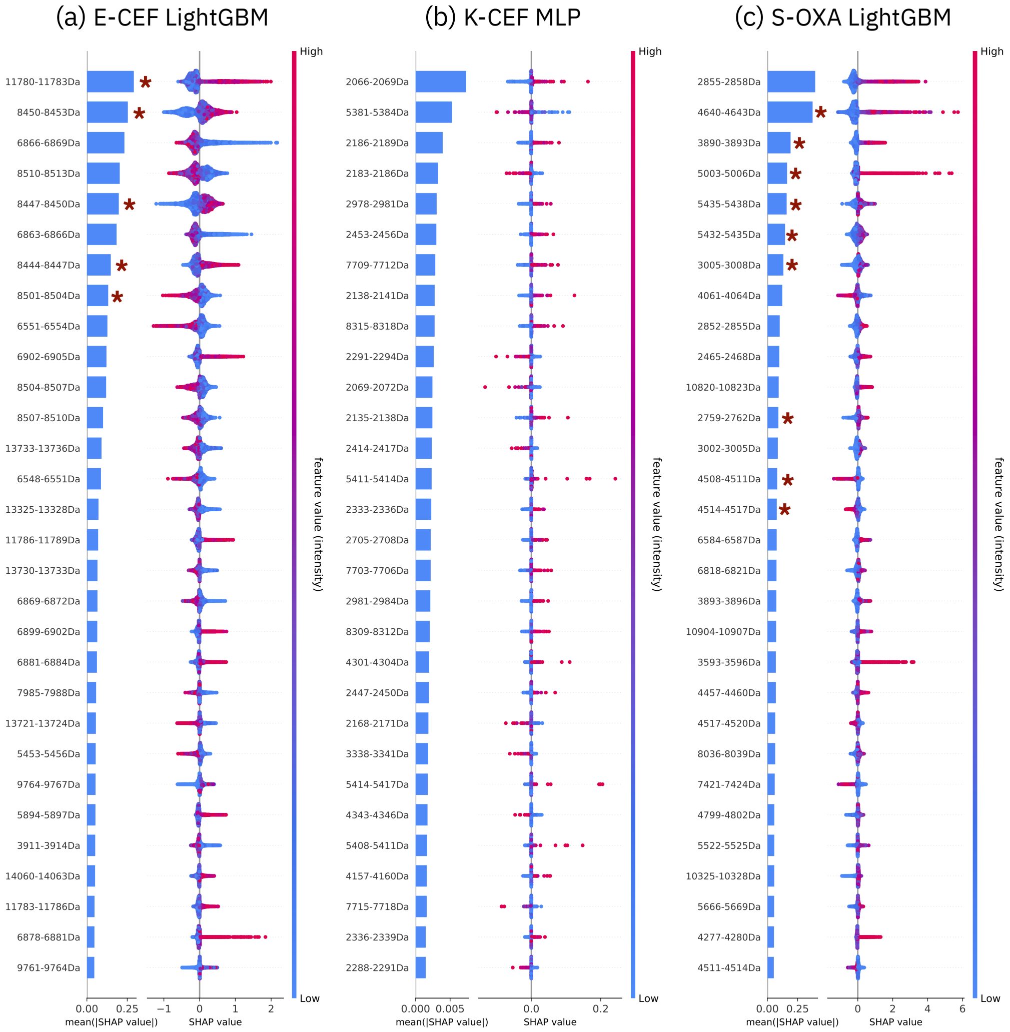
In the remaining nine patients, the presence of the machine learning prediction would have changed the treatment recommendation by the clinician at the time of MALDI-TOF MS: For seven patients (cases 3, 4, 5, 17, 18, 49 and 50 in Figure 4.10), considering the prediction would have caused the clinician to recommend a deescalation of therapy. In one other patient (case 48), the prediction led the clinician to keep the current antibiotic therapy, whilst the clinician would have suggested escalation to a broad-band antibiotic without the classifier. Together, these eight patients would have benefited from the presence of the MALDI-TOF MS based classifier. However, for the one remaining patient (case 19), a false-positive prediction of the resistance prediction would have led to an unnecessary escalation of antimicrobial therapy. In summary, in eight of the nine patients the machine learning guidance led to a better adjusted antimicrobial treatment, while for one patient an unnecessarily extensive antibiotic therapy would have been administered. This impact of the predictor is summarized in Figure 4.11.

4.5 Summary and discussion

This chapter introduced the first MALDI-TOF MS based antimicrobial resistance prediction study applying a comprehensive machine learning analysis pipeline on a large-scale clinical dataset. The results demonstrate that common machine learning models lead to accurate results on species-stratified single-antimicrobial resistance prediction tasks. Further, both focusing the task on species-specific prediction and increasing the number of samples leads to an improvement in predictive performance. Probability calibration was introduced to further improve prediction accuracy and steer the machine learning pipeline towards the development of a clinically-applicable tool. Lastly, Shapley values confirm that biologically meaningful signals are contributing to the predictions and indicating generalisability as the same MALDI-TOF mass peaks were detected to be associated to resistance in independent studies.

After establishing this baseline, several directions of research emerge or remain open. As good predictive accuracy could be reached with out-of-the-box machine learning models, development of MALDI-TOF MS specific classifiers hold the promise of capturing additional information and lead to further performance increases. Specifically the information loss during MALDI-TOF mass peak binning is of interest, and model working on a peak-based spectral representation should be investigated. Further, a comprehensive assessment of the generalisability of antimicrobial resistance prediction across hospital sites is necessary.

The Shapley values in Section 4.3.2 determined highly contributing feature bins for which the discriminatory potential has not been identified in previous studies. An investigation into the protein identity of these uninterpreted feature bins requires additional experimental research, but would be desirable in future work.



(a) E-CEF LightGBM

(b) K-CEF MLP

(c) S-OXA LightGBM

Figure 4.12: **MALDI-TOF MS feature bins with highest Shapley values** for the three antibiotic-species scenario and model selected in 4.1.3. The feature bins were cross-referenced with mass peaks described to be associated with the respective resistance in the literature. Bins corresponding to a known MALDI-TOF mass peak are marked with a red asterisk and can be found in Table 4.3. Note that there were no ceftriaxone resistance associated MALDI-TOF mass peaks for *K. pneumoniae* found in the literature. Scenario abbreviations follow Table 4.1. Figure adapted from Weis *et al.* [126].

scenario	feature bin m/z range [Da]	rank Shapley	m/z value reference	target as identified and stated in reference
S-OXA	2,759 to 2,762	12	2,762 Da	dominant lineages within MRSA (ST5) [139]
			2,760 Da	discrimination between MSSA and MRSA [111]
S-OXA	3,005 to 3,008	7	3,007 Da	main clonal lineages (CC1) [54]
S-OXA	3,890 to 3,893	3	3,891 Da	CC5, CC97 [80]
			3,891 Da	main clonal lineages (CC5, CC25) [54]
S-OXA	4,508 to 4,511	14	4,511 Da	major lineages within MRSA (CC45, CC30) [132]
			4,511 Da	CC30, CC45, CC398, ST88 [54]
S-OXA	4,514 to 4,517	15	4,514 Da	MRSA clonal complexes (CC398) [16]
S-OXA	4,640 to 4,643	2	4,641 Da	major lineages within MRSA (CC8, CC22) [132]
			4,641 Da	discrimination between MSSA and MRSA [111]
S-OXA	5,003 to 5,006	4	5,002 Da	major lineages within mMRSA (CC22) [132]
			5,004 Da	MRSA clonal complexes (CC22) [16]
			5,002 Da	CC22 [80]
			5,002 Da	main clonal lineages (CC22) [54]
S-OXA	5,432 to 5,435 5,435 to 5,438	6 5	5,437 Da	major lineages within MRSA [132] (CC5, CC45, CC22, CC8, ST1, ST15, ST80)
			5,440 Da	MSSA CC98 [101]
E-CEF	8,501 to 8,504	8	8,496 Da	ST131 [58]
E-CEF	8,444 to 8,447 8,447 to 8,450 8,450 to 8,453	7 5 2	8,448 Da	ST131 [73]
E-CEF	11,780 to 11,783	1	11,783 Da	ST131 [73]

Table 4.3: **MALDI-TOF mass peaks found to be correlated with resistance in literature can be attributed to highly contributing feature bins**. The column 'rank amongst Shapley' states the position among the 30 highest contributing feature bins regarding Shapley values in Figure 4.12. The rightmost column states the specific target, through which the MALDI-TOF mass peak is associated with the resistance. In most cases, this target is a species substrain that harbours a resistance mechanism. Note that only for *E. coli* (ceftriaxone resistance) and *S. aureus* (oxacillin resistance) relevant studies could be found in the literature. Abbreviations: Clonal complex is abbreviated with CC and sequence type with ST; scenario abbreviations follow Table 4.1.

Part III

Improving the predictive performance and transferability of MALDI-TOF MS based resistance prediction through kernel methods and representation learning

5 Kernel-based microbial phenotype prediction from MALDI-TOF mass spectra

C. Weis, M. Horn, B. Rieck, A. Cuénod, A. Egli, and K. Borgwardt. “Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra”. *OUP Bioinformatics* 36, 2020, pp. i30–i38.
doi: `https://doi.org/10.1093/bioinformatics/btaa429`

From this chapter onward, the nature of our analyses transitions towards leveraging MALDI-TOF mass spectra specific properties and tackling shortcomings hindering the clinical applicability of our resistance classifiers. We direct these efforts by referring to Chapter 2, which established the current state-of-the-art and research direction in MALDI-TOF MS based antimicrobial resistance prediction. The systematic review concluded that several important aspects of prediction pipelines are not addressed in any of the studies examined. These factors include the *preprocessing* of raw MALDI-TOF mass spectra and subsequent *feature representation*, the development MALDI-TOF MS specific machine learning models and confidence analysis of predicted resistance labels.

In this chapter, we take the first strides toward addressing the aforementioned shortcomings: (i) a new peak detection algorithm based on persistent topology is introduced, and different preprocessing techniques are compared and evaluated on several species and antibiotic resistance prediction scenarios in Section 5.1, (ii) a novel kernel, PIKE, specifically developed for MALDI-TOF MS based resistance classification is introduced in Section 5.2, and its application together with a Gaussian Process classifier is examined in Section 5.2.1, which enables (iii) reliable confidence estimates, which are compared of those of other classifiers when provided with out-of-distribution samples in Section 5.4. Contrasting the analysis of the previous chapter, MALDI-TOF mass spectra are represented by their individual mass peaks, not a vectorised representation. This has the advantage that the MALDI-TOF mass spectra representation retains its full accuracy in terms of $^m/_z$ and intensity value, which would be reduced during the vector binning step. As a downside, the mass peak representation is high-dimensional and not of fixed length, rendering many machine learning approaches infeasible. The code for the methods newly introduced in this chapter is publicly available. Please refer to Appendix V Software Availability for more details.

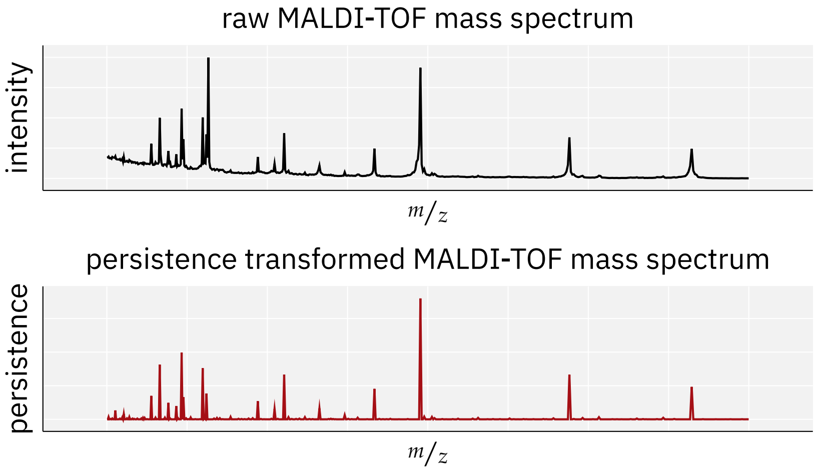


Figure 5.1: **Schematic illustration of the influence of the proposed persistence transformation on MALDI-TOF mass spectrum**. The top graph depicts a raw MALDI-TOF mass spectrum without any alignment or preprocessing. The bottom graph shows a persistence transformed spectrum produces a simplified and clean spectral representation, with the y-axis changing from an *intensity* to a *persistence*. Figure adapted from Weis *et al.* [128].

5.1 Topology-based peak detection

MALDI-TOF mass spectra preprocessing is commonly conducted using one of two softwares for this task, either the commercial software provided by MALDI-TOF MS manufacturers `ClinProTool` [12] or the open-source `R` software `MaldiQuant` [43]. Among the literature, a 'standard' preprocessing pipeline has transpired, consisting of several steps requiring a number of parameter choices. This pipeline is described in Chapter 3 and was employed to transform MALDI-TOF mass spectra for the preprocessed spectra in *DRIAMS*, which is used for all analysis in Chapter 4. In general, preprocessing and a subsequent binning steps produces feature vectors of fixed length, the required input format for the majority of standard machine learning techniques. However, during binning close-by MALDI-TOF mass peaks are summarized and information about precise m/z values is lost. This section introduces a new peak detection algorithm, based on the concept of persistence from computational topology. We compare the influence of the established preprocessing pipeline and the simpler persistence transformation on the task of antimicrobial resistance prediction, both employing logistic regression (requiring a binning before data read-in) and a Gaussian Process able to handle varying length inputs.

Peak calling employing persistence transformation. Inspired by the concept of *persistence* from computational topology [30], a simple peak detection algorithm determines the intensity of each peak above baseline signal. Formally, let $\mathbb{D} \subseteq \mathbb{R}^d$ be a compact domain and f a scalar function $f\colon \mathbb{D} \to \mathbb{R}$. Critical points of f, i.e. maxima, minima, and saddle points, are paired up with each other using the principle of persistence. Each maximum is paired with a minimum or saddle point, depending on the relation of the critical points to each other. The principle is best illustrated by determining the prominence of every mountain peak in a mountain range, where each peak is paired with the highest valley between the itself and an higher peak. In order to formally determine the pairing for each peak, the superlevel sets of f have to be analysed, i.e. sets of the form $\mathcal{L}_f^+(c) := \{x \in \mathbb{D} \mid f(x) \geq c\}$ for $c \in \mathbb{R}$. If the superlevel set $\mathcal{L}_f^+(c)$ is not empty, two points $(x, f(x))$ and $(x', f(x'))$ are said to be connected in $\mathcal{L}_f^+(c)$ if the path between them is a subset of $\mathcal{L}_f^+(c)$. The connection is denoted by $x \sim_c x'$. Considering that $\mathcal{L}_f^+(c) \subseteq \mathcal{L}_f^+(c')$ for $c' \leq c$, all points that satisfy $x \sim_c x'$ also satisfy $x \sim_{c'} x'$ for all $c' \leq c$. Therefore, it is sufficient to find the largest value of c that connects the two points, which is then referred to as the *partner* of x. Each point $(x, f(x))$ is paired to its partner by evaluating the following pairing function $\pi_f \colon \mathbb{D} \to \mathbb{R}$:

$$x \mapsto \sup\{c \leq f(x) \mid \exists x' \neq x \colon f(x') \geq f(x) \wedge x \sim_c x'\}. \tag{5.1}$$

While π_f maps each point $x \in \mathbb{D}$ to a c such that a point with a higher function value from x within $\mathcal{L}_f^+(c)$ exists, at the global maximum no such point can be reached and we set $\sup \emptyset := \min_x f(x)$. Circling back to the intuitive image of calculating the *topographic prominence* of a peak in mountaineering: a point x has a low prominence if $\pi_f(x) \approx f(x)$, and a point x has high prominence if $\pi_f(x) \ll f(x)$.

For $d = 1$, a prominence map $\mathcal{D}_f \colon \mathbb{D} \to \mathbb{R} \times \mathbb{R}$ can be constructed via

$$x \mapsto \big(f(x), f \circ \pi_f(x)\big), \tag{5.2}$$

mapping each $x \in \mathbb{D}$ to a point in the Euclidean plane. Letting x be a point in $\mathbb{D}$ and $\mathcal{D}_f(x) = (a, b)$, its persistence is defined through $|a - b|$ and we denote it by $\operatorname{pers}(x)$. The case for MALDI-TOF MS data (where $d = 1$) has two advantages, namely (i) the calculations are of computational complexity $\mathcal{O}(n \log n)$, with n denoting the number of measurement points in the MALDI-TOF mass spectrum, and (ii) the persistence values can be seen as a *direct* transformation of the MALDI-TOF mass spectrum. For any spectrum $f \colon \mathbb{R} \to \mathbb{R}$, most of the points will be mapped to the trivial values, i.e. directly to the diagonal by $\mathcal{D}_f$ — only critical points of f, i.e. maxima, minima, and saddle points, will receive non-trivial, i.e. non-zero, persistence values.

Persistence transformation. Let $x \in \mathbb{R}$ be a point in the domain of a (MALDI-TOF mass) spectrum f, we transform it to its persistence values so that we obtain a new transformed spectrum $\widetilde{f}$ with $\widetilde{f}(x) := \operatorname{pers}(x)$. A illustration of this *persistence transformation* is depicted in Figure 5.1. This process automatically produces a peak detection because local maxima get assigned a large persistence value. A sparse spectral representation can be constructed by considering only the k largest peaks and their m/z position. This representation forms a nested sequence of subsets for increasing values of k; each subset is a set of k tuples with values from $\mathbb{R}^2$. Sections 5.2 and 5.2.1 will introduce a classifier specifically developed for MALDI-TOF mass spectra that utilizes sparse input representations.

5.2 GP-PIKE: A kernel method designed for MALDI-TOF mass spectra

Representing the spectra by individual mass peaks, instead of a binned and vectorised representation, has the advantage of retaining its full accuracy in m/z and intensity values. Therefore, machine learning techniques that are capable of handling a sequence of mass peak pairs as input may lead to gains in classification performance. However, this spectral representation is of varying length, rendering many machine learning approaches non-applicable. This section introduces a new kernel specifically developed for MALDI-TOF mass peak input—PIKE, the Peak Information Kernel—which will be combined with a Gaussian Process classifier in a later section. To the best of our knowledge, this is the first machine learning method specifically developed for the task of antimicrobial resistance prediction from MALDI-TOF MS.

A kernel is a function that quantifies the similarity of objects by evaluating the inner product in a *reproducing kernel Hilbert space* (RKHS) [105]. Kernel methods are popular in numerous application domains, including computational biology [10, 106], due to their versatility and expressivity. Their application context ranges from classification to regression, data formats from graphs to text, since the infinite-dimensional RKHS is able to capture the nuances in the data.

While kernel approaches have been developed to be prominent in many domains, few kernel methods exploit information in mass spectrometry data. Several kernels designed for metabolomics information from mass spectrometry measurements exist [136], but their input format is restricted to structured feature vectors. Another kernel [11] was designed for comparing spectra, but requires additional information in the form of fragmentation trees, i.e. details about the molecule's mass spectrometry fragmentation process. Many other kernels rely on the existence of such a tree [29, 47, 108]—however, this information cannot be obtained in the domain of MALDI-TOF MS. The kernel introduced in this chapter can employ spectral peaks (or peak subsets) directly and does not require any additional information beyond the spectra themselves.

PIKE: the Peak Information Kernel. PIKE is motivated by heat diffusion on structured objects [6, 93] and is capable of capturing interactions between spectral peaks. The kernel is designed to process sets of tuples and does not required a fixed-length feature vector. Each (MALDI-TOF mass) spectrum is denoted by a set of tuples $S := \{(x_1, \lambda_1), (x_2, \lambda_2), \ldots\}$, where $x \in \mathbb{R}_{>0}$ is a m/z value, and $\lambda_i \in \mathbb{R}_{>0}$ an intensity. Let δ_x

be a Dirac delta function centred at x and $u(x,t)$, where $u\colon \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}$. The solution to the following heat diffusion partial differential equation is:

$$\frac{\partial u}{\partial t} = \nabla^2 u \tag{5.3}$$

$$\lim_{t \to 0} u(x,t) = \sum_i \lambda_i \delta_{x_i}, \tag{5.4}$$

While Dirac delta functions are not $L^2(\mathbb{R})$ functions, they can be approximated by them. Therefore, we write the boundary condition in Equation 5.4 as a limit, meaning that each spectrum is represented as a sum of Dirac delta functions, where the scale factors $\lambda_i \in \mathbb{R}_{>0}$ correspond to the intensity of a peak. This partial differential equation affords a closed-form solution [97]

$$u(x,t) = \frac{1}{2\sqrt{\pi t}} \sum_i \lambda_i \exp\left(-\frac{(x-x_i)^2}{4t}\right), \tag{5.5}$$

where $u(x,t) \in L^2(\mathbb{R})$ as each individual functions is square-integrable and $L^2(\mathbb{R})$ is a Hilbert space, closed with respect to addition of functions. $u(x,t)$ can also be seen as a *feature map* of the kernel, i.e. a map from a function space into $L^2(\mathbb{R})$. Let S be a (MALDI-TOF mass) spectrum and $t \in \mathbb{R}$, this feature map is described by $\Phi_t(S) := u_S(x,t)$, where the indexed S indicates that the spectrum was used as an input. The parameter t acts as a smoothing factor in $\Phi_t(S)$ that controls the influence peaks in the spectrum, as illustrated in Figure 5.2. With larger values of t, the spectrum becomes increasingly smooth, with individual measurements becoming less pronounced. The feature map is used as a kernel for calculating the similarity between two spectra by calculating the inner product of $L^2(\mathbb{R})$. For two given spectra S and S', potentially of different lengths, with m/z values x_i and x'_j and the respective intensities λ_i and λ'_j, the inner product is

$$\mathrm{k}_t(S,S') := \langle \Phi_t(S), \Phi_t(S') \rangle_{L^2(\mathbb{R})} := \int_{\mathbb{R}} \Phi_t(S)\Phi_t(S')\,\mathrm{d}x, \tag{5.6}$$

for which the closed-form solution is

$$\mathrm{k}_t(S,S') = \frac{1}{2\sqrt{2\pi t}} \sum_{i,j} \lambda_i \lambda'_j \exp\left(-\frac{\left(x_i - x'_j\right)^2}{8t}\right). \tag{5.7}$$

Equation 5.7 is a sum of exponential functions of a squared Euclidean distance with positive weights λ_i and λ'_j. It is positive definite, and thus a valid kernel [38]. Additionally, we need to ensure that each intensity $\lambda \geq 1$; otherwise, the value of $\lambda_i \lambda'_j$ decreases gradually, thereby decreasing the similarity between two spectra. In practise, a normalisation step can be applied to prevent this issue.

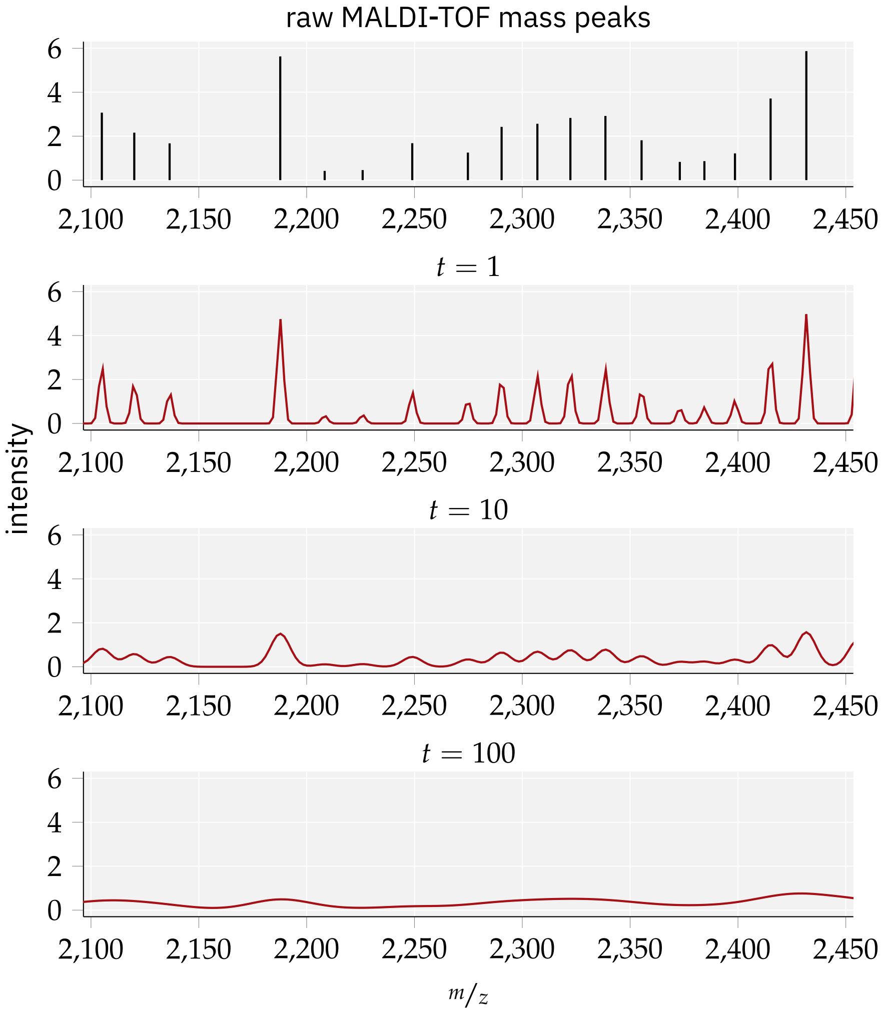


Figure 5.2: **Influence of kernel parameter** t **on feature map** $u(x, t)$ of a MALDI-TOF mass spectrum in its mass peak representation. The raw mass peaks are slowly diffused over the whole x-axis. With an increasing t the influence of a single peak is reduced. Figure adapted from Weis *et al.* [128].

A key property of PIKE is its capability to capture interactions between peaks in the spectrum. In Equation 5.7 the distances between all pairs of peaks, with one peak stemming from S and the other from S', are added. As a result, PIKE achieves the desired property of not requiring fixed-length feature vectors, but rather being able to operate on a flexible spectra representation of a set of tuples with potentially different cardinalities. This pairwise calculation reduces the kernel's scalability and PIKE cannot be readily applied to spectra consisting of thousands of mass peaks. However, this limitation does not arise in practice for MALDI-TOF mass spectra, as most of these spectra depict only hundreds of 'valid', i.e. non-noisy, mass peaks.

An additional advantage of PIKE is that it involves only one single parameter, the smoothing parameter t. Equation 5.7 is differentiable with respect to t. Therefore parameter t can be optimised by any classification model to obtain a kernel customized to the given problem domain. Keeping one spectrum S fixed, t can be optimised efficiently through

$$\frac{\partial\, \mathrm{k}_t(S,S')}{\partial t} = \sum_{i,j} \frac{\left(\left(x_i - x_j\right)^2 - 4t\right)}{8t^2}\, \mathrm{k}_t(S,S')_{[i,j]} \tag{5.8}$$

A step-by-step derivation of Equation 5.8 can be found in the next paragraph. In practice, having only a single parameter also simplifies the choice a final PIKE model: the overall model can be constructed by taking the mean t over all optimised t of the different training splits in a dataset. The benefit of this property is demonstrated in the next section.

Calculation of kernel derivative. The partial derivative Equation 5.8 is calculated by using the *product rule* formula. With definition of $f(t)$ and $g_{ij}(t)$ in Equation 5.7

$$\mathrm{k}_t(S,S') = \underbrace{\frac{1}{2\sqrt{2\pi t}}}_{f(t)} \sum_{i,j} \underbrace{\lambda_i \lambda'_j \exp\left(-\frac{\left(x_i - x'_j\right)^2}{8t}\right)}_{g_{ij}(t)}, \tag{5.9}$$

the product rule states the following:

$$\frac{\partial\, \mathrm{k}_t(S,S')}{\partial t} = \sum_{i,j}\left(\frac{\partial f(t)}{\partial t} g_{ij}(t) + f(t)\frac{\partial g_{ij}(t)}{\partial t}\right). \tag{5.10}$$

We calculate the partial derivatives

$$\frac{\partial f(t)}{\partial t} = -\frac{1}{4\sqrt{2\pi}t^{\frac{3}{2}}} \tag{5.11}$$

and

$$\frac{\partial g_{ij}(t)}{\partial t} = \frac{(x_i - x'_j)^2}{8t^2} g_{ij}(t). \tag{5.12}$$

By applying the product rule, we obtain

$$\frac{\partial\, \mathrm{k}_t(S,S')}{\partial t} = \sum_{i,j} \left(-\frac{g_{ij}(t)}{4\sqrt{2\pi} t^{\frac{3}{2}}} + \frac{g_{ij}(t)(x_i - x'_j)^2}{16\sqrt{2\pi} t^{\frac{5}{2}}} \right). \tag{5.13}$$

We now observe that $t^{\frac{3}{2}} = \sqrt{t}t$ and $t^{\frac{5}{2}} = \sqrt{t}t^2$, permitting us to rewrite (5.13) with the same denominator by multiplying the left term by $4t$:

$$\frac{\partial\, \mathrm{k}_t(S,S')}{\partial t} = \sum_{i,j} \frac{\left((x_i - x_j)^2 - 4t \right) g_{ij}(t)}{16\sqrt{2\pi} t^{\frac{5}{2}}}. \tag{5.14}$$

Notice how we switched the order of the two terms from (5.13) in order to obtain a nicer numerator. Finally, we split the denominator into a product containing $f(t) = 1/2\sqrt{2\pi t}$:

$$\frac{\partial\, \mathrm{k}_t(S,S')}{\partial t} = \sum_{i,j} \frac{\left((x_i - x_j)^2 - 4t \right) \cdot g_{ij}(t) \cdot 1}{8t^2 \cdot 2\sqrt{2\pi t}} \tag{5.15}$$

$$= f(t) \sum_{i,j} \frac{\left((x_i - x_j)^2 - 4t \right) g_{ij}(t)}{8t^2} \tag{5.16}$$

Equivalently, writing $\mathrm{k}(S,S')_{[i,j]}$ to denote only those terms of the kernel function that depend on i and j, we obtain the final derivative, as stated in Equation 5.8:

$$\frac{\partial\, \mathrm{k}_t(S,S')}{\partial t} = \sum_{i,j} \frac{\left((x_i - x_j)^2 - 4t \right)}{8t^2} \mathrm{k}_t(S,S')_{[i,j]}$$

5.2.1 The GP–PIKE method

Section 5.2 introduces a new kernel specifically designed to quantify the similarity between MALDI-TOF mass spectra, PIKE. We combine PIKE with a kernel based classification method, a Gaussian Process, to obtain a predictor tailored to MALDI-TOF MS based antimicrobial resistance prediction. The choice to employ a Gaussian Process for classification with the derived kernel is driven by its ability to (i) optimize the kernel hyperparameters through type II maximum likelihood, and (ii) recognize out-of-distribution samples due to well-calibrated confidence estimates. Flagging samples from an unknown population is particularly relevant in patient applications, where a method should notify practitioners if a prediction cannot be performed reliably.

Gaussian processes for classification. A Gaussian Process (GP) constitutes a stochastic process with every finite collection of variables following a multivariate Gaussian distribution. GPs can be applied as *lazy learners* for classification based on a kernel, i.e. a similarity measure between instances of the training data. Following Rasmussen *et al.* [92], a GP describes a distribution over functions $f(\mathbf{x})$ with $f\colon \mathcal{X} \to \mathcal{Y}$, the data domain being described by $\mathcal{X}$ and the prediction domain by $\mathcal{Y}$. A GP is completely characterized by its mean function $m(\mathbf{x})$ and its covariance (or kernel) function $\mathrm{k}(\mathbf{x}, \mathbf{x}')$. These functions are defined as:

$$\begin{aligned} m(\mathbf{x}) &:= \mathbb{E}[f(\mathbf{x})] \\ \mathrm{k}(\mathbf{x}, \mathbf{x}') &:= \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right] \end{aligned} \tag{5.17}$$

This definition can be represented by $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), \mathrm{k}(\mathbf{x}, \mathbf{x}'))$, defining a prior over functions, with a kernel k that captures functional variation over its domain. The conditional distributions in a GP are themselves Gaussian distributions and may thus be computed in a closed form. The goal is to compute the posterior distribution of function values f_* at test points X_* while conditioning on the training data X. In a regression task, the predictive distribution $\mathbf{f}_* | X_*, X, \mathbf{f}$ is computed. It can be can be written as a normal distribution, parametrised by a covariance matrix and described by a kernel function, between the samples in the training dataset and the test dataset, respectively.

This results in the Gaussian Process regression, where kernel and noise parameters are optimised according to the marginal likelihood of the model

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X) p(\mathbf{f}|X) \, \mathrm{d}\mathbf{f},$$

which can be computed analytically. Subsequently, the predicted mean and variance can be derived in closed form.

Next, we extend GPs to binary classification to be applicable in the given prediction scenario. Similarly to the logistic regression and MLP in Chapter 4, a sigmoidal function σ is laid over the latent function f_* in order to obtain class probability estimates. This results in a distribution of label predictions π_*. The prediction of a new sample is a two-step process. In the first step, the distribution of f_* at test points $\mathbf{x}_*$ is determined while conditioning on observed training data and labels

$$p(f_*|X, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|X, \mathbf{x}_*, \mathbf{f}) p(\mathbf{f}|X, \mathbf{y}) \, \mathrm{d}\mathbf{f}, \tag{5.18}$$

with $p(f|X, \mathbf{y})$ being the posterior probability over the latent variables. In the second step, the actual prediction will be obtained by passing the f_* values through σ and determining the expected label distribution, such that

$$p(y_* = 1|X, \mathbf{y}, \mathbf{x}_*) = \mathbb{E}[\pi_*] = \int \sigma(f_*) p(f_*|X, \mathbf{y}, \mathbf{x}_*) \, \mathrm{d}f_* \tag{5.19}$$

species	antibiotic	scenario	# samples	% positive
E. coli	amoxicillin / clavulanic acid	E-AMOXCLAV	1043	28.9
	ceftriaxone	E-CEF	1060	20.4
	ciprofloxacin	E-CIPRO	1051	29.7
K. pneumoniae	ceftriaxone	K-CEF	597	15.1
	ciprofloxacin	K-CIPRO	596	16.8
	piperacillin / tazobactam	K-PIPTAZO	576	13.9
S. aureus	amoxicillin / clavulanic acid	S-AMOXCLAV	973	13.7
	ciprofloxacin	S-CIPRO	987	14.7
	penicillin	S-PEN	941	71.4

Table 5.1: **MALDI-TOF mass peak dataset based on *DRIAMS-A*.** Nine species–antibiotic scenarios are included, and their scenario abbreviation, sample size and positive class ration stated. The positive class consists of resistant and intermediate samples.

However, these integrals cannot be computed in a closed form and therefore can only be approximated. Following the literature [92], we employ Laplace approximation, with the posterior $p(\mathbf{f}|X,\mathbf{y})$ in Eq. 5.18 being approximated by a Gaussian distribution around the posterior maximum. The GPs are trained through type II maximum likelihood optimisation on the training data, using the non-linear `L-BFGS-B` optimisation algorithm [82] for the kernel hyperparameters.

5.3 Evaluation of topological and kernel methods for MALDI-TOF mass-peak based resistance prediction

A sparse MALDI-TOF mass peak dataset. The analyses in this and the following chapter are based on a different data representation than the other ones in this thesis. This choice of data representation is based on several factors: (i) at a biological level the primary information in MALDI-TOF mass spectra is displayed in the peaks corresponding to proteins, (ii) the sparse spectral representation results in a compressed and memory efficient data format, while (iii) being more accurate as mass peaks tuples state the exact m/z–ratio value, as opposed to binned vectors. The created dataset is a subset of *DRIAMS-A*: 2676 MALDI-TOF mass spectra collected in the year 2018. It focuses on the same species as in previous experiments: *E. coli*, *K. pneumoniae* and *S. aureus*. Table 5.1 provides an overview of the dataset characteristics. The preprocessing followed the protocol described in Chapter 3. The antibiotic susceptibility phenotypes for the respective species are (i) amoxicillin/clavulanic acid, ceftriaxone, and ciprofloxacin resistance in *E. coli*, (ii) ceftriaxone, ciprofloxacin, and piperacillin/tazobactam resistance in *K. pneumoniae*, and (iii) amoxicillin/clavulanic acid, ciprofloxacin, and penicillin resistance in *S. aureus*. In this dataset, the positive (resistant) class constitutes the minority class for all species–phenotype combinations except for penicillin resistance in *S. aureus*. The list of antibiotics was expanded to obtain a comprehensive overview during the many performance comparisons in this chapter. Our aim is to provide a challenging classification scenario approximating real-world clinical applications as closely as possible.

Logistic regression and GP–PIKE experiments. All model performances are reported on a 5-fold cross validation using 80% training and 20% testing class-stratified splits. For logistic regression a further 5-fold cross validation is applied to the training split to determine the optimal model hyperparameters, which are then used to refit the model on the complete training data. This procedure is not necessary to optimise the GP hyperparameters, as they are derived through maximising the log marginal likelihood of the training data. For both methods, the class ratio differences are mitigated during training by oversampling the minority class, which in eight out of nine species–antibiotic scenarios is the positive class. Note that no oversampling is applied to the test data splits.

The logistic regression baseline requires a fixed-size feature vector for each MALDI-TOF mass spectrum. Similar to the full-spectrum binning described in Chapter 3.2, the feature vector is constructed through distribution of MALDI-TOF mass peaks into the bins of a histogram. In case two mass peaks are assigned to the same bin, their weight is accumulated. This results in a logistic regression classification pipeline consisting of peak binning, standardisation to zero mean and unit variance, followed by training the classifier using the following hyperparameter grid (i) number of bins (300, 600, 1800, and 3600), (ii) model regularisation (L_1, L_2, *elastic net*, and none), and (iii) regularisation penalty C (10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3} and 10^{4}). Our classification pipeline is implemented in `Python` using the `Pipeline` and `GridSearchCV` modules provided by the scikit-learn package [82].

5.3.1 MaldiQuant preprocessing pipeline performs more consistently than persistence transformation

We now compare the newly developed persistence transformation to the widely employed preprocessing pipeline implemented in `MaldiQuant`. The `MaldiQuant` pipeline is applied as described in Chapter 3.2. Additionally, we employ the provided peak detection algorithm with a signal-to-noise ratio of 2, with the noise estimated by the *MAD* method using a *half-window size* of $20\mathrm{Da}$.

The persistence transformation pipeline consists of applying the topology-based peak detection, a single normalisation step applying TIC normalisation (see Section 3.2), and subsequent extraction of the k largest peaks, with k as the only hyperparameter for the persistence transformation pipeline. We fixed the parameter to $k = 200$ peaks in this analysis for a fair comparison with the `MaldiQuant` preprocessing, which defines 216 peaks on average.

The quality of the preprocessing method is quantified through the predictive performance achieved by applying the same classification model to two differently preprocessed MALDI-TOF MS datasets. We infer that the dataset resulting in a higher pre-

scenario	PT / LR	MQ / LR	PT / GP–PIKE	MQ / GP–PIKE
E-AMOXCLAV	0.36±0.03	**0.41**±**0.07**	0.39±0.02	**0.47**±**0.04**
E-CEF	0.58±0.03	**0.63**±**0.06**	0.63±0.03	**0.71**±**0.03**
E-CIPRO	0.55±0.04	**0.61**±**0.09**	0.54±0.04	**0.68**±**0.03**
K-CEF	0.56±0.06	**0.58**±**0.10**	0.72±0.09	**0.77**±**0.07**
K-CIPRO	0.35±0.08	**0.42**±**0.10**	0.40±0.13	**0.55**±**0.10**
K-PIPTAZO	**0.39**±**0.09**	0.32±0.07	0.49±0.10	**0.56**±**0.10**
S-AMOXCLAV	**0.55**±**0.04**	0.53±0.04	0.61±0.12	**0.69**±**0.09**
S-CIPRO	0.26±0.06	**0.34**±**0.03**	0.31±0.03	**0.39**±**0.07**
S-PEN	**0.80**±**0.05**	0.80±0.03	0.81±0.02	**0.83**±**0.04**

Table 5.2: **Performance comparison of persistence transformation and tradition MALDI-TOF spectral preprocessing** through AUPRC±standard deviation using both logistic regression and GP–PIKE model. The predictive performance of predictions on `MaldiQuant`-preprocessed mass spectra is higher in most scenarios using the logistic regression model, and consistently higher for GP–PIKE. Scenario abbreviations follow Table 5.1. Other abbreviations: persistence transformation (PT), `MaldiQuant` [43] v1.19 (MQ), and logistic regression (LR).

dictive performance has been processed in a way, that the information contained in the spectrum is preserved and structured to a higher degree. The two models used for this assessment are logistic regression and GP–PIKE. The comparison of both preprocessing approaches are depicted in Table 5.2, reported by the mean and standard deviation AUPRC over the cross-validation splits. The results indicate that the conventional preprocessing pipeline produces spectra that are capable of high predictive performances on the task of antimicrobial resistance prediction. The logistic regression performance on the `MaldiQuant` preprocessed mass spectra (MQ–LR) leads to higher AUPRC values in most cases compared to our agnostic persistence transformation preprocessing (PT–LR). It is to be noted that in the majority of scenarios, the mean AUPRC of PT–LR is only slightly below that of MQ–LR. For S-AMOXCLAV and K-PIPTAZO, the topological method even outperforms MQ–LR. We note that the preprocessing by persistence transformation is conceptually simpler and may provide an alternative, also considering that additional steps can be combined with it. We conclude that the preprocessing procedure has a high influence on the results of the subsequent prediction task, and that the state-of-the-art pipeline *can* be outperformed. Further experiments on the influence of MALDI-TOF mass spectra preprocessing on applied machine learning pipelines are necessary.

5.3.2 Superior antimicrobial resistance prediction with GP–PIKE

The predictive power of the newly introduced kernel will now be evaluated on the task of MALDI-TOF MS based antimicrobial resistance prediction. Combined with a GP, the kernel forms a model referred to as GP–PIKE. We compare our model to both a non-

scenario	logistic regression	GP–RBF	GP–PIKE
E-AMOXCLAV	0.41±0.07	0.33±0.08	**0.47±0.04**
E-CEF	0.63±0.06	0.46±0.24	**0.71±0.03**
E-CIPRO	0.61±0.09	0.35±0.11	**0.68±0.03**
K-CEF	0.58±0.10	0.59±0.25	**0.77±0.07**
K-CIPRO	0.42±0.10	0.31±0.14	**0.55±0.10**
K-PIPTAZO	0.32±0.07	0.14±0.00	**0.56±0.10**
S-AMOXCLAV	0.53±0.04	0.14±0.00	**0.69±0.09**
S-CIPRO	0.34±0.03	0.23±0.12	**0.39±0.07**
S-PEN	0.80±0.03	0.74±0.03	**0.83**±**0.04**

Table 5.3: **Performance evaluation of GP–PIKE classifier** compared to two models: logistic regression and a Gaussian Process with an established kernel, the RBF kernel (GP–RBF). GP–PIKE outperforms both comparison models in all scenarios. Scenario abbreviations follow Table 5.1.

kernel model—logistic regression—and a GP with an established kernel, the radial basis function kernel (RBF). A comparison of all methods is depicted in Table 5.3.

Our kernel in combination with a GP outperforms all comparison models. Firstly, we trace this back to the capability of PIKE to compare non-linear interactions between peaks. The kernel was developed with this property, as some protein particles might receive a higher (i.e. larger than one) charge during the MALDI ionisation step and are detected at a smaller m/z–ratio. Secondly, we want to emphasize the compatibility of PIKE and the Gaussian Process, which performs a continuous (i.e. non-discrete) maximum likelihood hyperparameter optimisation, as opposed kernel-based methods such as SVMs, which optimise their hyperparameters by cross-validation over parameter grids. These grids have to be predefined and therefore do not allow to find the best values for continuous parameters.

A high variance (i.e. standard deviation) for the reported AUPRC values can be observed *independently* of the model, even though train–test splits are stratified by class. A possible explanation for these differences could lie in the underlying phylogenetic relatedness between microbial samples in the dataset. Microbial species undergo continuous evolutionary change, and whole branches in the evolutionary tree of a species can display an antibiotic resistance. If such latent structures are displayed in the MALDI-TOF mass spectra but are not accounted for in the stratified train–test splits, the results could be differences in the distribution of specific evolutionary branches associated with a specific resistance. We investigate these structures and possible improvements to the stratification in Chapter 6.

When we investigate the optimal hyperparameters chosen for each GP–PIKE scenario, we observe that t takes on similar parameter values and therefore we can construct a ‘common’ classifier by taking the mean of t for each split. The logistic regression optimisation results in diverging optimal parameters for each data split, with even the regularisation method differing. This behaviour does not directly affect the classification performance on the individual split, but we have observed that a set of optimal hyperparameters will lead to convergence errors when applied to a different split. Such incoherence between optimal hyperparameters can prevent training of a overall classifier with *one* final set of parameters.

The comparison of the two respective kernels combined with a GP illustrates the importance of a good kernel choice to reach high predictive performance. The GP–RBF is outperformed by GP–PIKE, and even the logistic regression, except in the case of K-CEF where the standard deviation is too high to allow for a conclusive assessment.

5.4 Confidence analysis of predictions with GP–PIKE

Machine learning applications that ultimately target decision processes influencing patient treatment are required to satisfy a more elaborate and stringent set of evaluations. When providing resistance prediction, the most crucial one being the need to provide uncertainty estimates with each predictions, and if necessary, reject label assignment in cases of high uncertainty. The reasoning is that any predictor will be faced with isolates stemming from microbial strain underrepresented in—or even completely absent from—the training dataset. The classification result must not give an uninformed prediction in that case, which cannot be recognized as such by the user. A classifier should therefore include the option to refuse the prediction if it cannot do so reliably.

5.4.1 Maximum class probability rejection

In addition to measures assessing the classification performance, such as AUPRC, we need a measure to estimate the confidence or reliability of each prediction. This assessment is crucial in two scenarios; (i) when samples fall close to the decision boundary of a classifier, and no unambiguous class decision can be performed, and (ii) for samples not included in the training dataset distribution, so-called out-of-distribution samples, for which no classification model can reliably predict a label and the desired behaviour would be to *reject* the prediction to notify the user, rather than performing an uninformed guess. The first scenario applies to every classifier. The second scenario is crucial in a settings where the input distribution cannot be controlled, e.g. clinical patient treatment and in the case of performing antimicrobial resistance predictions. Here, isolates collected from infected patients are not guaranteed to stem from the same strains included in the training distribution and could stem from an infectious strain that was picked up during travelling. In order to obtain a reliability estimate of the classifier's confidence, we employ the probabilities determined for the predicted class, i.e. $\max_c p(c|\mathbf{x})$, where c is the class label and $\mathbf{x}$ is the a sample, i.e. the MALDI-TOF mass spectrum. This value will be referred as the *maximum class probability* (MCP).

In principle, a well-trained classifier is highly confident for *all* for all samples stemming from training distribution, while out-of-distribution samples should be assigned a significantly lower probability in any prediction. In order to create a rejection scenario,

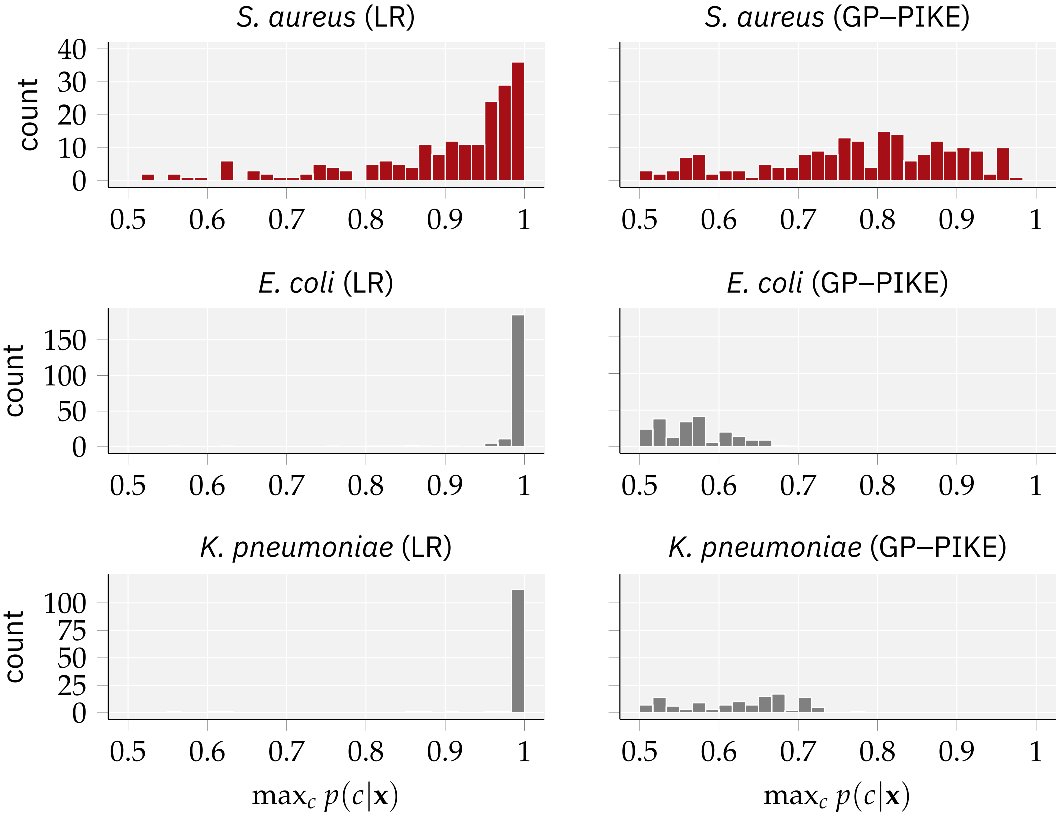


Figure 5.3: **Histogram depicting the maximum class probability distribution for in- and out-of-distribution samples** for logistic regression (LR) (left column) and GP–PIKE (right column) trained on *S. aureus*. The first row depicts the MCP in-training distribution from *S. aureus*, while the second and third rows depict the values for out-of-distributions samples from *E. coli* and *K. pneumoniae*. Figure adapted from Weis *et al.* [128].

a threshold $\theta \in [0.0, 1.0]$ is employed such that only predictions satisfying $\max_c p(c|\mathbf{x}) > \theta$ are kept. In the following the choice for threshold θ is motivated and analysed.

First, the distribution of MCP values in different test sample distribution scenarios is investigated in Figure 5.3, showing the difference in θ between *in-distribution* and *out-of-distribution* samples. The classifier is trained to predict resistance against the antibiotic amoxicillin-clavulanic acid in *S. aureus*. For in-distribution samples the test *S. aureus* dataset is used, and samples from the two other species, *E. coli* and *K. pneumoniae*, are used as out-of-distribution proxies. Two distributions of MCP values are depicted in Figure 5.3, LR (left column) and GP–PIKE (right column).

It can be observed that the MCP values in-training test sample received by the logistic regression classifier are distributed over the entire $[0.5, 1.0]$ range, with a visible skew towards values that are close to 1.0. However for out-of-distribution samples, the logistic regression also assigns values close to 1.0, indicating that the classifier is providing reliable predictions. The author attributes this strikingly incorrect behaviour to the linear decision boundary of the logistic regression classifier—samples close to the hyperplane are assigned $\max_c p(c|\mathbf{x})$ values ≈ 0.5. Out-of-distribution samples are likely to lie far away from the training data and therefore far away from the hyperplane, resulting in the assignment of a MCP value close to 1.0. We conclude that the MCP cannot be used for rejection of unreliable predictions in logistic regression, as $\max_c p(c|\mathbf{x})$ values ≈ 0.5 do not correspond to out-of-distribution samples. This behaviour also restricts the applicability of logistic regression classifiers in application for clinical treatment.

The right column in Figure 5.3 illustrates the behaviour of MCP values assigned to in- and out-of-distribution samples by the newly introduced GP–PIKE method. For in-training samples, the MCP values are evenly distributed over the $[0.5, 1.0]$ range, indicating that the GP–PIKE classifier does not report confidently on all samples of the test dataset. Experiments in the next subsection verify that low MCP values by GP–PIKE in fact correspond to less accurate predictions. Therefore, the classifier communicates if a reliable prediction could be made. Further, the MCP values of out-of-distribution samples are generally less than 0.7. We conclude that the GP recognises that no informed decision can be made on these samples and assigns values closer to 0.5. This behaviour can be explained with the non-linear decision boundary of GP–PIKE, caused by maximising the marginal likelihood of the data by optimising t in PIKE. The GP provides a probabilistic classification of unseen samples and is undecided about the class of out-of-distribution samples. We conclude that GP–PIKE MCP confidence estimates shows the desired behaviour for clinical applications. We thus take the MCP values to be suitable proxies for the confidence of a classifier and analyse the rejection rates in more detail.

5.4.2 Influence of rejection on predictive performance

The previous subsection illustrated that MCP values, i.e. $\max_c p(c|x)$, show the desired behaviour to be used as confidence estimates for GP–PIKE. While it is imperative to recognize and reject instances stemming from a different distribution, it is also important to reject in-distribution samples for which the classifier cannot give a reliable prediction. Rejection can lead to performance improvements in both cases, as a reliable classifier should increase its predictive performance if low-confidence samples are removed. We verify this assumption by varying the rejection threshold θ and analyse the development of the prediction accuracy on in-training test samples. The results are depicted in Figure 5.4. In the small θ value regime, both logistic regression and GP–PIKE improve in accuracy when the classifier can reject low-confidence samples. However, the logistic regression accuracy improves at a lower rate than the GP–PIKE and a sudden decrease in accuracy can be observed of $\theta > 0.95$. These results indicate that the samples receiving with the highest scores—closest to $\theta = 1.0$—in fact are assigned the wrong label, and rejecting all samples except the highest MCP results in only inaccurate predictions remaining. This expands the observation in 5.4.1 that in-

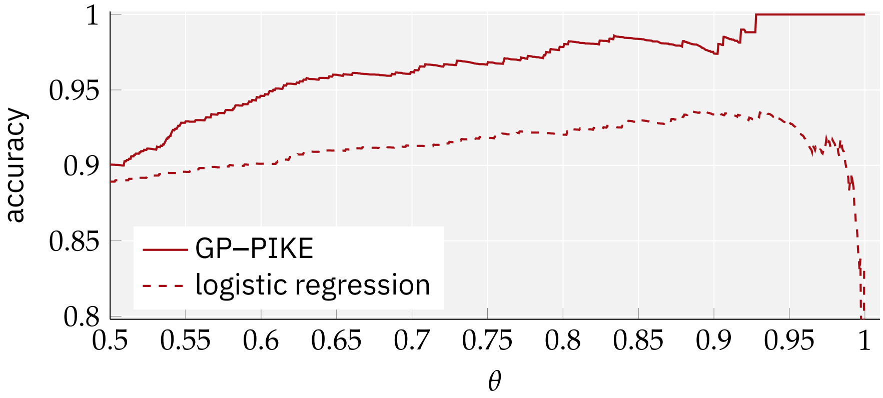


Figure 5.4: **Improvement of predictive accuracy with increasing rejection threshold θ** on the task of amoxicillin-clavulanic acid resistance prediction in *S. aureus*. To allow for comparability between test datasets with varying class ratios for each θ, the predictive accuracy is reported. The coarseness for larger θ is due to small sample size effects. Figure adapted from Weis *et al.* [128].

stances not following the training distribution receive a high MCP value, rendering the logistic regression unsuitable for clinical applications.

5.5 Summary and discussion

This chapter introduced a novel approach for classification from MALDI-TOF mass spectra. The new kernel—PIKE–in combination with a GP classifier outperforms logistic regression classifiers and traditional kernels on MALDI-TOF MS based antimicrobial resistance prediction. The method was evaluated to stay reliable in a realistic clinical application setting such as labelling unobserved bacterial strains. A newly developed, streamlined preprocessing method based on persistence challenged the commonly accepted pipeline.

Regarding the preprocessing pipeline, the introduced method did not improve predictive performance in the subsequent classification task. However, the complicated and underanalysed preprocessing pipeline remains in need for further assessment when it comes to the influence of parameters. A large disadvantage of GP–PIKE is its high computational complexity when computing the kernel. This complexity results from the pairwise comparison of each peak in either compared spectrum. This makes its application infeasible on large-scale data such as the full *DRIAMS-A* dataset, introduced in Chapter 3. A possible extension of PIKE could include confining the kernel to comparing a peak to its corresponding close-by m/z peaks. This would reduce the kernel's complexity but come at the cost of not comparing far-away MALDI-TOF mass peaks to each other. To avoid the loss of comparing peaks to its corresponding double-charge value, the 'close-by' region would need to include the m/z region at half the value.

Further studies requiring additional laboratory experiments could assess how isolates characterised by both DNA sequencing and MALDI-TOF MS provide insights into phylogenetic connections within the dataset and influence prediction depending on their distribution over train and test datasets. The gained insights could lead to new approaches for better train–test split choices or separating the prediction problem for different evolutionary strains. A computational approach to include phylogenetic information is introduced in Chapter 6.

6

Mitigating the effect of phylogenetic variance on MALDI-TOF MS based phenotype prediction through hierarchical stratification

C. Weis, B. Rieck, S. Balzer, A. Cuénod, A. Egli, and K. Borgwardt. “Improved MALDI-TOF MS based antimicrobial resistance prediction through hierarchical stratification”. *Unpublished*, 2020

In this chapter, we conduct a deeper exploration motivated by the results observed in experiments in the previous chapter. In Chapter 5, the results in Table 5.2 demonstrate how machine learning approaches tailored to MALDI-TOF mass spectra outperform established methods. However, we observe that the prediction results are stymied by a high standard deviation, regardless of the applied model. Therefore, the results indicate highly different predictive performance reached for each train–test split, despite having the same sample size and resistance label ratio. For ideal parameter optimisation, both train and test data should follow the structure of the total dataset. While the experiments of Chapter 5 stratify the train–test split with respect to resistance label, no stratification is performed for other structures within the train and test data, which could be differently distributed.

In this chapter, we hypothesise that the observed variation is caused by an underlying phylogenetic relatedness between microbial samples, affecting the sample distribution between the train and test dataset. If the evolutionary relationship between samples is implicitly reflected in their MALDI-TOF MS profiles, but not taken into account during the training of a model, this could potentially lead to highly dissimilar train–test splits. As a result, the phylogenetic structure of the test data might be underrepresented in the training data, causing lower predictive performance and high fluctuations between train–test split results. We conjecture that this information could be incorporated into the stratification process through an additional description step. In this chapter, we introduce an approach to infer this structure from the dataset through agglomerative hierarchical clustering and include the cluster information during the construction of our train and test set. Ideally, the relatedness between microbial probes could be determined through genetic information. However, genome information on *DRIAMS* (and generally for clinical MALDI-TOF MS datasets) is not available to infer the phylogeny in a dataset.

6.1 Hierarchical clustering

As the evolutionary process follows a hierarchical tree structure, the method selected to infer the underlying phylogenetic relatedness from the MALDI-TOF mass profiles is hierarchical clustering. The labels assigned by clustering will be used in combination with the resistance class labels during the train–test split for a stratification that conserves the ratio of of class–cluster labels in both datasets. This stratification ensures that the distribution of samples in the test data resembles that of the training samples, i.e. hindering the occurrence of *distribution shifts* between both, and that trained models will generalise to new datasets. The question of dealing with distribution shifts will be addressed in more detail in Chapter 7. The hierarchical clustering algorithm requires a dataset representation of either a distance matrix (between samples) or a feature matrix. In our approach, we implemented the latter representation, using fixed-size feature vectors by binning the MALDI-TOF mass peaks. For these experiments, we remain with this simplification, which merges peaks within a certain m/z–range, and analyse its power to infer the latent phylogenetic tree.

6.1.1 Clustering algorithm

We employed the hierarchical agglomerative clustering implementation provided by the `SciPy` package [119] for `Python`. The agglomerative clustering [98] belongs to the bottom-up clustering methods, meaning that at the start, each sample point belongs to its own cluster. With each iteration of the clustering process, the two nearest clusters, determined by a linkage and a distance measure, are merged into a single cluster. This process continues until only one cluster remains, containing all data points. By employing hierarchical clustering, we can capitalize on the high flexibility intrinsic to the approach: In a first step, the method constructs a tree capturing the inferred hierarchical structure between all samples (referred to as the *dendrogram*). This step does not require a fixed number of clusters k and therefore the dendrogram structure is independent of k. The number of clusters can be defined *a posteriori* by a distance threshold or by auxiliary visualisation. Figure 6.2 depicts the dendrogram for the dataset S-AMOXCLAV. Both the distance metric and the linkage method selected for the clustering algorithm influence the process and resulting cluster assignment. If not mentioned otherwise, we employ the Euclidean distance as a distance measure, $\mathrm{dist}(\cdot)$, as it is the established choice for numerical features. In the following paragraphs, we introduce all employed linkage criteria $\mathrm{d}(\cdot)$.

Ward's linkage. Ward's linkage is one of the most popular linkage methods. It employs the Ward's minimum variance method [124] and can only be used in combination with the Euclidean distance. The initial distances are calculated through squared Euclidean distances between the feature vectors u, v, i.e. $\mathrm{d}(u, v) = \mathrm{dist}(u, v)^2 = \|u - v\|_2^2$, as each cluster consists of only one sample. The two clusters with the lowest linkage criterion are merged and the new distance values are calculated in an iterative fashion by

$$\mathrm{d}(u,v) = \sqrt{\frac{|v|+|s|}{T}\,\mathrm{dist}(v,s)^2 + \frac{|v|+|t|}{T}\,\mathrm{dist}(v,t)^2 - \frac{|v|}{T}\,\mathrm{dist}(s,t)^2}, \tag{6.1}$$

where the number of samples within the cluster is denoted by $|\cdot|$, $T = |v| + |s| + |t|$, and u is the cluster resulting from combining clusters s and t. We repeat this process until all data points have been combined into one single cluster.

Average linkage. The average linkage value is calculated through the average distance between the data points of two clusters. The average linkage between clusters u and v is calculated by

$$\mathrm{d}(u,v) = \sum_{u_i \in u, v_j \in v} \frac{\mathrm{dist}(u_i, v_j)}{|u| \cdot |v|} \tag{6.2}$$

with $|a|$ referring to the number of points in cluster a. This approach is called UPGMA (unweighted pair group method with arithmetic mean).

Weighted linkage. Weighted linkage determines the weighted mean of average distances between all cluster members. Specifically, we calculate the distance between cluster u, formed through merging clusters s and t, and another cluster v by the following formula:

$$\mathrm{d}(u,v) = \frac{\mathrm{dist}(s,v) + \mathrm{dist}(t,v)}{2} \tag{6.3}$$

Initially, distances are determined between clusters containing only one data point using the selected distance measure, i.e. $\mathrm{d}(a,b) = \mathrm{dist}(a,b)$. The weighted linkage criterion is computationally more efficient than the average linkage criterion, with distances not contributing equally. This approach is referred to as WPGMA (weighted pair group method with arithmetic mean).

Single linkage. The most computationally efficient implementation for agglomerative hierarchical clustering is the single linkage criterion. The criterion is defined as the following for clusters u and v

$$\mathrm{d}(u,v) = \min_{u_i \in u, v_j \in v} \mathrm{dist}(u_i, v_j) \tag{6.4}$$

for all respective members of cluster u and v, i.e. u_i and v_j. As the single linkage criterion only considers the lowest distance between samples, its known to suffer from *chaining* [77], where cluster shapes form "chains" and lead to clusters with exceedingly imbalanced sizes. This opens up the possibility for two clusters to be considered very close by the single linkage due to a few close outliers, even though many cluster points are very far away from each other.

Complete linkage. The complete linkage criterion is designed to avoid the aforementioned shortcoming of single linkage. It quantifies the largest distance between points stemming from the two separate clusters, i.e.

$$\mathrm{d}(u,v) = \max_{u_i \in u, v_j \in v} \mathrm{dist}(u_i, v_j), \tag{6.5}$$

with u_i, v_j denoting points in cluster u and v, respectively. However, analogously to the single linkage, the complete linkage criterion also suffers from sensitivity to outliers [77].

6.1.2 Clustering output

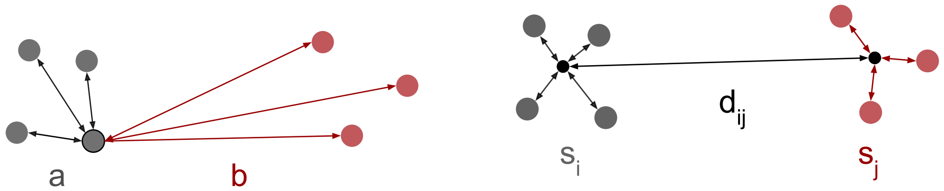
After describing the clustering algorithm, we detail its output and how it is incorporated into an enhanced train–test stratification scheme. The output of the clustering procedure is a linkage matrix Z of dimension $(n-1) \times 4$, with n referring to the number of samples. Each row in Z contains the clustering step of one iteration, i.e. the cluster connection formed in iteration i is stored in row i of Z. The indices of the clusters which are combined into the new cluster (with index $n+i$) are stated in the first and second column, $Z[i,0]$ and $Z[i,1]$. Cluster indices smaller than n indicate singleton clusters. The third and fourth columns, $Z[i,2]$ and $Z[i,3]$, state the distance between the two clusters which are combined and sample count in the newly-formed cluster [119].

Combined with a cut-off parameter, a cluster assignment for all samples can be determined via Z. For a predefined distance threshold, clusters are determined such that the distance between the samples in each cluster assessed through the linkage criterion is less than the threshold.

If we define the number of clusters a priori instead, we determine the distance threshold such that the wanted number is obtained. In this analysis, we define the number of clusters k beforehand; this enables an intuitive interpretation of the number of phylogenetic branches.

6.1.3 Clustering metrics

Arguably the bigger obstacle in analyses employing clustering algorithms is to decide on a number of clusters k. In our application—as in the case for most—no a priori information is available regarding the true number of clusters (here, number of phylogenetic branches) in the data. Numerous clustering validity measures have been developed to assess the quality of a cluster assignment, without the availability of the true cluster labels. We select clustering validity metrics under the assumption that the feature space does not construct highly-complex topological structures [95], e.g. cycles or voids. All clustering validity metrics employed in this analysis are unsupervised methods and do not incorporate resistance class labels. The two metrics employed are depicted in Figure 6.1 and discussed in the following paragraphs.



(a) Silhouette score (b) Davies–Bouldin index

Figure 6.1: **Illustration of the construction of Silhouette score and Davies-Bouldin index**. (6.1a) The mean distance between one point (marked by black edge) and all points in the same cluster is marked by a, while the mean distance to the points in the nearest cluster is denoted by b. (6.1b) The mean distance of all points in a cluster to its centroid point are denoted by s_i for cluster i and s_j for cluster j. These intra-cluster distances, e.g. s_i and s_j, are compared with the inter-cluster distance between the respective centroids d_{ij}.

Silhouette score. For the Silhouette score, the average distance between the data point itself and all points belonging to the same cluster (referred to as a), is compared to the average distance to all points from the nearest cluster (referred to as b). For a single data point, the Silhouette coefficient is determined by

$$s = \frac{b - a}{\max(a, b)}, \tag{6.6}$$

with the overall score determined by the mean of all individual Silhouette coefficients. This results in $s \in [-1, 1]$, with an overall Silhouette score close to -1 indicating low quality clustering, and an overall Silhouette score close to 1 signifying well-separated clusters.

Davies–Bouldin index. The Davies–Bouldin index quantifies the mean similarity between each cluster and the cluster with the largest ratio of intra-cluster distances to inter-cluster distances. With the cluster diameter, i.e. the mean distance between each point of the cluster and the respective centroid, denoted by s_i and the distance between centroids of two clusters i and j denoted by d_{ij}, the Davies–Bouldin index is determined by

$$\mathrm{DB} = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{s_i + s_j}{d_{ij}}, \tag{6.7}$$

with k referring to the number of clusters. DB is strictly positive, with lower values indicating a high quality clustering and 0 being the best score possible.

6.1.4 Incorporating cluster information to improve stratification

Regardless of the choice of linkage criterion, the hierarchical clustering algorithm will assign clusters that contain different MALDI-TOF mass spectra. We hypothesise that these clusters are able to capture phylogenetic structures between the MALDI-TOF mass spectra, i.e. two spectra that are assigned to the same cluster represent microbes that are more closely related to each other than those of any other cluster.

Each MALDI-TOF mass spectrum is assigned a meta-label (l_r, l_c), consisting of the antimicrobial resistance class l_r and the hierarchical cluster l_c. This meta-label prevalence can be distributed equally over both the train and test set during the stratified split. We conjecture that this meta-label (l_r, l_c) contains auxiliary knowledge compared to the resistance class information alone. Under the assumption that the cluster label is implicitly informed by the phylogeny relatedness, the train and test dataset have a similar distribution. This should result in (i) higher predictive performance, as the classifier will be faced with similar data distributions during training and evaluation, and (ii) lower variance of performances of different splits, as the classifiers will have been trained on enough representative samples to improve its generalisation capabilities to unseen test samples. The proposed method is easy to implement and incorporates directly into any classification workflow; only the clustering algorithm (incl. linkage criterion and distance metric) and the number of clusters are required. The class labels, i.e. the antimicrobial resistance labels, are not used for the clustering step, and are only required during the train–test stratification. This allows for the general application of our method, independent of the prediction task.

6.2 Data stratification based on inferred hierarchical structure

To allow for comparability with the results obtained in Chapter 5, we employ the logistic regression classifier (LR) as the baseline and the Gaussian Process classifier combined with PIKE (GP–PIKE) as the classification model tailored to sparse MALDI-TOF mass spectra representations. Further, the same dataset is employed (see Table 5.1).

The classifiers are evaluated over the same random seeds as in Section 5.3 to determine the stratified train (80% of samples) and test (the remaining 20% of samples) portions. On each split, a 5-fold cross-validation determines the optimal hyperparameters. Both the cross-validation and the overall predictive performance is reported by the average precision (AUPRC) metric, as we are working with heavily-imbalanced classes.

6.2.1 Inferring hierarchical structure

For each of the species, we cluster the dataset once with each linkage criterion introduced in Section 6.1.1, namely *ward*, *average*, *weighted*, *single*, and *complete*. We vary the number of clusters k from 1 (e.g. no clustering is performed) to 20, which should represent a good biologically plausible upper limit. Figure 6.2 illustrates an example output of hierarchical clustering with *ward's* linkage criterion, using S-AMOXCLAV as an example. For quality assessment, we report the two unsupervised clustering validity metrics—the *Silhouette score* and the *Davies–Bouldin index*—for each clustering. Please note again, there are no ground-truth strain-type labels measured and we

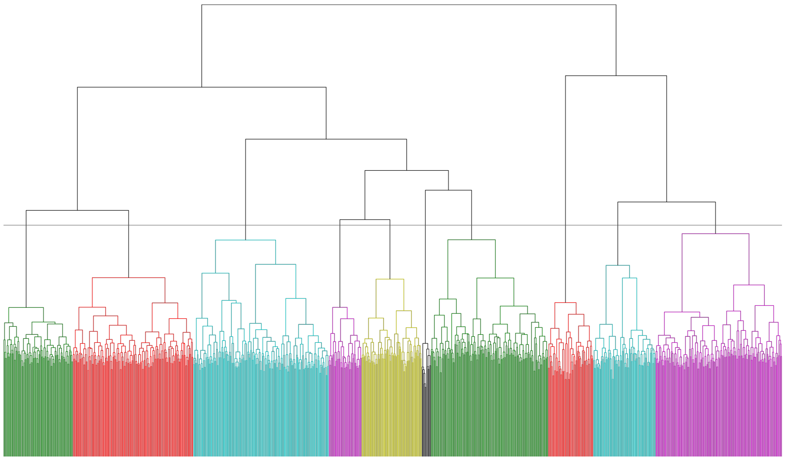


Figure 6.2: **Dendrogram of the complete hierarchical clustering tree** using ward's linkage to cluster the samples of the *S. aureus* (amoxicillin-clavulanic acid) dataset. All tree tips represent a single sample in the dataset. The colors indicated the clustering with choice $k^* = 10$. Note that the clustering does not depend on a k; the cluster labels are through a cut-off at the desired "height" of the determined dendrogram.

therefore require unsupervised validity metrics. The behaviour of the two cluster validity metrics with varying k is depicted in Figure 6.3. For each of the nine species-antibiotic scenarios, a curve is drawn, leading to three curves per species, as each species dataset consists of a varying list of samples depending on the antibiotic (see Table 5.1). Generally, the results indicate that the best clustering validity index values are reached with the lowest number of clusters, i.e. $k = 2$. Furthermore, different linkage criteria show markedly different behaviours for the different validity scores. The Silhouette score values when using single linkage are consistently decreasing; however they form a plateau for k larger than ten clusters with ward's linkage. The results indicate a high sensitivity of some linkage criteria to small differences in the dataset construction—i.e. within the same species but subsampled for different antibiotics—which can be observed for the Davies–Bouldin index with single linkage, or in both scores with weighted linkage.

6.2.2 Choosing the number of clusters

In order to obtain the cluster labels for the enhanced train–test, a single clustering model has to be selected. To define a model, an optimal number of clusters k^* has to be decided on. We make this choice in a unsupervised, data-driven fashion; no class labels or predictive performance are considered in the choice of k^*. The parameter k is evaluated merely on the clustering validity indices displayed in Figure 6.3, and one k^* is fixed for each species to use for hierarchical clustering. The final k^* values are $k^* = 8$ for *E. coli* and $k^* = 10$ for *S. aureus* and *K. pneumoniae*. For this decision, we focus on the commonly applied average and ward linkage criteria during subsequent analy-

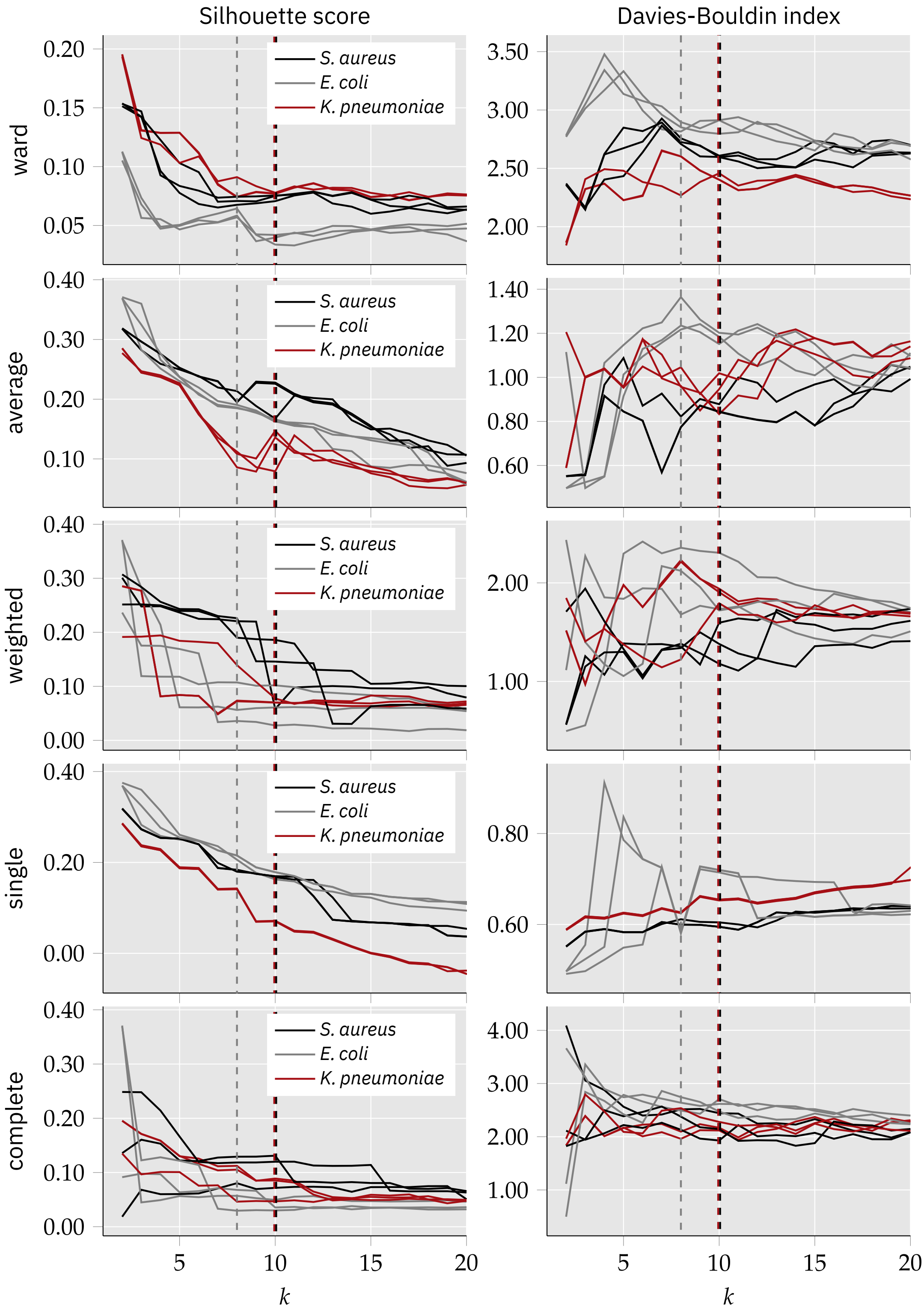


Figure 6.3: **Behaviour of cluster validity scores with an increasing number of clusters** using five different linkage criteria. No scores are reported for $k = 1$ as at least two clusters are required to compute these validity scores. For each of the nine species-antibiotic scenarios, a curve is drawn, leading to three curves per species. The dashed vertical line depicts the k^* value chosen for the respective species.

scenario	LR $k = 1$	LR k^* ward	LR k^* average	GP–PIKE $k = 1$	GP–PIKE k^* ward	GP–PIKE k^* average
E-AMOXCLAV	0.41±0.07	0.41±0.05	**0.45±0.03**	0.47±0.04	0.48±0.04	**0.52±0.04**
E-CEF	0.63±0.06	0.60±0.02	**0.64±0.05**	0.70±0.03	0.70±0.06	**0.72±0.04**
E-CIPRO	0.62±0.08	0.58±0.07	**0.67±0.03**	0.68±0.03	0.68±0.04	**0.72±0.03**
K-CEF	0.58±0.10	**0.64±0.08**	0.47±0.09	**0.77±0.07**	0.75±0.04	0.69±0.11
K-CIPRO	0.42±0.10	0.38±0.07	**0.48±0.15**	0.55±0.10	0.53±0.09	**0.55±0.11**
K-PIPTAZO	0.32±0.07	**0.41±0.11**	0.39±0.03	0.57±0.10	**0.61±0.05**	0.57±0.10
S-AMOXCLAV	0.53±0.04	**0.60±0.08**	0.55±0.03	0.69±0.09	0.73±0.06	**0.77±0.07**
S-CIPRO	**0.34±0.03**	0.34±0.02	0.31±0.07	0.40±0.07	**0.40±0.10**	0.40±0.08
S-PEN	0.80±0.03	0.82±0.03	**0.83±0.04**	0.83±0.04	**0.84±0.04**	0.83±0.03

Table 6.1: **Improved predictive performance employing the hierarchical train–test split** compared to a standard random train-test split. Results are reported through AUPRC mean ± standard deviation on five test datasets. The methods include $k^* = 8$ for *E. coli* and $k^* = 10$ for *S. aureus* and *K. pneumoniae*.

sis. For *E. coli*, a slight peak can be observed in the Silhouette score behaviour, while the Davies–Bouldin value flattens out once k increases past 8 (note that the Davies–Bouldin index optimal value is better the closer it is to 0, while the Silhouette score should be maximised). For *S. aureus* and *K. pneumoniae*, no clear message can be observed in the Silhouette score and the Davies–Bouldin values plateaus for $k > 10$. Therefore, we consider the average linkage criterion, and pick $k = 10$, where we observe slight peaks in the Silhouette coefficient.

6.2.3 Resistance prediction with enhanced train–test splits

Employing the chosen number of clusters k^*, we determine the clusterings and incorporate the cluster labels for the enhanced train–test split based on the meta-label (l_r, l_c) stratification. We then combine these enhanced train–test splits into two antimicrobial resistance classification scenarios: using both a logistic regression and a GP–PIKE model [128]. The results are reported by the AUPRC mean with standard deviation over all five random seeds. Results are reported with and without the hierarchical clustering-enhanced stratification in Table 6.1. The results table focuses on two of the linkage criteria, namely average and ward, as the average linkage criterion provides a good trade-off between the single and complete linkage criteria and their respective pitfalls, and ward is a frequently chosen criterion with Euclidean distance metric. For either logistic regression and GP–PIKE, eight out of the nine scenarios report increased performance with the novel hierarchical clustering-enhanced stratification. In logistic regression, improvements are as high as 9.7 percentage points for piperacillin-tazobactam resistance prediction in *K. pneumoniae*. GP–PIKE started at a higher baseline performance, with increases in predictive performance from 0.69 to 0.77 for amoxicillin-clavulanic acid resistance prediction in *S. aureus*. To illustrate the interaction between k, the clustering validity values, and AUPRC, we depict the predictive performance for all $k \in \{1, 2, \ldots, 20\}$, for logistic regression in Figure 6.4 and for GP–PIKE in Figure 6.5. It should be emphasised that k has to be determined prior to evaluating the predictive performance to avoid information leakage influencing the choice of the clustering parameters—as we have done by choosing k^* prior in this analysis. The results do not indicate that the stratification enhanced by hierarchical clustering results in robust results with a lower standard deviation. They do indicate, however, that the enhanced stratification can result in a higher predictive performance. In most of the scenarios, the curves display a similar behaviour for all five linkage criteria. An exception is the prediction of ceftriaxone resistance in *K. pneumoniae*, as ward linkage leads to high predictive performance for $k < 10$, with decreasing performance for $k > 10$. However, a seemingly opposite development can be observed for the weighted linkage criterion and average linkage criterion. Another exception can be observed for ciprofloxacin and amoxicillin-clavulanic acid resistance prediction in *S. aureus*—predictive performances are not consistent with changing k and large variations can be observed with no trend discernible. It can be observed that the choice of k has a large influence on whether the enhanced stratification will lead to improved predictive performance. For many scenarios, choosing a "bad" k leads to lower performance than the baseline employing no hierarchical clustering, $k = 1$. For only a few scenarios will any k lead to an increased predictive performance, e.g. amoxicillin-clavulanic acid resistance prediction in *E. coli* using ward or complete linkage criterion or penicillin prediction in *S. aureus* for all linkage criteria using logistic regression.

6.3 Summary and discussion

This chapter introduces a novel method to enhance the train–test stratification for MALDI-TOF MS based phenotype prediction tasks, based on inferring hierarchical connections within the dataset from the MALDI-TOF profile. This new stratification procedure was modeled on the hypothesis that the inferred hierarchical clusters depict phylogenetic branches and relationships between the bacteria contained in the dataset. We employ clustering validity scores to choose an optimal number of clusters k^* to be used as a cluster parameter. The results reveal that neither the Silhouette score nor the Davies-Bouldin index indicate a clear choice of k that leads to well-separated clusters. However, we can demonstrate a beneficial effect of the proposed enhanced train–test stratification for MALDI-TOF MS based antimicrobial resistance prediction.

One general observation from our experiments is that the train–test stratification technique affects the predictive performance considerably; despite the fact that for each k, the same number of spectra are in the train dataset (although the train and test distributions differ). Deviations of up to 20 % in AUPRC values between the lowest and the highest predictive performance were not expected, but they emphasise the potential of using auxiliary latent information (i.e. information that is not directly observable or measured) of MALDI-TOF mass spectra. High predictive performance—higher than employing no hierarchical stratification—can likely be explained by both the train and test dataset following the "true" structure of the data closely. Therefore, each strati-

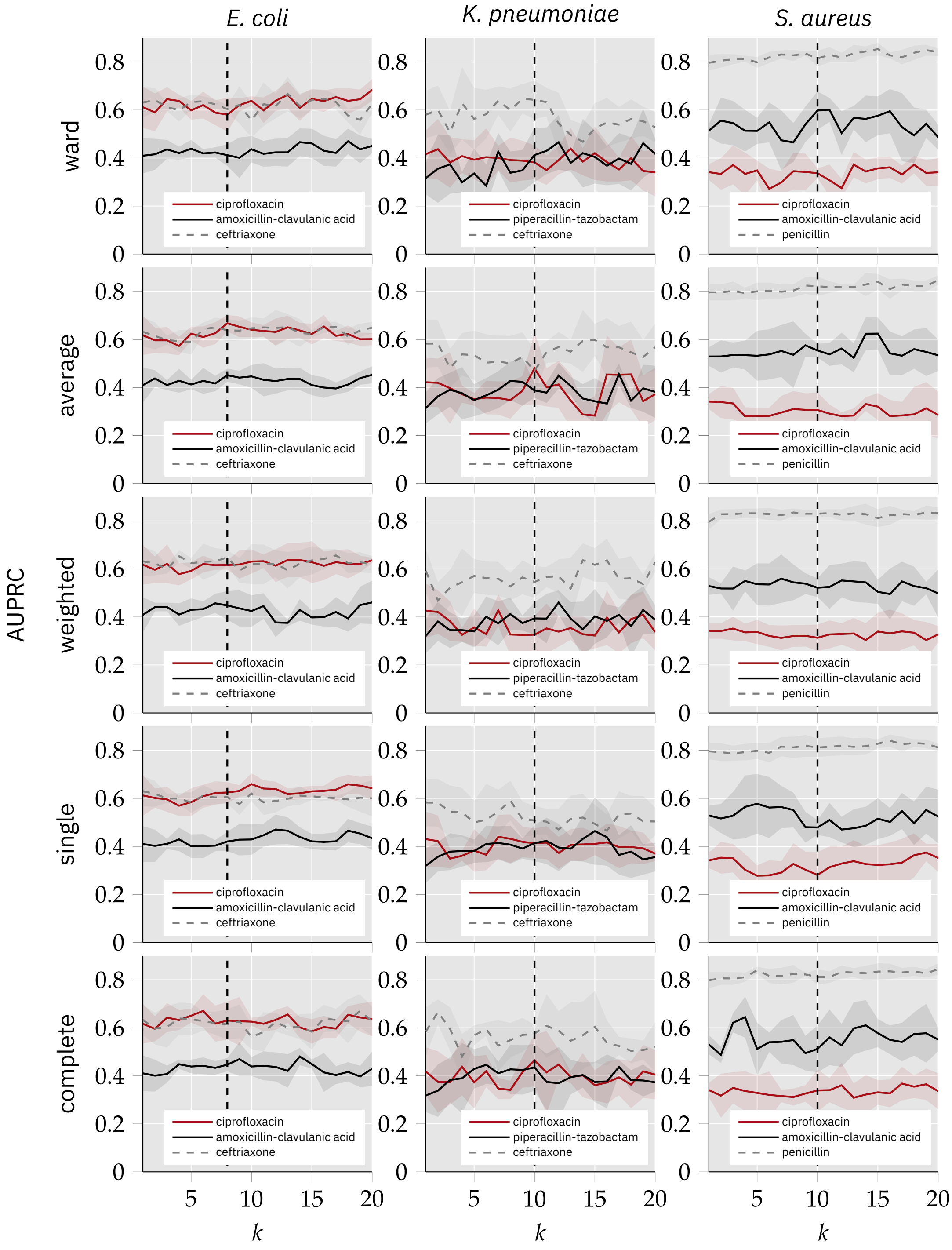


Figure 6.4: **Predictive performance using logistic regression illustrating the influence of different linkage criteria and the number of clusters** k of all nine antimicrobial resistance scenarios. The results are given in mean average precision (AUPRC) ± standard deviation on the test data over 5 random splits. The dashed vertical line depicts the k^* value chosen for the respective species.

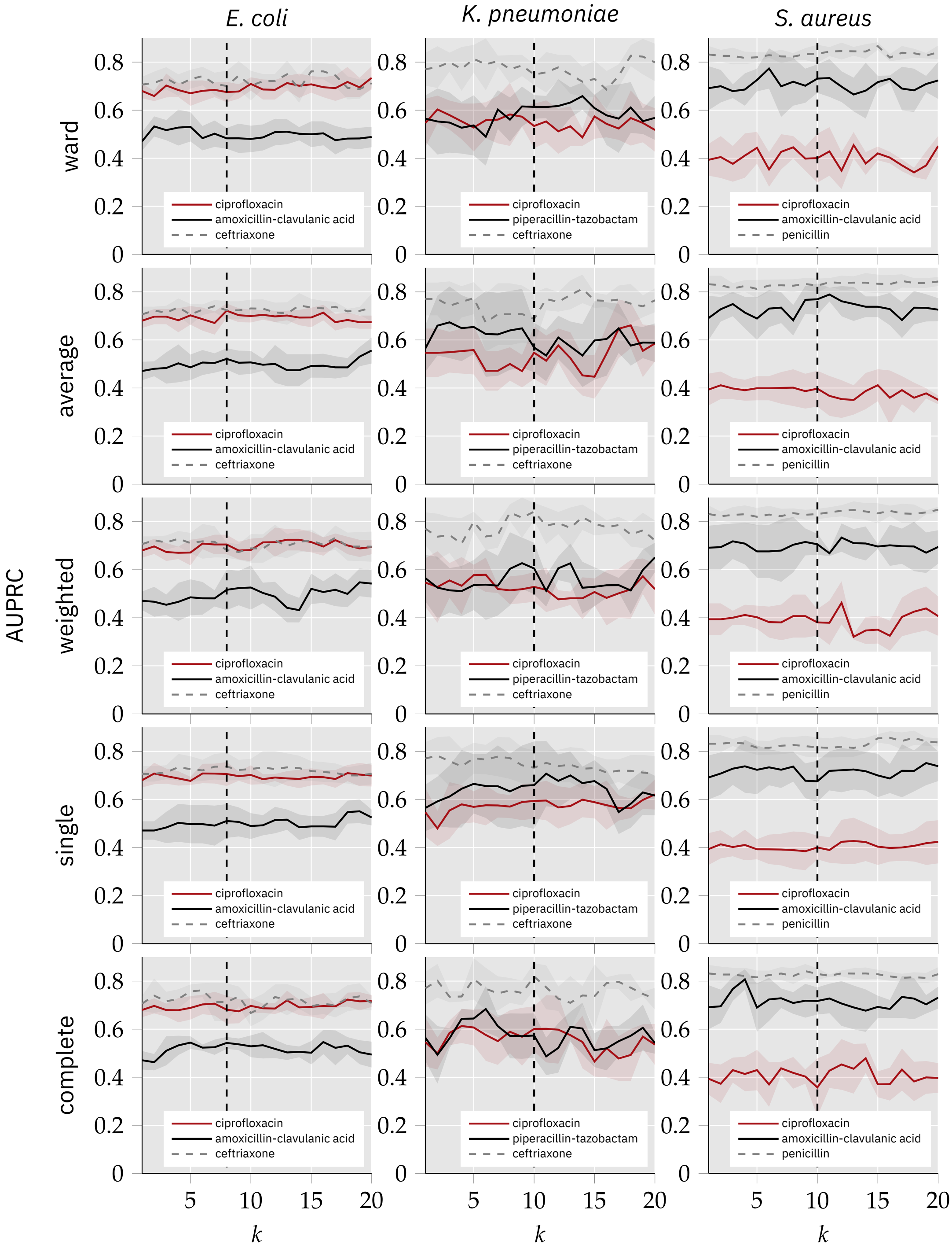


Figure 6.5: **Predictive performance using GP–PIKE illustrating the influence of different linkage criteria and the number of clusters** k of all nine antimicrobial resistance scenarios. The results are given in mean average precision (AUPRC) $\pm$ standard deviation on the test data over 5 random splits. The dashed vertical line depicts the k^* value chosen for the respective species.

fied split (induced by each seed) is capable of training on a dataset that closely follows the true distribution.

Albeit the observed benefit to predictive performance, the results do not support the hypothesis that the enhanced stratification can reduce the standard deviation between the prediction results of different splits. This observation could have several explanations: (i) the hierarchical clustering cannot capture the patterns that cause the high standard deviation, (ii) the stratification cannot mitigate a high classification complexity for certain parts of the data, leading to over- or underestimation of predictive performance, or (iii) the high variance between test data splits is due to small sample size, as *K. pneumoniae*—the species with the fewest samples—displays the largest standard deviation compared to *E. coli* and *S. aureus*. The results presented in this chapter highlight the potential of inferring and including phylogenetic structure into the stratification, but also indicate that such a method is fraught with obstacles to be overcome to obtain a stable method, e.g. (i) in a regular experiment, no ground-truth samples are available to validate the clustering, and (ii) obtaining a single optimal number of clusters k^* that is a good choice for all prediction tasks. The introduced hierarchical clustering based stratification should be seen as a step towards improved MALDI-TOF MS based resistance prediction—while easy to implement and conceptually simple, further improvement and research is necessary to increase the method's stability. We propose two directions of future research: First, collecting future MALDI-TOF MS datasets that include strain information about the microbial specimen measured would allow for analysing whether the inferred hierarchical tree in fact represents the phylogenetic relatedness or not. Other auxiliary latent information, such as the culture growth medium or measurement data, can be reflected in the MALDI-TOF MS profile. Comparison of the hierarchical tree and the aforementioned variables might provide additional insights into their influence on the clustering output. Secondly, the influence of the metric chosen for the clustering algorithm should be analysed. While the results in this chapter are restricted to the Euclidean distance for conceptual simplicity, the application of a MALDI-TOF MS-specific metric could be able to capture multi-scale nuances between spectra. Methods based on optimal transport [118] could be particularily suited for this application, as they have shown promising performance for classification tasks in recent years.

7 Learning domain independent MALDI-TOF mass spectrum representations

C. Weis, A. Cuénod, B. Rieck, F. Llinares-López, O. Dubuis, S. Graf, C. Lang, M. Oberle, K. K. Soegaard, M. Osthoff, M. Brackmann, K. Borgwardt, and A. Egli. "Direct Antimicrobial Resistance Prediction from clinical MALDI-TOF mass spectra using Machine Learning". *accepted in Nature Medicine*, 2021. doi: `https://doi.org/10.1101/2020.07.30.228411`

C. Weis, M. Horn, B. Rieck, A. Cuénod, A. Egli, and K. Borgwardt. "Domain adaptation for transferable antimicrobial resistance prediction from MALDI-TOF mass spectra". *Unpublished*, 2021

The previous chapters have focused on establishing the baseline performance, defining an optimal classification scenario, and improving prediction accuracy through new machine learning approaches. For all studies, the MALDI-TOF MS data originated from one collection site, *DRIAMS-A*. Chapter 2 concluded that the current literature on MALDI-TOF MS based phenotype prediction suffers from a lack of external validation of their results. Further, MALDI-TOF MS datasets are known to suffer from domain differences stemming from influences such as instrument settings and local laboratory procedures [79]. Mitigating these differences for MALDI-TOF MS based machine learning models to allow for transferability and robustness of prediction has never been addressed before.

Hence in this chapter, we first determine the decrease in predictive performances of predictors trained on the large number of samples contained in *DRIAMS* (see Section 4.1) on MALDI-TOF mass spectra collected at different medical institutions, i.e. *DRIAMS-B* to *DRIAMS-D*, in Section 7.1.1. Hereby the goal is to assess the generalisability of an antimicrobial resistance predictor trained at a single site and to gain insights into the decision making process. As this is an extension of the analysis in Chapter 4, to utilize the information along the entire m/z–axis and for state-of-the-art deep learning models to be applicable to the data, the full-spectrum binned feature vector representation is employed throughout this entire chapter. While this fulfills the requirement for a thorough machine learning validation, the need to obtain a predictor for sites with less available training data remains. To this end, we demonstrate an easy-to-implement approach for improving predictions on sites with few training samples by leveraging the vast amount of data in our dataset *DRIAMS* in Section 7.1.2.

Further, in Section 7.2 we investigate an approach aimed at learning a new data representation for MALDI-TOF MS that is independent from batch-effects stemming from its collection site. We base this approach on models used in adversarial deep learning, where two objectives—accurate resistance prediction and determining the collection site of MALDI-TOF mass spectrum—are pitted against each other to obtain a data representation that works well on the first objective, but fails on the second.

7.1 Leveraging large-scale multi-site data for improved local prediction

7.1.1 Direct transferability of an antimicrobial resistance predictor trained on external data

Three antibiotic resistance prediction scenarios were selected and evaluated on the *DRIAMS-A* subdataset of *DRIAMS* in Section 4.1 (see Table 4.1), namely (i) ceftriaxone resistance prediction in *E. coli* (referred to as E-CEF) using a LightGBM model, (ii) ceftriaxone resistance prediction in *K. pneumoniae* (K-CEF) using a MLP, and (iii) oxacillin resistance prediction in *S. aureus* (S-OXA), again using a LightGBM model. We continue with these classification scenarios to evaluate the general transferability of predictive performance from one *DRIAMS* subdataset to another. The added value of such an analysis is two-fold, as it allows us to (i) obtain information on the variance of predictive signals from different sites, and (ii) judge the feasibility of using a pretrained MALDI-TOF MS based resistance classifier for prediction at other sites. In light of that, each subdataset in *DRIAMS* is split into train and test, and a predictor is trained on each of the train datasets respectively. Then, we determine the predictive performances when testing each predictor on each test dataset. This process is carried out for each classification scenario and the results are depicted in Figure 7.1. Overall the result indicate that the best result on a test dataset is obtained when training on data collected at the same site. Direct application of a predictor on a non-training site can result in drastic decreases in performance to the point of hardly better-than-random predictions, such as the *DRIAMS-A* trained K-CEF predictor reporting AUROC values of 0.53 on *DRIAMS-B* and *DRIAMS-C*. Hence, additional investigation into the transferability of predictive power to new sites is essential. Amongst the results obtained through training and testing on the same site, *DRIAMS-A* performs consistently well, which can be attributed to its large number of samples. The variability within the site-specific-test results is quite high, ranging from 0.54 and 0.80, both for S-OXA on *DRIAMS-C*.

7.1.2 Empirical risk minimization on a union of datasets

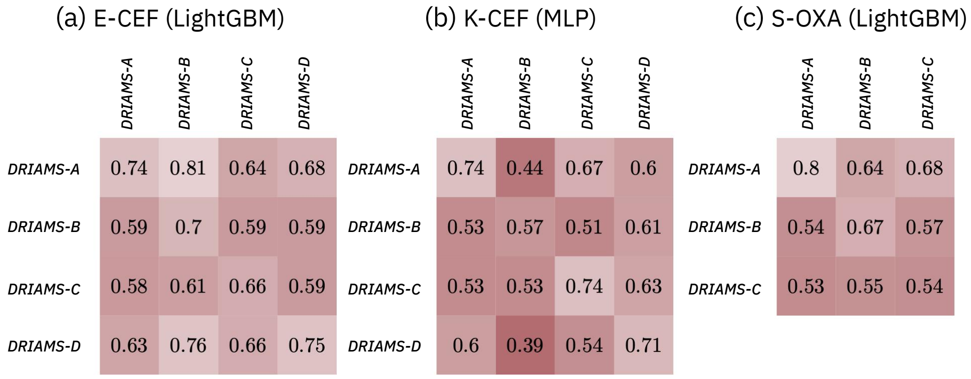
With a baseline on inter- and intra-site predictive performance established, we fully focus on the aim to improve prediction on a new site with few samples. Specifically, we are interested if the data contained in *DRIAMS*, which includes large subdatasets such as *DRIAMS-A* and represents MALDI-TOF mass spectra from multiple institutions, can be leveraged to improve predictions at a new prediction site. This problem statement mirrors the real-life objective of obtaining a new predictor for antimicrobial resistance

target site						******DRIAMS-D*
DRIAMS-B					**DRIAMS-D*	*DRIAMS-C*
				DRIAMS-B	*DRIAMS-C*	*DRIAMS-B*
scenario	model	*DRIAMS-A*	*DRIAMS-B*	*DRIAMS-A*	*DRIAMS-A*	*DRIAMS-A*
E-CEF	LightGBM	0.60±0.13	0.55±0.20	**0.66±0.12**	0.63±0.14	0.62±0.11
K-CEF	MLP	0.23±0.12	0.29±0.14	**0.37±0.17**	0.20±0.11	0.33±0.18
S-OXA	LightGBM	0.24±0.12	0.30±0.20	0.45±0.24	0.25±0.11	**0.48±0.18**

target site						**DRIAMS-D*
DRIAMS-C					**DRIAMS-D*	*DRIAMS-C*
				DRIAMS-C	*DRIAMS-B*	*DRIAMS-B*
scenario	model	*DRIAMS-A*	*DRIAMS-C*	*DRIAMS-A*	*DRIAMS-A*	*DRIAMS-A*
E-CEF	LightGBM	0.31±0.06	0.34±0.06	0.39±0.05	0.35±0.06	**0.40±0.06**
K-CEF	MLP	0.34±0.09	0.42±0.11	0.42±0.11	0.25±0.07	**0.44±0.09**
S-OXA	LightGBM	0.21±0.09	0.08±0.02	**0.27±0.12**	0.21±0.12	**0.27±0.14**

target site						**DRIAMS-D*
DRIAMS-D					**DRIAMS-C*	*DRIAMS-C*
				DRIAMS-D	*DRIAMS-B*	*DRIAMS-B*
scenario	model	*DRIAMS-A*	*DRIAMS-D*	*DRIAMS-A*	*DRIAMS-A*	*DRIAMS-A*
E-CEF	LightGBM	0.29±0.09	**0.48±0.05**	0.43±0.06	0.34±0.08	0.43±0.07
K-CEF	MLP	0.09±0.03	**0.23±0.06**	0.21±0.05	0.11±0.04	0.20±0.04

Table 7.1: **Combining data from several collection sites can improve predictions on a target site with few training samples** expressed through AUPRC. Predictive performance increases on *DRIAMS-A* and *DRIAMS-B* when training is performed on a union of the respective training data with other sites. *DRIAMS-D* does not benefit from an expanded training dataset with other collection sites.



(a) E-CEF (LightGBM)

	DRIAMS-A	DRIAMS-B	DRIAMS-C	DRIAMS-D
DRIAMS-A	0.74	0.81	0.64	0.68
DRIAMS-B	0.59	0.7	0.59	0.59
DRIAMS-C	0.58	0.61	0.66	0.59
DRIAMS-D	0.63	0.76	0.66	0.75

(b) K-CEF (MLP)

	DRIAMS-A	DRIAMS-B	DRIAMS-C	DRIAMS-D
DRIAMS-A	0.74	0.44	0.67	0.6
DRIAMS-B	0.53	0.57	0.51	0.61
DRIAMS-C	0.53	0.53	0.74	0.63
DRIAMS-D	0.6	0.39	0.54	0.71

(c) S-OXA (LightGBM)

	DRIAMS-A	DRIAMS-B	DRIAMS-C
DRIAMS-A	0.8	0.64	0.68
DRIAMS-B	0.54	0.67	0.57
DRIAMS-C	0.53	0.55	0.54

Figure 7.1: **Direct transferability of *DRIAMS* subdataset trained classifier** quantified by AUROC. The y–axis presents the train datasets of *DRIAMS-A*, *DRIAMS-B*, *DRIAMS-C* and *DRIAMS-D*, while the x–axis present the respective test datasets. Hence, the values on the diagonal correspond to intra-site training and testing. No *S. aureus* samples with oxacillin label are available in *DRIAMS-D*, causing gaps in 7.1c. Scenario abbreviations follow Table 4.1. Figure adapted from Weis *et al.* [126].

at a new medical institution, where too few training samples are available. We emulate this scenario by selecting one site from *DRIAMS-B*, *DRIAMS-C* and *DRIAMS-D*, as the target site. Subsequently, different combinations of train splits from all four datasets in *DRIAMS* are pooled, a predictor is trained and applied to the single target test dataset. Recently, it has been shown [137] such an approach of empirical risk minimization can outperform more complex models from the field of domain adaption.

We conduct this experiment for all three classification scenarios for each test site. The results in Table 7.1 show that a union of target-site training data with samples collected at external sites is beneficial for *DRIAMS-B* and *DRIAMS-C*. Predictions on both test sites improve by addition of the large *DRIAMS-A* dataset compared to training exclusively on target-site samples. *DRIAMS-C* further benefits from including additional collection sites as well. On test data from *DRIAMS-D*, the best performance was reached when training solely on target-site samples. In no scenario do unions exclusively containing external data outperform the combinations including target-site samples. These results indicate a higher similarity between the subdatasets *DRIAMS-A*, *DRIAMS-B* and *DRIAMS-C* than with *DRIAMS-D*. Combining several datasets suffering from large batch effects between them can lead to the dilution of relevant signal and no common information can be inferred. One explanation for this variance could be that *DRIAMS-D* is collected at a laboratory service provider, while the other stem from hospital clinical routine (see Section 3.1). Further, the results indicate that this approach of empirical risk minimization based on datasets union always requires site-specific samples.

7.2 Learning site-independent data representations

7.2.1 Domain adaptation

In real-world machine learning employment, we are often faced with the task of employing a predictor trained on one dataset—referred to as source domain—to another dataset—the target domain. In a realistic scenario a so-called *domain shift* is present, i.e., the samples from individual domains are not necessarily sampled from the same distribution. In the context of MALDI-TOF MS phenotype prediction, the reasons for domain shift include usage of different measurement instruments, or slight variations in laboratory protocols, for instance.

The general aim of domain adaptation is to use additional, possibly unlabelled, data from the target domain to boost the generalisation performance and thereby improve the predictive performance on the target domain. Following Farahani *et al.* [36], the problem of domain adaptation is defined by the following: A domain is composed of a feature space $\mathcal{X}$, a label space $\mathcal{Y}$, and a joint probability distribution over both $p(x, y)$, such that $\mathbb{D} = \{\mathcal{X}, \mathcal{Y}, p(x, y)\}$. The source domain is denoted by $\mathbb{D}_S = \{\mathcal{X}_S, \mathcal{Y}_S, p_S(x, y)\}$, and the target domain as $\mathbb{D}_T = \{\mathcal{X}_T, \mathcal{Y}_T, p_T(x, y)\}$.

We train a machine learning model $h\colon \mathcal{X} \to \mathcal{Y}$, taken from the hypothesis space $\mathcal{H}$, by minimising the expected risk on the source data $\mathcal{R}_S$ with respect to the loss function $\mathcal{L}\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

$$\mathcal{R}_S(h) = \mathop{\mathbb{E}}_{(x,y)\sim P_S(x,y)} [\mathcal{L}(h(x), y)] \tag{7.1}$$

Therefore, the expected risk on the target domain can be written as

$$\begin{aligned} \mathcal{R}_T(h) &= \mathop{\mathbb{E}}_{(x,y)\sim P_T(x,y)} [\ell(h(x), y)] \\ &= \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(x), y) p_T(x, y) dxdy \\ &= \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(x), y) p_T(x, y) \frac{p_S(x, y)}{p_S(x, y)} dxdy \\ &= \mathop{\mathbb{E}}_{(x,y)\sim P_S(x,y)} \left[\frac{p_T(x, y)}{p_S(x, y)} \mathcal{L}(h(x), y) \right] \end{aligned} \tag{7.2}$$

In the case where both data in the source and target domain follow the same distribution, the fraction in the last expression rescinds itself to

$$\frac{p_T(x, y)}{p_S(x, y)} = 1.$$

Therefore the risk on the target domain is equal to the risk on the source domain. This formulation provides the foundation for representation learning based domain adaptation as the essential goal is to determine the suitable upper bounds for $\mathcal{R}_T(h)$. We conclude that the target domain error can be bounded by the sum of source domain errors, for a specific task and a distance measure $\widehat{\mathrm{dist}}$ between the distributions of both domains. This distance measure was shown to be upper-bounded in Ben-David *et al.* [7]. A simplified version of this bound is given by

$$\mathcal{R}_T(h) \leq \mathcal{R}_S(h) + \widehat{\operatorname{dist}}(S, T) + C, \tag{7.3}$$

where $\mathcal{R}_S(h)$ is the empirical risk on the source domain S, $\widehat{\operatorname{dist}}(S, T)$ is the *domain divergence* between source domain and target domain and C is a set of data-specific constants. This formula includes all necessary parts for domain adaptation. The term $\mathcal{R}_S(h)$ is determined through training a classification model on source domain data. As an approximation for the domain divergence $\widehat{\operatorname{dist}}(S, T)$ a classification model can be trained to differentiate between source and target samples [7].

Since we are interested in ensuring that the performance on the target domain is similar to the performance on the source domain, our aim is to learn a model that minimises the divergence between source and target domain and maximises antimicrobial resistance prediction performance. In light of that, in the next chapter we design an adversarial representation learning setup, in which the input data is transformed based both on a classification and a domain discrimination objective.

7.2.2 Adversarial domain adaptation for MALDI-TOF mass spectra

We set out to develop a sample representation of MALDI-TOF mass spectra that (i) forms a basis for accurate antimicrobial resistance prediction, (ii) mitigates any domain shifts between source and target, and (iii) generalises to previously unseen domains. As described in the previous subsection, we aim to achieve such a representation by following an adversarial training setup, inspired by previous work [2, 40]. Figure 7.2 illustrates an overview of the approach. The individual components are explained in the following.

Data input and encoder network. The data input format are preprocessed MALDI-TOF mass spectra (see Chapter 3.2), as used for all analyses in Chapter 4. The MALDI-TOF mass spectra are given as input to an *encoder* neural network. This encoder network transforms the spectra into a new representation used for downstream classification and domain discrimination tasks. The aim is to obtain a representation from the encoder that fulfills all above-mentioned objectives. The encoder network is presently a multilayer perceptron (MLP) with a single hidden layer and a ReLU activation function. During implementation each batch consists of sample triples of the form (x, y_{amr}, y_d), where x is a preprocessed MALDI-TOF mass spectrum, $y_{amr} \in \mathcal{Y} = \{0, 1\}$ is the antimicrobial resistance label, and $y_d \in \{1, 2, \ldots\}$ is a domain label. For conducting the experiments, we require three domains; two domains for training, and one domain for the evaluation of our method.

Classifier and discriminator. To train the output of the encoder network towards both objectives, we employ both a *classifier* neural network and a *discriminator* neural network. They use the encoder transformed spectrum, i.e. a high-dimensional representation of a spectrum, as input to predict both the antimicrobial resistance phenotype and the domain, respectively. Both of these classifiers were chosen to be shallow logistic regression models. While more complex architectures, e.g. with convolutional layers, can be easily inserted, the model is intentionally kept simple in this study. The focus of model complexity on the encoder framework, allows for studying the potential benefits of domain adaptation for classification performance. A highly-complex classifier architecture would make it more difficult to disentangle the causes of any improvements.

Adversarial loss terms. The total loss term combines two components, the classifier loss $\mathcal{L}_c$ and the domain discriminator loss $\mathcal{L}_d$. Both components use a cross-entropy loss with a different target. For instance, the classifier employs a binary cross-entropy loss (i.e. a logistic regression loss) on the antimicrobial resistance labels $y_{amr} \in \mathcal{Y} = \{0, 1\}$, given by

$$\mathcal{L}_c(h(x), y) = -(y \log(h(x)) + (1 - y) \log(h(x))), \tag{7.4}$$

where $h(x)$ represents the predicted positive class probability of the model. The discriminator network calculates a (binary) cross-entropy loss according to the domain labels of samples. While we only employ two domains here during training, the framework is capable of training with more than two domains, in which case a multi-class cross-entropy loss would be applied. Note that during the training, the classifier is only aware of the antimicrobial resistance label (i.e. not the domain information), while the discriminator only has access to the domain label (i.e. not the antimicrobial resistance phenotype). This architecture also allows to take advantage of unlabelled samples—i.e. samples for which no resistance phenotype could be retrieved—as these sample could still inform the discriminator network. While we do not take advantage of this possibility in the experiments, we have laid the foundation for this direction for future studies. The joint loss term is of the form

$$\mathcal{L} := \mathcal{L}_c - \lambda \, \mathcal{L}_d \, . \tag{7.5}$$

$\lambda \in \mathbb{R}_{>0}$ controls the influence of the discriminator and $\lambda = 0$ results in no domain adaptation during optimization.

7.2.3 Defining domain dataset and distribution shift characterisation

We conduct the domain adaptation experiments on the resistance classification scenarios employed throughout this thesis, namely predicting (i) ceftriaxone resistance in *E. coli* (abbreviated to E-CEF), (ii) ceftriaxone resistance in *K. pneumoniae* (K-CEF), and (iii) oxacillin resistance in *S. aureus* (S-OXA). Further, we need to form three datasets representing real-world domain separation and suffering from distribution shifts to evaluate the adversarial deep learning model on. A domain should comprise a set of

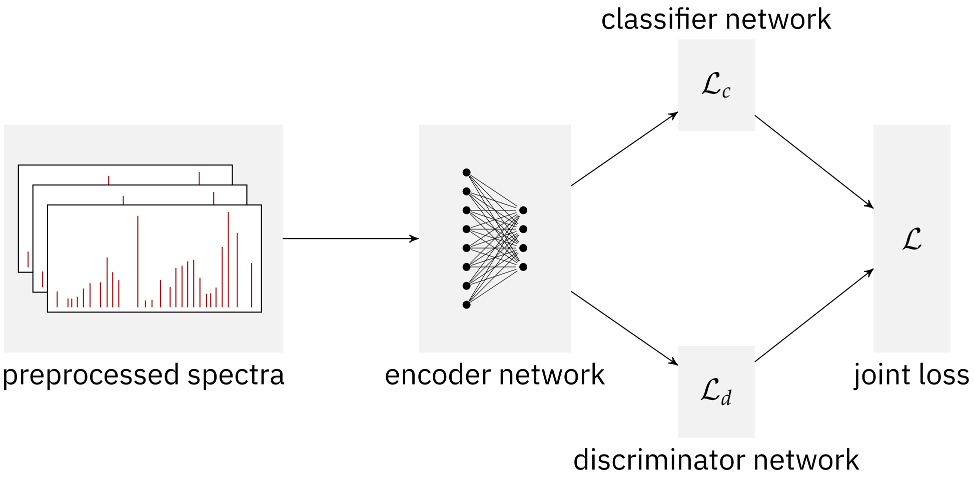


Figure 7.2: **Schematic illustration of adversarial domain adaptation framework**. Preprocessed MALDI-TOF mass spectra are transformed by the encoder network. The encoder is trained in an adversarial fashion, by alternating between minimising the joint loss (and training the classifier network) and minimising the discriminator loss. The framework therefore ensures good predictive performance on the resistance classification task, while learning spectral representations that are domain independent. Figure adapted from Weis *et al.* [127].

properties that allows for drawing samples from a reasonably fixed distribution. The vast amount of samples contains in *DRIAMS* are influenced by several characteristics that can vary over time. We define data domains based on two factors, specifically (a) the clinical routine mass spectra collection site, and (b) the MALDI-TOF MS instrument type. We employ data from both *DRIAMS-A* and *DRIAMS-B*, as these datasets represent different distributions but are still close in terms of geographical distance and inter-site transferability (as seen in Figure 7.1). Further, samples collected at *DRIAMS-A* are subdivided by the instrument type that was used to perform the MALDI-TOF MS measurement. Both the *Microflex Biotyper LT/SH System* and the *Microflex smart LS System* belong to the Bruker Daltonics [12] Microflex Biotyper System instruments. While MALDI-TOF mass spectra collected on both machines are very similar and are, in fact, analysed through the same reference database during the manufacturers species identification, we do see slight differences when comparing the mass spectra. The full extent of the domain shifts will be illustrated in the next subsection (see Figures 7.3 and 7.4). As *DRIAMS-A* and *DRIAMS-B* were collected over different time periods, to minimise confounding distribution shifts caused by biological variations over time, we restrict the samples to those collected in years 2017 and 2018. The rationale for these choices is that for an initial proof-of-concept of our model we want to obtain fairly similar datasets, for which the domain shift can be mitigated relatively easy. Note that the MALDI-TOF mass spectra are used in their preprocessed and binned, fixed-length feature vector representation as input. We subsequently describe each domain and its specific properties in more detail.

DRIAMS-A_mtI. This set of MALDI-TOF mass spectra is a subset of *DRIAMS-A*. All samples were collected in 2017 and 2018. The machine type used to acquire the spectra was the Microflex Biotyper LT/SH System, which will we referred to as *mtI*.

DRIAMS-A_mtII. The samples in this domain dataset are also part of *DRIAMS-A*, also collected in 2017 and 2018. The MALDI-TOF instrument used was a Microflex smart LS System, referred to as *mtII*. This systems differs in the laser gas used in the instrument, but as stated above, species identification can be performed using the *same* database as mtI, i.e. its decisions are based on the same underlying data. We note that the microbiology laboratory collecting both DRIAMS-A_mtI and DRIAMS-A_mtII spectra are the same—as spectra were collected and processed in the same hospital—and both dataset should only suffer domain shifts caused by different MALDI-TOF MS instrument type.

DRIAMS-B_mtI. This set of mass spectra was collected in 2018 at the hospital site of *DRIAMS-B*. *DRIAMS-B* employed a Microflex Biotyper LT/SH System, i.e. the MALDI-TOF MS machine type mtI.

7.2.4 Domain shifts

Generally, MALDI-TOF MS datasets exhibit certain shifts that motivate the use of domain adaptation methods to allow for model transferability. Figure 7.3 illustrates the differences between our defined MALDI-TOF domains in terms of hospitals and MALDI-TOF MS machines. Clear differences are observable, expressing themselves as variably-pronounced peaks and shifts along the m/z–ratio axis. Based on visual inspection, differences between mean MALDI-TOF mass spectra collected at different sites are larger than differences stemming from different MALDI-TOF instrument types. This observation is also supported by low-dimensional tSNE representations of the spectra, as depicted in Figure 7.4.

The visualisations inform our choice of which domains to use during training and evaluation, respectively. The training on one hospital site—regardless of machine type—is realistic in the sense that lab environments, including standard operating procedures and machine operator training, coincide. In this scenario, domain adaptation would be employed to mitigate differences stemming from different MALDI-TOF MS instruments. For model evaluation, we emulate applying the model to a previously unseen site, i.e. a different hospital. This scenario is aligned most realistically with the intent to roll out a trained prediction algorithm to a new hospital for application on local data.

7.2.5 Aligned distributions fail to retain their resistance information

In the following, we discuss our experimental set-up and the results achieved by our model. For model evaluation, we compare our model to two baselines. First, we establish a baseline using a logistic regression without any encoding of MALDI-TOF mass

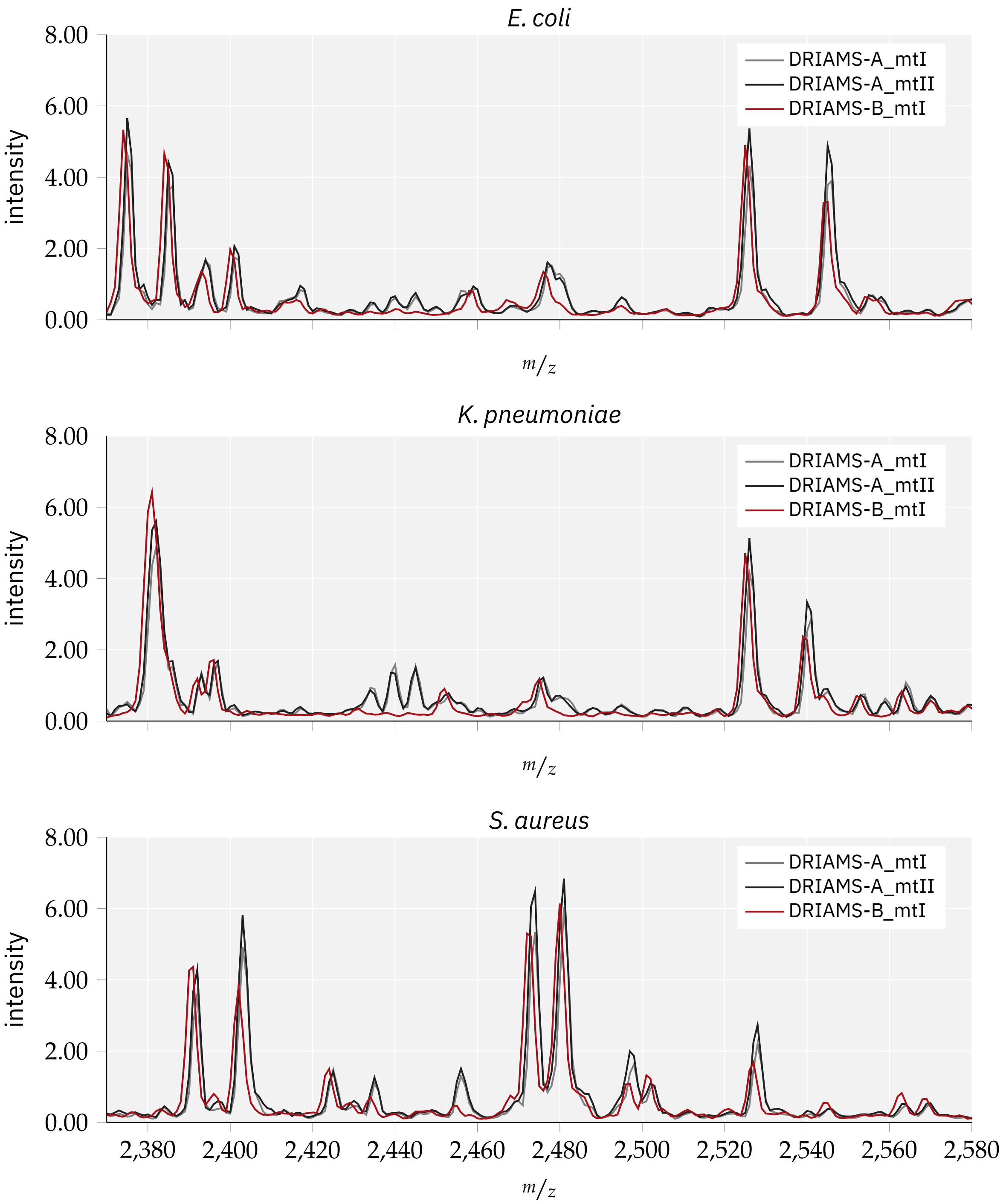


Figure 7.3: **Mean of all preprocessed MALDI-TOF mass spectra for each domain**. Differences in the mean spectra of different domains are clearly visible, expressing themselves in variably pronounced peaks and slight shifts along the m/z–ratio-axis. The m/z range of 2370 to 2580 m/z was chosen for illustrative purposes. Figure adapted from Weis *et al.* [127].

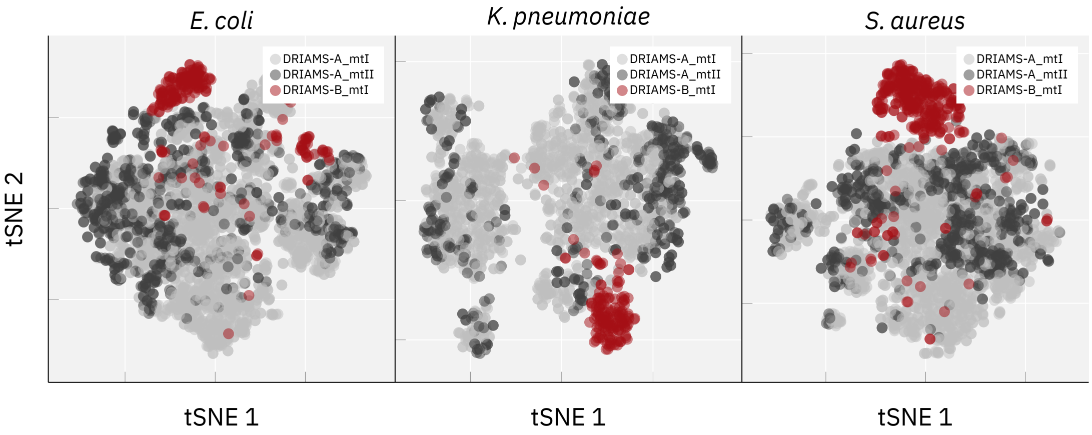


Figure 7.4: **Two-dimensional tSNE representation of MALDI-TOF mass spectra for each domain.** The tSNE was performed with a perplexity parameter of 30.0. Clear differences can be observed between datasets collected at sites *DRIAMS-A* (i.e. DRIAMS-A_mtI and DRIAMS-A_mtII) and *DRIAMS-B* (i.e. DRIAMS-B_mtI). Spectra collected at different instruments but at the same site, i.e. DRIAMS-A_mtI and DRIAMS-A_mtII, are more similar to each other and additional subgroups of datapoints can be observed for all three species. Figure adapted from Weis *et al.* [127].

experiment	LR (*src*)	$\lambda = 0$ (*src*)	$\lambda = 0.1$ (*src*)	LR (*trgt*)	$\lambda = 0$ (*trgt*)	$\lambda = 0.1$ (*trgt*)
E-CEF	72.3 ± 3.4	92.7 ± 1.7	92.8 ± 1.8	48.8 ± 5.8	48.0 ± 0.8	48.1 ± 1.5
K-CEF	72.8 ± 2.9	95.6 ± 2.5	94.5 ± 2.5	17.3 ± 5.7	17.1 ± 3.5	16.5 ± 4.1
S-OXA	77.8 ± 4.1	93.9 ± 0.6	94.2 ± 0.9	21.7 ± 2.8	20.9 ± 2.6	20.2 ± 2.7

Table 7.2: **Evaluation of two baselines and adversarial domain adaptation model on source and target** reported by AUPRC. We apply (i) a logistic regression (abbreviated by LR), (ii) our model with domain adaptation disabled ($\lambda = 0$), and (iii) our model with domain adaptation ($\lambda = 0.1$). The two evaluation domains are comprised of (i) both training domains DRIAMS-A_mtI and DRIAMS-A_mtII, as the source (abbreviated as *src*), and (ii) DRIAMS-B_mtI as target (*trgt*). Scenario abbreviations follow Table 4.1. Figure adapted from Weis *et al.* [127].

spectra based on domain knowledge. The samples from both training domains are combined into one training dataset to train the logistic regression. We then evaluated this non-domain adaptation model on both the test spectra from the train domain and on the independent evaluation domain. For the second baseline, we use our encoding model, but disable any form of domain adaptation by setting λ, and thereby the influence of the discriminator, to zero. Finally, we compare our domain adaptation model with $\lambda = 0.1$ to both of these baselines. The results of both baselines and the introduced domain adaptation framework are depicted in Table 7.2. Predictions reported by the domain adaptation framework are superior on the test dataset of the source domain. As indicated by both versions of the presented model, with and without domain adaptation ($\lambda = 0.1$ and $\lambda = 0$, respectively), classification improves over the logistic regression baseline. This highlights the predictive power of our encoder architecture. Moreover, including the domain adaptation element does not have any adverse effects on generalisation, i.e. the performance on source domain test samples does not decrease.

However, this predictive power does not translate to samples from the target domain. Unfortunately, we observe that both versions of our model are still performing comparative to a logistic regression on the target domain. While the models with either $\lambda = 0.1$ or $\lambda = 0$ exhibit a lower standard deviation and thus indicate a slightly better regularisation power, there are no significant differences in predictive performance between the introduced approach and a logistic regression baseline without any domain adaptation. However, as Figure 7.5 illustrated, the domain adaptation procedure results in a significantly better alignment of source and target domain samples in comparison to the raw tSNE depiction in Figure 7.4.

In conclusion, the experiments do not confirm a benefit of the introduced domain adaptation scenario. We infer that mitigating differences between MALDI-TOF MS instruments is insufficient to provide high predictive performances generalisable to new collection sites. These results match reports by another study [79], which observed a lack of technical and biological reproducibility of MALDI-TOF MS laboratory workflows. While the selection of collection sites mitigate some variance in biological variability—both sites will contain similar strains due to close spatial proximity—the decrease in target domain evaluation performance is driven by other factors that the proposed model cannot account for at present. In light of the distribution depicted in Figure 7.4, we hypothesise that defining more heterogeneous domain datasets will help to improve upon this issue in future work. Additionally, we see further insights into domain differences between datasets as critical for a large-scale inter-site application of MALDI-TOF MS based antimicrobial resistance prediction. One avenue of future research will therefore be an analysis of different types of shifts to shed some additional light on why target domain performance does not improve.

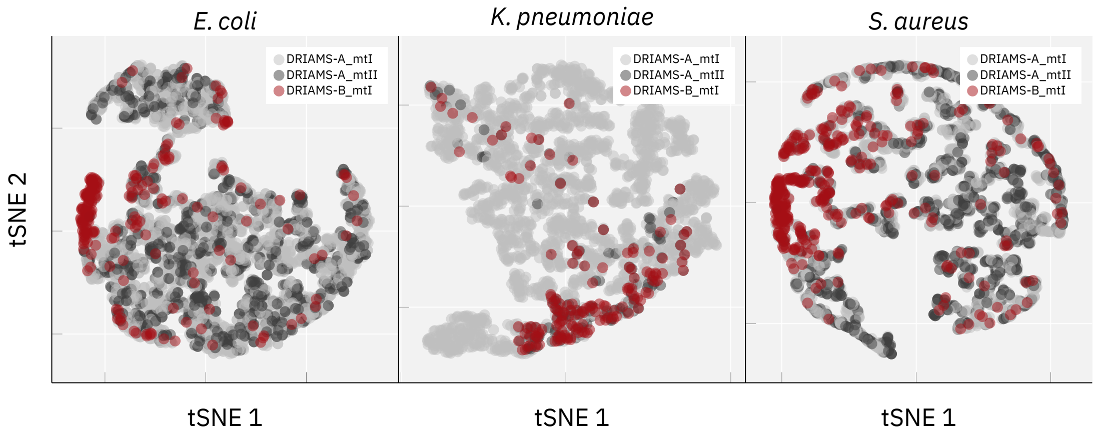


Figure 7.5: **Two-dimensional tSNE representation of the latent spectral representations**. The tSNE was performed with a perplexity parameter of 30.0. The distributions all datasets (i.e. DRIAMS-A_mtI, DRIAMS-A_mtII, and DRIAMS-B_mtI) are aligned to a higher degree compared to Figure 7.4. Figure adapted from Weis *et al.* [127].

7.3 Summary and discussion

In this chapter we have demonstrated that the transferability of an antimicrobial resistance predictor trained at one medical site to an unseen target site is limited, and comes at the cost of large decreases in predictive performance. This observation is in line with previous work [79]. Low transferability is a severe limitation to the application of such a predictor to a new medical site, where too little data is available to train a new classifier. We introduced two approaches to improve predictive performance on the target site.

The first technique boosts the model learning capabilities by combining target training data with external training data from other collection sites. We have demonstrated this approach can lead to large performance gains in comparison to training on the available target data alone, i.e. from 0.08 to 0.27 AUPRC for oxacillin resistance in *S. aureus* on target site *DRIAMS-C*, and from 0.30 to 0.48 AUPRC for oxacillin resistance in *S. aureus* on target site *DRIAMS-B*. These results demonstrate the power and benefits of public databases such as *DRIAMS*—not only as a reference dataset for future research, but also to improve predictions on datasets the users apply it to. While we have achieved great improvements on *DRIAMS-B* and *DRIAMS-C* through this approach, it did not benefit predictions on target site *DRIAMS-D*. We hypothesise that combining training data from several institutions only boosts predictive performance if the variance between individual collection sites is not too large. Shared underlying signal has to exist in the unified datasets that can be learned and used to improve the prediction. The degree to which the captured signal is similar will likely be influenced mostly by the implementation of laboratory protocols. For future work, we propose to investigate whether the usefulness of an external dataset can be predicted and explained through similarity measures, such as a difference in spectral mean or quantifying the distribution shift through Maximum Mean Discrepancy [45].

The second technique leverages the power of state-of-the-art machine learning in the field of adversarial domain adaptation to learn domain independent representations of spectra. This is achieved by training an encoder network in an adversarial fashion, to reach high performance in terms of resistance classification performance and low differentiation performance by a domain predictor. Unfortunately, our results indicate that the proposed framework is not able to cope with the distribution shifts between domains. While the learned representation mitigated differences between domains, it is not able to learn representations with a higher generalisability in resistance prediction. For future work, we aim to address several limitations in the current approach. The distributions shifts that affect MALDI-TOF mass spectra are highly complex and likely originate from a number of different sources that all influence the data in subtle ways. As the method in the current stage was developed to mitigate differences stemming from machine types, it is unlikely that more fundamental shifts—such as shifts arising from the fact that datasets are collected at different laboratories with slight variations in protocols—can be easily mitigated by our domain adaptation procedure. Covering these scenarios has to involve more sophisticated preprocessing and encoder architectures. Moreover, our approach is bound to fail in the presence of major systematic shifts, for instance a translation of the whole MALDI-TOF mass spectrum by 20 Da. In case of such a prominent shift, we expect these shifts to be mitigated better in preprocessing, as they can be corrected for quite easily. This architecture also allows to take advantage of unlabelled samples (i.e. samples for which no resistance phenotype could be retrieved), as these sample could still inform the discriminator network. While we do not take advantage of this possibility in the experiments, we have laid the foundation for this direction for future studies. Lastly—and not directly linking to transferability—the high prediction performance reached by the encoder architecture on the source data implies that more complex deep learning models than used in previous chapters (i.e. MLPs), harbour the potential to outperform the previously introduced methods. For future work, we propose to employ deep learning models with more layers, techniques to improve generalisation performance such as dropout and early stopping, and extend the baseline further.

Part IV

Outlook on the path to clinical MALDI-TOF MS based antimicrobial resistance prediction

8 Necessary steps to reach clinical applicability

Developing a resistance predictor based on MALDI-TOF mass profiles—that is suitable for application within the clinical routine—is a complex endeavor. It is forming its own field within *machine learning for healthcare* research: tackling MALDI-TOF based phenotype prediction, requiring close-nit collaboration with clinical practitioners to determine the model can be merged neatly into the treatment decision flow and consideration of a number of real-world practical aspects of algorithm deployment in a hospital. In this thesis, we have defined the shortcomings hindering the process, laid a broad foundation in data availability and model development, and took first strides to explore several aspects that we regard as essential for clinically-applied antimicrobial resistance models. This chapter summarises the collected experience and outlines the steps we deem necessary to obtain a predictor that is ready to be employed in a clinical setting.

Roadmap towards clinical applicability of a MALDI-TOF MS based resistance predictor. While the predictive performance reached shows the potential to utilize MALDI-TOF mass profiles for resistance prediction, we see many promising research directions unexplored and therefore believe that the upper limit of possible predictive performances is not yet reached. As this list is quite vast, we dedicate an entire paragraph to a detailed description of the individual research directions, which can be found below.

As explored in Chapter 5, a clinically applied prediction *must* be able to recognise out-of-distribution samples. As a results, the model chosen in the end for deployment has to provide reliable and well-calibrated confidence estimates along with its predictions. This safety mechanism can then ward off two potential sources of errors: (i) low quality MALDI-TOF MS measurements, i.e. owing to too little probe on the target plate, and (ii) out-of-distributions from MALDI-TOF mass profiles stemming non-local microbial populations (e.g. imported through travelling), where no informed prediction can be made. These reported confidences can the form the basis for an rejection option within the implementation of the machine learning model.

We propose to create an asymmetric rejection scenario following Weis *et al.* [125]: Two thresholds $\theta_0 \in [0.0, 0.5]$ and $\theta_1 \in [0.5, 1.0]$ are chosen to set the rejection boundaries for the negative and the positive class. If the predicted class score of a classified MALDI-TOF mass spectrum is higher than θ_1 the spectrum is assigned the positive class, while the negative class is assigned if the score is less than θ_0. In cases where the predicted score lies between θ_0 and θ_1, the model refuses any prediction and indicates that the predicted probability lies below the minimum confidence required by the classifier. A prediction model with this rejection option would correspond to the case of setting $\theta_0 = \theta_1 = 0.5$, i.e. none of the spectra are rejected by the algorithm. Sensitivity and specificity are two more metrics frequently-used in the clinic and for evaluating diagnostics tests, that are defined through the confusion matrix (Figure 4.2):

$$sensitivity = \frac{TP}{TP + FN} = \frac{TP}{all\ positives}, \tag{8.1}$$

and

$$specificity = \frac{TN}{TN + FP} = \frac{TN}{all\ negatives}. \tag{8.2}$$

Please note that the terms *recall*, *true positive rate* and *sensitivity* are all synonyms describing the same metric. The rejection thresholds θ_0 and θ_1 can be chosen to optimise the sensitivity and specificity values obtained on an internal validation dataset. Both metrics are connected through their underlying confusion matrix values; a more stringent threshold while predicting the negative class will be beneficial for the sensitivity value, but lower the specificity.

Another process in need of implementations is regular updating, retraining and reevaluating. Our results indicate that regular retraining of the machine learning model with the most recent data is necessary. We recommend a monthly updating protocol, constructed as the following: (i) define the data collected in the most recent month as the new evaluation dataset, (ii) use all data from the time window ending at the evaluation dataset as as training data to update and retrain the antimicrobial resistance predictor, and (iii) keep continuous monitoring of the performance of the predictor and distribution of confidence estimates on the most recent evaluation dataset. Particular care has to be taken when major changes to properties of the MALDI-TOF mass spectra are performed, i.e. recalibration of machine parameters through technicians from the manufacturing company, or changing the diode of the laser etc. Note the this protocol requires expert knowledge on machine learning and therefore, a person with the right expertise would need to be responsible for executing this protocol and supervising the quality of predictions. We also want to reiterate the classifier's sensitivity to small discrepancies in the spectra (see low transferability in Chapter 7) stemming from measurement on different MALDI-TOF MS instruments. As a result, the continued functionality of any predictor is highly dependent on the stability of the manufacturers spectral processing and software. For a long-term steady clinical machine learning model, collaboration with the manufacturers might be require to be warned of a changes head-on.

After implementation of this procedure, the usefulness of our approach for the patients can be evaluated through a prospective clinical study. These steps should cover then all major topic necessary to obtain a clinically-applicable antimicrobial resistance predictor.

Further MALDI-TOF MS tailored method development. We envision several lines of research to further increase the predictive performance: The high inconsistency intrinsic to MALDI-TOF MS measurements—which suffer from low peak reproducibility, high variance in the intensities along the y–axis, high discrepancies between spectra collected at different instruments—pose a large, and potentially the largest, challenge for MALDI-TOF MS based phenotype prediction. While this variation (depicted in Figure 7.4) can be handled by certain machine learning methods, and we will discuss approaches below, an improved preprocessing of MALDI-TOF mass spectra holds the highest potential to mitigate these differences. While we have seen superior performance for the well-established `MaldiQuant` preprocessing pipeline, we have not yet explored how warping to a reference spectrum would influence both the MALDI-TOF mass spectral distribution and subsequent prediction results. The challenge with this approach would be how to define the *golden spectrum*, i.e. the mass spectrum that serves as a reference for the warping algorithm. The golden spectrum has to be determined per species, which would then lead to a species-specific preprocessing method. In order to calculate the warping function, a significant overlap between peaks in the target spectrum and the golden spectrum is required. This could lead to the need for excluding spectra with too little matching peaks from spectral preprocessing and therefore prediction and inclusion into the dataset. While this could be seen as a pitfall of reference spectra calibration, this procedure could also provide an extra quality control step, by requiring that MALDI-TOF MS spectra with too little similarity to the reference spectrum is remeasured. While we see challenges with this approach for large-scale MALDI-TOF MS data analysis, the potential for improving downstream machine learning prediction warrants its exploration.

The results in Chapter 7 indicate that models including state-of-the-art modules with a higher complexity, e.g. architectures with more layers, including early stopping and dropout, could improve predictions. Further improvements could be reached from including other types of networks, such as convolutional layers instead of solely fully-connected layers. Convolutions are thought to be less sensitive to variations in peak position along the m/z–ratio depicted on the x–axis.

We further believe in the goal of learning site-independent representation through representation learning techniques. The experiments conducted in Chapter 7 indicate that real-world clinical dataset suffer from a distribution shift too large to be mitigated through our adversarial representation learning approach. However, we believe that collecting a new dataset directly designed for this task will allow for learning of latent encoding independent from collection site. We propose the collection of a biologically-stable MALDI-TOF MS dataset, that includes measurements covering the exact same set of microbial samples collected at several clinical laboratories. The dataset should be sufficiently large to allow for machine learning analysis after stratification splits. We estimate at least 2,500 MALDI-TOF MS are required (based on Figure 4.7) from each collection site; however representation learning models could require higher numbers of samples. Note that these measurements are still only approximations of the real domain shifts, as they only capture differences at the specific time when all samples are measured, and do not cover variances that span longer time periods, such as differences between laboratory staff.

Decreasing the time-to-MALDI-TOF MS using bacterial enrichment. Next, we briefly outline the current status of a research area that has the potential to further improve the speed and efficacy of MALDI-TOF MS based phenotype prediction—shortening the time until the MALDI-TOF mass profile can be determined, by employing bacterial enrichment techniques. Traditional MALDI-TOF MS requires single bacterial colonies, and thereby a culturing step that limits the speed of the bacterial identification [135]. Enrichment methods aim to speed up the analysis by shortening, or fully avoiding, the bacterial culturing step. These methods often utilise nano- or microparticles for affinity binding, e.g. Yi *et al.* [135] employ *rabbit immunoglobulin G* bound to Fe_3O_4 ($IgG@Fe_3O_4$). The immunoglobulin part of $IgG@Fe_3O_4$ binds to bacterial cells, after which the bacteria@$IgG@Fe_3O_4$ conjugate can be collected and washed [135]. Then, the material is directly applied to the MALDI-TOF MS target plate for measurement. The authors [135] report that by using their method, one can obtain a MALDI-TOF mass spectrum 40% faster than through the standard protocol for blood cultures and have demonstrated that it is able to identify six different species. However, a disadvantage of directly smearing the bacteria captured through enrichment materials onto the target plate is the confounding signal, produced by the enrichment molecules themselves. The resulting MALDI-TOF mass spectra will differ from current reference databases, which were measured through standard MALDI-TOF MS, thereby leading to a higher error rate during species identification. In light of that, Sun *et al.* [114] propose a protocol to release the enriched bacteria from the capturing molecules before measurement. Further reservations exist for implementing these approaches, as engineers have expressed concerns that exposing to these magnetic beads would cause damage to the MALDI-TOF MS detector over time [114]. Presently, none of these methods are ready to be employed to the clinical routine. Nonetheless, the active research and swift progress in the field of MALDI-TOF MS further increases the possibilities of MALDI-TOF mass spectra based phenotype prediction.

Genomics based antimicrobial resistance prediction. In this paragraph we explore the current state of antimicrobial resistance prediction based on another datatype. In 2017, the an EUCAST subcommittee report concluded that “for most bacterial species there is currently insufficient evidence to support the use of whole genome sequencing-inferred antimicrobial susceptibility testing to guide clinical decision making” [31]. Still, inferring of antimicrobial resistance phenotypes based on whole-genome and genotype sequencing using machine learning is gaining momentum, with many of the developments within the last two years. The field is currently at a more advanced stage of development than MALDI-TOF MS based AMR prediction, with public databases providing a number of datasets specifically collected for genome based resistance prediction in bacteria. The PATRIC database collects a multiplicity of data and a set of analysis tools to build a foundation for researchers to study antimicrobial resistance and causal genetic determinants. The data includes bacterial genomes and genome regions together with the corresponding antimicrobial resistance labels [24]. Single additions to the database are can reach substantial sizes, e.g. VanOeffelen *et al.* [117] provide a dataset comprised of 67,000 genomes from more than 100 bacterial species. Further software tools are provided for predicting antimicrobial resistance in bacteria, for example Chowdhury *et al.* [18] [19] employ a feature selection based on the Banzhaf power index. For its subsequent AMR classification they use a SVM classifier with the RBF kernel, reporting acetyltransferase, β-lactamase, and dihydrofolate reductase AMR protein sequences predictions for Gram-negative bacteria with an accuracy ranging from 0.93 to 0.99 [18] and classification accuracies between 0.87 and 0.90 for bacitracin and vancomycin resistance in Gram-positive bacteria [19].

Nevertheless, the process of implementing antimicrobial resistance prediction based on genomics into the clinical microbiology is progressing slowly. The technique is stymied by the slow turnaround time, added costs to the clinics, and the lack of robust results demonstrating effectivenes for patient treatment [91]. The time until a result is obtained varies between sequencing technologies, with Illumina sequencing taking at least 24 h and PacBio between 0.5 h to 24 h [91]. Rossen *et al.* [99] second this assessment for whole genome sequencing, estimating that measurement of 16 to 20 bacterial isolates would cost around 200 euros per isolate in the clinical routine setting and take 2.5 days to 3 days to report sequences. For comparison, MALDI-TOF MS measurements run within minutes after culture growth. Both studies [91] [99] see a big disadvantage in the fact that genome analysis is currently not part of the standard clinical laboratory.

In summary, antimicrobial resistance phenotyping based on whole-genome and genotype sequencing is an intriguing opportunity to provide early antimicrobial susceptibility labels. However, we conclude that MALDI-TOF MS based resistance prediction can be implemented much faster and at a much lower cost, due to the fact the MALDI-TOF MS is already the most widely-used technique for species identification.

Part V

Appendix

Data availability

DRIAMS

The data is publicly-available at the Dryad data repository[1]. For each site, the data consists of MALDI-TOF mass spectra in the form of `.txt` files and a meta-data file: (i) The meta-data, including species and antimicrobial resistance corresponding to each spectra, is part of the `id` folder, and (ii) the remaining folders store the MALDI-TOF mass spectra in various stages of preprocessing; `raw` all spectra as extracted from the MALDI-TOF MS instrument, `preprocessed` all spectra after the application of an established preprocessing pipeline and `binned_6000` all spectra after the application of an established preprocessing pipeline and binning along the mass-to-charge-ratio axis with a bin size of 3 Da, resulting in 6000 feature bins.

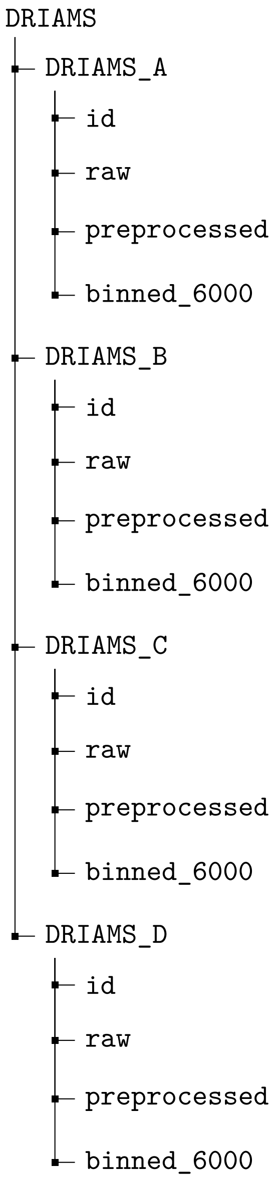
We recommend using our `Python` package maldi_learn[2], to load and analyse DRIAMS data for MALDI-TOF preprocessing and machine learning analysis. The github package comes with an elaborate README.md file, which gives details on installation and usage examples. In order to use this package the locations of data files and folder structure must be preserved. Please note that all four downloaded data packages should be kept in one folder, serving as the DRIAMS root folder, which then needs to be set as the DRIAMS_ROOT path in the .env file.

The folder structure obtained after download is depicted on the following page. Each folder—`id`, `raw`, `preprocessed`, `binned_6000`—contains folders according to the collection year, which then contain all `.csv` files.

---

[1] doi:10.5061/dryad.bzkh1899q

[2] https://github.com/BorgwardtLab/maldi_learn



Appendix Figure 1: ***DRIAMS* data file structure**

Software availability

In order to encourage method development by the community, we make the code used for the analysis and the `Python`-based MALDI-TOF MS processing libraries publicly available.

DRIAMS analysis

All code[1] relating to experiments in Weis *et al.* [126] was made publicly available.
A supporting software package for MALDI-TOF MS data processing was provided on GitHub[2] to facilitate *DRIAMS* read-in, exploration and filtering.

GP-PIKE

The code to replicate experiments in Weis *et al.* [128] is publicly available on GitHub[3].

---

[1] https://github.com/BorgwardtLab/maldi_amr
[2] https://github.com/BorgwardtLab/maldi_learn
[3] https://github.com/BorgwardtLab/maldi_PIKE

List of Figures

1.1 **Exemplary MALDI-TOF MS spectrum** depicting an *E.coli* sample. The spectrum is in its raw form as it was extracted from the MALDI-TOF MS instrument, i.e., no spectral preprocessing has been performed. The MALDI-TOF MS baseline signal is clearly observable up until 11.5 kDa. The mass spectrum was trimmed to 2 kDa to 12 kDa for illustration purposes. 5

1.2 **The current usage of MALDI-TOF MS in the clinic and the potential inclusion of a resistance predictor.** The upper timeline (following the black arrow) depicts the antimicrobial species and resistance workflow as it is currently implemented in many clinics, including the culture growth waiting times between steps and the point of MALDI-TOF MS measurement. The lower part (following from black to dashed and red arrow) depicts the vision of a machine learning driven prediction approach running in parallel to the established diagnostics path. Antimicrobial therapy guided by a MALDI-TOF MS based predictor could lead to an informed treatment decision 24 h to 72 h earlier than in the current workflow focused solely on diagnostics. All icons are listed in the Noun Project [1]. 7

2.1 **PRISMA flowchart** illustrating the complete literature search pipeline. In total, 960 studies were found using the search phrase and one additional publication was identified by manual search of references. 632 studies are excluded based on their article type or being duplicates in the search results. We exclude 267 studies as their research topic was not among the topic of antimicrobial phenotype prediction. Reading through the articles revealed that 26 studies do not apply machine learning methods. The remaining 36 studies are included in the review. Figure adapted from Weis *et al.* [129]. 15

3.1 ***DRIAMS* data collection and curation**. The upper row indicates the steps taken to collect and filter samples, consisting of MALDI-TOF mass spectra and AMR profiles. The lower row is a depiction of the creation of both the preprocessed and binned MALDI-TOF mass spectra, which (along with the raw spectra and AMR profiles) comprise all four subdatasets of *DRIAMS*. Figure adapted from Weis *et al.* [126]. 28

4.1 **Multi-layer perceptron model** (a) Activation functions used in neural networks, such as a heaviside step function $H(x)$ and rectified linear units $ReLU(x)$. (b) Structure of multiple layers of node in a multi-layer perceptron model, where all nodes in hidden layers take the output of previous layers as their input. 43

4.2 **Confusion matrix and resulting metrics** (a) two-by-two confusion matrix depicting the four possible outcomes in a binary classification, defined by the actual and predicted class (b) commonly reported machine learning metrics that are derived from the confusion matrix, including true and false positive rate (required for AUROC), precision and recall (AUPRC), and accuracy. 44

4.3 **Predictive performance increase through MALDI-TOF MS compared to species information**. For each antimicrobial drug resistance prediction task in *DRIAMS-A*, two logistic regression classifiers are trained: one predicting resistance for MALDI-TOF mass spectra and one based on species information alone. Note that for each drug, the dataset is comprised of all samples with a resistance label available for the respective drug and contains a number of different species. The red bars depict the AUROC when basing prediction on MALDI-TOF MS information, the grey bars indicating the predictive performance based on species information alone. The percentages of positive class (resistant/intermediate) samples in the training data are stated in brackets after the antibiotic name. The AUROC values and error bars depict the mean and standard deviation of ten random train-test-splits, while the asterisks indicate a significance level of less than 5% between the reported metrics of both approaches determined by a non-equal-variance Welch's t-test. Figure adapted from Weis *et al.* [126]. 46

4.4 **ROC and precision-recall curves** of three machine learning models for resistance in *E. coli* to various different antibiotics. Figure adapted from Weis *et al.* [126]. 48

4.5 **ROC and precision-recall curves** of three machine learning models for resistance in *K. pneumoniae* to various different antibiotics. Figure adapted from Weis *et al.* [126]. 49

4.6 **ROC and precision-recall curves** of three machine learning models for resistance in *S. aureus* to various different antibiotics. Figure adapted from Weis *et al.* [126]. 50

4.7 **Relationship between sample size and predictive performance** for training data consisting of a single species (dash-dotted line) vs. an ensemble (solid line) of species. The test data only includes MALDI-TOF mass spectra from the target species. The depicted species-antibiotic scenarios and models are selected based on Table 4.1. Curves trained on single-species data increase much more rapidly and often outperform even the full ensemble dataset. Figure adapted from Weis *et al.* [126]. . 53

4.8 **Sliding eight-month training window illustrates advantage of contemporary samples**. Each point depicts the performance of a predictor trained on all samples collected during the eight month window, ending on the date indicated on the x-axis. The test data is comprised of all instances sampled in the four months starting on the rightmost date on the x-axis. The dates are given in the format *dd.mm.yyyy*. The curves indicate a performance decrease with larger distance between the test and training window. All instances stem from *DRIAMS-A*. Scenario abbreviations in the legend follow Table 4.1. Figure adapted from Weis *et al.* [126]. . . 54

4.9 **Connection between number of training samples and predictive performance**. Each datapoint corresponds to a eight month training interval and value in Figure 4.8, with arrows indicating the direction of time passing. The progression towards the upper right corner indicates a correlation between increasing sample size and an increase in predictive performance. All instances stem from *DRIAMS-A*. Scenario abbreviations in the legend follow Table 4.1. Figure adapted from Weis *et al.* [126]. 55

4.10 **Retrospective clinical case study**. The 63 patient cases are grouped by species. Treatment recommendations are displayed in grey tones, with escalation (E) highlighted in black, deescalation (D) in light grey and continuation (K) of the current treatment in grey. The resistance class is indicated by red panels, with the positive class (R/I) is highlighted in dark red. Both the predicted resistance and the treatment decisions employing the prediction are highlighted in bold font. Figure adapted from Weis *et al.* [126]. 62

4.11 **Summary of clinical impact of predictor**. In 54 cases the algorithm would have made no difference to the treatment decision. For nine cases, the treatment recommendation by the specialist is influenced by the predictor: For seven patients the antibiotic therapy was correctly deescalated and in one clinical case the classifier caused a changed in treatment decision from escalation to continuation of therapy. For one patient however, the algorithm lead to an unnecessary escalation of antibiotic therapy. In summary, the retrospective clinical case study indicates that in eight out of nine cases (89%) the resistance predictor would have induced a beneficial treatment decision. 63

4.12 **MALDI-TOF MS feature bins with highest Shapley values** for the three antibiotic-species scenario and model selected in 4.1.3. The feature bins were cross-referenced with mass peaks described to be associated with the respective resistance in the literature. Bins corresponding to a known MALDI-TOF mass peak are marked with a red asterisk and can be found in Table 4.3. Note that there were no ceftriaxone resistance associated MALDI-TOF mass peaks for *K. pneumoniae* found in the literature. Scenario abbreviations follow Table 4.1. Figure adapted from Weis *et al.* [126]. 65

5.1 **Schematic illustration of the influence of the proposed persistence transformation on MALDI-TOF mass spectrum**. The top graph depicts a raw MALDI-TOF mass spectrum without any alignment or preprocessing. The bottom graph shows a persistence transformed spectrum produces a simplified and clean spectral representation, with the y-axis changing from an *intensity* to a *persistence*. Figure adapted from Weis *et al.* [128]. 70

5.2 **Influence of kernel parameter t on feature map $u(x, t)$** of a MALDI-TOF mass spectrum in its mass peak representation. The raw mass peaks are slowly diffused over the whole x-axis. With an increasing t the influence of a single peak is reduced. Figure adapted from Weis *et al.* [128]. 74

5.3 **Histogram depicting the maximum class probability distribution for in- and out-of-distribution samples** for logistic regression (LR) (left column) and GP–PIKE (right column) trained on *S. aureus*. The first row depicts the MCP in-training distribution from *S. aureus*, while the second and third rows depict the values for out-of-distributions samples from *E. coli* and *K. pneumoniae*. Figure adapted from Weis *et al.* [128]. 83

5.4 **Improvement of predictive accuracy with increasing rejection threshold θ** on the task of amoxicillin-clavulanic acid resistance prediction in *S. aureus*. To allow for comparability between test datasets with varying class ratios for each θ, the predictive accuracy is reported. The coarseness for larger θ is due to small sample size effects. Figure adapted from Weis *et al.* [128]. 85

6.1 **Illustration of the construction of Silhouette score and Davies-Bouldin index**. (6.1a) The mean distance between one point (marked by black edge) and all points in the same cluster is marked by a, while the mean distance to the points in the nearest cluster is denoted by b. (6.1b) The mean distance of all points in a cluster to its centroid point are denoted by s_i for cluster i and s_j for cluster j. These intra-cluster distances, e.g. s_i and s_j, are compared with the inter-cluster distance between the respective centroids d_{ij}. 91

6.2 **Dendrogram of the complete hierarchical clustering tree** using ward's linkage to cluster the samples of the *S. aureus* (amoxicillin-clavulanic acid) dataset. All tree tips represent a single sample in the dataset. The colors indicated the clustering with choice $k^* = 10$. Note that the clustering does not depend on a k; the cluster labels are through a cut-off at the desired "height" of the determined dendrogram. 93

6.3 **Behaviour of cluster validity scores with an increasing number of clusters** using five different linkage criteria. No scores are reported for $k = 1$ as at least two clusters are required to compute these validity scores. For each of the nine species-antibiotic scenarios, a curve is drawn, leading to three curves per species. The dashed vertical line depicts the k^* value chosen for the respective species. 94

6.4 **Predictive performance using logistic regression illustrating the influence of different linkage criteria and the number of clusters** k of all nine antimicrobial resistance scenarios. The results are given in mean average precision (AUPRC) $\pm$ standard deviation on the test data over 5 random splits. The dashed vertical line depicts the k^* value chosen for the respective species. 97

6.5 **Predictive performance using GP–PIKE illustrating the influence of different linkage criteria and the number of clusters** k of all nine antimicrobial resistance scenarios. The results are given in mean average precision (AUPRC) $\pm$ standard deviation on the test data over 5 random splits. The dashed vertical line depicts the k^* value chosen for the respective species. 98

7.1 **Direct transferability of *DRIAMS* subdataset trained classifier** quantified by AUROC. The y–axis presents the train datasets of *DRIAMS-A*, *DRIAMS-B*, *DRIAMS-C* and *DRIAMS-D*, while the x–axis present the respective test datasets. Hence, the values on the diagonal correspond to intra-site training and testing. No *S. aureus* samples with oxacillin label are available in *DRIAMS-D*, causing gaps in 7.1c. Scenario abbreviations follow Table 4.1. Figure adapted from Weis *et al.* [126]. 104

7.2 **Schematic illustration of adversarial domain adaptation framework**. Preprocessed MALDI-TOF mass spectra are transformed by the encoder network. The encoder is trained in an adversarial fashion, by alternating between minimising the joint loss (and training the classifier network) and minimising the discriminator loss. The framework therefore ensures good predictive performance on the resistance classification task, while learning spectral representations that are domain independent. Figure adapted from Weis *et al.* [127]. 108

7.3 **Mean of all preprocessed MALDI-TOF mass spectra for each domain**. Differences in the mean spectra of different domains are clearly visible, expressing themselves in variably pronounced peaks and slight shifts along the m/z–ratio-axis. The m/z range of 2370 to 2580 m/z was chosen for illustrative purposes. Figure adapted from Weis *et al.* [127]. . . . 110

7.4 **Two-dimensional tSNE representation of MALDI-TOF mass spectra for each domain**. The tSNE was performed with a perplexity parameter of 30.0. Clear differences can be observed between datasets collected at sites *DRIAMS-A* (i.e. DRIAMS-A_mtI and DRIAMS-A_mtII) and *DRIAMS-B* (i.e. DRIAMS-B_mtI). Spectra collected at different instruments but at the same site, i.e. DRIAMS-A_mtI and DRIAMS-A_mtII, are more similar to each other and additional subgroups of datapoints can be observed for all three species. Figure adapted from Weis *et al.* [127]. 111

7.5 **Two-dimensional tSNE representation of the latent spectral representations**. The tSNE was performed with a perplexity parameter of 30.0. The distributions all datasets (i.e. DRIAMS-A_mtI, DRIAMS-A_mtII, and DRIAMS-B_mtI) are aligned to a higher degree compared to Figure 7.4. Figure adapted from Weis *et al.* [127]. 113

List of Tables

2.1 **Machine learning algorithms applied in the 36 reviewed studies.** Multiple algorithms can be applied in a single study. 16

2.2 **Description of the most frequent machine learning algorithms** employed in the current literature [129]. 18

2.3 **Quality assessment of reviewed studies** All studies analysed in the systematic review are assessed with respect to nine quality criteria, which are described in more detail in Section 2.1. 22

2.4 **Summary of reviewed studies addressing species prediction.** Abbreviations: Algorithm abbreviations follow Table 2.1. 'ask' indicates that the study authors claim to give out code or data upon request. l.o.o. for 'leave-one-out'. The depicted columns address the limitations discussed in Section 2.3. A comprehensive version of this table can be found in Weis et al. [129]. 23

2.5 **Summary of reviewed studies addressing antimicrobial resistance prediction.** Abbreviations: Algorithm abbreviations follow Table 2.1. 'ask' indicates that the study authors claim to give out code or data upon request. l.o.o. for 'leave-one-out'. The depicted columns address the limitations discussed in Section 2.3. A comprehensive version of this table can be found in Weis et al. [129]. 24

3.1 **Distribution of samples over workstations** for three species–antibiotic datasets from *DRIAMS-A*. 33

3.2 ***DRIAMS-A* summary (Part 1)** n refers to the number of MALDI-TOF mass spectra with a phenotype for the respective drug. '% class' indicates the percentage of these mass spectra having a positive class, meaning either intermediate or resistant phenotype. Antimicrobial nomenclature and suffixes are described in detail in Section 3.2. 35

3.3 ***DRIAMS-A* summary (Part 2)** n refers to the number of MALDI-TOF mass spectra with a phenotype for the respective drug. '% class1' indicates the percentage of these mass spectra having a positive class, meaning either intermediate or resistant phenotype. Antimicrobial nomenclature and suffixes are described in detail in Section 3.2. 36

3.4 ***DRIAMS-B* summary** n refers to the number of MALDI-TOF mass spectra with a phenotype for the respective drug. '% class1' indicates the percentage of these mass spectra having a positive class, so either intermediate or resistant phenotype. Antimicrobial nomenclature and suffices are described in detail in Section 3.2. 37

3.5 ***DRIAMS-C* summary** n refers to the number of MALDI-TOF mass spectra with a phenotype for the respective drug. '% class1' indicates the percentage of these mass spectra having a positive class, so either intermediate or resistant phenotype. 38
3.6 ***DRIAMS-D* summary** n refers to the number of MALDI-TOF mass spectra with a phenotype for the respective drug. '% class1' indicates the percentage of these mass spectra having a positive class, so either intermediate or resistant phenotype. 39

4.1 **Species-stratified antimicrobial resistance predictors** for several antimicrobial drugs in *DRIAMS-A*. Three machine learning models are applied, with LightGBM performing consistently best for *E. coli* and *S. aureus* and MLP for *K. pneumoniae*. The scenarios selected based on predictive performance and clinical applicability are marked. 51
4.2 **Predictive performance comparison between non-calibrated and calibrated classifiers** reported in AUROC and accuracy. Accuracy values are improved in all cases through calibration. AUROC values are not affected by calibration (neither are AUPRC values — not depicted). Scenario abbreviations follow Table 4.1; logistic regression is abbreviated as LR. 57
4.3 **MALDI-TOF mass peaks found to be correlated with resistance in literature can be attributed to highly contributing feature bins**. The column 'rank amongst Shapley' states the position among the 30 highest contributing feature bins regarding Shapley values in Figure 4.12. The rightmost column states the specific target, through which the MALDI-TOF mass peak is associated with the resistance. In most cases, this target is a species substrain that harbours a resistance mechanism. Note that only for *E. coli* (ceftriaxone resistance) and *S. aureus* (oxacillin resistance) relevant studies could be found in the literature. Abbreviations: Clonal complex is abbreviated with CC and sequence type with ST; scenario abbreviations follow Table 4.1. 66

5.1 **MALDI-TOF mass peak dataset based on *DRIAMS-A***. Nine species–antibiotic scenarios are included, and their scenario abbreviation, sample size and positive class ration stated. The positive class consists of resistant and intermediate samples. 78
5.2 **Performance comparison of persistence transformation and tradition MALDI-TOF spectral preprocessing** through AUPRC±standard deviation using both logistic regression and GP–PIKE model. The predictive performance of predictions on `MaldiQuant`-preprocessed mass spectra is higher in most scenarios using the logistic regression model, and consistently higher for GP–PIKE. Scenario abbreviations follow Table 5.1. Other abbreviations: persistence transformation (PT), `MaldiQuant` [43] v1.19 (MQ), and logistic regression (LR). 80

5.3 **Performance evaluation of GP–PIKE classifier** compared to two models: logistic regression and a Gaussian Process with an established kernel, the RBF kernel (GP–RBF). GP–PIKE outperforms both comparison models in all scenarios. Scenario abbreviations follow Table 5.1. 81

6.1 **Improved predictive performance employing the hierarchical train–test split** compared to a standard random train-test split. Results are reported through AUPRC mean ± standard deviation on five test datasets. The methods include $k^* = 8$ for *E. coli* and $k^* = 10$ for *S. aureus* and *K. pneumoniae*. 95

7.1 **Combining data from several collection sites can improve predictions on a target site with few training samples** expressed through AUPRC. Predictive performance increases on *DRIAMS-A* and *DRIAMS-B* when training is performed on a union of the respective training data with other sites. *DRIAMS-D* does not benefit from an expanded training dataset with other collection sites. 103

7.2 **Evaluation of two baselines and adversarial domain adaptation model on source and target** reported by AUPRC. We apply (i) a logistic regression (abbreviated by LR), (ii) our model with domain adaptation disabled ($\lambda = 0$), and (iii) our model with domain adaptation ($\lambda = 0.1$). The two evaluation domains are comprised of (i) both training domains DRIAMS-A_mtI and DRIAMS-A_mtII, as the source (abbreviated as *src*), and (ii) DRIAMS-B_mtI as target (*trgt*). Scenario abbreviations follow Table 4.1. Figure adapted from Weis *et al.* [127]. 111

Acronyms

AMR	Antimicrobial resistance prediction
AUPRC	Area under the precision recall curve
AUROC	Area under the receiver operating characteristic
ESBL	extended-spectrum beta-lactamases
EUCAST	European Committee Antimicrobial Susceptibility Testing
GP	Gaussian Process
GP–PIKE	Gaussian Process employing the PIKE kernel
LightGBM	Light Gradient Boosting Machine
LR	logistic regression
MALDI-TOF	Matrix-assisted laser desorption ionization time of flight
MCP	maximum class probability
MLP	multi-layer perceptron
MRSA	methicillin-resistant *S. aureus*
MS	Mass spectrometry
MSSA	methicillin-susceptible *S. aureus*
PCR	Polymerase chain reaction
PIKE	peak information kernel
RBF	radial basis function
ReLU	rectified linear units
RKHS	reproducing kernel Hilbert space
SVM	support vector machine
TIC	total-ion-current
WHO	World Health Organization

Glossary

E. coli	Shortened naming of the bacterium *Escherichia coli*
K. pneumoniae	*Klebsiella pneumoniae*
S. aureus	*Staphylococcus aureus*
tSNE	Abbreviation for t-distributed stochastic neighbor embedding, a technique for visualizing high-dimensional data

Bibliography

1. Adrien Coquet, mark smith, Fredrik Edfors, sripfoto, and mungang kim. *Noun Project*. 2021.
2. H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand. *Domain-Adversarial Neural Networks*. 2015. arXiv: `1412.4446 [stat.ML]`.
3. K. Asakura, T. Azechi, H. Sasano, H. Matsui, H. Hanaki, M. Miyazaki, T. Takata, M. Sekine, T. Takaku, T. Ochiai, N. Komatsu, K. Shibayama, Y. Katayama, and K. Yahara. "Rapid and easy detection of low-level resistance to vancomycin in methicillin-resistant Staphylococcus aureus by matrix-assisted laser desorption ionization time-of-flight mass spectrometry". *PLOS ONE* 13:3, 2018. Ed. by B.-L. Lee, e0194212. doi: `10.1371/journal.pone.0194212`.
4. J. Bai, Z. C. Fan, L. P. Zhang, X. Y. Xu, and Z. L. Zhang. "Classification of Methicillin-Resistant and Methicillin-Susceptible Staphylococcus Aureus Using an Improved Genetic Algorithm for Feature Selection Based on Mass Spectra". In: *Proceedings of the 9th International Conference on Bioinformatics and Biomedical Technology - ICBBT '17*. ACM Press, 2017. doi: `10.1145/3093293.3093299`.
5. J. F. Banzhaf III. *Weighted voting doesn't work: A mathematical analysis*. Vol. 19. Rutgers Law Review, 1965, pp. 317–343.
6. M. Belkin and P. Niyogi. "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering". In: *Advances in Neural Information Processing Systems 14*. 2002, pp. 585–591.
7. S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. "Analysis of Representations for Domain Adaptation". In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press, 2007.
8. E. R. Bevan, A. M. Jones, and P. M. Hawkey. "Global epidemiology of CTX-M β-lactamases: temporal and geographical shifts in genotype". *Journal of Antimicrobial Chemotherapy* 72:8, 2017, pp. 2145–2155. doi: `10.1093/jac/dkx146`.
9. BioMérieux. `https://www.biomerieux.com`. 2018.
10. K. M. Borgwardt. "Kernel Methods in Bioinformatics". In: *Handbook of Statistical Bioinformatics*. Springer, 2011, pp. 317–334. doi: `10.1007/978-3-642-16345-6_15`.
11. C. Brouard, H. Shen, K. Dührkop, F. d'Alché-Buc, S. Böcker, and J. Rousu. "Fast metabolite identification with Input Output Kernel Regression". *Bioinformatics* 32:12, 2016, pp. i28–i36. doi: `10.1093/bioinformatics/btw246`.

12. Bruker Daltonics. `https://www.bruker.com`. 2018.
13. I. Burckhardt and S. Zimmermann. "Susceptibility Testing of Bacteria Using Maldi-Tof Mass Spectrometry". *Frontiers in Microbiology* 9, 2018. doi: `10.3389/fmicb.2018.01744`.
14. K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò. "Power failure: why small sample size undermines the reliability of neuroscience". *Nature Reviews Neuroscience* 14:5, 2013, pp. 365–376. doi: `10.1038/nrn3475`.
15. M. Camoez, J. M. Sierra, M. A. Dominguez, M. Ferrer-Navarro, J. Vila, and I. Roca. "Automated categorization of methicillin-resistant Staphylococcus aureus clinical isolates into different clonal complexes by MALDI-TOF mass spectrometry". *Clin. Microbiol. Infect.* 22:2, 2016, pp. 1–161.
16. M. Camoez, J. Sierra, M. Dominguez, M. Ferrer-Navarro, J. Vila, and I. Roca. "Automated categorization of methicillin-resistant Staphylococcus aureus clinical isolates into different clonal complexes by MALDI-TOF mass spectrometry". *Clinical Microbiology and Infection* 22:2, 2016, 161.e1–161.e7. doi: `10.1016/j.cmi.2015.10.009`.
17. J. H. K. Chen, K. K. K. She, O.-Y. Wong, J. L. L. Teng, W.-C. Yam, S. K. P. Lau, P. C. Y. Woo, V. C. C. Cheng, and K.-Y. Yuen. "Use of MALDI Biotyper plus ClinProTools Mass Spectra Analysis for Correct Identification of *Streptococcus pneumoniae* and *Streptococcus mitis/oralis*". en. *Journal of Clinical Pathology* 68:8, 2015, pp. 652–656. issn: 0021-9746, 1472-4146. doi: `10.1136/jclinpath-2014-202818`.
18. A. S. Chowdhury, D. R. Call, and S. L. Broschat. "Antimicrobial Resistance Prediction for Gram-Negative Bacteria via Game Theory-Based Feature Evaluation". 9:1, 2019. doi: `10.1038/s41598-019-50686-z`.
19. A. S. Chowdhury, D. R. Call, and S. L. Broschat. "PARGT: a software tool for predicting antimicrobial resistance in bacteria", 2020. doi: `10.1038/s41598-020-67949-9`.
20. C.-R. Chung, H.-Y. Wang, F. Lien, Y.-J. Tseng, C.-H. Chen, T.-Y. Lee, T.-P. Liu, J.-T. Horng, and J.-J. Lu. "Incorporating Statistical Test and Machine Intelligence Into Strain Typing of Staphylococcus haemolyticus Based on Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry". *Frontiers in Microbiology* 10, 2019. doi: `10.3389/fmicb.2019.02120`.
21. C.-R. Chung, H.-Y. Wang, F. Lien, Y.-J. Tseng, C.-H. Chen, T.-Y. Lee, T.-P. Liu, J.-T. Horng, and J.-J. Lu. "Incorporating Statistical Test and Machine Intelligence Into Strain Typing of Staphylococcus haemolyticus Based on Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry". *Frontiers in Microbiology* 10, 2019. doi: `10.3389/fmicb.2019.02120`.
22. Clinical and L. S. Institute. *Clinical and Laboratory Standards Institute (CLSI)*. `http://www.https://clsi.org/`.

23. A. Croxatto, G. Prod’hom, F. Faverjon, Y. Rochais, and G. Greub. “Laboratory automation in clinical bacteriology: what system to choose?” *Clinical Microbiology and Infection* 22:3, 2016, pp. 217–235. doi: 10.1016/j.cmi.2015.09.030.

24. J. J. Davis, A. R. Wattam, R. K. Aziz, T. Brettin, R. Butler, R. M. Butler, P. Chlenski, N. Conrad, A. Dickerman, E. M. Dietrich, J. L. Gabbard, S. Gerdes, A. Guard, R. W. Kenyon, D. Machi, C. Mao, D. Murphy-Olson, M. Nguyen, E. K. Nordberg, G. J. Olsen, R. D. Olson, J. C. Overbeek, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, C. Thomas, M. VanOeffelen, V. Vonstein, A. S. Warren, F. Xia, D. Xie, H. Yoo, and R. Stevens. “The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities”, 2020. doi: 10.1093/nar/gkz943.

25. K. DeBruyne, B. Slabbinck, W. Waegeman, P. Vauterin, B. DeBaets, and P. Vandamme. “Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning”. *Systematic and Applied Microbiology* 34:1, 2011, pp. 20–29. doi: 10.1016/j.syapm.2010.11.003.

26. M. Delavy, L. Cerutti, A. Croxatto, G. Prod’hom, D. Sanglard, G. Greub, and A. T. Coste. “Machine Learning Approach for Candida albicans Fluconazole Resistance Detection Using Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry”. *Frontiers in Microbiology* 10, 2020. doi: 10.3389/fmicb.2019.03000.

27. H. Desaire and D. Hua. “Adaption of the Aristotle Classifier for Accurately Identifying Highly Similar Bacteria Analyzed by MALDI-TOF MS”. *Analytical Chemistry* 92:1, 2019, pp. 1050–1057. doi: 10.1021/acs.analchem.9b04049.

28. A. Dierig, R. Frei, and A. Egli. “The Fast Route to Microbe Identification”. *Pediatric Infectious Disease Journal* 34:1, 2015, pp. 97–99. doi: 10.1097/inf.0000000000000601.

29. K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker. “Searching molecular structure databases with tandem mass spectra using CSI:FingerID”. *Proceedings of the National Academy of Sciences* 112:41, 2015, pp. 12580–12585. issn: 0027-8424. doi: 10.1073/pnas.1509788112.

30. H. Edelsbrunner and J. Harer. *Computational topology: An introduction*. American Mathematical Society, 2010.

31. M. Ellington, O. Ekelund, F. Aarestrup, R. Canton, M. Doumith, C. Giske, H. Grundman, H. Hasman, M. Holden, K. Hopkins, J. Iredell, G. Kahlmeter, C. Köser, A. MacGowan, D. Mevius, M. Mulvey, T. Naas, T. Peto, J.-M. Rolain, Ø. Samuelsen, and N. Woodford. “The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee”. 23:1, 2017, pp. 2–22. doi: 10.1016/j.cmi.2016.11.012.

32. N. Esener, M. J. Green, R. D. Emes, B. Jowett, P. L. Davies, A. J. Bradley, and T. Dottorini. “Discrimination of contagious and environmental strains of Streptococcus uberis in dairy herds by means of mass spectrometry and machine-learning”. *Scientific Reports* 8:1, 2018. doi: `10.1038/s41598-018-35867-6`.
33. N. Esener, M. J. Green, R. D. Emes, B. Jowett, P. L. Davies, A. J. Bradley, and T. Dottorini. “Discrimination of contagious and environmental strains of Streptococcus uberis in dairy herds by means of mass spectrometry and machine-learning”. *Scientific Reports* 8:1, 2018. doi: `10.1038/s41598-018-35867-6`.
34. EUCAST. *The European Committee on Antimicrobial Susceptibility Testing (EUCAST): Clinical breakpoints and dosing of antibiotics.* `http://www.eucast.org/clinical_breakpoints/`. 2018.
35. M.-S. Fangous, F. Mougari, S. Gouriou, E. Calvez, L. Raskine, E. Cambau, C. Payan, and G. Héry-Arnaud. “Classification Algorithm for Subspecies Identification within the *Mycobacterium abscessus* Species, based on Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry”. en. *Journal of Clinical Microbiology* 52:9, 2014. Ed. by G. A. Land, pp. 3362–3369. issn: 0095-1137, 1098-660X. doi: `10.1128/JCM.00788-14`.
36. A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia. “A Brief Review of Domain Adaptation”, 2020. arXiv: `2010.03978 [cs.LG]`. url: `http://arxiv.org/abs/2010.03978`.
37. T. Fawcett. “An introduction to ROC analysis”. *Pattern Recognition Letters* 27:8, 2006, pp. 861–874. doi: `10.1016/j.patrec.2005.10.010`.
38. A. Feragen, F. Lauze, and S. Hauberg. “Geodesic Exponential Kernels: When Curvature and Linearity Conflict”. In: *CVPR*. 2015, pp. 3032–3042.
39. W. Florio, A. Tavanti, S. Barnini, E. Ghelardi, and A. Lupetti. “Recent Advances and Ongoing Challenges in the Diagnosis of Microbial Infections by MALDI-TOF Mass Spectrometry”. *Frontiers in Microbiology* 9, 2018. doi: `10.3389/fmicb.2018.01097`.
40. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. “Domain-Adversarial Training of Neural Networks”. *Journal of Machine Learning Research* 17:59, 2016, pp. 1–35.
41. Geneva: World Health Organization. *Global antimicrobial resistance and use surveillance system (GLASS) report 2021*. 2017. url: `https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed`.
42. Geneva: World Health Organization. *Global antimicrobial resistance and use surveillance system (GLASS) report 2021*. 2021. url: `https://www.who.int/publications/i/item/9789240027336`.

43. S. Gibb and K. Strimmer. “MALDIquant: a versatile R package for the analysis of mass spectrometry data”. *Bioinformatics* 28:17, 2012, pp. 2270–2271.
44. K. O. Gradel, U. S. Jensen, H. C. Schønheyder, C. Østergaard, J. D. Knudsen, S. Wehberg, and M. Søgaard. “Impact of appropriate empirical antibiotic treatment on recurrence and mortality in patients with bacteraemia: a population-based cohort study”. *BMC Infectious Diseases* 17:1, 2017. doi: `10.1186/s12879-017-2233-z`.
45. A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. “A Kernel Two-Sample Test”. *Journal of Machine Learning Research* 13:25, 2012, pp. 723–773. url: `http://jmlr.org/papers/v13/gretton12a.html`.
46. T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2009. url: `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.
47. M. Heinonen, H. Shen, N. Zamboni, and J. Rousu. “Metabolite identification and molecular fingerprint prediction through machine learning”. *Bioinformatics* 28:18, 2012, pp. 2333–2341. doi: `10.1093/bioinformatics/bts437`.
48. F. Hillenkamp, M. Karas, R. C. Beavis, and B. T. Chait. “Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers”. *Analytical Chemistry* 63:24, 1991. PMID: 1789447, 1193A–1203A. doi: `10.1021/ac00024a716`.
49. P.-L. Ho, C.-Y. Yau, L.-Y. Ho, J. H. K. Chen, E. L. Y. Lai, S. W. U. Lo, C. W. S. Tse, and K.-H. Chow. “Rapid detection cfiA metallo-beta-lactamase-producing Bacteroides fragilis by the combination of MALDI-TOF MS and CarbaNP”. *Journal of Clinical Pathology* 70:10, 2017, pp. 868–873. doi: `10.1136/jclinpath-2017-204335`.
50. P.-L. Ho, C.-Y. Yau, L.-Y. Ho, J. H. K. Chen, E. L. Y. Lai, S. W. U. Lo, C. W. S. Tse, and K.-H. Chow. “Rapid Detection of cfiA Metallo-β-Lactamase-Producing *Bacteroides fragilis* by the Combination of MALDI-TOF MS and CarbaNP”. en. *Journal of Clinical Pathology* 70:10, 2017, pp. 868–873. doi: `10.1136/jclinpath-2017-204335`.
51. S.-Y. Hsieh, C.-L. Tseng, Y.-S. Lee, A.-J. Kuo, C.-F. Sun, Y.-H. Lin, and J.-K. Chen. “Highly Efficient Classification and Identification of Human Pathogenic Bacteria by MALDI-TOF MS”. *Molecular & Cellular Proteomics* 7:2, 2008, pp. 448–456. doi: `10.1074/mcp.m700339-mcp200`.
52. T.-S. Huang, S. S.-J. Lee, C.-C. Lee, and F.-C. Chang. “Detection of carbapenem-resistant Klebsiella pneumoniae on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using supervised machine learning approach”. *PLOS ONE* 15:2, 2020. Ed. by J. Banoub, e0228459. doi: `10.1371/journal.pone.0228459`.

53. T.-S. Huang, S. S.-J. Lee, C.-C. Lee, and F.-C. Chang. “Detection of carbapenem-resistant Klebsiella pneumoniae on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using supervised machine learning approach”. *PLOS ONE* 15:2, 2020. Ed. by J. Banoub, e0228459. doi: `10.1371/journal.pone.0228459`.
54. M. Josten, M. Reif, C. Szekat, N. Al-Sabti, T. Roemer, K. Sparbier, M. Kostrzewa, H. Rohde, H.-G. Sahl, and G. Bierbaum. “Analysis of the Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrum of Staphylococcus aureus Identifies Mutations That Allow Differentiation of the Main Clonal Lineages”. *Journal of Clinical Microbiology* 51:6, 2013, pp. 1809–1817. doi: `10.1128/jcm.00518-13`.
55. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. url: `https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf`.
56. P. D. Khot and M. A. Fisher. “Novel Approach for Differentiating Shigella Species and Escherichia coli by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry”. *Journal of Clinical Microbiology* 51:11, 2013, pp. 3711–3716. doi: `10.1128/jcm.01526-13`.
57. Y.-K. Kim, H. Pai, H.-J. Lee, S.-E. Park, E.-H. Choi, J. Kim, J.-H. Kim, and E.-C. Kim. “Bloodstream Infections by Extended-Spectrum β-Lactamase-Producing Escherichia coli and Klebsiella pneumoniae in Children: Epidemiology and Clinical Outcome”. *Antimicrobial Agents and Chemotherapy* 46:5, 2002, pp. 1481–1491. doi: `10.1128/aac.46.5.1481-1491.2002`.
58. J. Lafolie, M. Sauget, N. Cabrolier, D. Hocquet, and X. Bertrand. “Detection of Escherichia coli sequence type 131 by matrix-assisted laser desorption ionization time-of-flight mass spectrometry: implications for infection control policies?” *Journal of Hospital Infection* 90:3, 2015, pp. 208–212. doi: `10.1016/j.jhin.2014.12.022`.
59. P. Lasch, W. Beyer, H. Nattermann, M. Stämmler, E. Siegbrecht, R. Grunow, and D. Naumann. “Identification of Bacillus anthracis by Using Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry and Artificial Neural Networks”. *Applied and Environmental Microbiology* 75:22, 2009, pp. 7229–7242. doi: `10.1128/aem.00857-09`.
60. P. Lasch, M. Stämmler, and A. Schneider. *Version 3 (20181130) of the MALDI-TOF Mass Spectrometry Database for Identification and Classification of Highly Pathogenic Microorganisms from the Robert Koch-Institute (RKI)*. 2018. doi: `10.5281/zenodo.1880975`.

61. J. Lee, Y. Shin, S. Kim, K. Rho, and K. H. Park. “SVM Classification Model of Similar Bacteria Species using Negative Marker: Based on Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry”. In: *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2017. doi: `10.1109/bibe.2017.00-64`.
62. B. Li, T. Guo, F. Qu, B. Li, H. Wang, Z. Sun, X. Li, Z. Gao, C. Bao, C. Zhang, X. Li, and Y. Mao. “Matrix-Assisted Laser Desorption Ionization: Time of Flight Mass Spectrometry-Identified Models for Detection of ESBL-Producing Bacterial Strains”. *Medical Science Monitor Basic Research* 20, 2014, pp. 176–183. doi: `10.12659/msmbr.892670`.
63. J. Ling, H. Wang, G. Li, Z. Feng, Y. Song, P. Wang, H. Shao, H. Zhou, and G. Chen. “A novel short-term high-lactose culture approach combined with a matrix-assisted laser desorption ionization-time of flight mass spectrometry assay for differentiating Escherichia coli and Shigella species using artificial neural networks”. *PLOS ONE* 14:10, 2019. Ed. by J. Banoub, e0222636. doi: `10.1371/journal.pone.0222636`.
64. S. W. Long, R. J. Olsen, S. C. Mehta, T. Palzkill, P. L. Cernoch, K. K. Perez, W. L. Musick, A. E. Rosato, and J. M. Musser. “PBP2a Mutations Causing High-Level Ceftaroline Resistance in Clinical Methicillin-Resistant Staphylococcus aureus Isolates”. *Antimicrobial Agents and Chemotherapy* 58:11, 2014, pp. 6668–6674. doi: `10.1128/aac.03622-14`.
65. X. Lou, B. Li, B. F. de Waal, J. Schill, M. B. Baker, R. A. Bovee, J. L. van Dongen, L.-G. Milroy, and E. Meijer. “Fragmentation of organic ions bearing fixed multiple charges observed in MALDI MS”. *Journal of Mass Spectrometry* 53:1, 2018, pp. 39–47.
66. C. Ludden, A. G. Decano, D. Jamrozy, D. Pickard, D. Morris, J. Parkhill, S. J. Peacock, M. Cormican, and T. Downing. “Genomic surveillance of Escherichia coli ST131 identifies local expansion and serial replacement of subclones”. *Microbial Genomics* 6:4, 2020. doi: `10.1099/mgen.0.000352`.
67. C. A. Mather, B. J. Werth, S. Sivagnanam, D. J. Sengupta, and S. M. Butler-Wu. “Rapid Detection of Vancomycin-Intermediate *Staphylococcus aureus* by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry”. *Journal of Clinical Microbiology* 54:4, 2016, pp. 883–890. doi: `10.1128/jcm.02428-15`.
68. Microsoft. *LightGBM gradient boosting framework*. `https://github.com/microsoft/LightGBM`.
69. D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, and L. A. Stewart. “Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement”. *Systematic Reviews* 4:1, 2015. doi: `10.1186/2046-4053-4-1`.

70. C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. `https://christophm.github.io/interpretable-ml-book/`. 2019.
71. J. L. Moreno-Camacho, D. Y. Calva-Espinosa, Y. Y. Leal-Leyva, D. C. Elizalde-Olivas, A. Campos-Romero, and J. Alcántar-Fernández. “Transformation From a Conventional Clinical Microbiology Laboratory to Full Automation”. *Laboratory Medicine* 49:1, 2017, e1–e8. doi: `10.1093/labmed/lmx079`.
72. K. P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. url: `probml.ai`.
73. A. Nakamura, M. Komatsu, Y. Ohno, N. Noguchi, A. Kondo, and N. Hatano. “Identification of specific protein amino acid substitutions of extended-spectrum β-lactamase (ESBL)-producing Escherichia coli ST131: a proteomics approach using mass spectrometry”. *Scientific Reports* 9:1, 2019. doi: `10.1038/s41598-019-45051-z`.
74. S. Nakano, Y. Matsumura, Y. Ito, T. Fujisawa, B. Chang, S. Suga, K. Kato, T. Yunoki, G. Hotta, T. Noguchi, M. Yamamoto, M. Nagao, S. Takakura, M. Ohnishi, T. Ihara, and S. Ichiyama. “Development and evaluation of MALDI-TOF MS-based serotyping for Streptococcus pneumoniae”. *European Journal of Clinical Microbiology & Infectious Diseases* 34:11, 2015, pp. 2191–2198. doi: `10.1007/s10096-015-2468-9`.
75. S. Nakano, Y. Matsumura, K. Kato, T. Yunoki, G. Hotta, T. Noguchi, M. Yamamoto, M. Nagao, Y. Ito, S. Takakura, and S. Ichiyama. “Differentiation of vanA-positive Enterococcus faecium from vanA-negative E. faecium by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry”. *International Journal of Antimicrobial Agents* 44:3, 2014, pp. 256–259. doi: `10.1016/j.ijantimicag.2014.05.006`.
76. A. Niculescu-Mizil and R. Caruana. “Predicting good probabilities with supervised learning”. In: *Proceedings of the 22nd international conference on Machine learning - ICML ’05*. ACM Press, 2005. doi: `10.1145/1102351.1102430`.
77. F. Nielsen. “Hierarchical Clustering”. In: 2016, pp. 195–211. isbn: 978-3-319-21902-8. doi: `10.1007/978-3-319-21903-5_8`.
78. S. M. Novak and E. M. Marlowe. “Automation in the Clinical Microbiology Laboratory”. *Clinics in Laboratory Medicine* 33:3, 2013, pp. 567–588. doi: `10.1016/j.cll.2013.03.002`.
79. M. Oberle, N. Wohlwend, D. Jonas, F. P. Maurer, G. Jost, S. Tschudin-Sutter, K. Vranckx, and A. Egli. “The Technical and Biological Reproducibility of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) Based Typing: Employment of Bioinformatics in a Multicenter Study”. *PLOS ONE* 11:10, 2016. Ed. by F. Lisacek, e0164260. doi: `10.1371/journal.pone.0164260`.

80. C. Østergaard, S. G. Hansen, and J. K. Møller. “Rapid first-line discrimination of methicillin resistant Staphylococcus aureus strains using MALDI-TOF MS”. *International Journal of Medical Microbiology* 305:8, 2015, pp. 838–847. doi: `10.1016/j.ijmm.2015.08.002`.

81. C. Papagiannopoulou, R. Parchen, P. Rubbens, and W. Waegeman. “Fast Pathogen Identification Using Single-Cell Matrix-Assisted Laser Desorption/Ionization-Aerosol Time-of-Flight Mass Spectrometry Data and Deep Learning Methods”. *Analytical Chemistry* 92:11, 2020. PMID: 32330016, pp. 7523–7531. doi: `10.1021/acs.analchem.9b05806`.

82. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. “Scikit-learn: Machine Learning in Python ”. *Journal of Machine Learning Research* 12, 2011, pp. 2825–2830.

83. L. A. Pereira, G. B. Harnett, M. M. Hodge, J. A. Cattell, and D. J. Speers. “Real-Time PCR Assay for Detection of blaZ Genes in Staphylococcus aureus Clinical Isolates”. *Journal of Clinical Microbiology* 52:4, 2014, pp. 1259–1261. doi: `10.1128/jcm.03413-13`.

84. Y. Perez-Riverol, E. Alpi, R. Wang, H. Hermjakob, and J. A. Vizcaíno. “Making proteomics data accessible and reusable: Current state of proteomics databases and repositories”. *PROTEOMICS* 15:5-6, 2015, pp. 930–950. doi: `10.1002/pmic.201400302`.

85. M. Pérez-Sancho, A. I. Vela, P. Horcajo, M. Ugarte-Ruiz, L. Domínguez, J. F. Fernández-Garayzábal, and R. de la Fuente. “Rapid differentiation of Staphylococcus aureus subspecies based on MALDI-TOF MS profiles”. *Journal of Veterinary Diagnostic Investigation* 30:6, 2018, pp. 813–820. doi: `10.1177/1040638718805537`.

86. M. Pietsch, C. Eller, C. Wendt, M. Holfelder, L. Falgenhauer, A. Fruth, T. Grössl, R. Leistner, G. Valenza, G. Werner, and Y. Pfeifer. “Molecular characterisation of extended-spectrum β-lactamase (ESBL)-producing Escherichia coli isolates from hospital and ambulatory patients in Germany”. *Veterinary Microbiology* 200, 2017, pp. 130–137. doi: `10.1016/j.vetmic.2015.11.028`.

87. J. C. Platt. “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods”. *ADVANCES IN LARGE MARGIN CLASSIFIERS*, 1999, pp. 61–74.

88. A. Potron, L. Poirel, E. Rondinaud, and P. Nordmann. “Intercontinental spread of OXA-48 beta-lactamase-producing Enterobacteriaceae over a 11-year period, 2001 to 2011”. *Eurosurveillance* 18:31, 2013. doi: `10.2807/1560-7917.es2013.18.31.20549`.

89. C. Proteomics, 2020. url: `https://www.creative-proteomics.com/technology/maldi-tof-mass-spectrometry.htm`.

90. N. Qiao. “A systematic review on machine learning in sellar region diseases: quality and reporting items”. *Endocrine Connections* 8:7, 2019, pp. 952–960. doi: `10.1530/ec-19-0156`.
91. E. M. Ransom, R. F. Potter, G. Dantas, and C.-A. D. Burnham. “Genomic Prediction of Antimicrobial Resistance: Ready or Not, Here It Comes!” 66:10, 2020, pp. 1278–1289. doi: `10.1093/clinchem/hvaa172`.
92. C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for machine learning*. MIT Press, 2006. isbn: 9780262182539.
93. J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt. “A Stable Multi-scale Kernel for Topological Machine Learning”. In: *CVPR*. 2015, pp. 4741–4748. doi: `10.1109/CVPR.2015.7299106`.
94. P. Retamar, M. M. Portillo, M. D. López-Prieto, F. Rodríguez-López, M. de Cueto, M. V. García, M. J. Gómez, A. del Arco, A. Muñoz, A. Sánchez-Porto, M. Torres-Tortosa, A. Martín-Aspas, A. Arroyo, C. García-Figueras, F. Acosta, J. E. Corzo, L. León-Ruiz, T. Escobar-Lara, and J. R.-B. and. “Impact of Inadequate Empirical Therapy on the Mortality of Patients with Bloodstream Infections: a Propensity Score-Based Analysis”. *Antimicrobial Agents and Chemotherapy* 56:1, 2011, pp. 472–478. doi: `10.1128/aac.00462-11`.
95. B. Rieck and H. Leitte. “Exploring and comparing clusterings of multivariate data sets using persistent homology”. *Computer Graphics Forum* 35:3, 2016, pp. 81–90. doi: `10.1111/cgf.12884`.
96. C. Rodrigues, V. Passet, A. Rakotondrasoa, and S. Brisse. “Identification of Klebsiella pneumoniae, Klebsiella quasipneumoniae, Klebsiella variicola and Related Phylogroups by MALDI-TOF Mass Spectrometry”. *Frontiers in Microbiology* 9, 2018. doi: `10.3389/fmicb.2018.03000`.
97. J. Roe. *Elliptic operators, topology and asymptotic methods*. Second. Chapman & Hall/CRC, 1988.
98. L. Rokach and O. Maimon. “Clustering Methods”. In: *Data Mining and Knowledge Discovery Handbook*. Ed. by O. Maimon and L. Rokach. Springer, 2005, pp. 321–352. isbn: 978-0-387-25465-4. doi: `10.1007/0-387-25465-X_15`.
99. J. Rossen, A. Friedrich, and J. Moran-Gilad. “Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology”. 24:4, 2018, pp. 355–360. doi: `10.1016/j.cmi.2017.11.001`.
100. G. A. Satten, S. Datta, H. Moura, A. R. Woolfitt, M. d. G. Carvalho, G. M. Carlone, B. K. De, A. Pavlopoulos, and J. R. Barr. “Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens”. *Bioinformatics* 20:17, 2004, pp. 3128–3136. doi: `10.1093/bioinformatics/bth372`.
101. M. Sauget, N. van der Mee-Marquet, X. Bertrand, and D. Hocquet. “Matrix-assisted laser desorption ionization-time of flight Mass spectrometry can detect Staphylococcus aureus clonal complex 398”. *Journal of Microbiological Methods* 127, 2016, pp. 20–23. doi: `10.1016/j.mimet.2016.05.010`.

102. R. Schaumann, K. Dallacker-Losensky, C. Rosenkranz, G. H. Genzel, C. S. Stîngu, W. Schellenberger, S. Schulz-Stübner, A. C. Rodloff, and K. Eschrich. "Discrimination of Human Pathogen Clostridium Species Especially of the Heterogeneous C. sporogenes and C. botulinum by MALDI-TOF Mass Spectrometry". *Current Microbiology* 75:11, 2018, pp. 1506–1515. doi: `10.1007/s00284-018-1552-7`.

103. R. Schaumann, N. Knoop, G. H. Genzel, K. Losensky, C. Rosenkranz, C. S. Stîngu, W. Schellenberger, A. C. Rodloff, and K. Eschrich. "A step towards the discrimination of β-lactamase-producing clinical isolates of Enterobacteriaceae and Pseudomonas aeruginosa by MALDI-TOF mass spectrometry". *Medical Science Monitor* 18:9, 2012, MT71–MT77. doi: `10.12659/msm.883339`.

104. A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrné, B. Volkmer, L. Callipo, K. Knoops, M. Bauer, R. Aebersold, and M. Heinemann. "The quantitative and condition-dependent Escherichia coli proteome". *Nature Biotechnology* 34:1, 2016, pp. 104–110. doi: `10.1038/nbt.3418`.

105. B. Schölkopf and A. Smola. *Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond.* Vol. 98. MIT Press, Cambridge, MA, 2001.

106. B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel Methods in Computational Biology*. MIT Press, 2004.

107. L. S. Shapley. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA, 1952. doi: `10.7249/P0295`.

108. H. Shen, K. Dührkop, S. Böcker, and J. Rousu. "Metabolite identification through multiple kernel learning on fragmentation trees". *Bioinformatics* 30:12, 2014, pp. i157–i164. doi: `10.1093/bioinformatics/btu275`.

109. N. Singhal, M. Kumar, P. K. Kanaujia, and J. S. Virdi. "MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis". *Frontiers in Microbiology* 6, 2015. doi: `10.3389/fmicb.2015.00791`.

110. K. Sogawa, M. Watanabe, T. Ishige, S. Segawa, A. Miyabe, S. Murata, T. Saito, A. Sanda, K. Furuhata, and F. Nomura. "Rapid Discrimination between Methicillin-Sensitive and Methicillin-Resistant *Staphylococcus aureus* using MALDI-TOF Mass Spectrometry". en. *Biocontrol Science* 22:3, 2017, pp. 163–169. issn: 1342-4815, 1884-0205. doi: `10.4265/bio.22.163`.

111. K. Sogawa, M. Watanabe, K. Sato, S. Segawa, C. Ishii, A. Miyabe, S. Murata, T. Saito, and F. Nomura. "Use of the MALDI BioTyper system with MALDI-TOF mass spectrometry for rapid identification of microorganisms". *Analytical and Bioanalytical Chemistry* 400:7, 2011, pp. 1905–1911. doi: `10.1007/s00216-011-4877-7`.

112. P. Sonthayanon, J. Jaresitthikunchai, S. Mangmee, T. Thiangtrongjit, V. Wuthiekanun, P. Amornchai, P. Newton, R. Phetsouvanh, N. P. Day, and S. Roytrakul. "Whole cell matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) for identification of Leptospira spp. in Thailand and Lao PDR". *PLoS Negl Trop Dis* 13:4, 2019, e0007232.

113. P. Sonthayanon, J. Jaresitthikunchai, S. Mangmee, T. Thiangtrongjit, V. Wuthiekanun, P. Amornchai, P. Newton, R. Phetsouvanh, N. P. Day, and S. Roytrakul. "Whole cell matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) for identification of Leptospira spp. in Thailand and Lao PDR". *PLOS Neglected Tropical Diseases* 13:4, 2019. Ed. by T. Lin, e0007232. doi: `10.1371/journal.pntd.0007232`.

114. J. Sun, H. Shi, Y. Xue, W. Cheng, M. Yu, C. Ding, F. Xu, and S. Yu. "Releasing bacteria from functional magnetic beads is beneficial to MALDI-TOF MS based identification". 225, 2021, p. 121968. doi: `10.1016/j.talanta.2020.121968`.

115. W. Tang, N. Ranganathan, V. Shahrezaei, and G. Larrouy-Maumus. "MALDI-TOF mass spectrometry on intact bacteria combined with a refined analysis framework allows accurate classification of MSSA and MRSA". *PLOS ONE* 14:6, 2019. Ed. by J. M. Koomen, e0218951. doi: `10.1371/journal.pone.0218951`.

116. S. Y. C. Tong, J. S. Davis, E. Eichenberger, T. L. Holland, and V. G. Fowler. "Staphylococcus aureus Infections: Epidemiology, Pathophysiology, Clinical Manifestations, and Management". *Clinical Microbiology Reviews* 28:3, 2015, pp. 603–661. doi: `10.1128/cmr.00134-14`.

117. M. VanOeffelen, M. Nguyen, D. Aytan-Aktug, T. Brettin, E. M. Dietrich, R. W. Kenyon, D. Machi, C. Mao, R. Olson, G. D. Pusch, M. Shukla, R. Stevens, V. Vonstein, A. S. Warren, A. R. Wattam, H. Yoo, and J. J. Davis. "A genomic data resource for predicting antimicrobial resistance from laboratory-derived antimicrobial susceptibility phenotypes". *Briefings in Bioinformatics*, 2021. bbab313. issn: 1477-4054. doi: `10.1093/bib/bbab313`.

118. C. Villani. *Optimal Transport: Old and New*. Springer, 2009.

119. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. 1. 0. Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". *Nature Methods* 17, 2020, pp. 261–272. doi: `https://doi.org/10.1038/s41592-019-0686-2`.

120. H.-Y. Wang, C.-H. Chen, T.-Y. Lee, J.-T. Horng, T.-P. Liu, Y.-J. Tseng, and J.-J. Lu. "Rapid Detection of Heterogeneous Vancomycin-Intermediate Staphylococcus aureus Based on Matrix-Assisted Laser Desorption Ionization Time-of-Flight:

Using a Machine Learning Approach and Unbiased Validation". *Frontiers in Microbiology* 9, 2018. doi: `10.3389/fmicb.2018.02393`.

121. H.-Y. Wang, T.-Y. Lee, Y.-J. Tseng, T.-P. Liu, K.-Y. Huang, Y.-T. Chang, C.-H. Chen, and J.-J. Lu. "A New Scheme for Strain Typing of Methicillin-Resistant *Staphylococcus aureus* on the Basis of Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry by Using Machine Learning Approach". *PLOS ONE* 13:3, 2018. Ed. by K. Becker, e0194289. doi: `10.1371/journal.pone.0194289`.

122. H.-Y. Wang, W.-C. Li, K.-Y. Huang, C.-R. Chung, J.-T. Horng, J.-F. Hsu, J.-J. Lu, and T.-Y. Lee. "Rapid classification of group B Streptococcus serotypes based on matrix-assisted laser desorption ionization-time of flight mass spectrometry and machine learning techniques". *BMC Bioinformatics* 20:S19, 2019. doi: `10.1186/s12859-019-3282-7`.

123. H.-Y. Wang, F. Lien, T.-P. Liu, C.-H. Chen, C.-J. Chen, and J.-J. Lu. "Application of a MALDI-TOF analysis platform (ClinProTools) for rapid and preliminary report of MRSA sequence types in Taiwan". *PeerJ* 6, 2018, e5784. doi: `10.7717/peerj.5784`.

124. J. H. Ward Jr. "Hierarchical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association* 58:301, 1963, pp. 236–244. doi: `10.1080/01621459.1963.10500845`.

125. C. Weis, A. Cuénod, B. Rieck, F. Llinares-López, O. Dubuis, S. Graf, C. Lang, M. Oberle, K. K. Soegaard, M. Osthoff, K. Borgwardt, and A. Egli. "Direct Antimicrobial Resistance Prediction from MALDI-TOF mass spectra profile in clinical isolates through Machine Learning". *bioRxiv*, 2020. doi: `https://doi.org/10.1101/2020.07.30.228411`.

126. C. Weis, A. Cuénod, B. Rieck, F. Llinares-López, O. Dubuis, S. Graf, C. Lang, M. Oberle, K. K. Soegaard, M. Osthoff, M. Brackmann, K. Borgwardt, and A. Egli. "Direct Antimicrobial Resistance Prediction from clinical MALDI-TOF mass spectra using Machine Learning". *accepted in Nature Medicine*, 2021. doi: `https://doi.org/10.1101/2020.07.30.228411`.

127. C. Weis, M. Horn, B. Rieck, A. Cuénod, A. Egli, and K. Borgwardt. "Domain adaptation for transferable antimicrobial resistance prediction from MALDI-TOF mass spectra". *Unpublished*, 2021.

128. C. Weis, M. Horn, B. Rieck, A. Cuénod, A. Egli, and K. Borgwardt. "Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra". *OUP Bioinformatics* 36, 2020, pp. i30–i38. doi: `https://doi.org/10.1093/bioinformatics/btaa429`.

129. C. Weis, C. R. Jutzeler, and K. Borgwardt. "Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review". *Clinical Microbiology and Infection* 26:10, 2020, pp. 1310–1317. doi: `https://doi.org/10.1016/j.cmi.2020.03.014`.

130. C. Weis, B. Rieck, S. Balzer, A. Cuénod, A. Egli, and K. Borgwardt. "Improved MALDI-TOF MS based antimicrobial resistance prediction through hierarchical stratification". *Unpublished*, 2020.

131. R. Wise, T. Hart, O. Cars, M. Streulens, R. Helmuth, P. Huovinen, and M. Sprenger. "Antimicrobial resistance". *BMJ* 317:7159, 1998, pp. 609–610. doi: `10.1136/bmj.317.7159.609`.

132. M. Wolters, H. Rohde, T. Maier, C. Belmar-Campos, G. Franke, S. Scherpe, M. Aepfelbacher, and M. Christner. "MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant Staphylococcus aureus lineages". *International Journal of Medical Microbiology* 301:1, 2011, pp. 64–68. doi: `10.1016/j.ijmm.2010.06.002`.

133. D. Xiao, F. Zhao, H. Zhang, F. Meng, and J. Zhang. "Novel Strategy for Typing Mycoplasma pneumoniae Isolates by Use of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry Coupled with ClinProTools". *Journal of Clinical Microbiology* 52:8, 2014, pp. 3038–3043. doi: `10.1128/jcm.01265-14`.

134. D. Xiao, F. Zhao, H. Zhang, F. Meng, and J. Zhang. "Novel Strategy for Typing Mycoplasma pneumoniae Isolates by Use of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry Coupled with ClinProTools". *Journal of Clinical Microbiology* 52:8, 2014, pp. 3038–3043. doi: `10.1128/jcm.01265-14`.

135. J. Yi, Q. Qin, Y. Wang, R. Zhang, H. Bi, S. Yu, B. Liu, and L. Qiao. "Identification of pathogenic bacteria in human blood using IgG-modified Fe3O4 magnetic beads as a sorbent and MALDI-TOF MS for profiling". 185:12, 2018. doi: `10.1007/s00604-018-3074-1`.

136. X. Zhan, A. D. Patterson, and D. Ghosh. "Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data". *BMC Bioinformatics* 16:1, 2015, p. 77. doi: `10.1186/s12859-015-0506-3`.

137. H. Zhang, N. Dullerud, L. Seyyed-Kalantari, Q. Morris, S. Joshi, and M. Ghassemi. *An Empirical Framework for Domain Generalization in Clinical Settings*. 2021. arXiv: `2103.11163 [cs.LG]`.

138. T. Zhang, J. Ding, X. Rao, J. Yu, M. Chu, W. Ren, L. Wang, and W. Xue. "Analysis of methicillin-resistant Staphylococcus aureus major clonal lineages by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS)". *J. Microbiol. Methods* 117, 2015, pp. 122–127.

139. T. Zhang, J. Ding, X. Rao, J. Yu, M. Chu, W. Ren, L. Wang, and W. Xue. "Analysis of methicillin-resistant Staphylococcus aureus major clonal lineages by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry (MALDI–TOF MS)". *Journal of Microbiological Methods* 117, 2015, pp. 122–127. doi: `10.1016/j.mimet.2015.08.002`.

140. Z. Zhang, D. Wang, P. de B Harrington, K. J. Voorhees, and J. Rees. “Forward selection radial basis function networks applied to bacterial classification based on MALDI-TOF-MS”. *Talanta* 63:3, 2004, pp. 527–532. doi: `10.1016/j.talanta.2003.11.034`.

CAROLINE VIKTORIA WEIS

Curriculum vitæ

✉ carolinevweis@gmail.com github.com/cvweis
weis.ml Google Scholar

My research interests lie in the development of *personalized healthcare* through data analysis and machine learning on medical and biological data. In my PhD, I develop models predicting *antimicrobial resistance* from MALDI-TOF mass spectrometry data, through *kernel methods*, *domain adaptation* and *topological data analysis*. Additionally, I work on applying *survival analysis* to assess disease risk from human genotype data, and topological data analysis for single-cell cancer data. Beyond my own work, I like to stay up-to-date on developments in time series methods with electronic health records and in graph kernels, due to ongoing research in my lab.

Keywords: Machine learning, Antimicrobial Resistance Prediction, Personalized Medicine, Topological Data Analysis, Domain Adaptation, Kernel Methods

SKILLS

- Strong knowledge of `Python` for data analysis (`numpy`, `scipy`, `pandas`, `scikit-learn`) and experience with deep learning frameworks (`PyTorch`, `TensorFlow` and `Keras`).
- Working knowledge of `R` and `MATLAB`. Knowledge of `SQL`.
- Strong knowledge of data visualization tools in `Python`. Working knowledge of `ggplot2` in `R`, plus `Bokeh` and `Rshiny` for interactive visualization.
- Knowledge of digital typesetting language LaTeXand of the `Git` revision control system.
- Strong writing and public speaking skills.

EDUCATION

2017–present **Ph.D. candidate** in **Machine Learning for Healthcare** at **ETH Zurich**, Switzerland
Thesis: *MALDI-TOF MS based clinical antimicrobial resistance prediction using machine learning*
Machine Learning and Computational Biology group
Adviser: Prof. Dr. Karsten Borgwardt

2014–2016 M.Sc. in Biotechnology at ETH Zurich, Switzerland, final grade **5.55** (very good)[1]
Thesis: *Assessing the Potential of Feature-pairs in Predicting the Impact of Missense Variants*
Advisers: Dr. L. Folkman, Dr. D. Grimm, Prof. Dr. Karsten Borgwardt

2010–2014 B.Sc. in Integrated Life Sciences at FAU Erlangen, Germany, final grade **1.6** (excellent)[2]
Thesis: *Analysis of zinc-oxide particle growth under influence of triethylamine by using Small Angle X- Ray Scattering and UV/Vis Spectroscopy*
Adviser: Prof. Dr. Tobias Unruh

2001–2010 *Abitur*[3], Alexander-von-Humboldt Gymnasium Schweinfurt[4], Germany, final grade **1.8** (very good)

---

[1] ETH Zurich's Grading System: https://ethz.ch/content/dam/ethz/special-interest/itet/department/Studies/Forms/20160112_Grading_System.pdf
[2] German Grading System
[3] General qualification for university entrance
[4] Secondary school

Publications

In the following list of publications, equal first-author contributions are indicated using a superscript ‘dagger’ symbol, i.e. †, while joint supervision is denoted by a ‘double-dagger’, i.e. ‡.

2021 **Caroline Weis**, Aline Cuenod, Bastian Rieck, Felipe Llinares-López, Olivier Dubuis, Susanne Graf, Claudia Lang, Michael Oberle, Maximilian Brackmann, Kirstine K. Soegaard, Michael Osthoff, Karsten Borgwardt‡, Adrian Egli‡. *Direct Antimicrobial Resistance Prediction from MALDI-TOF mass spectra using Machine Learning*. bioRxiv, July 2021.
https://doi.org/10.1101/2020.07.30.228411

2020 Stefan Groha†, **Caroline Weis**†, Alexander Gusev, Bastian Rieck. *Topological Data Analysis of copy number alterations in cancer*. Learning Meaningful Representations of Life Workshop. Neural Information Processing Systems (NeurIPS) 2020
arXiv:2011.11070

Catherine R. Jutzeler†, Lucie Bourguignon†, **Caroline Weis**, Bobo Tong, Cyrus Wong, Bastian Rieck, Hans Pargger, Sarah Tschudin-Sutter, Adrian Egli, Karsten Borgwardt‡ and Matthias Walter‡. *Comorbidities, clinical signs and symptoms, laboratory findings, imaging features, treatment strategies, and outcomes in adult and pediatric patients with COVID-19: A systematic review and meta-analysis*. Travel Medicine and Infectious Disease, September 2020.
https://doi.org/10.1101/2020.05.20.20103804

Caroline Weis†, Max Horn†, Bastian Rieck†, Aline Cuenod, Adrian Egli, Karsten Borgwardt. *Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra*. OUP Bioinformatics, accepted at ISMB 2020.
https://doi.org/10.1093/bioinformatics/btaa429

Caroline Weis†, Catherine Jutzeler†, Karsten Borgwardt. *Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review*. Clinical Microbiology and Infection, March 2020.
https://doi.org/10.1016/j.cmi.2020.03.014

2017 Jamie R. Wallen, Hao Zhang, **Caroline Weis**, Weidong Cui, Brittni M. Foster, Chris M. W. Ho, Michal Hammel, John A. Tainer, Michael L. Gross, Tom Ellenberger. *Hybrid Methods Reveal Multiple Flexibly Linked DNA Polymerases within the Bacteriophage T7 Replisome*. Structure, 25. 157–166., 2017.
https://doi.org/10.1016/j.str.2016.11.019

Oliver Ratmann, Emma B. Hodcroft, Michael Pickles, Anne Cori, Matthew Hall, Samantha Lycett, Caroline Colijn, Bethany Dearlove, Xavier Didelot, Simon Frost, A.S. Md Mukarram Hossain, Jeffrey B. Joy, Michelle Kendall, Denise Kühnert, Gabriel E. Leventhal, Richard Liang, Giacomo Plazzotta, Art F.Y. Poon, David A. Rasmussen, Tanja Stadler, Erik Volz, **Caroline Weis**, Andrew J. Leigh Brown, Christophe Fraser, on behalf of the PANGEA-HIV Consortium. *Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison*. Molecular Biology and Evolution, Volume 34, Issue 1, January 2017, Pages 185–203, 2017.
https://doi.org/10.1093/molbev/msw217

Invited Talks

2021 *Case Study: MALDI-TOF Mass Spectrometry for Antimicrobial Resistance Prediction*. Invited talk. Workshop on Geometrical and Topological Representation Learning. International Conference of Machine Learning (ICML) 2021.

Kernel-based microbial phenotype prediction from MALDI-TOF mass spectra. Invited talk. Luxembourg Institute of Science and Technology (LIST) symposium on MALDI-TOF mass spectrometry 2021.

2020 *Topological Data Analysis of Copy Number Alterations in cancer*. Abstract chosen for oral presentation. Learning Meaningful Representations of Life Workshop. Neural Information Processing Systems (NeurIPS) 2020.

Kernel-based antimicrobial resistance prediction from MALDI-TOF mass spectra. Abstract chosen for oral presentation. Machine Learning for Global Health Workshop. International Conference of Machine Learning (ICML) 2020.

Proceedings Presentation: Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra. Invited talk. Intelligent Systems for Molecular Biology (ISMB) 2020.

Posters and other publishing

2020 Mandana Samiei, **Caroline Weis**, Larissa Schiavo, Tatjana Chavdarova, Fariba Yousefi. *Convening during COVID-19: Lessons learnt from organizing virtual workshops in 2020*. Whitepaper.
arXiv:2012.01191

2019 **Caroline Weis**, Max Horn, Bastian Rieck, Karsten Borgwardt. *Sparse representations for MALDI-TOF based microbial classification*. Poster and peer-reviewed abstract. 14th Machine Learning in Computational Biology Meeting (MLCB).

Professional experience

10/2015–03/2016 **Industrial research intern** at *Genedata AG*, Basel, Switzerland.

As a Machine Learning research intern in the Screener Business Unit I assessed different algorithms for the task of classifying screening images. Images depicting bacteria treated with new compounds of unknown effect were classified into effect category. I developed a pipeline in R that has the potential to speed-up screening for drugs with a desired effect. Classical approaches such as support vector machine algorithms and t-SNE clustering provided good results, and a Genedata business poster I prepared for SLAS conference 2016 in San Diego set a new record for requested poster downloads at Screener Business Unit. Additionally I implemented a GUI for active learning in MATLAB, allowing for images classified with low probability to be presented to a human expect to be classified.

10/2014–03/2016 **Research assistant** at *Control Theory and Systems Biology group*, ETH Zurich, Basel, Switzerland.

As a wetlab research assistant I performed standard tasks – such as Minipreps, PCRs, gel electrophoresis etc. – to assist projects lead by Post-Doctoral researchers.

09/2013–07/2014 **Academic research intern** at *Lawrence Berkeley National Laboratory*, Berkeley, USA.

I worked at the SIBYLS beamline of the Physical Biosciences Division. I performed Protein Crystallography and Small Angle X-Ray Scattering experiments and subsequent data analysis. I worked on project, which eventually led to a publication in *Structure* while also providing guidance on data analysis to SIBYLS beamline users.

Supervisor: Dr. Michal Hammel

04/2012–07/2012 **Summer research assistant** at *Chair of Crystallography and Structural Physics*, FAU Erlangen, Germany.

In this summer internship I grew zinc oxide crystals, mounted them to a plate and aligned them using a goniometer.

Thesis supervision

2020 Sebastian Balzer. *Improved MALDI-TOF based antimicrobial phenotype prediction through incorporating phylogenetic structure*
Bachelor thesis, ETH Zurich

2019 Lucie Bourguignon. *Mortality prediction using self-reported health records and large scale genomic data*
Master thesis, ETH Zurich

Teaching experience

In each of the following courses, I have served as a teaching assistant. Duties are listed for each course individually.

2018 + 2020 Exercises *Data Mining II*, ETH Zurich
The course duties consisted of programming exercises in Python, creation and grading of bi-weekly exercises, as well as as the development and correction of exam questions.

2012 – 2013 Exercises *Mathematical modeling and statistics for scientists*, Department Mathematics, FAU Erlangen
This computer science tutorial consisted of modeling and statistical analyses programming exercises in R which I had to present and supervise. In addition I supervised and corrected the exam.

2012

Exercises *Structural physics*, Chair of Crystallography and Structural Physics, FAU Erlangen

The course duties consisted of presenting written exercises in crystallography and structural physics, as well as as the development and correction of exam questions.

2011 – 2012

Microscopy course *Biology for physicians*, Animal Physiology, Department of Biology, FAU Erlangen

In this microscopy course I facilitated laboratory exercises about plant biology and basic physiology for first-year physiology students.

LANGUAGES

German	native speaker
English	fluent
French	basic knowledge

SERVICE TO THE COMMUNITY

05/2021–
07/2021

'Computational Approaches to Mental Health' (CA2MH) workshop organizer, ICML 2021 virtual conference

I co-organized the first installment of the CA2MH workshop at ICML 2021, including reviewing the workshop application draft, communicating with speakers and recruiting programme committee members.

10/2020–
12/2020

'Topological Data Analysis and Beyond' workshop program committee member, NeurIPS 2020 virtual conference

I reviewed several workshop paper contributions and shared my previous experience organizing a virtual workshop at ICML 2020 with the workshop organizers.

05/2020–
09/2020

Women in Machine Learning (WiML) un-workshop organizer, ICML 2020 virtual conference

As the *Finance and Sponsorship Chair* I was solely responsible for the representation and communication with industry partners of the WiML organization.

02/2020–
05/2020

Academic Jury member, St. Gallen Symposium, St. Gallen, Switzerland

As Academic Jury member I evaluate essays on the topic of *Freedom revisited* and identify the top 100 contributions which receive an all-expenses-covered invitation to the 2020 St. Gallen Symposium. I will also participate in the three-day St. Gallen Symposium, to discuss the most urging problems threatening our freedom at this time and what actions should be taken.[5]

06/2019–
06/2020

Founding member of peer mentoring group 'Women in Data Science'

Along with two colleagues I successfully applied for a grant of 5'000 CHF provided by ETH Zurich's Fix-the-Leaky-Pipeline program. Among my tasks were to approach industry partners who serve as mentors for our peer groups, as well as leading peer meetings with fellow Data Science PhD students and Post-Doctoral researchers.

2018

ETH representative, ETH Zurich pavillion, World Economic Forum, Davos, Switzerland

I represented ETH Zurich at the universities pavillion in Davos during the 2018 World Economic Forum. Over the course of four days I made the research topic of 'significant pattern mining' accessible to visitors, which ranged from Switzerland's leading research representatives and honorary guests to several high-school groups.

02/2015–
02/2016

Student and Academic Affairs Commissioner, Biotechnology Student Association, ETH Zurich, Switzerland

Among my duties as Commissioner was preparing, holding and interpreting the lecture evaluations each semester. I was also the primary contact person for communication between student and professor concerning teaching matters. I served as the student body contact person to admitted students before their arrival, organized welcome events and helped them settle in at our department at ETH Zurich.

02/2013–
08/2014

Board Member, Biotechnology Student Initiative Erlangen, Germany

At the Biotechnology Student Initiative I organized talks of professors and company representatives, soft skill workshops and company tours, e.g. at Siemens Healthcare in Erlangen.

---

[5] symposium cancelled after evaluation round due to *SARS-CoV-2* pandemic

REFERENCES

PROF. DR. KARSTEN BORGWARDT
Machine Learning and Computational Biology Lab
ETH Zurich
karsten.borgwardt@bsse.ethz.ch

DR. BASTIAN RIECK
Senior Assistant
ETH Zurich
bastian.rieck@bsse.ethz.ch

DR. CATHERINE JUTZELER
SNSF Ambizione Group Leader
ETH Zurich
catherine.jutzeler@bsse.ethz.ch

Further references and credentials are available on request.

Last updated on 27th October 2021.