THE UNIVERSITY OF EXETER

CENTRE FOR ENVIRONMENTAL MATHEMATICS

A NOVEL PIPELINE FOR MASS SPECTRAL BIOMARKER DISCOVERY

BY

MATT PRILL

Submitted to the University of Exeter as a dissertation for the degree of Master of Science in Applied Data Science (Environment & Sustainability)

August 2025

Abstract

Pathogens present a severe and ongoing threat to global public health. In response, the clinical utility of diagnostic tools such as matrix assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry are being explored through sophisticated analytical workflows capable of characterising pathogens beyond species identification. This study designed and sequentially evaluated two pre-processing pipelines, the first for selecting meaningful peaks from MALDI-TOF mass spectral data, and the second for reducing their dimensionality for downstream analysis. The pipelines were applied to 278 *Staphylococcus aureus* bacteraemia isolates to investigate the associations between spectral features and toxicity profiles, as well as infection outcomes. Non-metric multidimensional scaling and hierarchical clustering revealed peak composition to be primarily driven by data source (hypothesised as a proxy for clonal complex), while feature selection and modelling identified several important biomarkers. Random forest identified multiple peaks around 3000 m/z to be associated with high toxicity, while Bayesian inference identified three peaks (2304, 6404, and 6943 m/z) to be significantly associated with an increased risk of 30-day *S. aureus* bacteraemia mortality. The findings demonstrate the pipelines' merit in facilitating biomarker discovery which simultaneously illuminates biologically significant peaks warranting further investigation.

Key words: biomarker identification, MALDI-TOF, Staphylococcus aureus bacteraemia, machine learning, Bayesian inference

Acknowledgements

I am very thankful to Professor Mario Recker for supervising this research and offering insightful discussion and feedback throughout its formation.

Contents

Abbreviations

AUROC - Area Under the Receiver Operator Characteristic

CCI – Charlson Comorbidity Index

CI – Credibility Interval

CWT – Continuous Wavelet Transformation

Da – Dalton

HDSS – High-Dimensionality-Small-Sample

MALDI-TOF MS – Matrix Assisted Laser Desorption/Ionisation Time of Flight Mass Spectrometry

MS - Modified Sinc

MSSA – Methicillin-susceptible *Staphylococcus aureus*

MRSA – Methicillin-resistant *Staphylococcus aureus*

m/z – Mass-to-charge ratio (Daltons/charge)

NMDS – Non-Metric Multidimensional Scaling

PCR – Polymerase Chain Reaction

PERMANOVA – Permutational Multivariate Analysis of Variance

RFECV – Recursive Feature Elimination with Cross-Validation

SAB – *Staphylococcus aureus* Bacteraemia

SMOTE – Synthetic Minority Oversampling Technique

SNR – Signal-to-Noise Ratio

UMAP – Uniform Manifold Approximation and Projection

1. Introduction

1.1 Bacterial Infections & Staphylococcus aureus

Bacterial infections' threat to global public health is persistent and growing; 13.7 million deaths in 2019 are attributed to 33 bacterial pathogens (Ikuta *et al.*, 2022). Of those, *Staphylococcus aureus* was the leading cause of mortality across 135 countries and resulted in over one million deaths worldwide (Ikuta *et al.*, 2022), prompting multiple strains to be included in the World Health Organisation's 'Bacterial Priority Pathogen List' (WHO, 2024).

Primarily a commensal, *S. aureus* is a gram-positive bacterium that commonly colonises the human skin and intestines. However, as an opportunistic pathogen, infections of varying natures can arise depending on the strain (Bai *et al.*, 2022), site of infection (Pollitt *et al.*, 2018), and host's immune system (Tong *et al.*, 2015; Howden *et al.*, 2023). Particularly catastrophic infections can occur when *S. aureus* penetrates the epithelial layer, reaching tissues and/or the bloodstream.

Bloodstream infections, known as *S. aureus* bacteraemia (SAB), can have disastrous complications, such as septic shock, and are associated with poor prognoses. Bai *et al.*'s (2022) meta-analysis found that, since 2011, 30-day mortality following SAB was 18.1%. Furthermore, the threat posed by SAB is compounded by the rise in antibiotic resistant bacteria (Urban-Chmiel *et al.*, 2022), largely caused by an overreliance on, and misuse of, antibiotics (Tarrant & Krockow, 2022). The growing concern surrounding antibiotic resistance extends to a number of *S. aureus* strains (John, 2020).

Following the discovery of penicillin, 95% of all *S. aureus* strains became penicillin-resistant which, in turn, encouraged the development of methicillin as an alternative anti-biotic (Guo *et al.*, 2020). Consequently, methicillin-resistant *S. aureus* (MRSA) has become the most clinically significant group of anti-biotic resistant *S. aureus* (Gajdács, 2019) and exhibits higher virulence than methicillin-susceptible *S. aureus* (MRSA). This was showcased by the aforementioned meta-analysis which observed a higher 30-day mortality in SAB cases caused by MRSA (23.4%) than those of MSSA infections (18.6%) (Bai *et al.*, 2022).

Infection outcome following SAB is also largely dependent on the aptness (David and Daum, 2017) and timeliness of treatment (Corl *et al.,* 2020). Therefore,

identifying and characterising *S. aureus* infections when they occur is vital for heightening patient prospects and thus emphasises the need for rapid and accurate means of pathogen identification.

1.2 Conventional Pathogen Identification and Profiling Techniques

With the advancement of methods used for differentiating between pathogenic species, the fields of genomics and proteomics have expanded rapidly in recent years to better understand pathogen virulence (Pérez-Llarena & Bou, 2016), anti-biotic resistance (Sulaiman & Lam, 2022), transmission (Zubair *et al.*, 2022), and infection outcome (Jean Beltran *et al.*, 2017), thereby informing more effective treatment strategies (Horvatić *et al.*, 2016). Genetic mutations form the basis for changes in protein synthesis which are key determinants of pathogen virulence (Arenas *et al.*, 2018; Sukumaran *et al.*, 2021).

A recent genome wide enrichment analysis by Coll *et al.* (2025) revealed the extent of protein-altering mutation rates in *S. aureus* upon colonisation. They observed more pronounced increases in mutation rates across multiple accessory gene regulators (*agr*), relative to the rest of the genome. These included *agrA* and *agrC* , both which are known to influence virulence (Jenul & Horswill, 2019). Furthermore, genes encoding anti-biotic targets showed marginally insignificant increases in mutations (Coll *et al.*, 2025). Today, a myriad of techniques are used to identify pathogens and their phenotypes, extending to those of *S. aureus*.

Conventional approaches to pathogen identification include culture-based methods (Lazcka *et al*., 2007), microscopy (Golding *et al.*, 2016; Müller *et al.*, 2018), electrophoresis (Buszewski *et al.*, 2021), fluorescence *in situ* hybridization (Levsky & Singer, 2003), and polymerase chain reaction (PCR) followed by sequencing (Levi *et al.*, 2003; Frye *et al.*, 2020). Early sequencing technology included 16S rRNA (Patel, 2001) and 18S rRNA gene sequencing (Embong *et al.*, 2008). Then, at its advent, next-generation sequencing became a primary method of pathogen identification owing to its higher throughput and resolution (Hu *et al.*, 2021).

Sequencing techniques involve the mapping and analysis of genetic information which offers insights into the presence and composition of particular genes, including those associated with virulence. For example, in clinical settings, genes of interest often include biofilm layer-encoding genes (Vestby *et al.*, 2020) or those tied to anti-biotic resistance (Banin *et al.*, 2017). Continued innovation has reduced the cost, increased the practicality, and encouraged the uptake of next-generation sequencing considerably (Gu *et al*., 2019). Alongside sequencing, mass spectrometry is another well-established method for pathogen identification and profiling, although it varies in its approach.

Mass spectrometers quantify molecular mass-to-charge ratios (m/z) which, in a biological context, permits the identification and characterisation of microbes through the analysis of biomolecules such as proteins and lipoproteins (Sabbagh *et al.*, 2016). Such information can be useful for evaluating microbial pathogenicity (Grenga *et al.,* 2019) and virulence (Man *et al.*, 2021) which can then inform treatment plans (Sukumaran *et al.*, 2021; Zubair *et al.*, 2022). There are a range of mass spectrometers each with nuanced mechanisms, but all are based on common principles.

First, a portion of a sample is ionised by an ion source (Garg & Zubair, 2023). Next, a mass analyser separates the ions based on their m/z (Haag, 2016). Finally, the ions travel through a detector which quantifies their charge. The data is digitised, calibrated and presented, usually as a mass spectrum chart with ion m/z on the x-axis, and their respective relative abundances on the y-axis. (Aebersold & Mann, 2003; Yucel & Smith, 2024). Amongst the nuanced mass spectrometry techniques, matrix assisted laser desorption ionisation-time of flight mass spectrometry (MALDI-TOF MS) has emerged as a favourable method for identifying pathogens, their phenotypes and, in turn, their virulence (Calderaro & Chezzi, 2024).

1.3 The Diagnostic Power of MALDI-TOF MS

MALDI-TOF MS is most commonly used to identify and characterise biological samples, namely microbes (Haider *et al.*, 2023) (Fig. 1). Initially, samples are applied to a matrix on the mass spectrometer's target plate. The matrix absorbs energy and facilitates desorption so that, once dry, a pulsed laser can

simultaneously vaporise and ionise the sample (Smith, 2004). The ionisation process is ‘soft’, meaning that samples’ molecular structure is largely retained; direct exposure to the laser would fragment the sample. Hence, as discussed, MALDI-TOF MS is appropriate for identifying and characterising biological specimens from a range of substrates (Doern & Butler-Wu, 2016).

Upon ionisation, the ions accelerate under an electrostatic field and pass through focusing lenses, directional deflectors and finally through a time-of-flight analyser vacuum until they collide with a linear detector nanoseconds later (Kovtoun & Cotter, 2000). All ions are exposed to the same electrical potential energy (voltage) during acceleration, and so their kinetic energy (eV) is a function of their m/z. Therefore, lighter laser-desorbed ions of a given charge travel faster than their heavier counterparts along the distance of the analyser (d) i.e. they exhibit shorter ‘times-of-flight’ (t) (Equation 1).

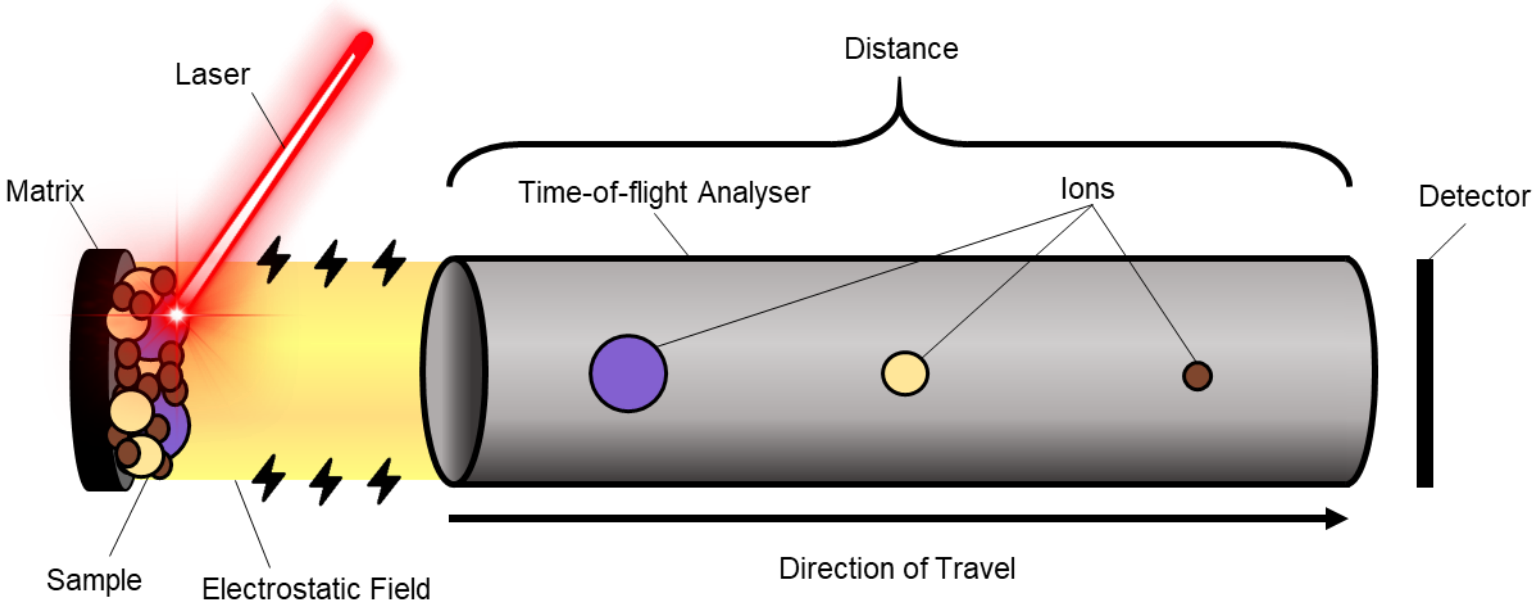


Figure 1. Not-to-scale diagram of the MALDI-TOF MS process. The schematic illustrates the key processes including the desorption/ionisation of the analytes, the transfer of electrical potential to kinetic energy (eV) and the movement and separation of ions within the time-of-flight analyser.

Equation 1: $$t = d\sqrt{\frac{m}{2eV}}$$

Recorded flight times are used to determine molecular m/z which are subsequently translated into mass spectra. Usually, these are compared against reference databases to identify biomolecules based on their signature peaks, fragmentation patterns and relative abundances (Smith, 2004). Today, MALDI-TOF has become the standard for microbial identification in clinical laboratories (Greco *et al.*, 2018), owing to a range of qualities separating it from alternate techniques.

Rapid pathogen identification is imperative in clinical contexts; treatment efficacy is often dependent on the rate of administration (Kaur *et al.*, 2013; Siewers *et al.*, 2021). This is true to *S. aureus* infections resulting in septic shock; Corl *et al.* (2020) found that the probability of 30-day mortality increased by 1.3% for every hour of delayed anti-biotic administration. Yet, culture and sequencing techniques often include lengthy incubation, DNA extraction, library preparation and PCR procedures which, cumulatively, can last weeks (Loman *et al.*, 2012; Pfyffer & Wittwer, 2012).

Favourably, MALDI-TOF MS maintains rapid turnover rates while simultaneously detecting low m/z which has cemented it a fast and reliable pathogen-profiling method (Haider *et al.*, 2023). Another contributing factor for its practical turnover time is that many of the preparatory steps associated with sequencing, such as digestion, are often unnecessary (Váradi *et al.*, 2017). Instead, MALDI-TOF MS can identify pathogens from whole-cell complexes within minutes (Quéro *et al.*, 2020). This has allowed MALDI-TOF MS to be used to detect range of bacterial (Seng *et al.*, 2009) and fungal pathogens (Gautier *et al.*, 2014; Patel, 2019) from various environments. In the case of *S. aureus*, MALDI-TOF MS has been used to characterise samples from food (Taban & Cevik, 2021), blood (Nix *et al.*, 2020), urine (Pinault *et al.*, 2019) and even the periprosthetic fluid associated with prosthetic joints (Peel *et al.*, 2015).

Importantly, the high resolution data yielded by MALDI-TOF MS allows microbes to be detected from low-concentration samples which is particularly crucial for SAB where even low bacterial-cell counts can have life threatening implications (Abraham & Bamberger, 2020). This also allows subtle differences in bimolecular make-ups to be detected, including those associated with virulence (Culebras, 2018) and antibiotic resistance (Kempf *et al.*, 2012; Sparbier *et al.*, 2012). For example, Hrabák *et al.* (2012) used meropenem hydrolysis assays in tandem with MALDI-TOF MS to assess β-lactamase activity (β-lactam antibiotic resistance-conferring enzymes) in multiple bacterial species. This highlights MALDI-TOF MS's potential for informing specialised treatment plans that recognise pathogens' intrinsic resistances and susceptibilities.

Finally, MALDI-TOF MS is relatively inexpensive compared to other methods. While mass spectrometer units are costly, cheaper sample processing makes MALDI-TOF MS more cost-efficient over time in comparison to many other methods (El-Bouri *et al.*, 2012). Many commercial processing kits exist (Perše *et al.*, 2022) but many 'in-house' techniques have been shown to reduce costs further while simultaneously retaining accuracy such as Zhou *et al.'s* (2017) method which cost just $1.5 per sample. However, the need for pre-processing in some contexts, in itself, highlights a limitation of MALDI-TOF MS.

As discussed, MALDI-TOF MS can be used to characterise unprocessed samples. However, detecting the relevant biomarkers used for protein profiling can often require preparatory steps to enhance analytical performance (Jakovljev & Bergh, 2015; Topić Popović *et al.*, 2023). There are a range of nuanced methods with varying levels of accuracy, but all increase overall turnover times to some degree (Haider *et al.*, 2023). As a result, the high resolution data associated with strain typing and, in turn, tailored patient treatments may require lengthier pipelines to process, thus presenting a trade-off between treatment specificity and timeliness (Wang *et al.*, 2021). That said, MALDI-TOF MS processing remains faster than sequencing and, in addition, microbe identification with mass spectra also precludes the human error associated with interpreting samples from microscopy (Cheng *et al.*, 2016). For spectra where a match cannot be found, other *de novo* investigative tools and algorithms are used to deduce biomolecules (Ng *et al.*, 2023).

MADLI-TOF MS's potential for diagnosing infections has become widely recognised (Singhal *et al.*, 2015). Yet, despite its diagnostic power, many analytic pipeline outputs are limited to species identification (Li *et al.*, 2022). As discussed, rapid pathogen identification is of the utmost importance and remains a primary objective in clinical settings, particularly for treating *S. aureus* infections (Corl *et al.,* 2020). However, MALDI-TOF mass spectra are rich in information and contain multitudes of deeper insights relating to virulence and transmission factors (Man *et al.*, 2021). Therefore, conventionally discarded mass spectra could otherwise be used to characterise pathogen phenotypes. Not only could this inform medical interventions to minimise mortality, but it could also contribute to more accurate infection outcome models. This foresight is invaluable for preparing patients and their next of kin for probable prognoses and for offering them the fitting support.

MALDI-TOF MS's applicability in clinical settings is primarily limited by incomplete reference spectral libraries. Without comprehensive reference databases, profiling pathogens is difficult, especially where subtle, but potentially clinically significant characteristics exist (Rychert, 2019). As a result, current research is largely focused on the identification of clinically significant pathogen biomarkers i.e. those that have a bearing on infection outcome. Therefore, revealing predictive biomarkers from high-dimensional spectral data remains a research priority. Already, significant advancements have been made in mapping and analysis of *S. aureus* MALDI-TOF mass spectra.

1.4 Staphylococcus aureus Biomarkers

1.4.1 Antibiotic Resistance

Biomarkers associated with both virulence and antibiotic resistance are crucial for characterising *S. aureus,* and exhibit substantial intraspecific variation (Pérez-Sancho *et al.*, 2018). *S. aureus'* most documented proteomic biomarkers relate to methicillin resistance which have been used to identify MRSA/ MSSA for over a decade. For example MALDI-TOF mass spectrum peaks at m/z 2,415, and doublet peaks m/z 4,594, are biomarkers indicative of MRSA (Rhoads *et al.*, 2016; Flores-Treviño *et al.*, 2019). Yu *et al.* (2022) also discovered a peak at m/z 6,590 - 6,599 which they used for identifying the presence of MRSA across more

than 20,000 isolates. Similarly, Kim *et al.*, (2019) identified 21 peaks that significantly differed in their frequency between MRSA and MSSA.

MALDI-TOF MS has also been used to identify biomarkers for resistance to other antibiotics such as ciprofloxacin and clindamycin (Garbacz *et al.*, 2021). However, knowledge surrounding the peptide biomarkers associated with *S. aureus* resistance to 'last resort' antibiotics, namely vancomycin and linezolid, is sparse. Vancomycin resistance biomarkers in bacteria such as *Enterococcus faecium* have been identified (Candela *et al.*, 2022), but in *S. aureus*, the credibility of such biomarkers has been inconclusive (Drummelsmith *et al.*, 2007). In one study, peaks at 1835 and 1863 m/z were observed in the majority of vancomycin-intermediate *S. aureus* (VISA) infections (Lu *et al.*, 2012) but no other studies have built on these findings. Yet, this is an undoubtedly important area of research given the aforementioned rise in anti-biotic resistance.

1.4.2 Virulence Factors

In addition to antibiotic resistance, virulence factors are often extremely relevant and complex predictors of mortality. Toxicity biomarkers are perhaps the most widely researched genre of *S. aureus* virulence, in regards to MALDI-TOF MS. *S. aureus* produces many different toxins such as leukotoxins and hemolysins which can result in cell lysis (Tong *et al.*, 2015). A peak at 3000 Daltons (Da) is known to indicate the presence of δ-toxin and is thus a proxy for cytotoxicity and, in turn, virulence (Gagnaire *et al.*, 2012; Brignoli *et al.*, 2022), although this largely depends on other factors, as will be discussed. Aside from the biomarkers directly indicative of toxicity, Josten *et al.* (2014) identified and attributed peaks at 2411 – 2419 m/z to a peptide produced exclusively by *agr*-posititve strains, which regulate toxin production.

In summary, it is evident that numerous biomarkers have clear links to antibiotic resistance and virulence factors. Equally, biomarkers are also useful when they can be used to predict infection outcome. In turn, many of the aforementioned biomarkers' influence on SAB mortality has been assessed and/ or incorporated into existing models.

1.5 Predictors of SAB Infection Outcome

1.5.1 Host & Contextual Predictors

The predictors of SAB infection outcome are multifaceted and depend on a multitude of host- and pathogen-features. Before highlighting the predictive biomarkers revealed by MALDI-TOF MS, it is important to first recognise the host- and contextual-factors that influence infection outcome.

Firstly, age is considered to be the most important predictor of SAB mortality (van Hal *et al.*, 2012), even amongst those with similar levels of general health (Tacconelli *et al.*, 2006), meaning that SAB mortality is influenced by the ageing process irrespective of host comorbidity (van Hal *et al.*, 2012). That said, comorbidity has been increasingly cited as a confounding factor in SAB infection outcome (van der Vaart *et al.*, 2022). A number of comorbidities have been identified as independent determinants of SAB infection outcome such as immunosuppression (Kaech *et al.*, 2006), alcoholism (Kessel *et al.*, 2024), and heart disease (Kwiecinski & Horswill, 2020). This has seen the Charlson Comorbidity Index (CCI), a score based on predicted mortality over a given timeframe, be adopted as a composite predictor (Charlson *et al.*, 2022; Russell *et al.*, 2024).

The site of infection acquisition has been also been a cited as a predictor of SAB mortality for over two decades (Fowler *et al.*, 2003). Community-acquired SAB is consistently associated with lower mortality than hospital-acquired SAB (Abdollahi *et al.*, 2024; Yang *et al.*, 2024). This is thought to be for a range of reasons including that MRSA strains are more common in hospitals (and are associated with poorer prognoses). Additional contextual predictors include the duration of SAB (Fowler *et al.*, 2003), the time taken for treatment administration (Lee *et al.*, 2017; Corl *et al.*, 2020), and the presence of particular symptoms/ complications such as a fever (Fowler *et al.*, 2003) and sepsis (García Fenoll *et al.*, 2022). In contrast, the influence of sex on infection outcome remains unclear with contrasting findings between studies (Mansur *et al.*, 2012; van Hal *et al.*, 2012; Westgeest *et al.*, 2024; Yang *et al.*, 2024). That said, one of the most obvious gaps in the literature regarding any predictor of SAB mortality is the effect of co-infection.

Of course, the effect of co-infection is likely to be a highly complex, and difficult to study given the low incidence of specific co-infections, which also restricts studies to being retrospective. That said, co-infections are commonly associated with high mortality (Liu *et al.*, 2021) which also extends to *S. aureus* and COVID-19 co-infections (Adalbert *et al.*, 2021). Therefore, while co-infections' influence on SAB infection outcome is largely unknown, it is important to recognise that it could be an important predictor of mortality in any given co-infection. Aside from host- and contextual-factors, a number of *S. aureus* strain-specific features with corresponding biomarkers are proven predictors of SAB mortality.

1.5.2 Pathogen-specific Predictors

MALDI-TOF MS biomarkers are predominantly used for species identification or to characterise strains' resistance e.g. MRSA/ MSSA, which can subsequently be used to inform treatment and/or predict mortality (Yu *et al.*, 2022). That said, the value of distinguishing between MRSA and MSSA as a predictor of mortality has been contested and is thought to be biased by the time patients spend in hospital before the onset of SAB (Wolkewitz *et al.*, 2011; van Hal *et al.*, 2012). For example, Bai *et al.'s* (2022) meta-analysis found an increased risk of 30-day SAB mortality in MRSA bacteraemia than MSSA bacteraemia since 2011 (odds ratio = 1.44; 95% CI, 1.28–1.63), although they acknowledge that those with MRSA usually have higher comorbidities.

As previously discussed, many *S. aureus* biomarkers have been identified with MALDI-TOF MS. Consequently, their utility for predicting infection outcome directly has also been assessed in some cases. For example, the aforementioned *S. aureus* peak found at 3000 Da is telling of cytotoxicity (δ-toxin) and its inclusion into Brignoli *et al.*'s (2022) predictive model of 30-day SAB mortality increased the mean AUC by 0.06. That said, this analysis only extended to infections stemming from the CC22 clonal complex (Brignoli *et al.*, 2022), and it should be noted that the modes of *S. aureus* pathogenicity vary between clonal complex (Recker *et al.*, 2017). While toxicity and biofilm formation biomarkers are proven predictors of SAB mortality, their predictive power also depends on the clonal type responsible for infection, namely CC22/ CC30 (Recker *et al.*, 2017). In

addition, toxicity is known to vary widely between MRSA strains (Laabei *et al.*, 2014, 2021).

Even if MALDI-TOF MS-identified biomarkers' underlying biological mechanisms are unknown, they can still be used to predict mortality directly. In turn, this offers opportunities for identifying biomarkers that improve SAB infection outcome models, even without complete reference mass spectral libraries. In addition to pathogen-specific biomarkers, the choices of pre-processing pipeline and model framework are foundational considerations for any prospective model. In turn, a number of studies have compared the accuracy and applicability of different pre-processing and machine learning methods in predicting *S. aureus* phenotype SAB infection outcome.

1.6 MALDI-TOF MS Analytic Pipelines

1.6.1 Pre-processing

In order to interpret mass spectra and withdraw meaningful inferences, they must first go through pre-processing. Ultimately, an effective pre-processing workflow capable of robust feature extraction is an integral aspect of any predictive pipeline. A several different approaches have been adopted for various stages of the pipeline, each considering the trade-offs between computational efficiency and information retention while simultaneously accounting for the influence of noise in MALDI-TOF mass spectra.

1.6.2 Baseline Correction

Noise is particularly common in mass spectra's low m/z which results in a upwards baseline drift (Hilario *et al.*, 2006). To limit the effect of such noise, an array of baseline correction techniques have been presented. These include those based on mathematical morphology such as Top-hat filtering which involves estimating a mass spectrum's true baseline and subtracting this from the original signal (Conrad *et al.*, 2006). Similarly, Brignoli *et al.* (2022) used moving windows from which a minimum was quantified and subtracted from the same window.

Alternative approaches focus on optimising loss functions, such as asymmetric least squares (Ruckstuhl *et al.*, 2001), or combining least squares fitting with morphological weighting (Li *et al.*, 2013). Finally, Statistics-sensitive Non-linear Iterative Peak-clipping (SNIP) (Morháč & Matoušek, 2008) has also been developed for estimating baselines and does so by iteratively clipping peaks with intensities higher than their neighbours average. While these methods all reduce the effect of background noise at low m/z, they do not nullify noise which can cause artifacts be carried on into subsequent analyses (Picaud *et al.*, 2018). In turn, a range of smoothing techniques have been developed, often for the purpose of noise reduction.

1.6.3 Smoothing

Spectral smoothing serves many purposes, one of which is stabilising signals for peak detection while also preserving peak shape. The Savitzky-Golay (SG) filter (Savitzky & Golay, 1964), Kaiser digital filter (Mantini *et al.*, 2007), and moving average smoothing (López-Fernández *et al.*, 2015) are examples of commonly used techniques in which intensity values in sliding windows are adjusted based on neighbouring intensities (Yu *et al.*, 2022). That said, SG filters have been criticised for having poor noise suppression at high frequencies (Schmid *et al.*, 2022). Instead, other novel methods such as Modified Sinc (MS) Kernels have been proven to separate signal from noise more effectively, thus preserving important spectral features (Schmid *et al.*, 2022), although these have never been applied to MALDI-TOF mass spectra.

Additionally, wavelet transformations are a well-documented approach used for spectral noise reduction and biomarker identification (Alexandrov *et al.*, 2009; Wijetunge *et al.*, 2015; Deng *et al.*, 2021; Zhou *et al.*, 2022). For smoothing, the Undecimated Wavelet Transform (UDWT) has been shown to facilitate more effective peak extraction compared to other methods (Coombes *et al.*, 2005; Picaud *et al.*, 2018). Furthermore, Yang *et al.* (2009) show the benefit of the Continuous Wavelet Transform (CWT) for distinguishing important peaks; it performed better than five other peak detection algorithms due to its ability to accurately characterise peak shape.

1.6.4 Signal Normalisation

Each m/z value of each mass spectrum has an (absolute) signal intensity which reflects the abundance of a given biomolecule. In order to facilitate comparison between spectra, normalising signal intensity is necessary. Deininger *et al.* (2011) demonstrate how different normalisation methods can result in contrasting results. They find that while normalising by total ion count is a common technique for normalising MALDI-TOF MS, it is particularly sensitive to high-abundance peaks, which can distort the relative intensities of other features (Deininger *et al.*, 2011). This is especially problematic where a small number of intense signals are not necessarily representative of the overall biochemical profile. Instead, they found median- and noise-based normalisation methods to be more robust. Another method is to normalise spectra based on their minimum and maximum intensity values so that they range from 0 to 1 (Yu *et al.*, 2022).

Cross-normalisation of MALDI-TOF spectra could be useful where data stemming from different sources are combined since they are likely to exhibit inter-spectrum variation from the differing procedures and equipment involved in their respective collections. To date, this has only been demonstrated by Boskamp *et al.*, (2021) for MALDI spectrometry imaging data.

1.6.5 Peak Detection & Extraction

Identifying and extracting peaks from MALDI-TOF mass spectra to form a matrix of meaningful features is the basis for analysis and there are an abundance of existing methods proposed in the literature. Peak detection methods revolving around signal-to-noise ratios (SNR) and continuous wavelet transformations (CWT) (Du *et al.*, 2006) are particularly common (Bauer *et al.*, 2011). SNR methods, which detect peaks based on local maxima and surrounding noise, only account for peak intensity whereas CWT also accounts for peak shape (Zheng *et al.*, 2016) and is therefore thought to offer a more nuanced results (Du *et al.*, 2006). Some pre-processing methods have also been shown to tackle baseline noise and peak detection in parallel such as Picaud *et al.*'s (2018) peak deconvolution method which outperformed a conventional two-step process by reducing the error stemming from conventional baseline corrections.

Peak extraction methods are necessary for reducing dimensionality to an appropriate level for training classifiers (Manchanda *et al.*, 2018). One method used to prevent the extraction and overrepresentation of multiple biologically equivalent peaks from signal-dense regions is to extract a local maximum from disjoint windows across the m/z scale of a given range. Windows spanning many lengths have been used, from 5 Da (López-Cortés *et al.*, 2025) to 100 Da (Brignoli *et al.*, 2022) which presents an important trade-off between computational load and potential information loss.

Small windows may retain more information but are more susceptible to noise and results in a lengthier computational process. In contrast, larger windows undoubtedly mitigate more noise stemming from sub-peaks surrounding local maxima, but they could also cause important distinctions between spectra to be missed. For example, Flores-Treviño *et al.* (2019) found that a single 4,594 m/z peak is indicative of MSSA, but that MRSA strains can be characterised by doublet split between 4,594 and 4,605 m/z. Therefore, an intermediate approach that compares the performance of models using different window sizes could be a well-rounded approach. For example, Yu *et al.* (2022) used 5 different window lengths (1–Da, 5–Da, 10–Da, 15–Da, and 20–Da) and compared corresponding models' performance metrics. However, this method does not resolve a key underlying flaw: where a true peak lies at a window edge, equivalent peaks in other spectra may fall into different windows, leading to inconsistent representation and potential misclassification. Regardless, these considerations highlight the complex and persisting trade-off between feature extraction simplification and resolution.

Another common method used during peak extraction to reduce the number of uninformative peaks is intensity thresholding. This is where peaks are retained only if their relative intensity exceeds a given threshold such as 0.01 (Candela *et al.*, 2022) or 0.02 (Brignoli *et al.*, 2022). Doing so results in a more practical number of peaks which are likely to contain a higher proportion of relevant biomarkers. While theoretically, peaks below the threshold could be important biomarkers, this is extremely unlikely since the majority of peaks falling below 2% of the maximum relative intensity are thought to be caused by background noise (Krutchinsky & Chait, 2002).

1.6.6 Peak Alignment

Peaks caused by a given biomolecule can exhibit marginally different m/z values across different isolates' mass spectra. This is partly due to instrumentation misalignment, but is also confounded by a pipeline's peak extraction method. In order to ensure that biologically equivalent peaks are recognised as such, thereby preserving the validity of future analyses, peaks must be aligned.

Pre-existing methods have addressed m/z misalignments through parameters used to merge peaks located within a given proportion of their m/z. For example, the ±0.2% threshold presented by Yasui *et al.* (2003), has been adopted in other studies as it is thought to balance the trade-off between preserving distinct biological signals and grouping equivalent ones (Manchanda *et al.*, 2018). However, the optimal threshold will vary depending on the m/z region and, as before, this method this could cause adjacent but biologically distinct peaks to be merged e.g. doublets.

1.6.7 Dimensionality Reduction

While peak extraction methods limit the number of chosen peaks, the resultant data is still high-dimensional. Coupled with the low sample sizes associated with clinical data, this presents the high-dimensionality-small-sample (HDSS) problem (Hilario *et al.*, 2006). As a result, reducing the number of biomarkers used to predict target variables is a necessary aspect of the pipeline. Aside from the computational benefit offered dimensionality reduction, reducing hundreds of biomarkers down to a select few is essential for accelerating their integration into clinical pipelines. This is because newly discovered biomarkers must be characterised and have their proteomic mechanisms understood by biomedical researchers. That said, dimensionality reduction techniques often fall victim to overfitting despite efforts at over-, under- and synthetic-sampling (López-Cortés *et al.*, 2025).

As discussed, moving windows are often used during peak detection/extraction to limit the number of identified biomarkers (Brignoli *et al.*, 2022). However, further dimensionality reduction methods often involve removing features based on their ranked importance (Wang *et al.*, 2022). These methods include principal component analysis (PCA) (Shao *et al.*, 2012), ant colony optimisation (Ressom

et al., 2007) and recursive feature elimination (Liu *et al.*, 2009). Additionally, some methods filter peaks through statistical filtering whereby the occurrence ratio of meaningful peaks are quantified and ranked before an upper quartile is calculated and the corresponding peaks are retained Zhang *et al.*, (2023). The resultant subset of peaks is then ready to be used in the next stage of the diagnostic/ predictive pipeline: modelling.

1.7 Model Frameworks for SAB Infection Outcome

Models that use MALDI-TOF mass spectra to predict pathogen characteristics and/ or infection outcome could be invaluable for clinical management. In the context of SAB, the majority of these models are binary classification problems e.g. 'Is the infection caused by MRSA/ MSSA?', or 'Will the infection end in mortality/ survival?'. Recent studies have compared the performance of various models.

Unsupervised learning algorithms are a relevant group of machine learning frameworks in the context of MALDI-TOF MS since they do not require labelled data, as is the reality in clinical contexts. Methods such as hierarchical cluster analysis could, in theory, help identify clusters of peaks associated with either MRSA or MSSA, thus providing valuable insights into strain-specific virulence. However, the discriminatory power of unsupervised learning algorithms has been called into question. Lasch *et al.* (2014) show that hierarchical cluster analysis could not successfully identify methicillin-resistance biomarkers from *S. aureus* MALDI-TOF MS data. In contrast, supervised learning algorithms are more widely used for predicting strain characteristics and infection outcome.

Bayesian inference is a hopeful, yet underexplored, genre of model for predicting infection outcome with MALDI-TOF mass spectra. Bayesian architectures incorporate prior distributions (priors) into analysis to output a posterior distribution. The addition of prior beliefs makes this framework stand out as a promising method capable of offering better-informed predictions, especially given the increasing number of studies on the predictors of SAB mortality (Yang *et al.*, 2024). However, to date, no research has evaluated this or any other Bayesian framework in the context of SAB infection outcomes. This may be partly due to its higher computational demands and, therefore, longer run times which

can limit its applicability in clinical settings. Equally though, posterior distributions permit 95% credibility intervals (CI) which would offer more nuanced assessments of mortality risk (Goligher *et al.*, 2024). Furthermore, Bayesian frameworks handle smaller samples sizes well which could make them a preferred method where data is sparse, as is often the case in clinical contexts.

1.8 Research Objectives

Given the importance of tackling the burden posed by *S. aureus*, this study aims to explore a number of important relevant topics. The study's primary objective is to design a robust pre-processing pipeline that transforms raw mass spectra into a cleaned and aligned binary peak matrix.

After creating and implementing the feature selection pipeline, the second objective is to devise a dimensionality reduction pipeline that uses the resultant peak matrix from the first pipeline as an input, and creates a subset of important binary peaks as the output. This will facilitate biomarker discovery during downstream analysis.

Following the extraction of mass spectral peaks, their collective structure will be explored through clustering to gain insights into the data's underlying patterns. Then, given the importance of biomarker discovery, the next objective is to identify and explore the spectral peaks associated with both toxicity and infection outcome, using case study data.

The final study objective is to create and compare predictive models of 30-day SAB infection outcome that incorporate metadata, priors and/or relevant peaks in parallel. This objective is an underpinning motivation for the field's research as it may offer key insights into how to best predict clinical outcomes.

2. Data

2.1 Sources

To evaluate the prospective pipelines and illustrate their applications through case studies, this study used MALDI-TOF mass spectra from two independent sources stemming from isolates of patients with SAB.

2.1.1 CC22

The CC22 dataset, described by Recker *et al.* (2017), consists of MALDI-TOF mass spectra for 137 SAB isolates, all of which were caused by the same *S. aureus* clonal type: CC22.

2.1.2 Cork

The Cork data, undescribed formally, contains the absolute signal intensities for 181 isolates, the majority of which (178) stemmed from patients with SAB. Unlike the CC22 data, the clonal background(s) for the Cork isolates are unknown.

2.2 Merging & Harmonisation

2.2.1 Mass Spectra

The CC22 mass spectral data (.txt) was conformed to the structure of the Cork data (.RData) so that each element contained a dataframe with an isolate's m/z and relative/ absolute intensities. In some cases, the Cork data contained repeat mass spectra e.g. 'A095 #2'. It was inferred, after visual inspection, that such repeats were to be treated as the best quality mass spectra and so they were renamed and matched with the corresponding metadata entry.

Three infections from the Cork dataset were not isolated from bacteraemic patients, and so these isolates were omitted for modelling infection outcome. Additionally, isolates with no corresponding mass spectra or metadata were also omitted meaning that all redundant data was removed. Finally, the sampling rate along the m/z axis was standardised by resampling the data onto a uniform m/z axis by interpolating data in the overlapping regions. This was necessary for

future pre-processing steps requiring uniformly spaced data, namely convolution-based smoothing. Interpolation is common in MALDI-TOF pre-processing (Tong *et al.*, 2011) and the integrity of the interpolated spectra was evidenced by comparing it to the raw spectra (Appendix 1). The new step size for all spectra became 0.84 m/z. In total, the data merging and cleaning amounted to a total of 278 isolates suitable for analysis with 110 and 168 stemming from the CC22 and Cork datasets, respectively, which permitted a unified modelling workflow encompassing data from two independent sources.

2.2.2 Metadata

The mass spectral isolates were accompanied by corresponding metadata which would be used to model and investigate the pipelines outputs in case studies. The metadata from the two datasets were combined and harmonised where possible for downstream analysis. This was limited by the lack of annotations/description and features for the Cork data. For example, while features such as 30-day mortality, age, sex, methicillin resistance (MRSA/ MSSA) and toxicity (high (60% >) /low) were included in both datasets, other potentially informative features such as CCI, toxicity, and clonal complex remained unknown for the Cork data.

3. Methods & Results

All analyses were performed in R (version 4.4.0) with integration of Python scripts via the reticulate package (Navarro *et al.*, 2024), including pre-processing, baseline correction, and statistical modelling.

3.1 Pipeline 1: Peak Selection

3.1.1 Smoothing

While a baseline correction is the first step of many pre-processing pipelines, Picaud *et al.*, (2018) argue this is better applied after smoothing. This is because

smoothing particularly reduces the impact of sharp peaks which, in mass spectra, are often high frequency noise. Conversely, ‘real’ peaks are often wider and therefore preserved after smoothing. The suppression of noise then facilitates more accurate estimations of baselines.

SG filters are commonly used to smooth mass spectra but their poor noise suppression at high frequencies, as well as their propensity to introduce boundary artifacts through windowing has been scrutinised (Picaud *et al.*, 2018; Schmid *et al.*, 2022). In turn, alternative smoothing methods with superior performances such as the convolution-based Modified Sinc (MS) kernel have been presented (Schmid *et al.*, 2022).

The SG filter and MS Kernel were applied to isolate A060 and visually compared. The MS kernel appeared to balance smoothing, and better retain seemingly important peaks in comparison to the SG filter, such as at ~3400 m/z (Fig. 2, Appendix 2). Therefore, MS smoothing was applied to all spectra. Since MS kernels have never been applied to MALDI-TOF data in the literature, user-defined values were informed by manual testing on multiple example spectra, and by Schmid *et al.*’s (2022) recommended values for spectral data more generally, to balance the trade-off between smoothing intensity and peak preservation.

A Gaussian weighting window ($\alpha = 4$) was multiplied with a sinc function to create a smoother, faster-decaying kernel with improved frequency selectivity and reduced ringing artifacts during convolution (Appendix 3). This hybrid design allowed the kernel to focus smoothing around the centre of each peak (Schmid *et al.*, 2022) and supress the extended side lobes associated with the sinc function. The kernel had a window size (m) of ±30 and rate of oscillation per kernel (n) of 6 (Appendix 3). To reduce edge artefacts, linear extrapolation was applied to extend the spectrum before convolution. Furthermore, a small sinusoidal correction was added to flatten the passband response; the relevant correction coefficient values were taken and fitted empirically from Schmid *et al.* (2022) who found them to be most optimal where $n = 6$. Finally, the convolution output was centred along the spectrum to preserve the spatial location of peaks.

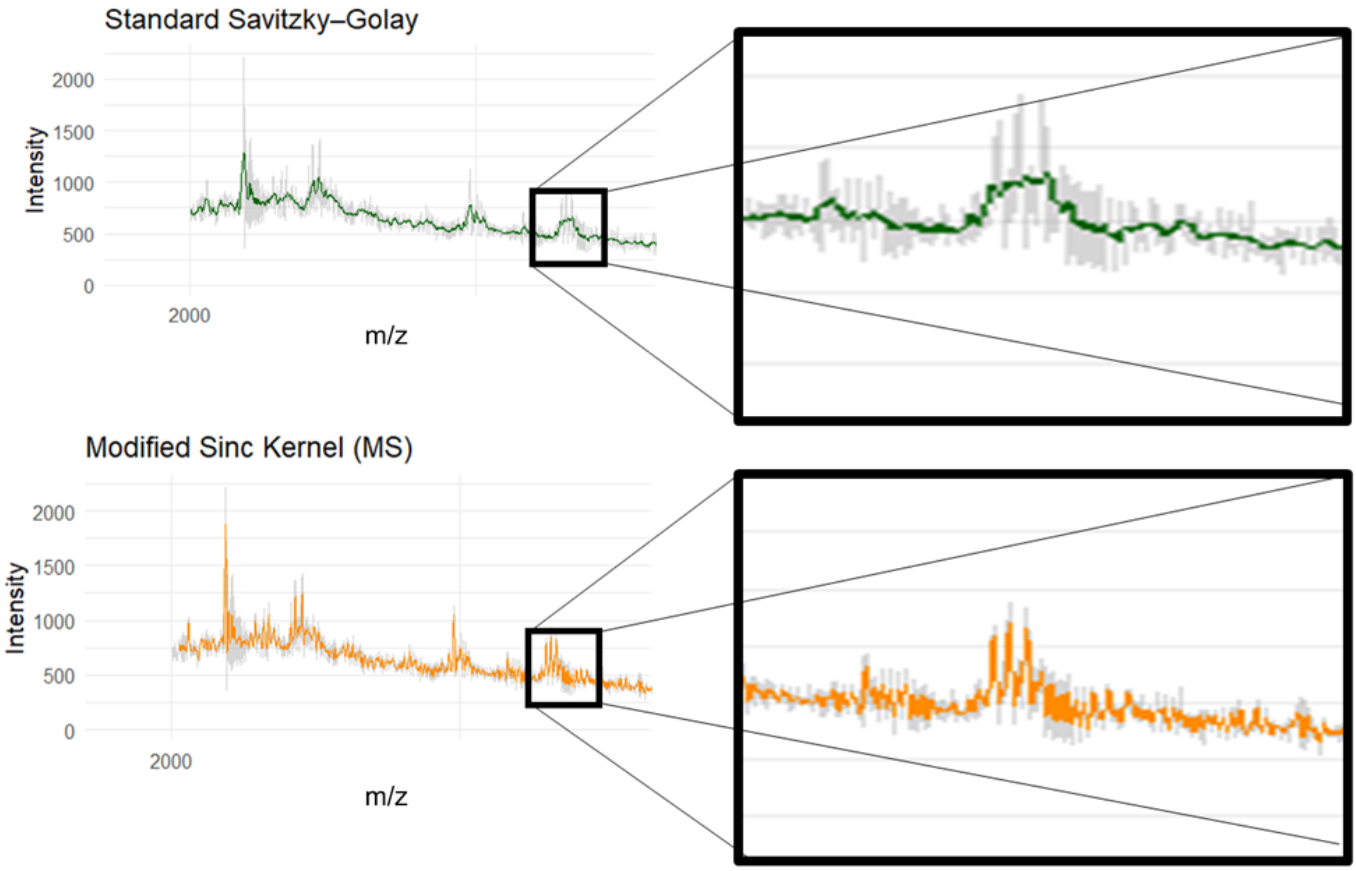


Figure 2. Side-by-side comparison of a spectrum smoothed by the Standard Savitzky-Golay filter and a Modified Sinc Kernel, overlaid on the raw spectrum (Isolate A060). The areas of clear differences (~3400 m/z) are expanded in the boxes.

3.1.2 Baseline Correction

3.1.2.1 Statistics-sensitive Non-linear Iterative Peak-clipping

The noise associated with MALDI-TOF mass spectra's low m/z (Hilario *et al.*, 2006) was apparent across many isolates (Fig. 2). In turn, two separate baseline correction methods were applied to a spectrum with obvious noise (A060) and the outputs were visually compared.

First, Statistics-sensitive Non-linear Iterative Peak-clipping (SNIP) (Morháč & Matoušek, 2008) was applied. To do this, the intensity values were first transformed to stabilise the variance associated their exponential peak shapes. Two transformations were exercised and compared, the first being a log transformation, and the second a Log-Log-Square-root (LSS) (Equation 2) (Morháč & Matoušek, 2008). The SNIP function was created manually to give

control over the parameters, a freedom not afforded by the 'MALDIquant' package (Gibb & Strimmer, 2012).

Equation 2:

$$v(i) = \log\{\log\left[\sqrt{y(i)+1}+1\right]+1\}$$

Next, each intensity value (*i*) was iteratively (*p*) compared to a local background estimate calculated within a defined dynamic window that changed with each iteration ($i \pm p$) (Equation 3) (Morháč & Matoušek, 2008).

Equation 3:

$$v_p(i) = \min\{v_{p-1}(i), \frac{1}{2}\left[v_{p-1}(i+p) + v_{p-1}(i-p)\right]\}$$

Peaks above the local background estimate were clipped and the process was repeated over 100 iterations to estimate a baseline. This number of iterations was chosen on the basis that it could reach stable conclusions while maintaining computational efficiency. The intensities were then back transformed with an exponent function/ inverse LSS operator (depending on the baseline estimation function) before the resultant estimated baseline was subtracted from the smoothed mass spectra.

A visual comparison of the baseline corrected spectrum stemming from the log and LLS functions for isolate 'A060' revealed near identical spectra (Fig. 3) and it was inferred that any differences would likely have negligible effects on downstream analyses. As a result, the more literature-prominent LSS variation of the function was carried forward into the successive baseline-correction method comparisons.

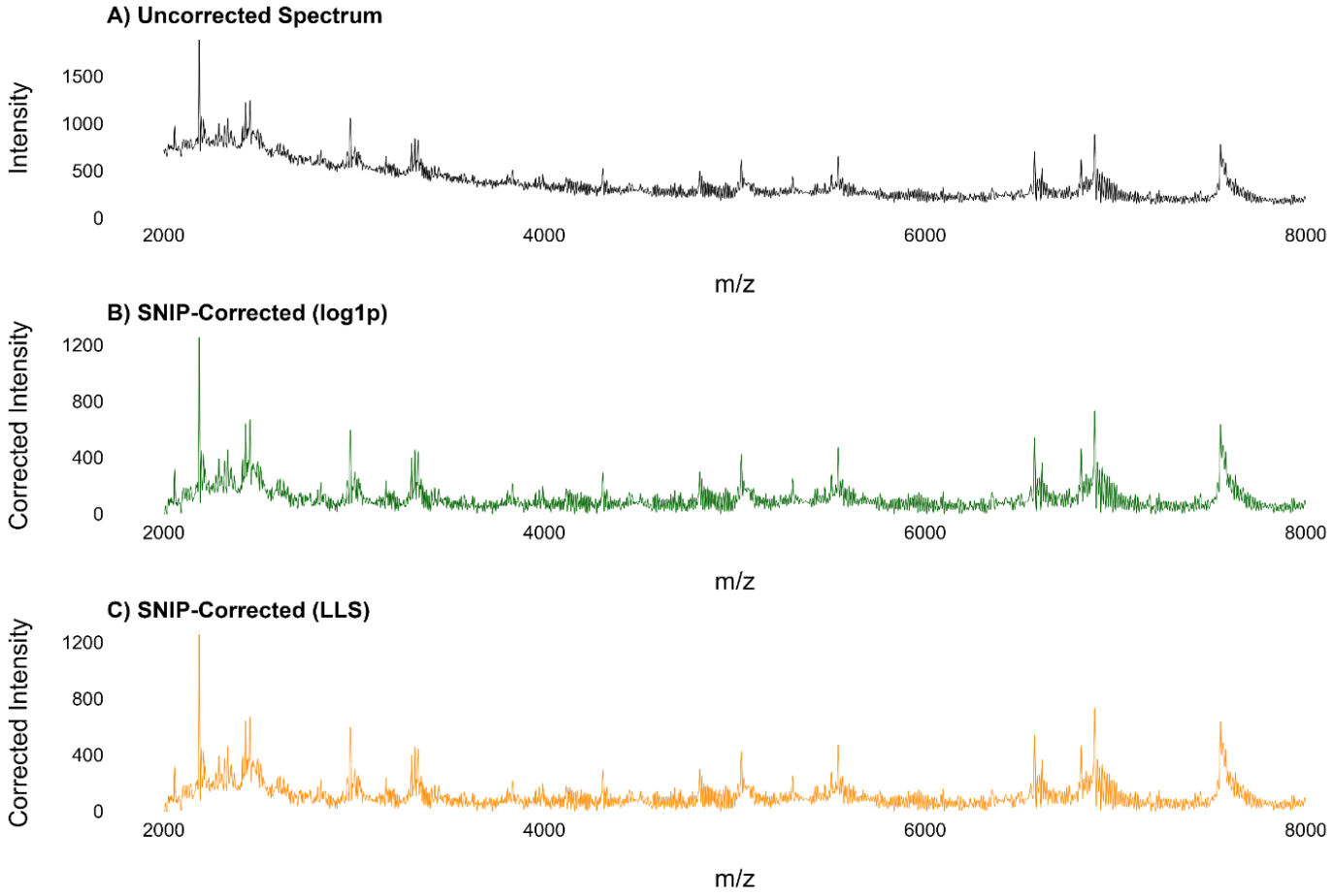


Figure 3. An example of an uncorrected spectrum (A) (Isolate A060), and the effect of two transformations during SNIP baseline correction: log1p (B) and LLS (C).

3.1.2.2 Top-hat

The second explored baseline correction framework was the top-hat baseline correction which estimates a baseline based on the spectral morphology (Conrad *et al.*, 2006). Two functions were created that used rolling, overlapping half windows spanning ±20 m/z from which the minimum and maximum intensity values were taken and assigned to the centre point sequentially. The resultant estimated baseline was subtracted from the original mass spectra.

The SNIP and Top-hat methods were then visually compared against each other and their corresponding raw spectrum (isolate A060) (Fig. 4). The comparison revealed the Top-hat to be far more aggressive with a baseline consistently close to zero. Conversely, the SNIP was characterised by a smoother, curving baseline that also seemed to retain peaks more clearly in some places, such as at around 5000 m/z in isolate A060 (Fig. 4). While the peaks at lower intensities may appear

to be less clean, they are more likely to contain preserved signal in comparison to Top-hat transformed spectra which could have low/ broad peaks removed entirely.

Two other baseline correction methods (Convex Hull and a Rolling Median Filter) were tested using the 'MALDIquant' package (Gibb & Strimmer, 2012) but these had poorer performances and so are not included here. Furthermore, while convolutional neural networks have been cited as a method of baseline correction (Schmidt *et al.*, 2019), this requires thousands of samples for training. Therefore, SNIP was chosen to baseline correct the MS kernel-smoothed spectra (Fig. 4). The baseline correction was applied before normalisation to prevent any baseline noise responsible for minimum or maximum intensity values from being relayed in the subsequent pre-processing steps.

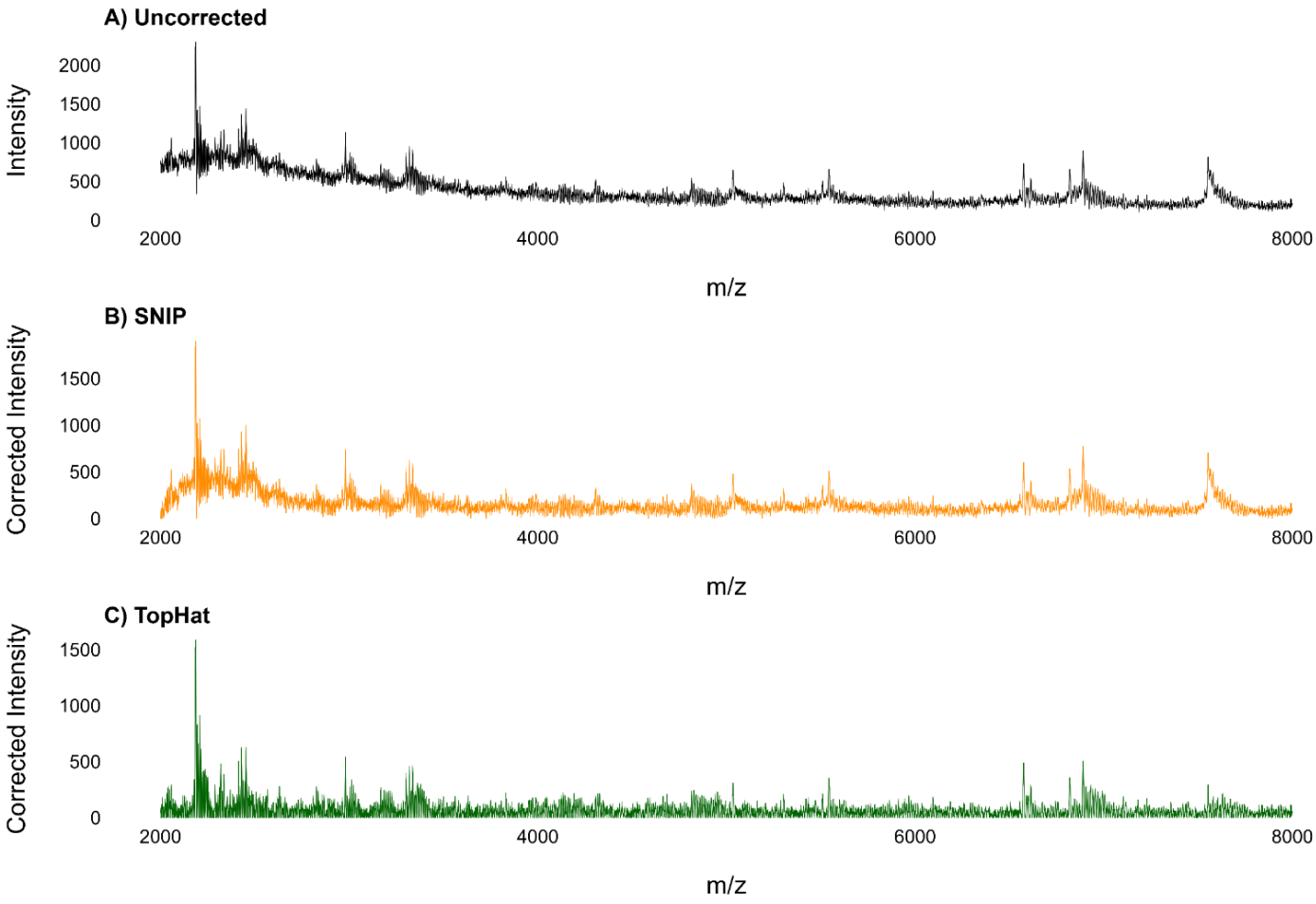


Figure 4. Comparison between an uncorrected spectrum (A), and the two baseline correction methods: SNIP and TopHat baseline corrections (Isolate A060).

3.1.2 Signal Intensity Normalisation

Each spectrums' absolute signal intensities were made relative to facilitate comparison and improve computational efficiency for modelling (Equation 4) (Yu *et al.*, 2022). As a result, each spectrums' intensity ranged from 0 – 1.

Equation 4:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

3.1.3 Peak Detection & Extraction

Following normalisation, a number of peak detection methods were explored. It was evident that methods based on intensity thresholding would be unsuitable because the spectra displayed non-uniform noise whereby some spectra had a smoother baseline curves at low m/z compared to others (Fig 5). Instead, peak detection based on local maxima was adopted.

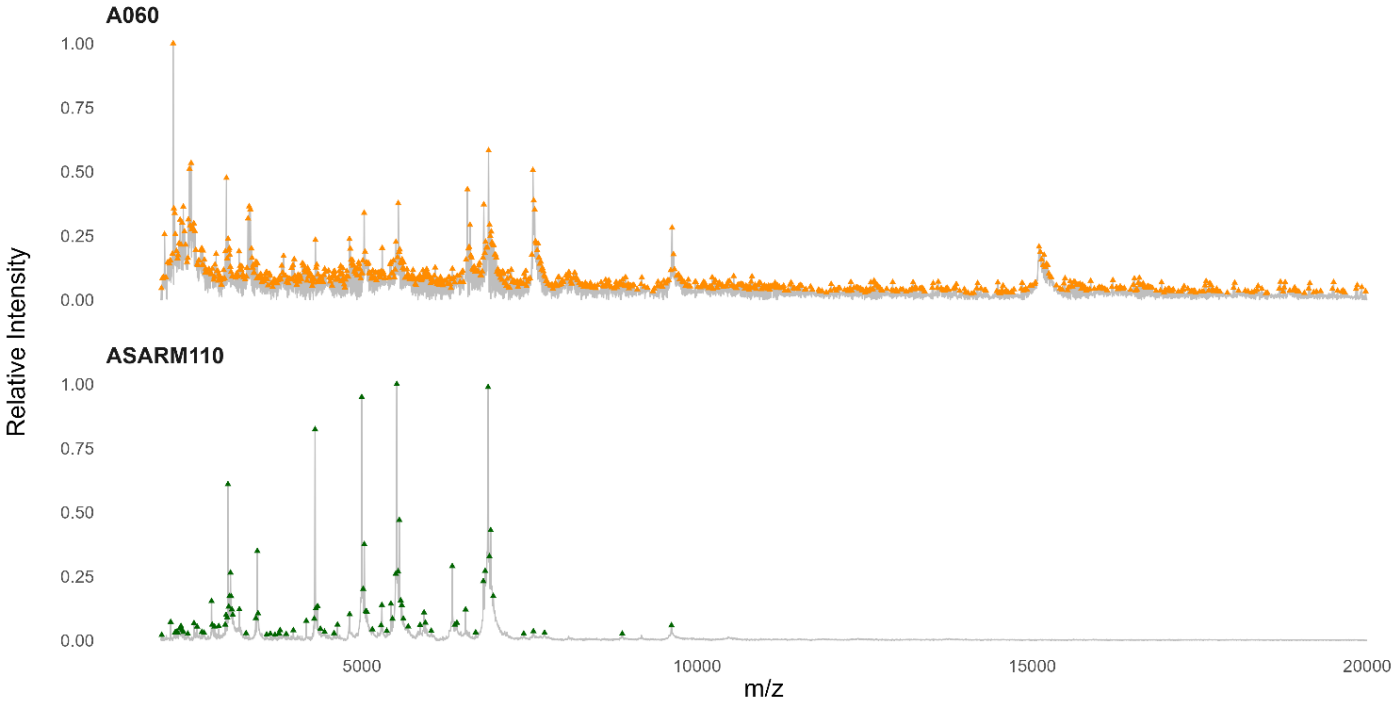


Figure 5. Visual demonstration of the differences in baseline curves and peak prominence between different spectra (Isolates A060 & ASARM110)

Local maxima were identified in the spectra from which peaks were detected based on user-defined criteria encoded by SciPy's 'find_peaks()' function (Virtanen *et al.*, 2020). Peak prominence was calculated relative to neighbouring local minima rather than other spectra which prevents the heterogeneous noise between spectra from influencing peak selection. Furthermore, the method overcame the issues of peak misinterpretation at window edges and the overly selective nature of peak extraction where fixed windows are used (Flores-Treviño *et al.*, 2019).

The 'prominence' argument dictated how high a m/z should rise above its surrounding baseline to be considered an independent peak. This customisation allowed the trade-off between information retention and computational efficiency to be carefully considered. After manually testing and comparing different values, a prominence value of 0.02 was chosen (Fig. 6). Peaks with a relative intensity below 0.02 were discarded ('height = 0.02'), in line with other methodologies (Brignoli *et al.*, 2022; Candela *et al.*, 2022). Finally, arguments for the thresholds for the 'distance' and 'width' between peaks were made to be zero since no reliable assumptions could be made regarding peak morphology.

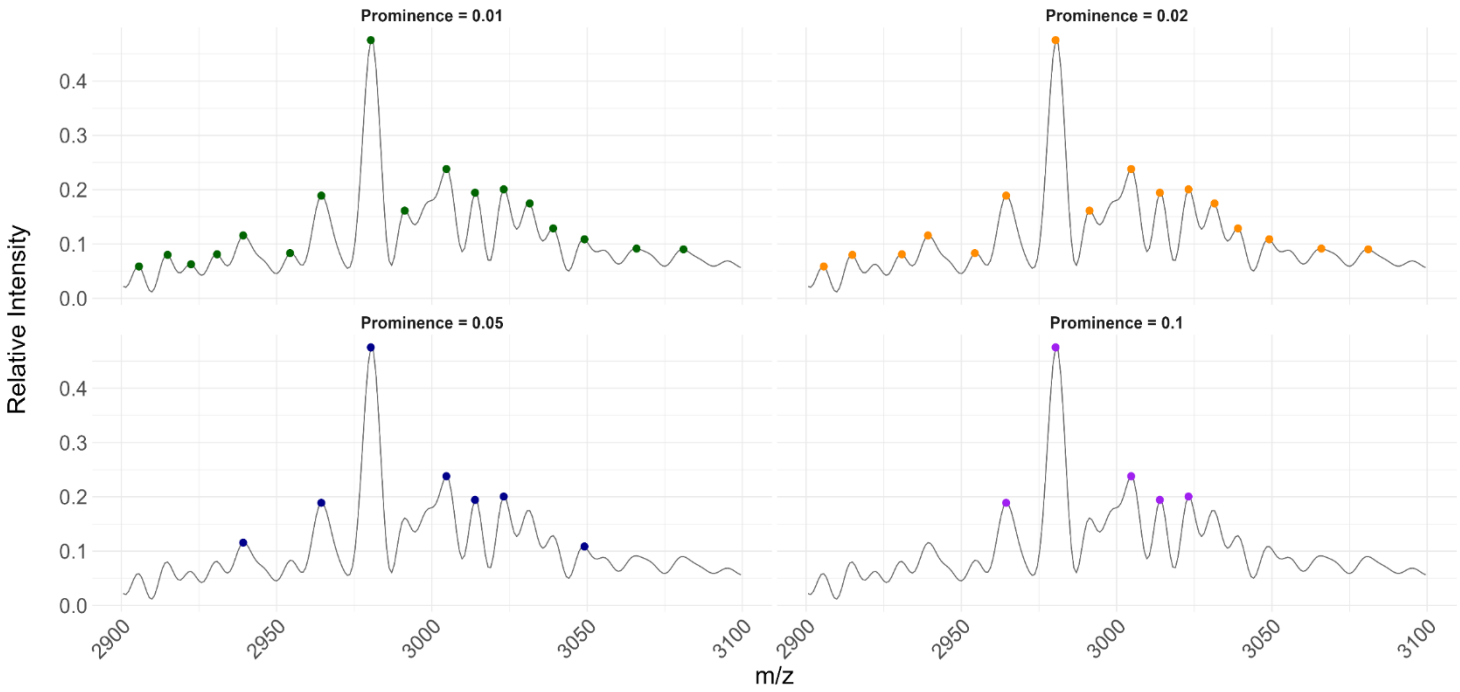


Figure 6. The effect of different peak prominence values on peak selection for an example spectrum (Isolate A060).

3.1.4 Peak Alignment

As expected, selected peaks exhibited marginally different m/z values across different isolates' mass spectra. So, to allow biologically equivalent peaks to be treated as such for analysis, a function was built to align peaks by grouping those within a given distance threshold (5 m/z). To do this, all peaks were collated and listed in ascending order before they were sequentially grouped with neighbouring peaks. The mean m/z of each group's peaks was then calculated to define the bin value (Fig. 7). Each isolate's peaks were then assigned to the closest bin, as long as they were within the distance threshold.

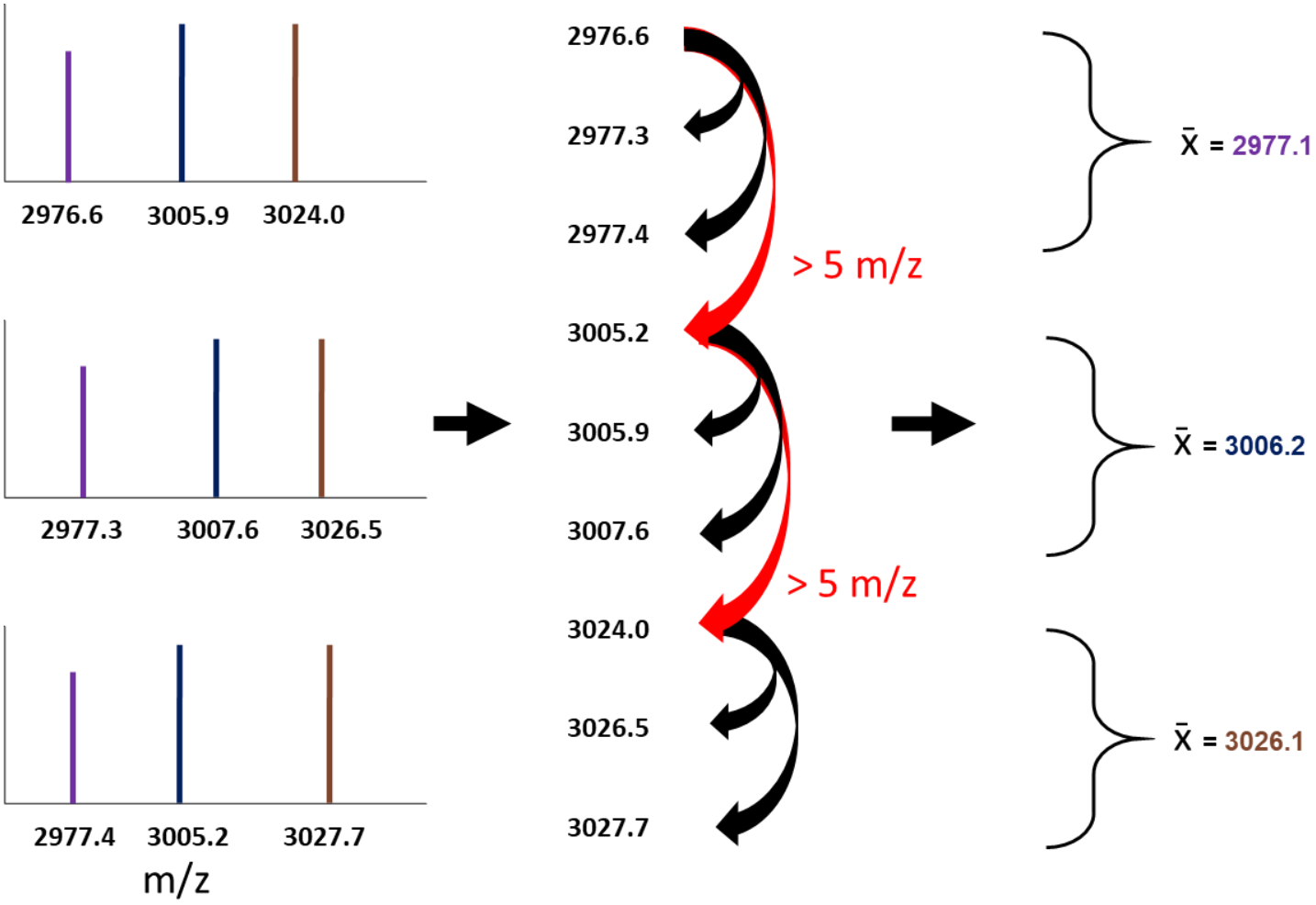


Figure 7. Not-to-scale schematic illustrating the peak alignment process. Selected peaks from different spectra are listed in ascending order. The first peak becomes the reference and all peaks within 5 m/z (black arrows) are grouped to form a reference bin. The first peak outside this range (red arrows) is the next reference peak and so on. Each bins m/z is defined as the mean of its constituent peaks.

The 5 m/z alignment threshold was chosen by manually inputting different values and visually comparing their resultant alignments across multiple sections of m/z. First, the effects on alignment were gleaned for the whole spectra (Fig. 8). Then, the effects between 2950 and 3050 m/z was visually assessed within and between spectra (Fig. 9 and 10, respectively) since this m/z range has been associated with peaks linked to toxicity and SAB mortality (Recker *et al.*, 2017; Bringoli *et al.*, 2022). This qualitative evaluation meant that 5 m/z was deemed to be the most appropriate threshold. This resulted in 33,167 peaks being assigned to 1,970 independent bins (reference peaks).

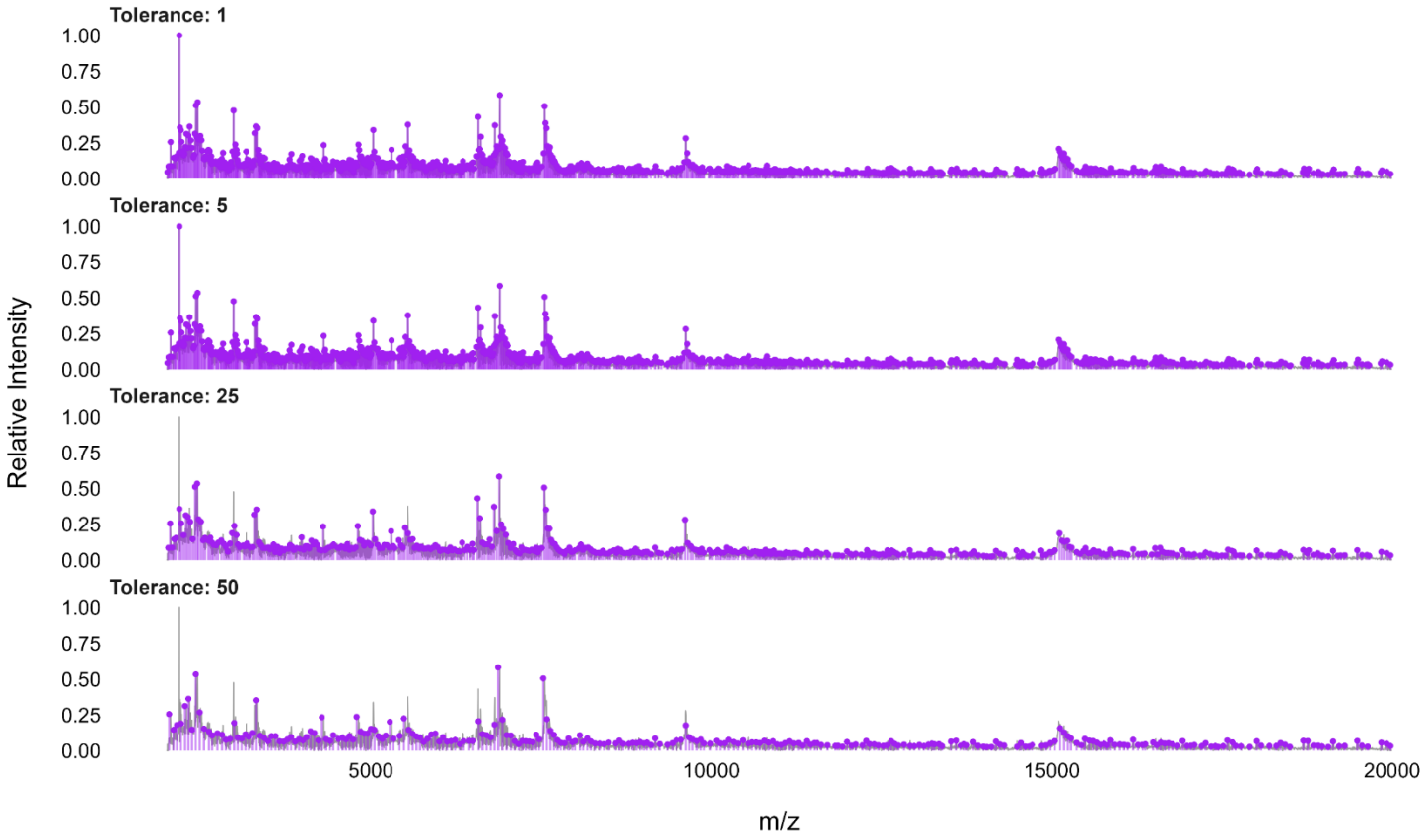


Figure 8. The effect of different tolerance values on peak alignment on a mass spectrum (Isolate A060).

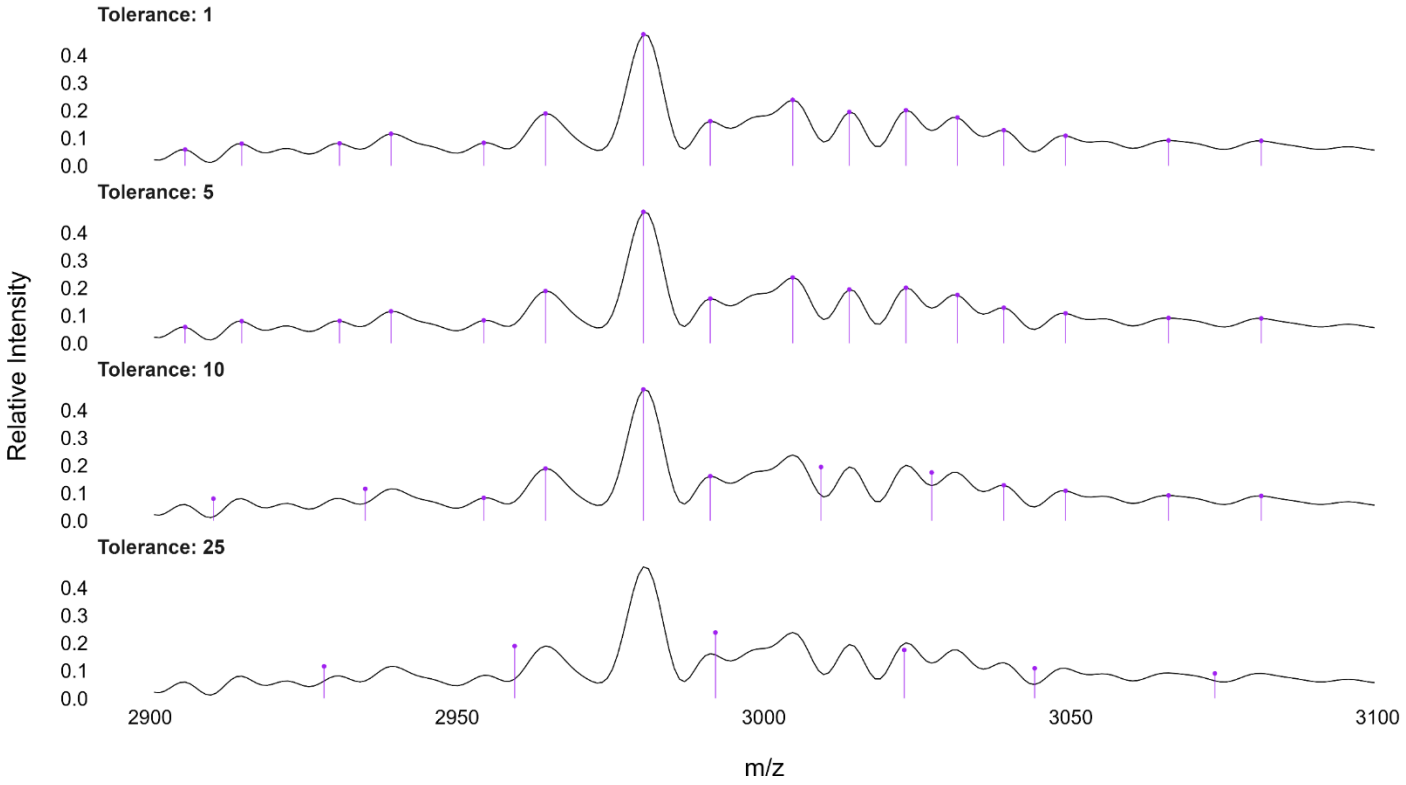


Figure 9. The effect of different tolerance values on peak alignment between 2900 and 3100 m/z on a mass spectrum (Isolate A060).

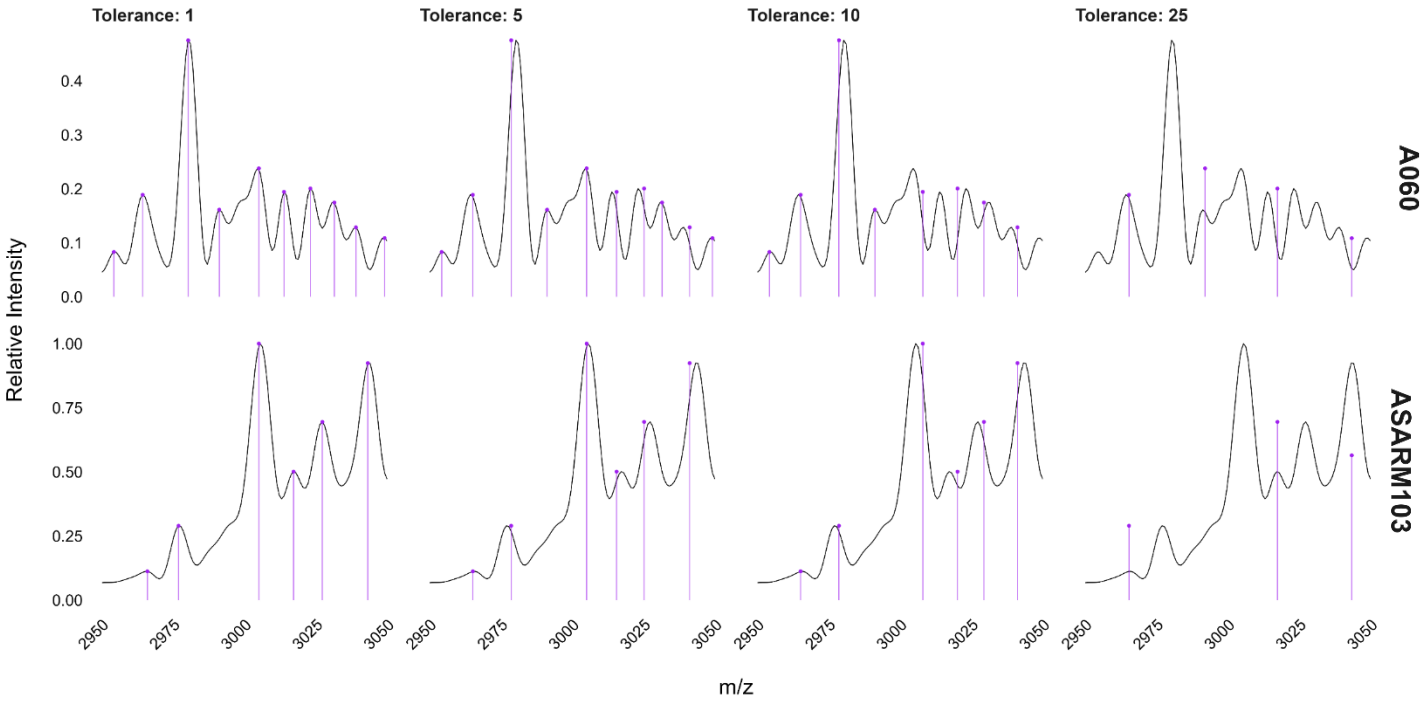


Figure 10. The effect of different tolerance values on peak alignment between 2900 and 3100 m/z on mass spectra from two sources (Isolates A060 (CC22) and ASARM103 (Cork)).

It should be noted that peak relative intensities can fluctuate between isolates for a range of technical reasons such as differences in ionisation efficiency, matrix effects, or instrument variability (Hilario *et al.*, 2006). This was apparent when comparing spectra across the CC22 and Cork datasets (Fig. 5). While variations can reflect biology (Aguiar *et al.*, 2021), the max-min normalised data could not be used to infer this, especially given the noise introduced using data from two different sources. Furthermore, since the study was focused on the presence/ absence of peaks, peak relative intensity was not calibrated. This step completed the pre-processing pipeline (Fig. 11), resulting in a matrix of 1970 unique reference peaks where each row represented an isolate, and each column a bin.

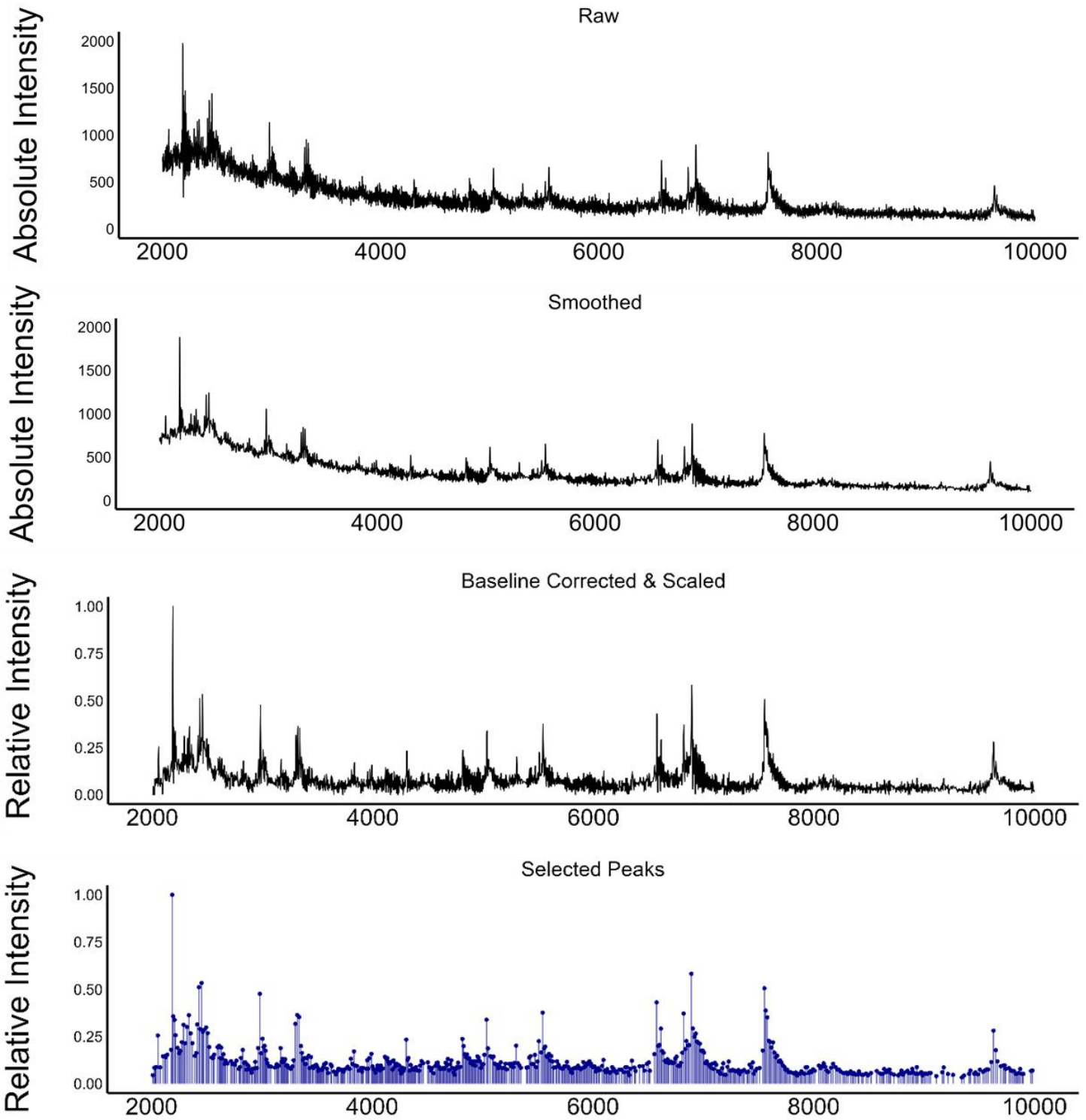


Figure 11. The stages of the peak selection pipeline on an example spectrum (Isolate A060). The stages from top to bottom are the raw, smoothed, baseline corrected & scaled spectra, and aligned & selected peaks.

3.2 Pipeline 2: Dimensionality Reduction

3.2.1 Peak Frequency Filtering

Before investigating the peaks associated with toxicity and SAB infection outcome, candidate peaks were identified and shortlisted to reduce the peak matrix's high-dimensionality. The first step in reducing the number of uninformative peaks involved applying a minimum peak frequency threshold, as proposed by Tibshirani *et al.* (2004). This was to omit reference peaks whose importance would be limited by their low abundance. A frequency value of 0.15 was chosen based on existing literature (Manchanda *et al.*, 2018) and by considering the number of isolates available for analysis. This resulted in the removal of 1,731 reference peaks (88%); there were 239 reference peaks remaining after frequency filtering (Fig. 12).

3.2.2 Correlated Peak Filtering

The correlation between peaks was determined using Pearson's coefficient. For pairs with correlation coefficients of ≥ 0.7, the peak with the largest mean absolute correlation was removed. The removed peaks were compiled to allow them to be referred to where necessary. The threshold of 0.7 was chosen to adequately reduce both co-linearity and dimensionality (Fig. 12). This resulted in the removal of 68 peaks, reducing the number of binarised candidate peaks from 239 to 171.

3.2.3 RFECV

Next, Recursive Feature Elimination with Cross-Validation (RFECV) was used to determine two final peak subsets for modelling toxicity (high/ low) and infection outcome (López-Cortés *et al.*, 2025). It should be noted that this aspect is dependent on the response variable e.g. binary vs non-binary peak data, and so it should be considered as analysis rather than pre-processing. RFECV iteratively fits a model, ranks each features' importance based on their contribution to classification performance and subsequently removes the most redundant feature (Liu *et al.*, 2009). Model performance is evaluated at each iteration with stratified cross-validation which preserved class balance within folds.

RFECV was implemented with the scikit-learn package via reticulate (Navarro *et al.*, 2024), using a logistic regression model with L2 regularisation for its robustness to high-dimensional data and sensitivity to collinearity. At each iteration, model performance was evaluated using stratified 10-fold cross-validation with AUROC. RFECV removed features until an optimal AUC was realised which resulted in final subsets of 53 and 9 peaks for modelling toxicity and infection outcome, respectively (Fig. 12).

Steps were taken to address each of the core logistic regression assumptions. Firstly, multi-collinearity was mitigated by the aforementioned filtering of highly correlated peaks. Secondly, class imbalance was handled through stratified 10-fold cross-validation, ensuring proportionate representation of outcome classes in each fold. Thirdly, the linearity of the logit was ensured by the binary nature of the response data, which aligns with the model's assumption of a linear relationship between predictors and the log-odds of the outcome. This concluded the dimensionality reduction pipeline (Fig. 12), The outputs of both pipelines were next applied to case studies involving bacterial structure, toxicity, and infection outcome to illustrate the practical utility of the pipeline outputs.

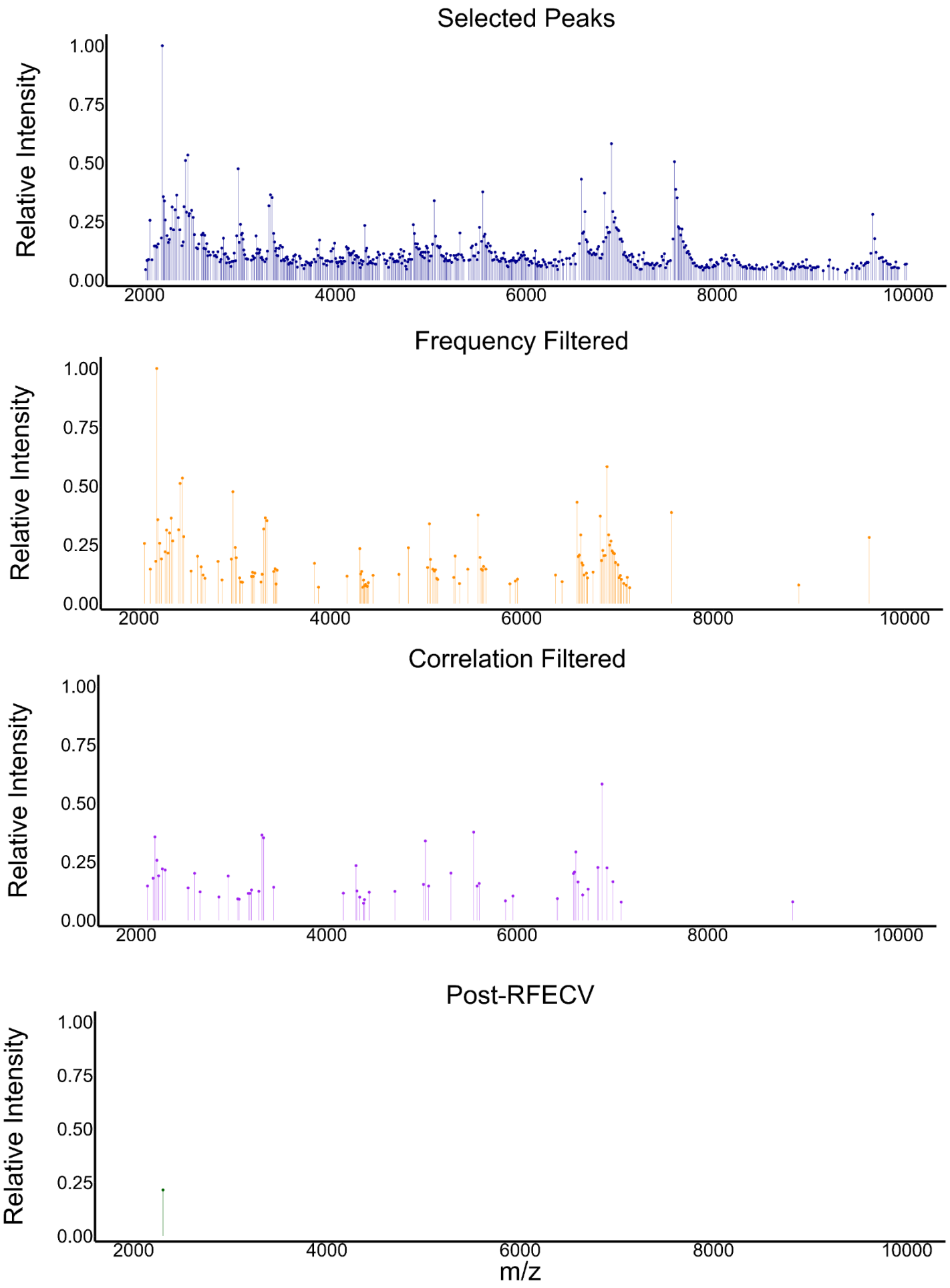


Figure 12. The stages of the peak dimensionality reduction pipeline for modelling infection outcome on an example spectrum (Isolate A060). The facets from top to bottom show the peaks after selection, frequency filtering, correlation filtering and RFECV. The final facet shows the single peak retained for this isolate. The number of peaks varied between isolates, all were treated as binary features.

3.3 Case Study 1: Structure

3.3.1 NMDS

Before modelling toxicity and infection outcome, the collective structure of the selected peaks was gleaned. This was based on the outputs of the peak selection pipeline, a peak matrix for each isolate. The drivers of peak matrix composition were investigated with non-metric multidimensional scaling (NMDS) and, subsequently, Permutational Multivariate Analysis of Variance (PERMANOVA). The NMDS calculates a distance-matrix and preserves rank order to map the samples into a low-dimensional space. This method was chosen over other methods such as Uniform Manifold Approximation and Projection (UMAP) because in contrast to UMAP which amplifies structures, NMDS forms more conservative, reliable clusters owing to distance matrices which preserve relative distances. This was confirmed by visually comparing the raw clusters formed by the two methods; the UMAP resulted in a higher and more distinct number of clusters (Fig. 13). While NMDS is more commonly used in ecology, it has been used in proteomics/ genomics to explore gene expression (Taguchi & Oono, 2005). Therefore, NMDS was chosen for investigating the data's structure using the 'vegan' package (Dixon, 2003).

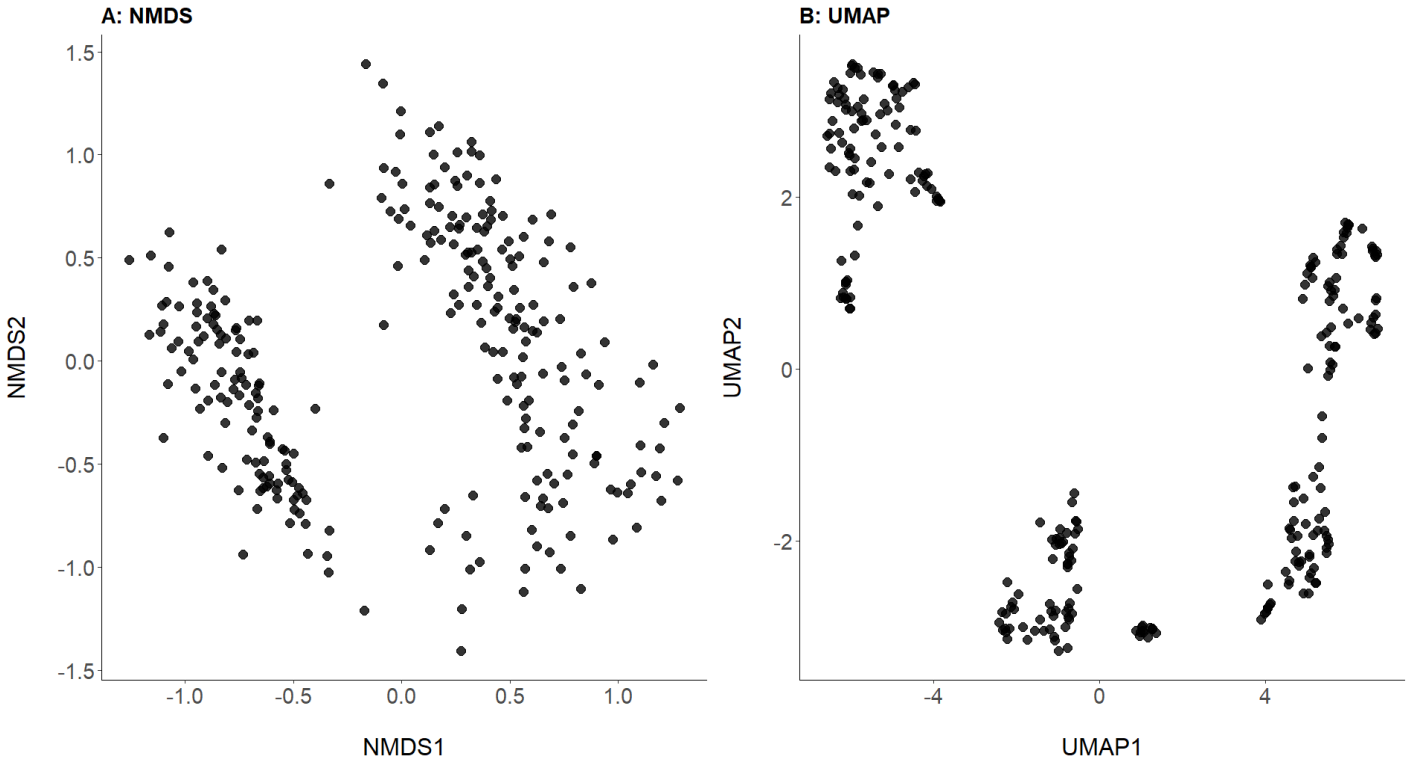


Figure 13. Dimensionality reduction maps for isolate peak matrices via A) NMDS and B) UMAP (n = 278).

The peak matrix was binarised before a Jaccard distance matrix (well-suited for binary data) was calculated and inputted into the NMDS. The resulting ordination was visualised in two dimensions and annotated by three metadata variables: data source, methicillin resistance, and 30-day infection outcome. Additionally, hierarchal clustering was applied to the distance matrix and visualised. The optimal number of clusters was determined by using silhouette analysis (k = 2 - 5) to evaluate how many clusters best reflected the data's structure.

3.3.2 Findings

The NMDS and PERMANOVA gave clear insights into the factors driving the peak matrix's structure. While infection outcome did not explain a significant portion of the variation in peak matrix composition (PERMANOVA, $R^2 = 0.006$, $F_{1,\ 276} = 1.6$, $p = 0.095$; Fig. 14a), both methicillin-resistance status (PERMANOVA, $R^2 = 0.218$, $F_{1,\ 276} = 45.9$, $p = 0.001$; Fig. 14b) and source (PERMANOVA; $R^2 = 0.143$, $F_{1,\ 276} = 77.1$ $p = 0.001$; Fig. 14c) did show significant relationships. However, there was apparent co-linearity between these two variables; colouring the NMDS showed that the clusters aligned entirely with source (Fig. 14c) suggesting that either technical variation or clonal differences between the datasets likely drove the structure. Despite re-pre-processing the spectra with different peak alignment tolerance values, similar findings persisted. Finally, the hierarchal clustering paired with silhouette analysis showed the optimal number of clusters to be 3 (Fig. 14d & Appendix 4, respectively). As will discussed, these results raise important questions surrounding the variability between sources. Namely, if clonal complex is the main driver of the data's structure, this would highlight the extent of inter-clonal complex diversity which may be reflected in their phenotypes – an important consideration in clinical settings.

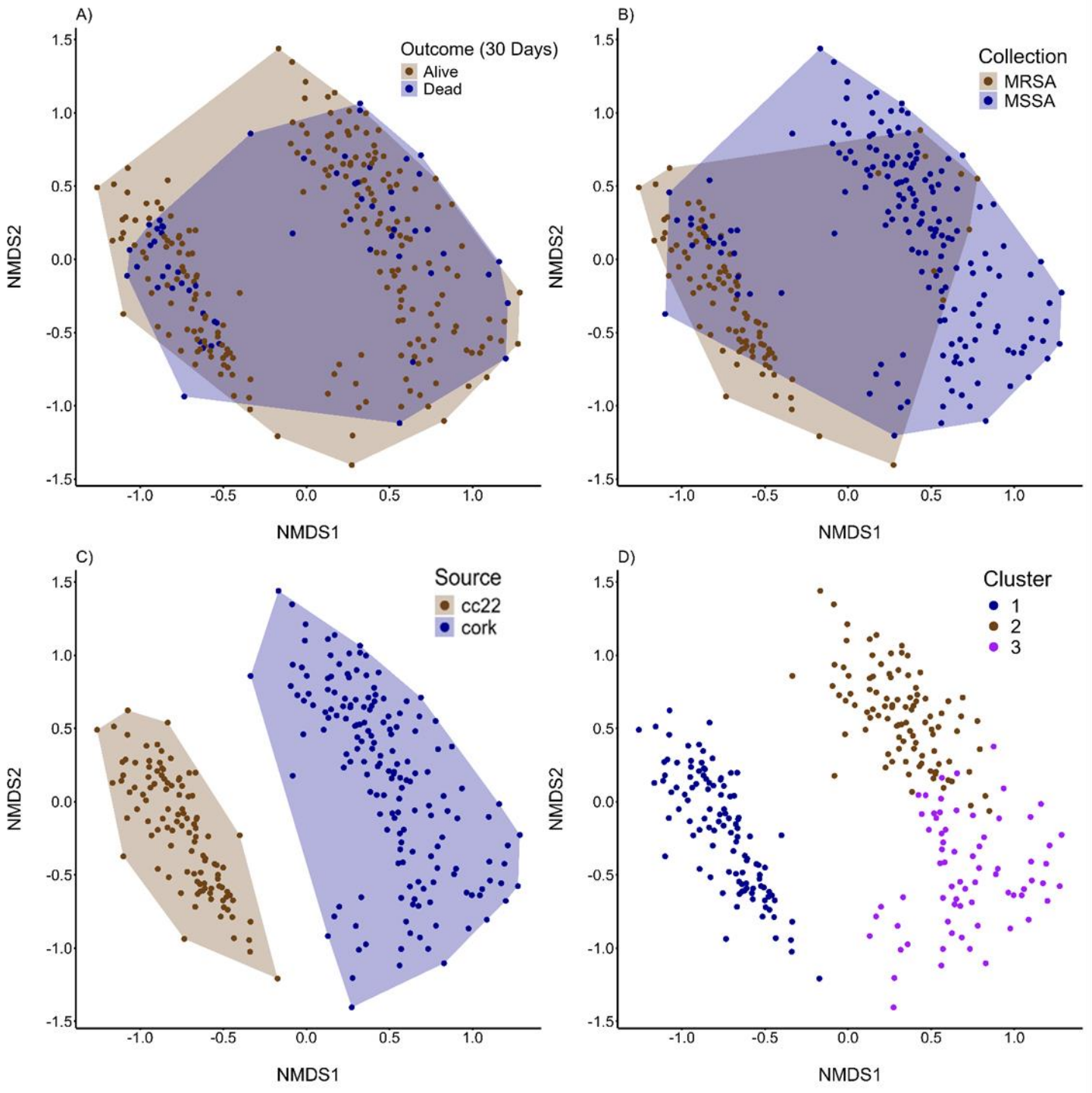


Figure 14. Non-metric multidimensional scaling (NMDS) plots for binarised peak matrix composition, coloured by infection outcome (A), methicillin resistance (B) and data source (C) (n = 278, 2-D stress = 0.15). Facet D represents the 3 clusters identified by hierarchal clustering.

3.4 Case Study 2: Toxicity

3.4.1 Addressing Class Imbalance

The binary target variables, toxicity and 30-day mortality exhibited class imbalances with 17% of the (CC22) isolates having ‘low’ toxicity, and 21% of all infections ending in mortality. Previous studies have sought to tackle such imbalances issue through data augmentation methods such as Synthetic Minority Oversampling Technique (SMOTE). However, López-Cortés *et al.* (2025) found this to reduce their model’s ability to predict MRSA/MSSA by introducing noise and clouding biological patterns. In light of these findings, neither SMOTE nor other resampling techniques were applied. In fact, no method capable of overcoming the bias presented by class imbalance was used. However, instead, stratified cross-validation was applied during modelling infection outcome to at least maintain the original class distribution in each fold.

3.4.2 Random Forest

To identify the peaks associated with bacterial toxicity, a random forest model was created. This framework was chosen since it is applicable for high dimensional data and thus suited the dataset which consisted of 53 features. The data was subsetted to isolate the CC22 data since toxicity data was available for these isolates only (n = 110) (Recker *et al.*, 2017). The model ran 500 trees using a binary target variable for toxicity (high/ low). Importance was inferred through the mean decrease in accuracy metric before the 10 most important peaks were plotted. These peaks were cross referenced with the correlated peak list to understand whether any had an associated peak(s).

3.4.3 Findings

The random forest model revealed three peaks within ± 100 of 3000 m/z to be in the top four predictors of toxicity: 3006, 2977, and 3026 (rounded to the nearest whole number; Fig. 15).

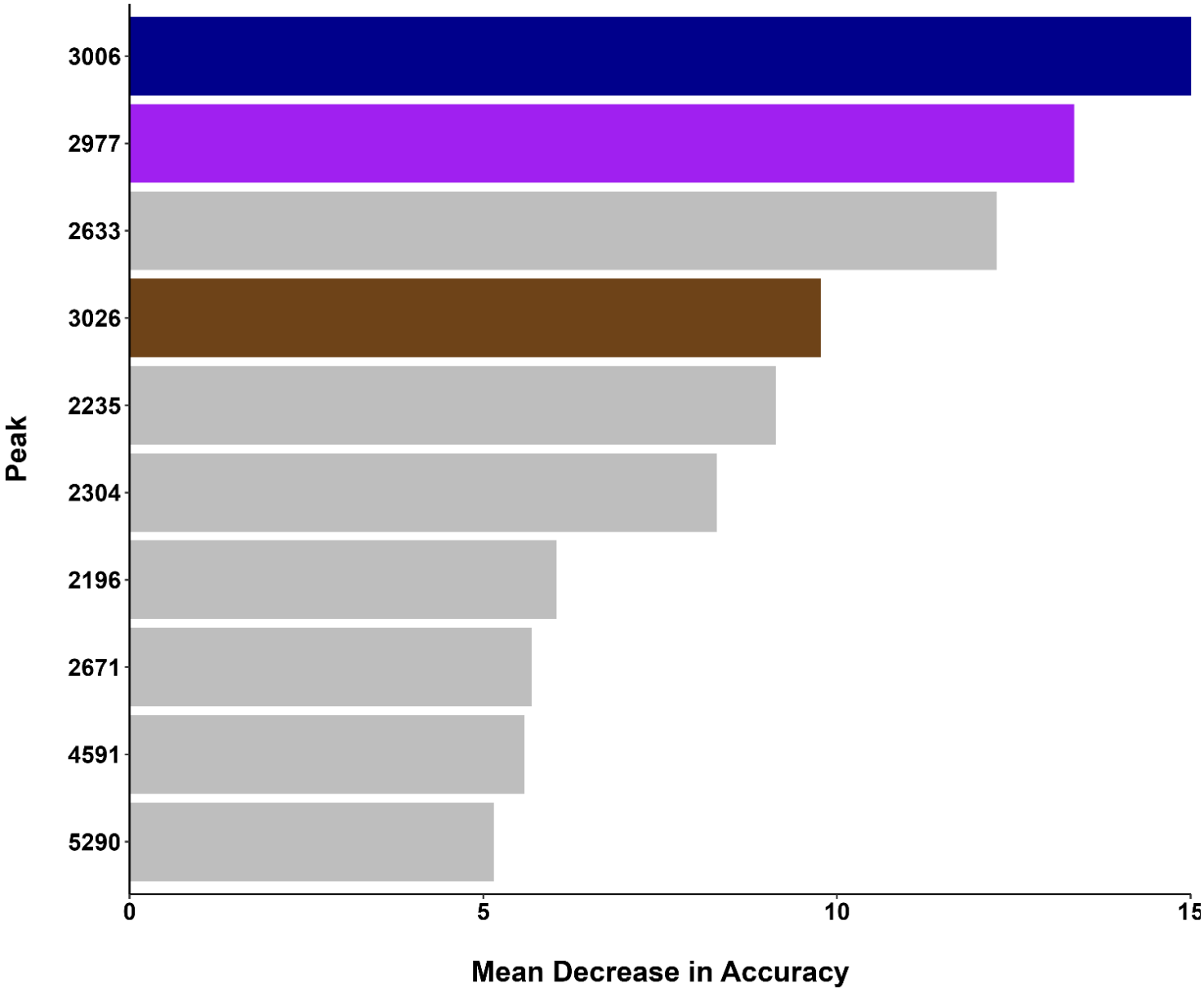


Figure 15. Importance plot for the 10 peaks found by the random forest to be most associated with *S. aureus* toxicity (n = 110). Three of the top four peaks are within ± 100 of 3000 m/z (coloured).

Next, the co-occurrence counts of important peaks within ± 100 of 3000 m/z (the toxicity-associated peak identified by Brignoli *et al.* (2022)) were assessed. This was to evaluate whether these peaks were mutually exclusive, and therefore likely biologically equivalent, or instead truly distinct and evidence of a peak selection pipeline capable of higher resolution. Of the 92 isolates with at least one of these peaks, 77 (84%) had all three peaks, while two (2%) had just one (Fig. 16). Further visual inspections of the the smoothed and raw spectra in this region showed that all three peaks were independent (Fig. 17 & Appendix 5, respectively).

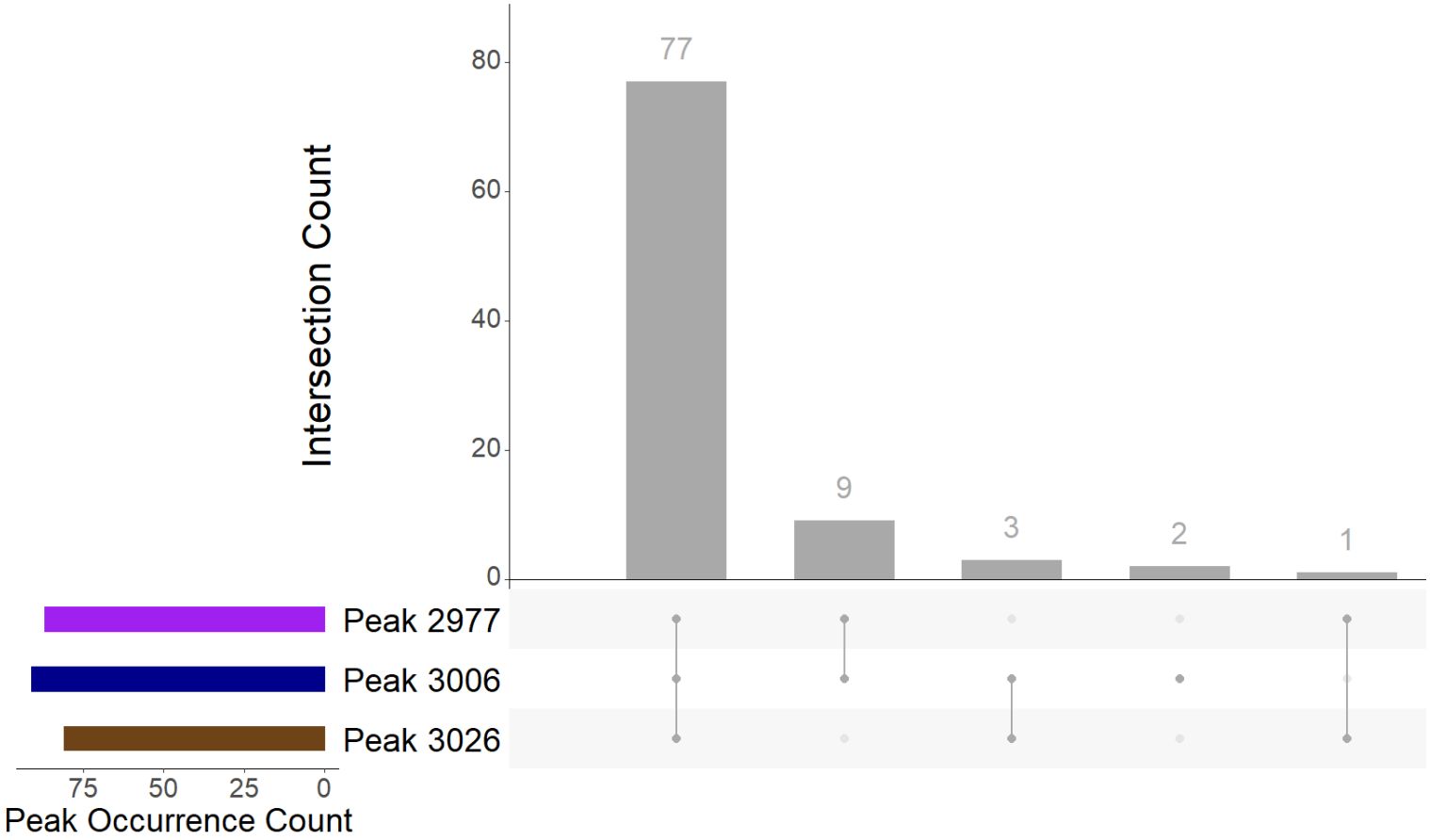


Figure 16. Co-occurrence plot for *S. aureus* spectra containing peaks 2977, 3006 and/ or 3026 (n = 92).

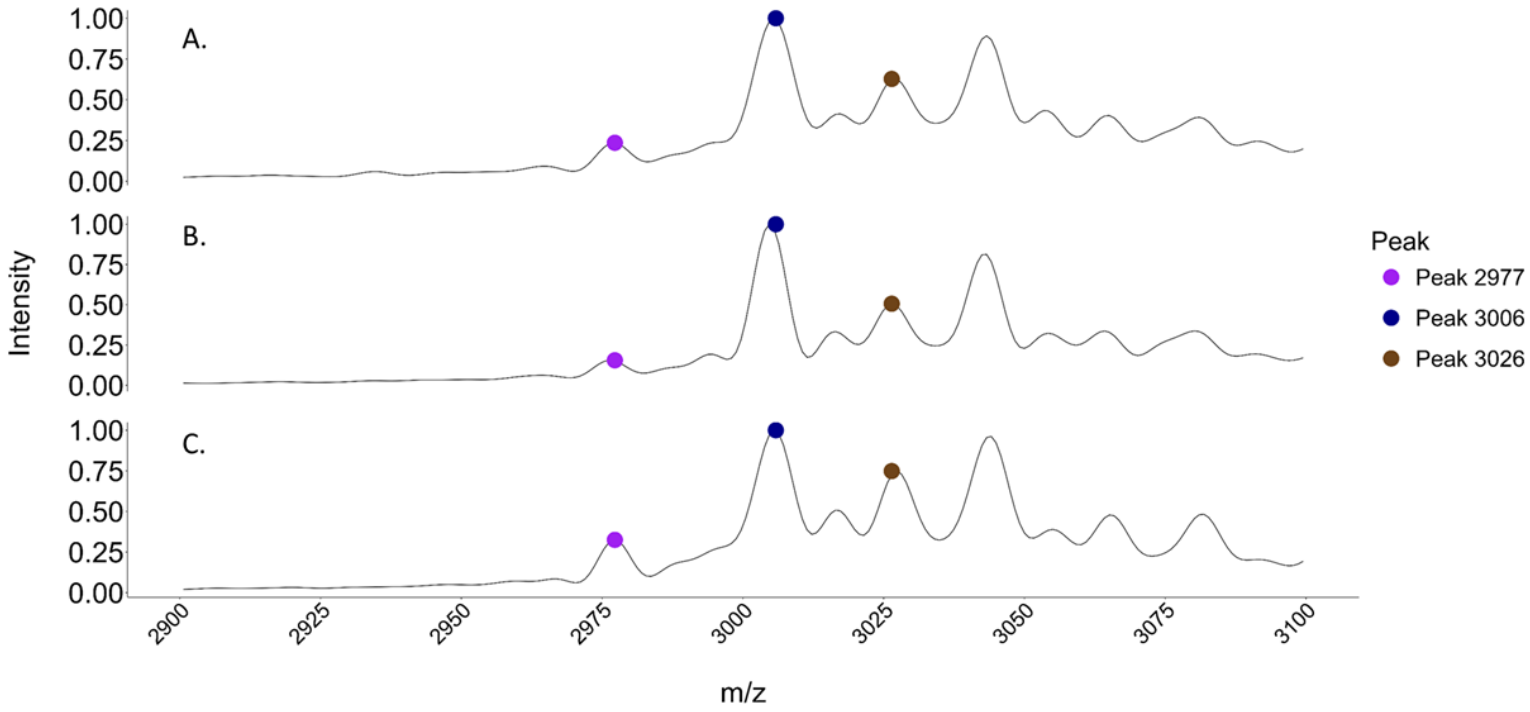


Figure 17. Spectral plots illustrating the prominence of peaks 2977, 3006 and 3026 in three example spectra (smoothed): ASARM102 (A), ASASM12 (B), and ASASM181 (C).

Finally, the toxicity values between each combination of these peaks were compared and statically tested with ANOVA. The different combinations of peaks 3026, 3006 and 2977 had no signifcant impact on toxicity (ANOVA, $F_{4, 87} = 0.15$, $p = 0.96$; see Fig. 18). 100% of the isolates for each combination of peaks had 'high' toxicities, except those where all three peaks were present (97.4%).

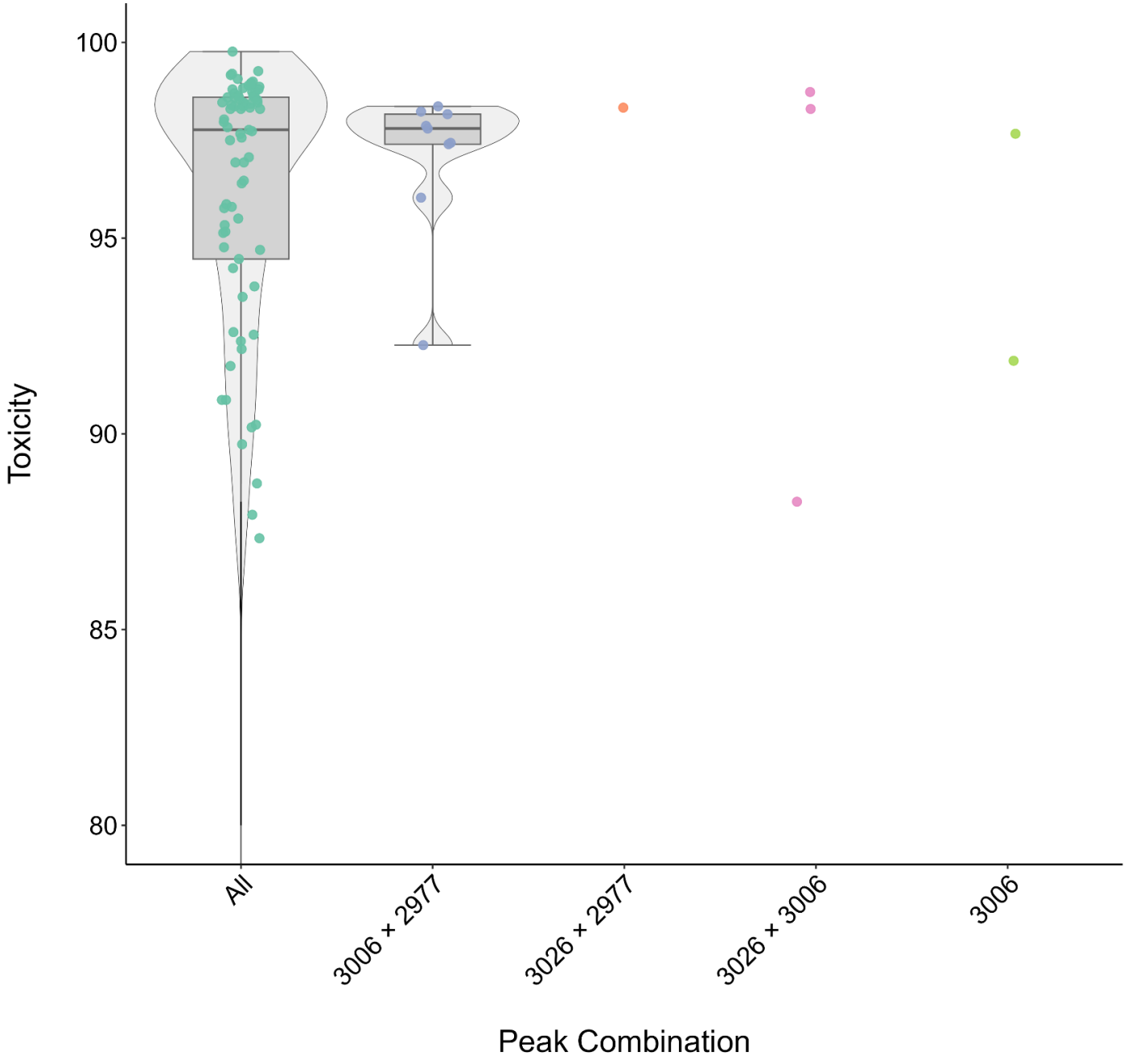


Figure 18. The different combinations of peaks 3026, 3006 and 2977's associated toxicities (n = 92).

3.5 Case Study 3: Infection Outcome

3.5.1 Bayesian Models

A Bayesian framework was chosen to model infection outcome. This was chosen over other methods, such as the aforementioned random forest architecture because it allowed prior distributions for metadata features, reflecting existing knowledge, to be incorporated into more nuanced and holistic models. Furthermore, such models inherently quantify uncertainty through the posterior distributions which is particularly relevant for modelling infection outcome; communicating uncertainty is invaluable for informing effective clinical management.

Four separate Bayesian models were created using the 'brms' package (Bürkner, 2017) to predict 30-day mortality in patients with *S. aureus* bacteraemia (SAB). The first used the clinical metadata that would be expected to be available at the point of care (age and gender) but with uninformative priors. The second model used the same predictors but had informative priors for each of the metadata. Third was a model with the addition of the nine peaks found to be associated with infection outcome (via RFECV), but with uninformative priors. Finally, the fourth model combined the metadata, informative priors, and the nine peaks. These priors were carefully curated using the available literature.

3.5.2 Priors

3.5.2.1 Age

van Hal *et al.* (2012) found that for every 10 years of life, there is a 1.3-fold increase in SAB mortality risk. This equates to a log-odds ratio of ~0.03 per year. Accordingly, 0.03 was used for this prior's coefficient. The corresponding standard deviation was set to 0.01, based on an estimated 95% confidence interval (1.01–1.05 odds ratio) informed by van Hal *et al.*'s summary table. This interval was converted to the log-odds scale (0.00995 – 0.0488) and used to derive a standard deviation estimate using Equation 5.

Equation 5:

$$\sigma = \frac{Upper\ CI\ - Lower\ CI}{2\ \times\ z} = \frac{0.0488 - 0.00995}{2\ \times\ 1.96} \approx 0.01$$

This informative prior reflects the strong tendency towards increased risk with age, and that this is of the strongest and most consistent predictors of SAB mortality (Tacconelli *et al.*, 2006; van Hal *et al.*, 2012).

3.5.2.2 Gender (Female = 0, Male = 1)

Yang et al.'s (2024) multivariate analysis found that being male was associated with lower risk of 30-day mortality, with an odds ratio of 0.26 (95% CI = 0.10 – 0.72), corresponding to a log-odds of approximately −1.35 (95% CI = -0.33 – -2.30). In turn, this value was used as the prior's coefficient. The prior's standard deviation was estimated to be 0.5 using Equation 5.

3.5.2.3 Intercept

Bai *et al.*'s (2022) meta-analysis found that, since 2011, 30-day mortality following SAB was 18.1% (95% CI = 16.3% – 20.0%). This was used to inform the intercept prior i.e. the baseline mortality risk. Converting to a log odds coefficient (logit) scale gave a mean log-odds of -1.51 (95% CI = -1.63 – -1.39). Therefore, the prior's coefficient was calculated to be -1.51 and, using Equation 5, its standard deviation 0.061.

3.5.2.4 Methicillin Resistance

Methicillin resistance (MRSA/MSSA) was not included as a predictor in the Bayesian model. As discussed, although methicillin resistance has been associated with increased mortality in some studies (van Hal *et al.*, 2012), this is contested and is likely to be correlated with the time patients spend in hospital

before the onset of SAB (Wolkewitz *et al.*, 2011; van Hal *et al.*, 2012). Furthermore, while *S. aureus'* methicillin-resistance may be gleaned in clinical settings through one of the recently developed MALDI-TOF MS-based models (Yu *et al.*, 2022), this was not assumed for this model because such tools are not yet integrated into routine clinical workflows.

3.5.2.5 Data Source

Data source (CC22/ Cork) was not used as a predictor in the models. This was because it was because it could not be confidently assumed that data source corresponded to any biological differences, such as clonal complex. Treating data source as a random effect was considered, as it may account for much of the variability in the dataset. However, there was an insufficient number of levels (< 5) to execute this.

3.5.3 Hyperparameters & Convergence Checks

Each Bayesian model ran 2000 iterations including 200 warm-up iterations to allow them to converge at stable and accurate posterior distributions. Stratified 5-fold cross-validation was used rather than 10-fold to reduce computational demand. A feature's effect was deemed to be 'significant' if its credibility intervals did not span 0. All models used a Bernoulli distribution with a logit link to model the binary outcome, with death encoded as the positive class. The convergence of each model was checked by visually assessing each models trace- and posterior distribution-plots, and ensuring that all Rhat values equalled 1. Finally, model performance was compared and evaluated through AUC, F1-scores and balanced accuracy.

3.5.4 Findings

The stratified 5-fold cross-validation Bayesian models predicting 30-day mortality in SAB had a range of performances for the test data. The uninformative and informative prior models, which used only age and gender as predictors, displayed similar performances with AUCs of 0.71 and 0.72, and balanced accuracies of 0.51 (Table 1). The inclusion of the 9 peaks outputted by RFECV

greatly improved model performance, with the model with peaks and uniformative priors reaching an AUC of 0.83 and a balanced accuracy of 0.67. The complete model which incorporated metadata with informative priors and peaks performed similarly (AUC = 0.82, balanced accuracy = 0.68). Furthermore, F1 scores were highest for the models that included peaks (0.50 (uninformative) and 0.51 (informative) compared to 0.09 for the peak-less models), highlighting their improved ability for predicting infection outcome (Table 1).

For the final model (peaks, informative), three peaks were signficanttly associated with a higher 30-day SAB mortality. These were peaks 6943 (β = 1.73, 95% CI = 0.84 to 2.64; see Fig. 19), 6404 (β = 1.50, 95% CI = 0.64 to 2.38; see Fig. 19) and 2304 (β = 1.25, 95% CI = 0.51 to 2.01; see Fig. 19). The latter was also the sixth most important peak for determining toxicity ((CC22 data), Fig. 15). Conversely, two peaks were associated with a negative effect on mortality. These were peaks 2689 mortality (β = -1.35, 95% CI = -2.48 to -0.28; see Fig. 19) and 7422 (β = -1.70, 95% CI = -3.01 to -0.49; see Fig. 19).

For the metadata, age (years) was associated with an increase in the proability of mortality (β = 0.03, 95% CI = 0.02 to 0.05; see Fig. 19) whereas being Male showed the inverse (β = -0.79, 95% CI = -1.31 to -0.05; see Fig. 19). Finally, as expect, the intercept was significantly negative (β = -2.89, 95% CI –4.12 to –1.65; see Fig. 19), in line with the prior distribution corresponding to a low baseline 30-day SAB mortality risk.

All four infection outcome models successfully converged with Rhat values of 1 and trace- and posterior distribution-plots which showed that stable conclusions had been reached (Appendix 6 & 7, respectively).

Table 1. Performance metrics (AUC, F1 score, and balanced accuracy) from stratified 5-fold cross-validation of four Bayesian models predicting 30-day mortality in SAB (n = 278, ± standard deviation).

	Bayesian Model			
Metric	No Peaks, Uninformative	No Peaks, Informative	Peaks, Uninformative	Peaks, Informative
AUC	0.71 ± 0.05	0.72 ± 0.04	0.83 ± 0.04	0.83 ± 0.05
F1 Score	0.09 ± 0.19	0.09 ± 0.19	0.50 ± 0.13	0.51 ± 0.07
Balanced Accuracy	0.51 ± 0.07	0.51 ± 0.07	0.67 ± 0.07	0.68 ± 0.03

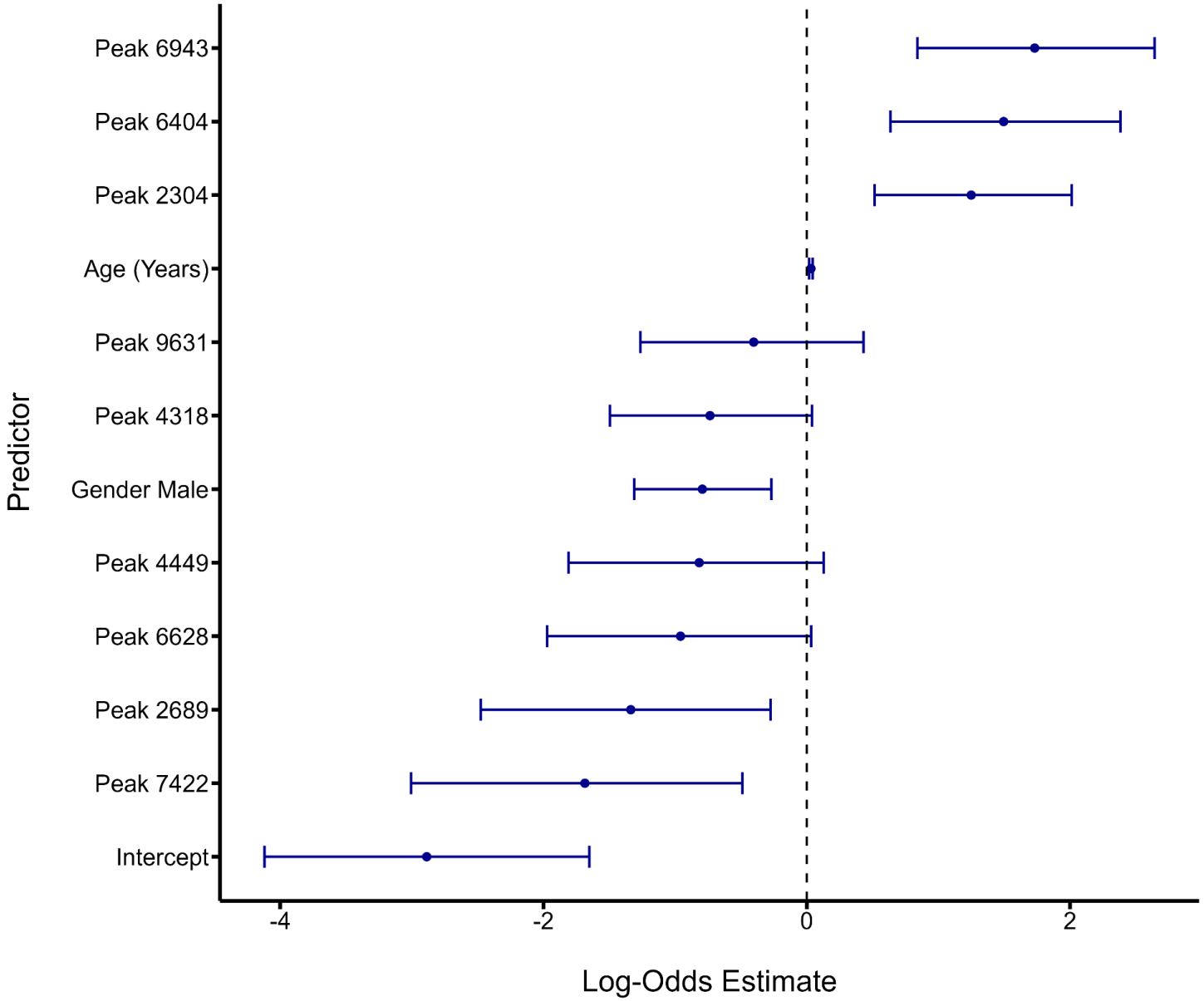


Figure 19. 95% Credibility intervals for the final Bayesian model's coefficients representing the effect on the log-odds of 30-day mortality following SAB (n = 278).

4. Discussion

4.1 Summary

In light of the threats posed by *S. aureus* infections (Ikuta *et al.*, 2022), and the potential of MALDI-TOF MS for navigating these (Haider *et al.*, 2023), this study sought to design robust pre-processing pipelines for extracting reference peaks from MALDI-TOF mass spectra, and for filtering selected peaks for downstream analyses. After evaluating prospective methods for each stage of the pipelines, they were applied to 278 mass spectra stemming from patient isolates with SAB, before the outputs were used to explore the data. First, NMDS and hierarchal clustering gave insights into the drivers of the data's underlying structure. Then, the peaks associated with toxicity were identified with RFECV and a random forest model before three peaks ~3000 m/z were investigated further in terms of their co-occurrence and visual prominence. Finally, peaks associated infection outcome were identified through RFECV and their effects were modelled through Bayesian inference. Three peaks were significantly associated with 30-day mortality: 6943, 6404 and 2304 m/z (Fig. 19).

4.2 Peak Selection Pipeline

When identifying biomarkers from MALDI-TOF mass spectra, pre-processing is unequivocally at the foundation of the workflow. Many pre-processing pipelines have been developed and presented in the literature, each with their own strengths and flaws (Brignoli *et al.*, 2022; Yu *et al.*, 2022). This study successfully constructed a novel pre-processing pipeline robust to the addition of new raw MALDI-TOF spectra by drawing on a number of different methods and combining them sequentially (Fig. 11). While there is no true ground truth for which spectral data is noise versus true signal, each step of the pipeline was well-justified resulting in a novel method for identifying biomarkers from MALDI-TOF MS data to higher resolutions than previous methodologies.

4.2.1 Smoothing

After combining and harmonising the spectra from two different sources, they were interpolated onto a uniform x-axis to allow for convolution-based smoothing (Appendix 1). The MS kernel smoothing technique, adapted from Schmid *et al.*, (2022), appeared to better balance smoothing and peak preservation in comparison to the conventional SG filter (Fig. 2); it limited over-smoothing and retained sharp and high-intensity features likely to reflect meaningful peaks. As the first documented application of MS kernel's to MALDI-TOF data, this offers encouraging evidence that the method may help overcome the insufficient suppression of high-frequency noise and boundary artifacts associated with SG filters which have long been a conventional approach within the field (Schmid *et al.*, 2022).

4.2.2 Baseline Correction

The next step of the pipeline was baseline correction. Comparison between the prospective methods, TopHat and SNIP, revealed the former to be more aggressive than the latter (Fig. 4). It was inferred that this could result in signal underestimates leading to impeded reliability of downstream results. Furthermore, given that the subsequent peak selection method was based on peak signal relative to its own spectra, there was no issue in having different baseline morphologies between spectra such as the curve seen in Fig. 4. That said, the peak selection method entailed omitting peaks below a given relative intensity which could mean that spectra with curved baselines are more likely to output irrelevant peaks than those from spectra with baselines that run along relative intensity 0. Additionally, differences in baseline curves would be problematic if the pipeline were to use absolute intensities rather than presence/ absence, as will be discussed. Regardless, the comparison facilitated the well-informed decision to proceed with the more nuanced and iterative SNIP method.

4.2.3 Peak Selection & Alignment

Peak selection and alignment were the final steps in the peak selection pipeline and were key for producing the high resolution peaks seen in Fig. 11.

Implementing peak selection based on peaks' prominence relative to the surrounding spectral data was a well-suited approach for this pipeline. Namely, it complemented the SNIP baseline correction because the bias introduced by differing baseline morphology on peak selection was nullified i.e. higher baseline intensities did not inherently mean a higher number of peaks would be selected. This is only the case since the pipeline sought to identify peak presence, not absolute intensity values. The prominence value of 0.2 was informed by comparing manually selected values effect on peak selection to discern the threshold that best balanced the trade-off between to best balance the inclusion of relevant peaks and the exclusion of those induced by noise (Fig. 6). This also contributed to computational efficiency and limited overfitting of downstream analysis caused by excessive dimensionality. In this study, other peak selection methods such as CWT were not explored. However, CWT has been quoted as a nuanced method since it accounts for spectral morphology in addition to prominence (Du *et al.*, 2006). Therefore, a future study may seek to apply CWT and compare its findings against those yielded by this pipeline.

Selected peaks require alignment since their m/z vary slightly between samples (Tibshirani *et al.*, 2004). On one hand, peak alignment methods that group peaks over large m/z ranges, such as the 100 m/z value used by Brignoli *et al.* (2022), help evade the risk of treating biologically equivalent peaks at independent (unless boundary edge effects split equivalent peaks into two bins). On the other hand, they inflate the risk of losing valuable distinctions between peaks upon their merger. Therefore, an effective pre-processing pipeline should balance and evaluate this trade-off. Again, while no ground truth exists for the spectra's true signal versus noise, the peak selection and alignment methods used in this study were well justified and key for the discovery of important peaks. By comparing the effect of different tolerance thresholds on peak alignment (Fig. 9 & 10), it was inferred that a tolerance of 5 was best suited to meeting this trade off. This value appeared to be capable of grouping peaks with marginally different m/z values (likely equivalent), and of preserving more disjoint peaks. That said, the reference peak bin values were determined by the mass spectra used in the pipeline. This is perhaps the pipeline's primary limitation in regards to applicability; newly identified peaks cannot be directly compared to a reference standard upon the addition of new spectral data.

4.2.4 Limitations & Further Directions

Peak alignment bins were formed by sorting all peaks in order, grouping those within a 'tolerance' value of each other (5 m/z), and calculating a mean for each bin (Fig. 7). Therefore, while the pipeline is robust to the addition of new data, the exact peak alignment bin values are a function of the input data. The consequence of this is that if a new set of isolates were to be used in the pipeline, they would create slightly different reference peaks. In practise, this would still be useful for identifying which peaks are present. However, it would not be possible to automate the identification of peaks against those in a clinical database. Thus, this step should be revised so that additional data can be compared to a set reference peaks created beforehand. This could be done by first generating a stable reference peak list (e.g. from a large representative dataset), which is then fixed and used as a reference for all future incoming spectra. This would mean that instead of recalculating bins for every new batch of data, peaks would be matched to the closest reference bin within a set tolerance, to allow consistent downstream interpretation and comparison.

A second topic of research that should be pursued following this study is the development of a pipeline that accounts for peak intensity. This study is limited to peak presence/ absence. While this undoubtedly important for understanding pathogen virulence and phenotype, peak intensity is also relevant. For example subtle differences in intensity values are known to reflect whether fungal- (Aguiar *et al.*, 2021) and non-*S. aureus* bacterial-species (Hadpanus *et al.*, 2019) can form biofilms. In turn, the creation of a pipeline capable of quantifying peak intensity would allow for higher resolution pathogen characterisation. One aspect of current pipeline that would be revised is the choice of baseline correction. The SNIP method produced some curved spectra but this did not impede the collection of binary peak data (Fig. 4). However, it could be insufficiently aggressive for gathering continuous intensity data since its results in differing baseline morphology between spectra. In this case, a more aggressive method such as TopHat would likely be more applicable (Fig. 4).

4.3 Dimensionality Reduction Pipeline

4.3.1 Frequency Filtering

Following the application of the first pipeline, a pre-requisite for dimensionality reduction, the second pipeline was formed and evaluated. High dimensionality presents an ongoing challenge for the analysis of MALDI-TOF MS data (Manchanda *et al.*, 2018); the first pipeline resulted in 239 reference peaks. The first method used to reduce dimensionality was frequency filtering (Fig. 12). This is an established and logical approach, especially with the goal of identifying important peaks in mind Tibshirani *et al.* (2004). However, it is possible that rare, but biologically significant, peaks may be omitted from analysis as a result of this filtering. For example, such peaks could be associated with the emergence of resistance to various anti-biotics such as vancomycin. While resistance is uncommon (McGuinness *et al.*, 2017), evidence for potentially responsible biomarkers is ambiguous (Drummelsmith *et al.*, 2007). Additionally, the frequency threshold (0.15) was based on literature (Manchanda *et al.*, 2018) and by considering the number of isolates available for analysis. In future, the pipeline could benefit from a more quantitative method for determining the optimal frequency filtering threshold, but this should also be independent of input data to ensure robustness to the addition of new data.

4.3.2 Correlated Peak Filtering

The Pearson correlation coefficient filtering was necessary for removing multi-collinear features and for preventing the inflation of the perceived feature importance. However, as will be discussed, it fails to account for non-linear multi-feature dependencies.

4.3.3 RFECV

The use of RFECV is an intuitive way of determining the most important features for modelling since the least informative peaks are iteratively omitted until an optimal AUC is reached. The RFECV was also beneficial for subsequent modelling as it, for example, encouraged downstream Bayesian model convergence (Appendix 6 & 7). However, the RFECV model was still prone to

overfitting given the number of input features (151). It also contributed to a key limitation of the dimensionality reduction pipeline; the full complexity of the relationships between biomarkers was not fully accounted for.

4.3.4 Limitations and Future Directions

The primary shortcoming of the dimensionality reduction pipeline is that it does not account for non-linear relationships between peaks. For example, the Pearson coefficient filtering is based on linear relationships and is pairwise meaning that multi-feature dependencies could be overlooked. Similarly, the logistic regression chosen as the basis for RFECV does not automatically account for interaction effects, despite their existence in such contexts (Howden *et al.*, 2023).

The RFECV fitted a logistic regression which is prone to overfitting, particularly for imbalanced datasets (Hawkins, 2004). While the cross-validation preserves class imbalances, it does not mitigate the potential for overfitting and bias towards the majority class. This was evidenced by the low F1 scores for the peak-less models (Table 1). Other methods face similar issues; the HDSS problem is an ongoing issue in the field that is pending further investigation. In turn, a future study should seek to compare the effect of different models in this part of the pipeline to evaluate which architectures offer the more reliable results. Alternatively, a mixture of over- and under-sampling or methods for replacing correlation filtering and RFECV could be explored to streamline dimensionality reduction and account for complex relationships.

4.4 Case Studies

4.4.1 Clustering

The NMDS and hierarchal clustering gave valuable insights into peak profile community composition. Firstly, the NMDS plots made it evident that the data-source (CC22/ Cork) was a key driver of the data's structure (Fig 14c). On one hand, these results could be as a result of true biological variation. While the CC22 data was known to come from a single clonal background, this information was unknown for the Cork data and so this may have driven the groupings. If

true, it also highlights the merit of the peak selection pipeline for revealing important biomarkers that can be used to understand immediate *S. aureus* infections. Moreover, the three clusters revealed by the hierarchal clustering further evidence this. Two clusters formed amongst the cork data (Fig. 14d suggesting that the groups may reflect two different *S. aureus* clonal backgrounds. While this cannot be currently verified, it does highlight the potential of clonal backgrounds for driving phenotypic variation, which could render them important predictors of infection outcome or informative for treatment. This is in alignment with the findings of Recker *et al.,* (2017) who show that *S. aureus* virulence and pathogenicity profiles significantly vary between clonal backgrounds.

On the other hand, the NMDS findings could reflect how deep-rooted technical variability presents a real challenge for pipelines and models attempting to use data from different sources. This is because mass spectra are largely a function of sample preparation which can also present a challenge for reproducibility. For example, the type of matrix used for spotting underpins the desorption/ionisation process, and yet matrices differ between sources, as does the frequency of laser pulsing which also drives desorption/ ionisation (Boskamp *et al.*, (2021). Consequently, there should be continued efforts to minimise potential batch effects and discern noise from true biological signal. Characterising different clonal backgrounds in laboratory settings to determine their virulence profiles would provide more evidence surrounding whether these clustering patterns are as a result of noise or true signal.

4.4.2 Toxicity

The study revealed a number of peaks associated with high toxicity in *S. aureus.* Interestingly, the random forest found multiple peaks to be within ±100 of 3000 (Fig. 15). The importance of a peak in this region had already been identified by Brignoli *et al.*, (2022). Not only are this study's findings consistent with theirs, but they also provide evidence for newly identified biomarkers important for toxicity. This also evidences the pipelines' biomarker identification capabilities; the increased number of peaks reflect the pipelines high resolution discoveries.

These could be particularly valuable for future research and for informing management in clinical scenarios.

While there is no ground truth as to which spectral peaks are noise versus true signal, visual scrutiny of the peaks overlaid on the spectra showed that the three most important peaks around 3000 m/z (2977, 3006 and 3026 m/z) did indeed appear to be distinct (Fig. 17). Furthermore, their lack of mutually exclusivity (Fig. 16) implies that they are likely caused by different compounds as opposed to a single biomolecule. Despite the peak combinations' corresponding high toxicities (Fig. 18), small sample sizes for combinations where at least one of the three peaks were not present (n = 1 - 9) means that any effect that different combinations may have on toxicity remain unproven (ANOVA, $F_{4, 87} = 0.15$, $p = 0.96$; see Fig. 18).

Each peak's causal biomolecule is also unknown and should be characterised in a laboratory. For example, it could that the peaks are caused by similar proteins with various mutations. It is known that a variant of the delta toxin at 3000 m/z instead occurs at a 3035 m/z (Gagnaire *et al.*, 2012). Equally, smaller peaks around a more prominent peak may be caused by polypeptide adducts bound to the matrix (Du *et al.*, 2006). Given that the majority of the CC22 samples with a peak around 3000 m/z had multiple closely positioned peaks, it is plausible that these represent polypeptide adducts or related forms of the same protein, rather than entirely distinct biomolecules.

Whether peaks around 3000 m/z confer toxicity in *S. aureus* isolates from other clonal backgrounds, such as those of the Cork dataset, remains unknown. However, such knowledge will be important for developing a comprehensive understanding of peak importance in clinical contexts. It is possible that in other clonal backgrounds, different peaks may be more associated with toxicity e.g. they could produce an entirely different toxin with a m/z that reflects this. Ultimately, understanding this could help guide future research into potential biomarkers until they are fully characterised at which point clinicians can make them a point of interest when characterising future SAB infections.

4.4.3 Infection Outcome

Timely treatment is paramount for heightening patient survival prospects following SAB (Corl *et al.*, 2020). In turn, a model that can effectively predict infection outcome is an invaluable facet of diagnostic pipelines. Four Bayesian models were created to predict 30-day SAB infection outcome with 278 data points. The models that incorporated spectral peaks, identified through the successive application of the pre-processing pipelines, had better predictive abilities that those did not.

The model that used peaks and uninformative priors had an AUC of 0.83 which was 0.12 higher than its peak-less counterpart (0.71) (Table 1). The model with peaks and uninformative priors also had a far superior F1 score (0.50) in comparison to the peak-less model (0.09) (Table 1). This exhibits yet another piece of evidence that study's pipelines are capable of facilitating the discovery of new and important biomarkers. This was confounded by the posterior distributions and CIs of multiple features. Three peaks were significantly associated with an increased risk of 30-day mortality: 6943, 6404 and 2304 m/z (CIs did not span 0; see (Fig. 19)). Interestingly, the peak at 2304 m/z was also one of the most important predictors of toxicity (Fig. 15). Conversely, two peaks were associated with a decreased risk: 2689 and 7422 m/z (CIs did not span 0; see Fig. 19). While unrelated to the pipeline, a future study should characterise these biomolecules to help understand why they may be associated with their respective effects on mortality.

The results also emphasise that the Bayesian inference is a particularly exciting model architecture for being integrated into clinical workflows. Importantly, Bayesian models account for uncertainty through posterior distributions; this is an important aspect of any model with prospects of being deployed in clinical settings (Goligher *et al.*, 2024). Furthermore, the inclusion of priors offers a more holistic approach for informing predictions (Consonni *et al.*, 2018); the priors used in the final model were well justified using a range of literature. Conflicting results between existing studies surrounding particular priors were also taken into consideration upon deciding the error to be associated with them. That said, priors are inherently subjective and should be continually scrutinised and updated to ensure that they remain well informed, in line with the changing beliefs around them as more research is published. While the priors used in this study made

little impact on the predictions made by model (Table 1), this is unsurprising given that the metadata priors were associated with relatively large errors.

The complete models' performance would have likely been stronger if the CCI data had been available for all patient isolates, since it has such a large bearing on infection outcome (Charlson *et al.*, 2022; Russell *et al.*, 2024). Similarly, it should be noted the study assumes that all isolates stem from infections where *S. aureus* occurred in isolation. However, this assumption may have significant implications for model outcomes where misinformed. Adalbert *et al.* (2021) observed death in 61.7% of patients co-infected with *S. aureus* and COVID-19. While the data in this study was collected pre-pandemic, various co-infections are still likely to be a source of underlying variation in the data, especially since the patients were mostly elderly and had a range of comorbidities. Given that co-infections often increase patient mortality (Limoli *et al.*, 2016; Liu *et al.*, 2021), future research around optimising diagnostic pipelines should consider the effect of co-infection on infection outcome and patient treatment requirements.

5. Conclusion

This study designed and applied two distinct analytical pipelines to MALDI-TOF spectral data from SAB isolates, before exploring their utility in three case studies. The pipelines, as the first of their kind, illuminated several valuable methodological insights. Their capacity to facilitate downstream machine learning models was further evidenced by the identification of clinically relevant *S. aureus* toxicity and SAB mortality biomarkers. The toxicity case study found several peaks around 3000 m/z to be associated with high toxicity, while Bayesian inference exposed three peaks significantly associated with an increased risk of 30-day SAB mortality, namely 2304, 6404, and 6943 m/z. These peaks enhanced the models' ability to distinguish between survival and mortality outcomes following SAB.

This cemented the success of the peak selection pipeline which combined MS kernel smoothing, SNIP baseline correction, prominence-based peak selection, and tolerance-based alignment to produce a clear, binary peak matrix. Despite

the resultant alignment bins being data-dependent, the pipelines were robust in identifying high resolution, biologically meaningful signal amongst complex spectral noise. Not only do these results reinforce the power of MALDI-TOF data beyond species identification and serve as a foundation for future MALDI-TOF studies. They also set a precedent for the shifting mindset focused on the continued innovation and prioritisation of robust signal extraction, clarity of interpretation, and predictive performance in real-world contexts. Ultimately, MALDI-TOF MS-based diagnostics continue to demonstrate their potential for supporting clinical management, which with continued research, promises to alleviate pathogens' global threat.

6. References

Abdollahi, A. *et al.* (2024) 'Mortality patterns in patients with Staphylococcus aureus bacteremia during the COVID-19 pandemic: Predictors and insights', *Heliyon*, 10(2), p. e24511. Available at: https://doi.org/10.1016/j.heliyon.2024.e24511.

Abraham, L. and Bamberger, D.M. (2020) 'Staphylococcus aureus Bacteremia: Contemporary Management', *Missouri Medicine*, 117(4), pp. 341–345.

Adalbert, J.R. *et al.* (2021) 'Clinical outcomes in patients co-infected with COVID-19 and Staphylococcus aureus: a scoping review', *BMC Infectious Diseases*, 21(1), p. 985. Available at: https://doi.org/10.1186/s12879-021-06616-4.

Aebersold, R. and Mann, M. (2003) 'Mass spectrometry-based proteomics', *Nature*, 422(6928), pp. 198–207. Available at: https://doi.org/10.1038/nature01511.

Aguiar, P.A.D.F. *et al.* (2021) 'Rapid detection of biofilm-producing Candida species via MALDI-TOF mass spectrometry', *Journal of Applied Microbiology*, 131(4), pp. 2049–2060. Available at: https://doi.org/10.1111/jam.15066.

Alexandrov, T. *et al.* (2009) 'Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation', *Bioinformatics*, 25(5), pp. 643–649. Available at: https://doi.org/10.1093/bioinformatics/btn662.

Arenas, M. *et al.* (2018) 'Mutation and recombination in pathogen evolution: Relevance, methods and controversies', *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 63, pp. 295–306. Available at: https://doi.org/10.1016/j.meegid.2017.09.029.

Bai, A.D. *et al.* (2022) 'Staphylococcus aureus bacteraemia mortality: a systematic review and meta-analysis', *Clinical Microbiology and Infection*, 28(8), pp. 1076–1084. Available at: https://doi.org/10.1016/j.cmi.2022.03.015.

Banin, E., Hughes, D. and Kuipers, O.P. (2017) 'Editorial: Bacterial pathogens, antibiotics and antibiotic resistance', *FEMS Microbiology Reviews*, 41(3), pp. 450–452. Available at: https://doi.org/10.1093/femsre/fux016.

Bauer, C., Cramer, R. and Schuchhardt, J. (2011) 'Evaluation of Peak-Picking Algorithms for Protein Mass Spectrometry', in M. Hamacher, M. Eisenacher, and C. Stephan (eds) *Data Mining in Proteomics: From Standards to Applications*. Totowa, NJ: Humana Press, pp. 341–352. Available at: https://doi.org/10.1007/978-1-60761-987-1_22.

Boskamp, T. *et al.* (2021) 'Cross-Normalization of MALDI Mass Spectrometry Imaging Data Improves Site-to-Site Reproducibility', *Analytical Chemistry*, 93(30), pp. 10584–10592. Available at: https://doi.org/10.1021/acs.analchem.1c01792.

Brignoli, T. *et al.* (2022) 'Diagnostic MALDI-TOF MS can differentiate between high and low toxic Staphylococcus aureus bacteraemia isolates as a predictor of patient outcome', *Microbiology (Reading, England)*, 168(8), p. 001223. Available at: https://doi.org/10.1099/mic.0.001223.

Bürkner, P.-C. (2017) 'brms: An R Package for Bayesian Multilevel Models Using Stan', *Journal of Statistical Software*, 80, pp. 1–28. Available at: https://doi.org/10.18637/jss.v080.i01.

Buszewski, B. *et al.* (2021) 'A new approach to identifying pathogens, with particular regard to viruses, based on capillary electrophoresis and other analytical techniques', *TrAC Trends in Analytical Chemistry*, 139, p. 116250. Available at: https://doi.org/10.1016/j.trac.2021.116250.

Calderaro, A. and Chezzi, C. (2024) 'MALDI-TOF MS: A Reliable Tool in the Real Life of the Clinical Microbiology Laboratory', *Microorganisms*, 12(2), p. 322. Available at: https://doi.org/10.3390/microorganisms12020322.

Candela, A. *et al.* (2022) 'Rapid and Reproducible MALDI-TOF-Based Method for the Detection of Vancomycin-Resistant Enterococcus faecium Using Classifying Algorithms', *Diagnostics*, 12(2), p. 328. Available at: https://doi.org/10.3390/diagnostics12020328.

Charlson, M.E. *et al.* (2022) 'Charlson Comorbidity Index: A Critical Review of Clinimetric Properties', *Psychotherapy and Psychosomatics*, 91(1), pp. 8–35. Available at: https://doi.org/10.1159/000521288.

Cheng, K. *et al.* (2016) 'Recent development of mass spectrometry and proteomics applications in identification and typing of bacteria', *Proteomics. Clinical Applications*, 10(4), pp. 346–357. Available at: https://doi.org/10.1002/prca.201500086.

Coll, F. *et al.* (2025) 'The mutational landscape of Staphylococcus aureus during colonisation', *Nature Communications*, 16(1), p. 302. Available at: https://doi.org/10.1038/s41467-024-55186-x.

Conrad, T.O.F. *et al.* (2006) 'Beating the Noise: New Statistical Methods for Detecting Signals in MALDI-TOF Spectra Below Noise Level', in M. R. Berthold, R.C. Glen, and I. Fischer (eds) *Computational Life Sciences II*. Berlin, Heidelberg: Springer, pp. 119–128. Available at: https://doi.org/10.1007/11875741_12.

Consonni, G. *et al.* (2018) 'Prior Distributions for Objective Bayesian Analysis', *Bayesian Analysis*, 13(2), pp. 627–679. Available at: https://doi.org/10.1214/18-BA1103.

Coombes, K.R. *et al.* (2005) 'Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform', *PROTEOMICS*, 5(16), pp. 4107–4117. Available at: https://doi.org/10.1002/pmic.200401261.

Corl, K.A. *et al.* (2020) 'Delay in Antibiotic Administration Is Associated With Mortality Among Septic Shock Patients With Staphylococcus aureus Bacteremia*', *Critical Care Medicine*, 48(4), p. 525. Available at: https://doi.org/10.1097/CCM.0000000000004212.

Culebras, D.E. (2018) 'Chapter Fifteen - Application of MALDI-TOF MS in Bacterial Strain Typing and Taxonomy', in F. Cobo (ed.) *The Use of Mass Spectrometry Technology (MALDI-TOF) in Clinical Microbiology*. Academic Press, pp. 213–233. Available at: https://doi.org/10.1016/B978-0-12-814451-0.00015-0.

David, M.Z. and Daum, R.S. (2017) 'Treatment of Staphylococcus aureus Infections', in *Staphylococcus aureus*. Springer, Cham, pp. 325–383. Available at: https://doi.org/10.1007/82_2017_42.

Deininger, S.-O. *et al.* (2011) 'Normalization in MALDI-TOF imaging datasets of proteins: practical considerations', *Analytical and Bioanalytical Chemistry*, 401(1), pp. 167–181. Available at: https://doi.org/10.1007/s00216-011-4929-z.

Deng, F. *et al.* (2021) 'An improved peak detection algorithm in mass spectra combining wavelet transform and image segmentation', *International Journal of Mass Spectrometry*, 465, p. 116601. Available at: https://doi.org/10.1016/j.ijms.2021.116601.

Dixon, P. (2003) 'VEGAN, a package of R functions for community ecology', *Journal of Vegetation Science*, 14(6), pp. 927–930. Available at: https://doi.org/10.1111/j.1654-1103.2003.tb02228.x.

Doern, C.D. and Butler-Wu, S.M. (2016) 'Emerging and Future Applications of Matrix-Assisted Laser Desorption Ionization Time-of-Flight (MALDI-TOF) Mass Spectrometry in the Clinical Microbiology Laboratory: A Report of the Association for Molecular Pathology', *The Journal of molecular diagnostics: JMD*, 18(6), pp. 789–802. Available at: https://doi.org/10.1016/j.jmoldx.2016.07.007.

Drummelsmith, J. *et al.* (2007) 'Comparative Proteomics Analyses Reveal a Potential Biomarker for the Detection of Vancomycin-Intermediate

Staphylococcus aureus Strains', *Journal of Proteome Research*, 6(12), pp. 4690–4702. Available at: https://doi.org/10.1021/pr070521m.

Du, P., Kibbe, W.A. and Lin, S.M. (2006) 'Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching', *Bioinformatics*, 22(17), pp. 2059–2065. Available at: https://doi.org/10.1093/bioinformatics/btl355.

El-Bouri, K. *et al.* (2012) 'Comparison of bacterial identification by MALDI-TOF mass spectrometry and conventional diagnostic microbiology methods: agreement, speed and cost implications', *British Journal of Biomedical Science*, 69(2), pp. 47–55. Available at: https://doi.org/10.1080/09674845.2012.12002436.

Embong, Z. *et al.* (2008) 'Specific detection of fungal pathogens by 18S rRNA gene PCR in microbial keratitis', *BMC Ophthalmology*, 8(1), p. 7. Available at: https://doi.org/10.1186/1471-2415-8-7.

Flores-Flores, A.S. *et al.* (2025) 'MALDI-TOF MS profiling to predict resistance or biofilm production in gram-positive ESKAPE pathogens from healthcare-associated infections', *Diagnostic Microbiology and Infectious Disease*, 111(1), p. 116562. Available at: https://doi.org/10.1016/j.diagmicrobio.2024.116562.

Flores-Treviño, S. *et al.* (2019) 'Screening of biomarkers of drug resistance or virulence in ESCAPE pathogens by MALDI-TOF mass spectrometry', *Scientific Reports*, 9, p. 18945. Available at: https://doi.org/10.1038/s41598-019-55430-1.

Fowler, V.G., Jr *et al.* (2003) 'Clinical Identifiers of Complicated Staphylococcus aureus Bacteremia', *Archives of Internal Medicine*, 163(17), pp. 2066–2072. Available at: https://doi.org/10.1001/archinte.163.17.2066.

Frye, A.M. *et al.* (2020) 'Clinical Impact of a Real-Time PCR Assay for Rapid Identification of Staphylococcal Bacteremia', *Journal of Clinical Microbiology*, 50(1), pp. 127–133. Available at: https://doi.org/10.1128/jcm.06169-11.

Gagnaire, J., Dauwalder, O., Boisset, S., Khau, D., Freydière, A.-M., Ader, F., Bes, M., Lina, G., Tristan, A., Reverdy, M.-E., Marchand, A., Geissmann, T., Benito, Y., Durand, G., Charrier, J.-P., Etienne, J., Welker, M., Belkum, A.V., *et al.* (2012) 'Detection of Staphylococcus aureus Delta-Toxin Production by Whole-Cell MALDI-TOF Mass Spectrometry', *PLOS ONE*, 7(7), p. e40660. Available at: https://doi.org/10.1371/journal.pone.0040660.

Gajdács, M. (2019) 'The Continuing Threat of Methicillin-Resistant Staphylococcus aureus', *Antibiotics*, 8(2), p. 52. Available at: https://doi.org/10.3390/antibiotics8020052.

Garbacz, K. *et al.* (2021) 'Distribution and antibiotic-resistance of different Staphylococcus species identified by matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) isolated from the oral cavity', *Journal of Oral Microbiology*, 13(1), p. 1983322. Available at: https://doi.org/10.1080/20002297.2021.1983322.

García Fenoll, R. *et al.* (2022) '[Clinical characteristics and prognosis of Staphylococcus aureus bacteremia]', *Revista Espanola De Quimioterapia:*

Publicacion Oficial De La Sociedad Espanola De Quimioterapia, 35(6), pp. 544–550. Available at: https://doi.org/10.37201/req/035.2022.

Garg, E. and Zubair, M. (2025) 'Mass Spectrometer', in *StatPearls*. Treasure Island (FL): StatPearls Publishing. Available at: http://www.ncbi.nlm.nih.gov/books/NBK589702/ (Accessed: 23 January 2025).

Gautier, M. *et al.* (2014) 'Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: revolutionizing clinical laboratory diagnosis of mould infections', *Clinical Microbiology and Infection*, 20(12), pp. 1366–1371. Available at: https://doi.org/10.1111/1469-0691.12750.

Gibb, S. and Strimmer, K. (2012) 'MALDIquant: a versatile R package for the analysis of mass spectrometry data', *Bioinformatics*, 28(17), pp. 2270–2271. Available at: https://doi.org/10.1093/bioinformatics/bts447.

Golding, C.G. *et al.* (2016) 'The scanning electron microscope in microbiology and diagnosis of infectious disease', *Scientific Reports*, 6(1), p. 26516. Available at: https://doi.org/10.1038/srep26516.

Goligher, E.C., Heath, A. and Harhay, M.O. (2024) 'Bayesian statistics for clinical research', *The Lancet*, 404(10457), pp. 1067–1076. Available at: https://doi.org/10.1016/S0140-6736(24)01295-9.

Greco, V. *et al.* (2018) 'Applications of MALDI-TOF mass spectrometry in clinical proteomics', *Expert Review of Proteomics*, 15(8), pp. 683–696. Available at: https://doi.org/10.1080/14789450.2018.1505510.

Grenga, L., Pible, O. and Armengaud, J. (2019) 'Pathogen proteotyping: A rapidly developing application of mass spectrometry to address clinical concerns', *Clinical Mass Spectrometry*, 14, pp. 9–17. Available at: https://doi.org/10.1016/j.clinms.2019.04.004.

Gu, W., Miller, S. and Chiu, C.Y. (2019) 'Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection', *Annual Review of Pathology: Mechanisms of Disease*, 14(Volume 14, 2019), pp. 319–338. Available at: https://doi.org/10.1146/annurev-pathmechdis-012418-012751.

Guo, Y. *et al.* (2020) 'Prevalence and Therapies of Antibiotic-Resistance in Staphylococcus aureus', *Frontiers in Cellular and Infection Microbiology*, 10. Available at: https://doi.org/10.3389/fcimb.2020.00107.

Haag, A.M. (2016) 'Mass Analyzers and Mass Spectrometers', *Advances in Experimental Medicine and Biology*, 919, pp. 157–169. Available at: https://doi.org/10.1007/978-3-319-41448-5_7.

Hadpanus, P. *et al.* (2019) 'Biomarker discovery in the biofilm-forming process of *Burkholderia pseudomallei* by mass-spectrometry', *Journal of Microbiological Methods*, 159, pp. 26–33. Available at: https://doi.org/10.1016/j.mimet.2019.02.011.

Haider, A. *et al.* (2023) 'The Current Level of MALDI-TOF MS Applications in the Detection of Microorganisms: A Short Review of Benefits and Limitations',

Microbiology Research, 14(1), pp. 80–90. Available at: https://doi.org/10.3390/microbiolres14010008.

van Hal, S.J. *et al.* (2012) 'Predictors of Mortality in Staphylococcus aureus Bacteremia', *Clinical Microbiology Reviews*, 25(2), pp. 362–386. Available at: https://doi.org/10.1128/CMR.05022-11.

Hawkins, D.M. (2004) 'The Problem of Overfitting', *Journal of Chemical Information and Computer Sciences*, 44(1), pp. 1–12. Available at: https://doi.org/10.1021/ci0342472.

Hilario, M. *et al.* (2006) 'Processing and classification of protein mass spectra', *Mass Spectrometry Reviews*, 25(3), pp. 409–449. Available at: https://doi.org/10.1002/mas.20072.

Horvatić, A. *et al.* (2016) 'High-throughput proteomics and the fight against pathogens', *Molecular BioSystems*, 12(8), pp. 2373–2384. Available at: https://doi.org/10.1039/C6MB00223D.

Howden, B.P. *et al.* (2023) 'Staphylococcus aureus host interactions and adaptation', *Nature Reviews Microbiology*, 21(6), pp. 380–395. Available at: https://doi.org/10.1038/s41579-023-00852-y.

Hrabák, J. *et al.* (2012) 'Detection of NDM-1, VIM-1, KPC, OXA-48, and OXA-162 carbapenemases by matrix-assisted laser desorption ionization-time of flight mass spectrometry', *Journal of Clinical Microbiology*, 50(7), pp. 2441–2443. Available at: https://doi.org/10.1128/JCM.01002-12.

Hu, T. *et al.* (2021) 'Next-generation sequencing technologies: An overview', *Human Immunology*, 82(11), pp. 801–811. Available at: https://doi.org/10.1016/j.humimm.2021.02.012.

Ikuta, K.S. *et al.* (2022) 'Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019', *The Lancet*, 400(10369), pp. 2221–2248. Available at: https://doi.org/10.1016/S0140-6736(22)02185-7.

Jakovljev, A. and Bergh, K. (2015) 'Development of a rapid and simplified protocol for direct bacterial identification from positive blood cultures by using matrix assisted laser desorption ionization time-of- flight mass spectrometry', *BMC Microbiology*, 15(1), p. 258. Available at: https://doi.org/10.1186/s12866-015-0594-2.

Jean Beltran, P.M. *et al.* (2017) 'Proteomics and integrative omic approaches for understanding host–pathogen interactions and infectious diseases', *Molecular Systems Biology*, 13(3), p. 922. Available at: https://doi.org/10.15252/msb.20167062.

Jenul, C. and Horswill, A.R. (2019) 'Regulation of Staphylococcus aureus Virulence', *Microbiology Spectrum*, 7(2), p. 10.1128/microbiolspec.gpp3-0031–2018. Available at: https://doi.org/10.1128/microbiolspec.gpp3-0031-2018.

John Jr, J. (2020) ‘The treatment of resistant staphylococcal infections’, *F1000Research*, 9, p. F1000 Faculty Rev-150. Available at: https://doi.org/10.12688/f1000research.17718.1.

Josten, M. *et al.* (2014) ‘Identification of agr-positive methicillin-resistant Staphylococcus aureus harbouring the class A mec complex by MALDI-TOF mass spectrometry’, *International journal of medical microbiology: IJMM*, 304(8), pp. 1018–1023. Available at: https://doi.org/10.1016/j.ijmm.2014.07.005.

Kaech, C. *et al.* (2006) ‘Course and outcome of *Staphylococcus aureus* bacteraemia: a retrospective analysis of 308 episodes in a Swiss tertiary-care centre’, *Clinical Microbiology and Infection*, 12(4), pp. 345–352. Available at: https://doi.org/10.1111/j.1469-0691.2005.01359.x.

Kaur, G. *et al.* (2013) ‘Timing is important in medication administration: a timely review of chronotherapy research’, *International Journal of Clinical Pharmacy*, 35(3), pp. 344–358. Available at: https://doi.org/10.1007/s11096-013-9749-0.

Kempf, M. *et al.* (2012) ‘Rapid Detection of Carbapenem Resistance in Acinetobacter baumannii Using Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry’, *PLOS ONE*, 7(2), p. e31676. Available at: https://doi.org/10.1371/journal.pone.0031676.

Kessel, K. *et al.* (2024) ‘Staphylococcus aureus bacteremia in alcoholics’, *PloS One*, 19(5), p. e0298612. Available at: https://doi.org/10.1371/journal.pone.0298612.

Kim, J.-M. *et al.* (2019) ‘Rapid Discrimination of Methicillin-Resistant Staphylococcus aureus by MALDI-TOF MS’, *Pathogens (Basel, Switzerland)*, 8(4), p. 214. Available at: https://doi.org/10.3390/pathogens8040214.

Kovtoun, S.V. and Cotter, R.J. (2000) ‘Mass-correlated pulsed extraction: theoretical analysis and implementation with a linear matrix-assisted laser desorption/ionization time of flight mass spectrometer’, *Journal of the American Society for Mass Spectrometry*, 11(10), pp. 841–853. Available at: https://doi.org/10.1016/S1044-0305(00)00165-3.

Krutchinsky, A.N. and Chait, B.T. (2002) ‘On the mature of the chemical noise in MALDI mass spectra’, *Journal of the American Society for Mass Spectrometry*, 13(2), pp. 129–134. Available at: https://doi.org/10.1016/S1044-0305(01)00336-1.

Kwiecinski, J.M. and Horswill, A.R. (2020) ‘Staphylococcus aureus bloodstream infections: pathogenesis and regulatory mechanisms’, *Current Opinion in Microbiology*, 53, pp. 51–60. Available at: https://doi.org/10.1016/j.mib.2020.02.005.

Laabei, M. *et al.* (2014) ‘Predicting the virulence of MRSA from its genome sequence’, *Genome Research*, 24(5), pp. 839–849. Available at: https://doi.org/10.1101/gr.165415.113.

Laabei, M. *et al.* (2021) ‘Significant variability exists in the cytotoxicity of global methicillin-resistant Staphylococcus aureus lineages’, *Microbiology*, 167(12), p. 001119. Available at: https://doi.org/10.1099/mic.0.001119.

Lasch, P. *et al.* (2014) 'Insufficient discriminatory power of MALDI-TOF mass spectrometry for typing of *Enterococcus faecium* and *Staphylococcus aureus* isolates', *Journal of Microbiological Methods*, 100, pp. 58–69. Available at: https://doi.org/10.1016/j.mimet.2014.02.015.

Lazcka, O., Campo, F.J.D. and Muñoz, F.X. (2007) 'Pathogen detection: A perspective of traditional methods and biosensors', *Biosensors and Bioelectronics*, 22(7), pp. 1205–1217. Available at: https://doi.org/10.1016/j.bios.2006.06.036.

Lee, C.-C. *et al.* (2017) 'Timing of appropriate empirical antimicrobial administration and outcome of adults with community-onset bacteremia', *Critical Care (London, England)*, 21(1), p. 119. Available at: https://doi.org/10.1186/s13054-017-1696-z.

Levi, K., Smedley and, J. and Towner, K.J. (2003) 'Evaluation of a real-time PCR hybridization assay for rapid detection of *Legionella pneumophila* in hospital and environmental water samples', *Clinical Microbiology and Infection*, 9(7), pp. 754–758. Available at: https://doi.org/10.1046/j.1469-0691.2003.00666.x.

Levsky, J.M. and Singer, R.H. (2003) 'Fluorescence in situ hybridization: past, present and future', *Journal of Cell Science*, 116(14), pp. 2833–2838. Available at: https://doi.org/10.1242/jcs.00633.

Li, D. *et al.* (2022a) 'MALDI-TOF Mass Spectrometry in Clinical Analysis and Research', *ACS Measurement Science Au*, 2(5), pp. 385–404. Available at: https://doi.org/10.1021/acsmeasuresciau.2c00019.

Li, D. *et al.* (2022b) 'MALDI-TOF Mass Spectrometry in Clinical Analysis and Research', *ACS Measurement Science Au*, 2(5), pp. 385–404. Available at: https://doi.org/10.1021/acsmeasuresciau.2c00019.

Limoli, D.H. *et al.* (2016) 'Staphylococcus aureus and Pseudomonas aeruginosa co-infection is associated with cystic fibrosis-related diabetes and poor clinical outcomes', *European Journal of Clinical Microbiology & Infectious Diseases*, 35(6), pp. 947–953. Available at: https://doi.org/10.1007/s10096-016-2621-0.

Liu, Q. *et al.* (2009) 'Comparison of feature selection and classification for MALDI-MS data', *BMC Genomics*, 10(1), p. S3. Available at: https://doi.org/10.1186/1471-2164-10-S1-S3.

Liu, Y. *et al.* (2021) 'Outcomes of respiratory viral-bacterial co-infection in adult hospitalized patients', *eClinicalMedicine*, 37. Available at: https://doi.org/10.1016/j.eclinm.2021.100955.

Loman, N.J. *et al.* (2012) 'Performance comparison of benchtop high-throughput sequencing platforms', *Nature Biotechnology*, 30(5), pp. 434–439. Available at: https://doi.org/10.1038/nbt.2198.

López-Cortés, X.A. *et al.* (2025) 'Integrating Machine Learning with MALDI-TOF Mass Spectrometry for Rapid and Accurate Antimicrobial Resistance Detection

in Clinical Pathogens', *International Journal of Molecular Sciences*, 26(3), p. 1140. Available at: https://doi.org/10.3390/ijms26031140.

López-Fernández, H. *et al.* (2015) 'Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery', *BMC Bioinformatics*, 16(1), p. 318. Available at: https://doi.org/10.1186/s12859-015-0752-4.

Lu, J.-J. *et al.* (2012) 'Peptide Biomarker Discovery for Identification of Methicillin-Resistant and Vancomycin-Intermediate Staphylococcus aureus Strains by MALDI-TOF', *Analytical Chemistry*, 84(13), pp. 5685–5692. Available at: https://doi.org/10.1021/ac300855z.

Man, L. *et al.* (2021) 'Integrated mass spectrometry-based multi-omics for elucidating mechanisms of bacterial virulence', *Biochemical Society Transactions*, 49(5), pp. 1905–1926. Available at: https://doi.org/10.1042/BST20191088.

Manchanda, S. *et al.* (2018) 'On Comprehensive Mass Spectrometry Data Analysis for Proteome Profiling of Human Blood Samples', *Journal of Healthcare Informatics Research*, 2(3), pp. 305–318. Available at: https://doi.org/10.1007/s41666-018-0022-0.

Mansur, N. *et al.* (2012) 'Does Sex Affect 30-Day Mortality in *Staphylococcus Aureus* Bacteremia?', *Gender Medicine*, 9(6), pp. 463–470. Available at: https://doi.org/10.1016/j.genm.2012.10.009.

Mantini, D. *et al.* (2007) 'LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise', *BMC Bioinformatics*, 8(1), p. 101. Available at: https://doi.org/10.1186/1471-2105-8-101.

McGuinness, W.A., Malachowa, N. and DeLeo, F.R. (2017) 'Vancomycin Resistance in Staphylococcus aureus', *The Yale Journal of Biology and Medicine*, 90(2), pp. 269–281.

Morháč, M. and Matoušek, V. (2008) 'Peak Clipping Algorithms for Background Estimation in Spectroscopic Data', *Applied Spectroscopy*, 62(1), pp. 91–106. Available at: https://doi.org/10.1366/000370208783412762.

Müller, V. *et al.* (2018) 'Identification of pathogenic bacteria in complex samples using a smartphone based fluorescence microscope', *RSC Advances*, 8(64), pp. 36493–36502. Available at: https://doi.org/10.1039/C8RA06473C.

Navarro, A.L.G. *et al.* (2024) 'A Comprehensive Guide to Combining R and Python code for Data Science, Machine Learning and Reinforcement Learning'. arXiv. Available at: https://doi.org/10.48550/arXiv.2407.14695.

Ng, C.C.A., Zhou, Y. and Yao, Z.-P. (2023) 'Algorithms for de-novo sequencing of peptides by tandem mass spectrometry: A review', *Analytica Chimica Acta*, 1268, p. 341330. Available at: https://doi.org/10.1016/j.aca.2023.341330.

Nix, I.D. *et al.* (2020) 'Detection of Methicillin Resistance in Staphylococcus aureus From Agar Cultures and Directly From Positive Blood Cultures Using MALDI-TOF Mass Spectrometry-Based Direct-on-Target Microdroplet Growth

Assay', *Frontiers in Microbiology*, 11. Available at: https://doi.org/10.3389/fmicb.2020.00232.

WHO (2024) *WHO bacterial priority pathogens list, 2024: bacterial pathogens of public health importance, to guide research, development, and strategies to prevent and control antimicrobial resistance*. World Health Organization.

Patel, J.B. (2001) '16S rRNA Gene Sequencing for Bacterial Pathogen Identification in the Clinical Laboratory', *Molecular Diagnosis*, 6(4), pp. 313–321. Available at: https://doi.org/10.1007/BF03262067.

Patel, R. (2019) 'A Moldy Application of MALDI: MALDI-ToF Mass Spectrometry for Fungal Identification', *Journal of Fungi*, 5(1), p. 4. Available at: https://doi.org/10.3390/jof5010004.

Peel, T.N. *et al.* (2015) 'Matrix-assisted laser desorption ionization time of flight mass spectrometry and diagnostic testing for prosthetic joint infection in the clinical microbiology laboratory', *Diagnostic Microbiology and Infectious Disease*, 81(3), pp. 163–168. Available at: https://doi.org/10.1016/j.diagmicrobio.2014.11.015.

Pérez-Llarena, F.J. and Bou, G. (2016) 'Proteomics As a Tool for Studying Bacterial Virulence and Antimicrobial Resistance', *Frontiers in Microbiology*, 7. Available at: https://doi.org/10.3389/fmicb.2016.00410.

Pérez-Sancho, M. *et al.* (2018) 'Rapid differentiation of Staphylococcus aureus subspecies based on MALDI-TOF MS profiles', *Journal of Veterinary Diagnostic Investigation*, 30(6), pp. 813–820. Available at: https://doi.org/10.1177/1040638718805537.

Perše, G. *et al.* (2022) 'Sepsityper® Kit versus In-House Method in Rapid Identification of Bacteria from Positive Blood Cultures by MALDI-TOF Mass Spectrometry', *Life*, 12(11), p. 1744. Available at: https://doi.org/10.3390/life12111744.

Pfyffer, G.E. and Wittwer, F. (2020) 'Incubation Time of Mycobacterial Cultures: How Long Is Long Enough To Issue a Final Negative Report to the Clinician?', *Journal of Clinical Microbiology*, 50(12), pp. 4188–4189. Available at: https://doi.org/10.1128/jcm.02283-12.

Picaud, V. *et al.* (2018) 'Linear MALDI-ToF simultaneous spectrum deconvolution and baseline removal', *BMC Bioinformatics*, 19(1), p. 123. Available at: https://doi.org/10.1186/s12859-018-2116-3.

Pinault, L. *et al.* (2019) 'Direct Identification of Pathogens in Urine by Use of a Specific Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectrum Database', *Journal of Clinical Microbiology*, 57(4), p. 10.1128/jcm.01678-18. Available at: https://doi.org/10.1128/jcm.01678-18.

Pollitt, E.J.G. *et al.* (2018) 'Staphylococcus aureus infection dynamics', *PLOS Pathogens*, 14(6), p. e1007112. Available at: https://doi.org/10.1371/journal.ppat.1007112.

Quéro, L. *et al.* (2020) ‘Application of MALDI-TOF MS to species complex differentiation and strain typing of food related fungi: Case studies with *Aspergillus* section *Flavi* species and *Penicillium roqueforti* isolates’, *Food Microbiology*, 86, p. 103311. Available at: https://doi.org/10.1016/j.fm.2019.103311.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Recker, M. *et al.* (2017) ‘Clonal differences in Staphylococcus aureus bacteraemia-associated mortality’, *Nature Microbiology*, 2(10), pp. 1381–1388. Available at: https://doi.org/10.1038/s41564-017-0001-x.

Ressom, H.W. *et al.* (2007) ‘Peak selection from MALDI-TOF mass spectra using ant colony optimization’, *Bioinformatics*, 23(5), pp. 619–626. Available at: https://doi.org/10.1093/bioinformatics/btl678.

Rhoads, D.D. *et al.* (2016) ‘The presence of a single MALDI-TOF mass spectral peak predicts methicillin resistance in staphylococci’, *Diagnostic Microbiology and Infectious Disease*, 86(3), pp. 257–261. Available at: https://doi.org/10.1016/j.diagmicrobio.2016.08.001.

Russell, C.D. *et al.* (2024) ‘Distinct Clinical Endpoints of Staphylococcus aureus Bacteraemia Complicate Assessment of Outcome’, *Clinical Infectious Diseases*, 79(3), pp. 604–611. Available at: https://doi.org/10.1093/cid/ciae281.

Rychert, J. (2019) ‘Benefits and Limitations of MALDI-TOF Mass Spectrometry for the Identification of Microorganisms’, *Journal of Infectiology and Epidemiology*, 2(4). Available at: https://www.infectiologyjournal.com/articles/benefits-and-limitations-of-malditof-mass-spectrometry-for-the-identification-of-microorganisms.html (Accessed: 1 February 2025).

Sabbagh, B. *et al.* (2016) ‘Clinical applications of MS-based protein quantification’, *Proteomics. Clinical Applications*, 10(4), pp. 323–345. Available at: https://doi.org/10.1002/prca.201500116.

Savitzky, Abraham. and Golay, M.J.E. (1964) ‘Smoothing and Differentiation of Data by Simplified Least Squares Procedures.’, *Analytical Chemistry*, 36(8), pp. 1627–1639. Available at: https://doi.org/10.1021/ac60214a047.

Schmid, M., Rath, D. and Diebold, U. (2022) ‘Why and How Savitzky–Golay Filters Should Be Replaced’, *ACS Measurement Science Au*, 2(2), pp. 185–196. Available at: https://doi.org/10.1021/acsmeasuresciau.1c00054.

Schmidt, M.N. *et al.* (2019) ‘Peak Detection and Baseline Correction Using a Convolutional Neural Network’, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2757–2761. Available at: https://doi.org/10.1109/ICASSP.2019.8682311.

Seng, P. *et al.* (2009) 'Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry', *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 49(4), pp. 543–551. Available at: https://doi.org/10.1086/600885.

Shao, C. *et al.* (2012) 'The Use of Principal Component Analysis in MALDI-TOF MS: a Powerful Tool for Establishing a Mini-optimized Proteomic Profile', *American journal of biomedical sciences*, 4(1), pp. 85–101. Available at: https://doi.org/10.5099/aj120100085.

Siewers, K. *et al.* (2021) 'Time to administration of antibiotics and mortality in sepsis', *Journal of the American College of Emergency Physicians Open*, 2(3), p. e12435. Available at: https://doi.org/10.1002/emp2.12435.

Singhal, N. *et al.* (2015) 'MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis', *Frontiers in Microbiology*, 6, p. 791. Available at: https://doi.org/10.3389/fmicb.2015.00791.

Smith, R.M. (2004) *Understanding Mass Spectra: A Basic Approach*. John Wiley & Sons.

Sparbier, K. *et al.* (2020) 'Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry-Based Functional Assay for Rapid Detection of Resistance against β-Lactam Antibiotics', *Journal of Clinical Microbiology*, 50(3), pp. 927–937. Available at: https://doi.org/10.1128/jcm.05737-11.

Sukumaran, A. *et al.* (2021) 'Proteomics of host-bacterial interactions: new insights from dual perspectives', *Canadian Journal of Microbiology*, 67(3), pp. 213–225. Available at: https://doi.org/10.1139/cjm-2020-0324.

Sulaiman, J.E. and Lam, H. (2022) 'Proteomics in antibiotic resistance and tolerance research: Mapping the resistome and the tolerome of bacterial pathogens', *PROTEOMICS*, 22(8), p. 2100409. Available at: https://doi.org/10.1002/pmic.202100409.

Taban, B.M. and Numanoglu Cevik, Y. (2021) 'The efficiency of MALDI-TOF MS method in detecting Staphylococcus aureus isolated from raw milk and artisanal dairy foods', *CyTA - Journal of Food*, 19(1), pp. 739–750. Available at: https://doi.org/10.1080/19476337.2021.1977392.

Tacconelli, E., Pop-Vicas, A.E. and D'Agata, E.M.C. (2006) 'Increased mortality among elderly patients with meticillin-resistant Staphylococcus aureus bacteraemia', *Journal of Hospital Infection*, 64(3), pp. 251–256. Available at: https://doi.org/10.1016/j.jhin.2006.07.001.

Taguchi, Y. -h. and Oono, Y. (2005) 'Relational patterns of gene expression via non-metric multidimensional scaling analysis', *Bioinformatics*, 21(6), pp. 730–740. Available at: https://doi.org/10.1093/bioinformatics/bti067.

Tarrant, C. and Krockow, E.M. (2022) 'Antibiotic overuse: managing uncertainty and mitigating against overtreatment', *BMJ Quality & Safety*, 31(3), pp. 163–167. Available at: https://doi.org/10.1136/bmjqs-2021-013615.

Tibshirani, R. *et al.* (2004) ‘Sample classification from protein mass spectrometry, by “peak probability contrasts”’, *Bioinformatics*, 20(17), pp. 3034–3044. Available at: https://doi.org/10.1093/bioinformatics/bth357.

Tong, D.L. *et al.* (2011) ‘A simpler method of preprocessing MALDI-TOF MS data for differential biomarker analysis: stem cell and melanoma cancer studies’, *Clinical Proteomics*, 8(1), p. 14. Available at: https://doi.org/10.1186/1559-0275-8-14.

Tong, S.Y.C. *et al.* (2015) ‘Staphylococcus aureus infections: epidemiology, pathophysiology, clinical manifestations, and management’, *Clinical Microbiology Reviews*, 28(3), pp. 603–661. Available at: https://doi.org/10.1128/CMR.00134-14.

Topić Popović, N. *et al.* (2023) ‘Sample preparation and culture condition effects on MALDI-TOF MS identification of bacteria: A review’, *Mass Spectrometry Reviews*, 42(5), pp. 1589–1603. Available at: https://doi.org/10.1002/mas.21739.

Urban-Chmiel, R. *et al.* (2022) ‘Antibiotic Resistance in Bacteria—A Review’, *Antibiotics*, 11(8), p. 1079. Available at: https://doi.org/10.3390/antibiotics11081079.

van der Vaart, T.W. *et al.* (2022) ‘All-Cause and Infection-Related Mortality in Staphylococcus aureus Bacteremia, a Multicenter Prospective Cohort Study’, *Open Forum Infectious Diseases*, 9(12), p. ofac653. Available at: https://doi.org/10.1093/ofid/ofac653.

Váradi, L. *et al.* (2017) ‘Methods for the detection and identification of pathogenic bacteria: past, present, and future’, *Chemical Society Reviews*, 46(16), pp. 4818–4832. Available at: https://doi.org/10.1039/C6CS00693K.

Vestby, L.K. *et al.* (2020) ‘Bacterial Biofilm and its Role in the Pathogenesis of Disease’, *Antibiotics*, 9(2), p. 59. Available at: https://doi.org/10.3390/antibiotics9020059.

Virtanen, P. *et al.* (2020) ‘SciPy 1.0: fundamental algorithms for scientific computing in Python’, *Nature Methods*, 17(3), pp. 261–272. Available at: https://doi.org/10.1038/s41592-019-0686-2.

Wang, C. *et al.* (2022) ‘Large-Scale Samples Based Rapid Detection of Ciprofloxacin Resistance in Klebsiella pneumoniae Using Machine Learning Methods’, *Frontiers in Microbiology*, 13. Available at: https://doi.org/10.3389/fmicb.2022.827451.

Wang, J. *et al.* (2021) ‘Evaluation of three sample preparation methods for the identification of clinical strains by using two MALDI-TOF MS systems’, *Journal of Mass Spectrometry*, 56(2), p. e4696. Available at: https://doi.org/10.1002/jms.4696.

Wang, J. *et al.* (2022) ‘Rapid Detection of Carbapenem-Resistant Klebsiella pneumoniae Using Machine Learning and MALDI-TOF MS Platform’, *Infection and Drug Resistance*, 15, pp. 3703–3710. Available at: https://doi.org/10.2147/IDR.S367209.

Westgeest, A.C. *et al.* (2024) ‘Female Sex and Mortality in Patients with Staphylococcus aureus Bacteremia: A Systematic Review and Meta-analysis’, *JAMA network open*, 7(2), p. e240473. Available at: https://doi.org/10.1001/jamanetworkopen.2024.0473.

Wijetunge, C.D. *et al.* (2015) ‘A new peak detection algorithm for MALDI mass spectrometry data based on a modified Asymmetric Pseudo-Voigt model’, *BMC Genomics*, 16(12), p. S12. Available at: https://doi.org/10.1186/1471-2164-16-S12-S12.

Wolkewitz, M. *et al.* (2011) ‘Mortality associated with in-hospital bacteraemia caused by Staphylococcus aureus: a multistate analysis with follow-up beyond hospital discharge’, *Journal of Antimicrobial Chemotherapy*, 66(2), pp. 381–386. Available at: https://doi.org/10.1093/jac/dkq424.

Yang, C., He, Z. and Yu, W. (2009) ‘Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis’, *BMC Bioinformatics*, 10(1), p. 4. Available at: https://doi.org/10.1186/1471-2105-10-4.

Yang, E. *et al.* (2024) ‘Clinical and microbiological characteristics of persistent Staphylococcus aureus bacteremia, risk factors for mortality, and the role of CD4+ T cells’, *Scientific Reports*, 14, p. 15472. Available at: https://doi.org/10.1038/s41598-024-66520-0.

Yasui, Y. *et al.* (2003) ‘A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection’, *Biostatistics*, 4(3), pp. 449–463. Available at: https://doi.org/10.1093/biostatistics/4.3.449.

Yu, J. *et al.* (2022) ‘Rapid Identification of Methicillin-Resistant Staphylococcus aureus Using MALDI-TOF MS and Machine Learning from over 20,000 Clinical Isolates’, *Microbiology Spectrum*, 10(2), p. e0048322. Available at: https://doi.org/10.1128/spectrum.00483-22.

Yucel, U. and Smith, J.S. (2024) ‘Mass Spectrometry’, in B.P. Ismail and S.S. Nielsen (eds) *Nielsen’s Food Analysis*. Cham: Springer International Publishing, pp. 147–164. Available at: https://doi.org/10.1007/978-3-031-50643-7_11.

Zhang, Y.-M. *et al.* (2023) ‘Rapid identification of carbapenem-resistant Klebsiella pneumoniae based on matrix-assisted laser desorption ionization time-of-flight mass spectrometry and an artificial neural network model’, *Journal of Biomedical Science*, 30(1), p. 25. Available at: https://doi.org/10.1186/s12929-023-00918-2.

Zheng, Y. *et al.* (2016) ‘An improved algorithm for peak detection in mass spectra based on continuous wavelet transform’, *International Journal of Mass Spectrometry*, 409, pp. 53–58. Available at: https://doi.org/10.1016/j.ijms.2016.09.020.

Zhou, M. *et al.* (2017) ‘An Improved In-house MALDI-TOF MS Protocol for Direct Cost-Effective Identification of Pathogens from Blood Cultures’, *Frontiers in Microbiology*, 8. Available at: https://doi.org/10.3389/fmicb.2017.01824.

Zhou, Y. *et al.* (2022) 'An Improved Algorithm for Peak Detection Based on Weighted Continuous Wavelet Transform', *IEEE Access*, 10, pp. 118779–118788. Available at: https://doi.org/10.1109/ACCESS.2022.3220640.

Zubair, M. *et al.* (2022) 'Proteomics approaches: A review regarding an importance of proteome analyses in understanding the pathogens and diseases', *Frontiers in Veterinary Science*, 9. Available at: https://doi.org/10.3389/fvets.2022.1079359.

Generative AI (GenAI) Statement

AI-supported use is permitted in this dissertation. I acknowledge the use of GenAI tools in this assessment to:

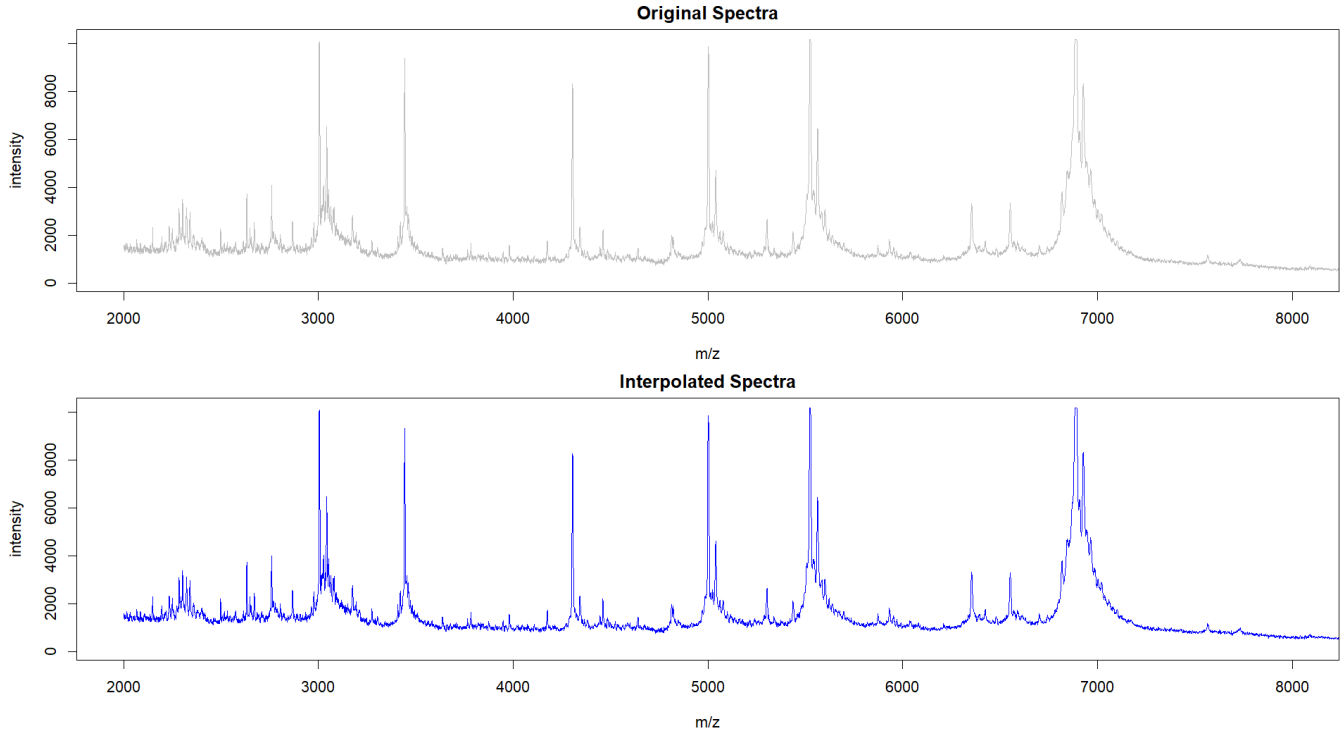
i. Help me understand key theories and concepts.
ii. Generate ideas and code.
iii. Refine and proofread prose.

OpenAI. (2025). ChatGPT (July 21 version) Accessible at: https://chat.openai.com/ (Last accessed: 21 July 2025).

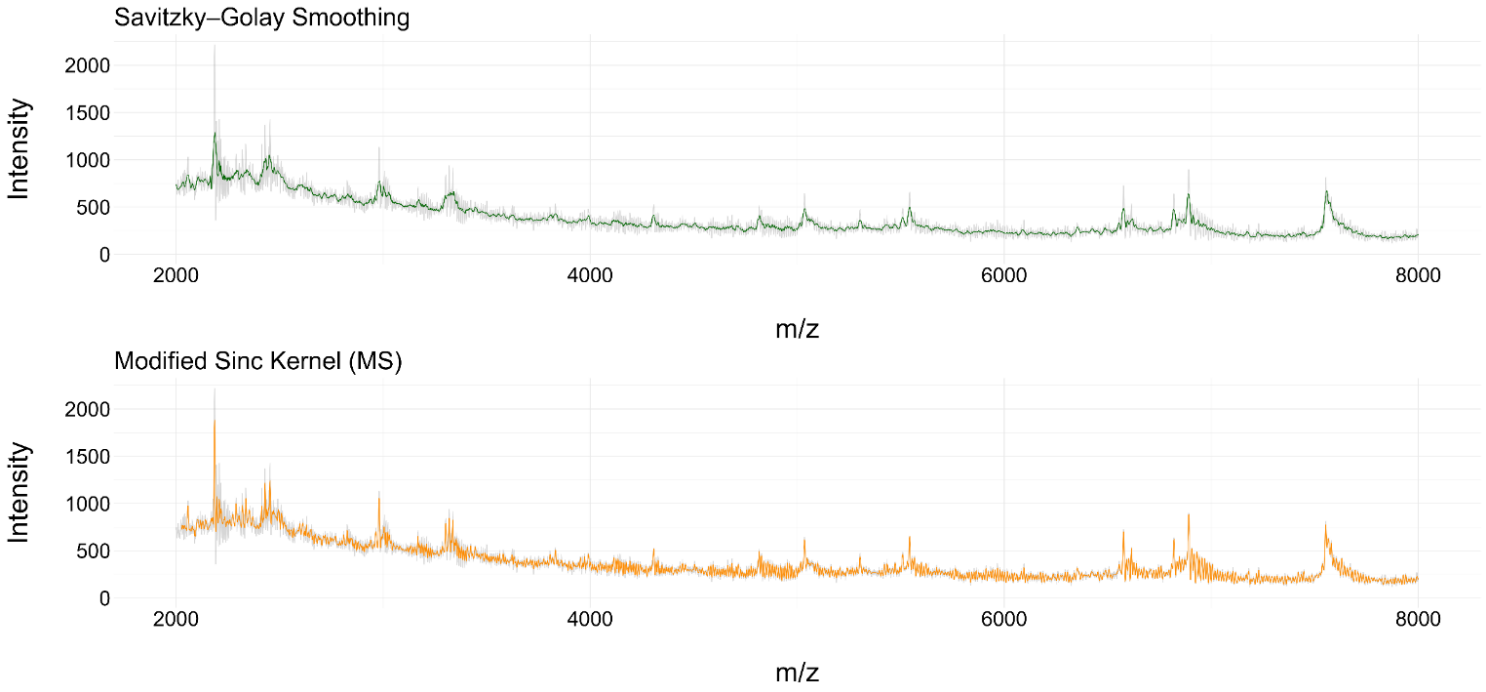
7. Appendix

Code

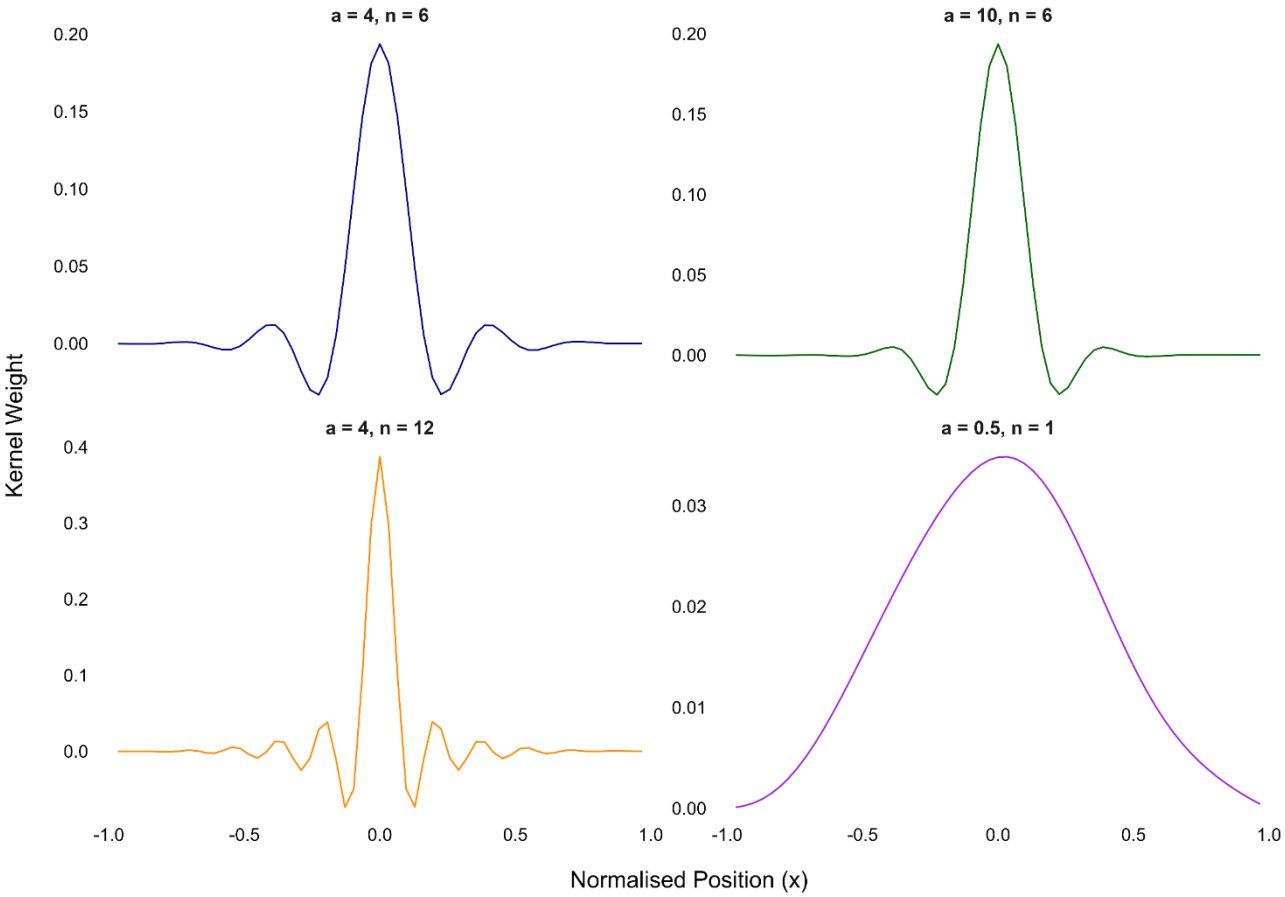
All code used in the creation of this study can be found in the publically accessible GitHub repository linked here. The data used at each stage of the report can be found here.



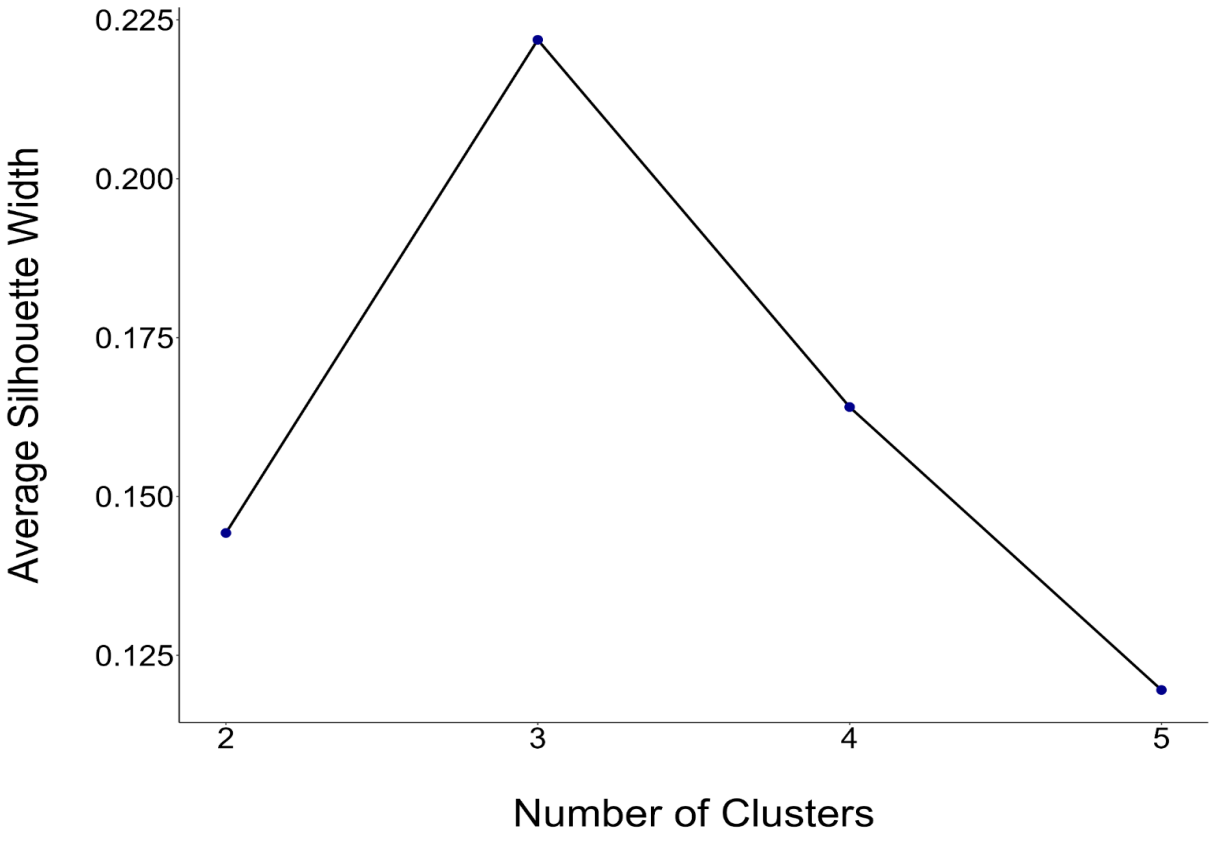
Appendix 1. A visualisation of an example mass spectrum before and after interpolation (Isolate A060).



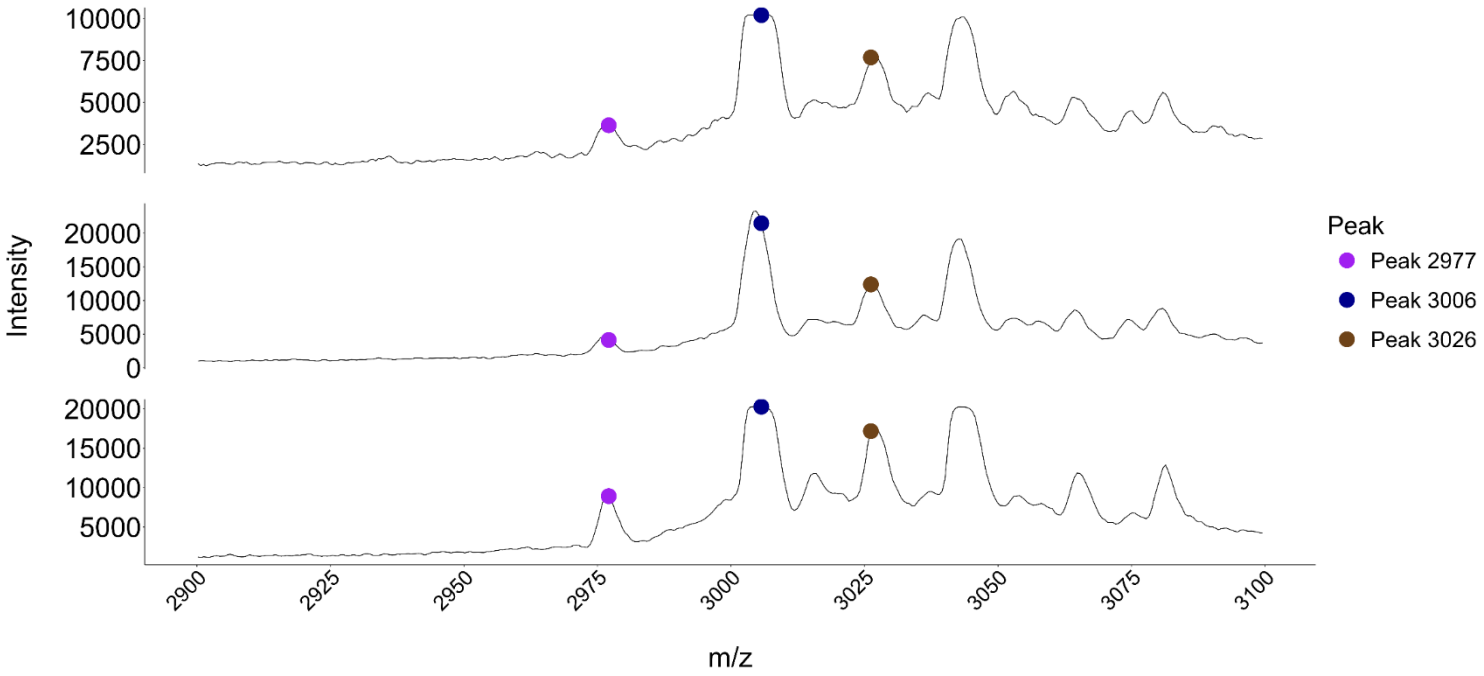
Appendix 2. Side-by-side comparison of a spectrum smoothed by the Standard Savitzky-Golay filter and a Modified Sinc Kernel, overlaid on the raw spectrum (Isolate A060).



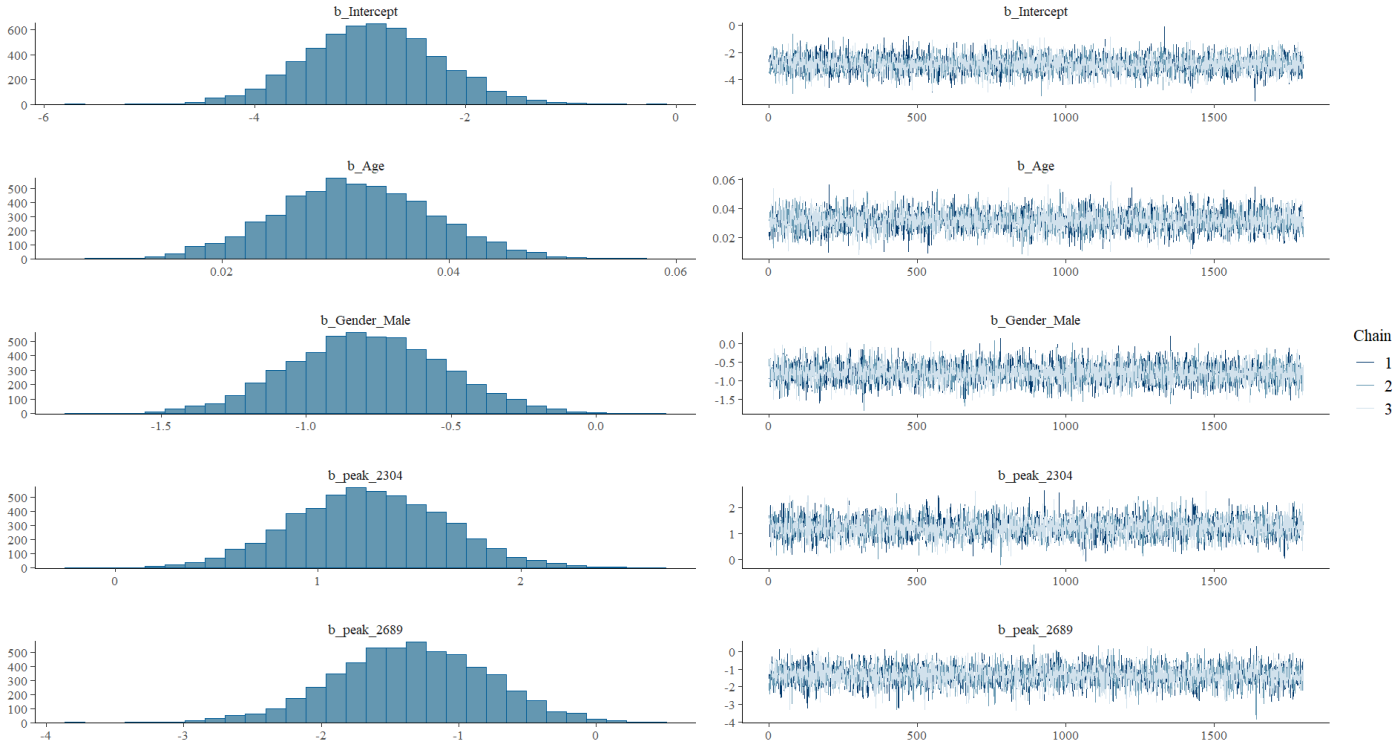
Appendix 3. The effect of different alpha (a) and frequency (n) values on kernel morphologies. The final kernel is shown in the top-left facet.



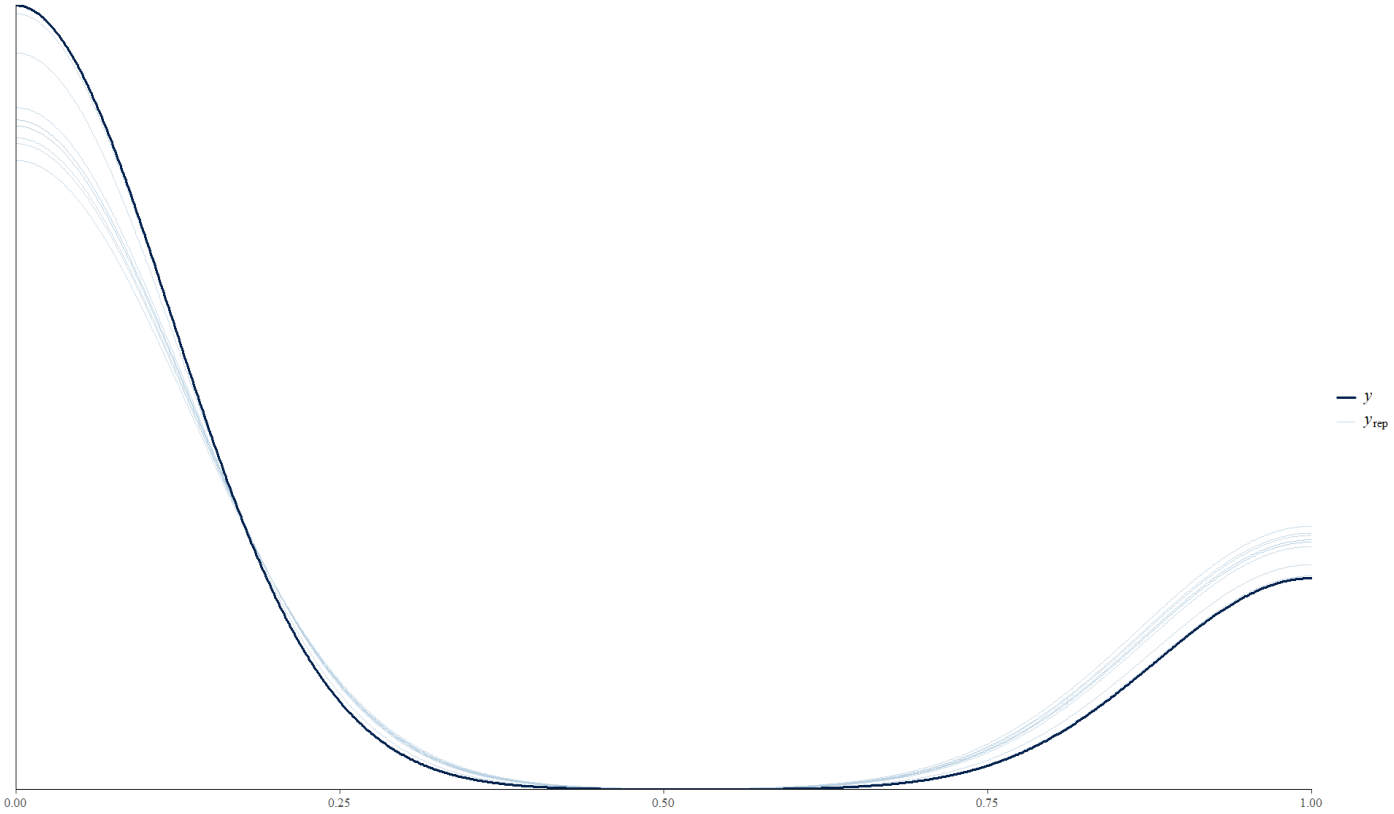
Appendix 4. Silhouette analysis (Jaccard) for the NMDS clusters (n = 278).



Appendix 5. Spectral plots illustrating the prominence of peaks 2977, 3006 and 3026 in three example spectra (raw): ASARM102 (A), ASASM12 (B), and ASASM181 (C).



Appendix 6. Posterior distributions and trace plots showing successful convergence for the final Bayesian model.



Appendix 7. Posterior predictive plot distributions illustrating the quality of fit between the statistical model and the observed data.