Development of a Brain Tissue Classifier Using Atmospheric Solids Analysis Probe Mass Spectrometry

Liwen Song

Hertford College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2025

Abstract

Recent advances in compact, low-cost, and rapid analytical platforms have enabled direct measurement of molecular composition within complex biological materials. Clinical disciplines such as neurosurgery and neuropathology could benefit from analytical frameworks that combine fast chemical profiling with machine-learning-based interpretation. This thesis presents a brain tissue classification framework based on atmospheric solids analysis probe mass spectrometry (ASAP–MS) and supervised machine learning, and evaluates its potential to support molecular discrimination between brain tissue types in neurosurgical and neuropathological contexts.

To obtain high-quality data from biological samples, several sources of technical variability in ASAP–MS were systematically investigated. These included background contamination by residual calibrant, probe cooling after background acquisition, probe cleaning protocols, contamination from consumables, inter-user variability, and batch effects. For each factor, optimisation strategies were developed to improve data quality.

Beyond technical considerations, the influence of molecular class and sample composition on ASAP mass spectra was examined. Small polar metabolites in biological samples were found to undergo extensive in-source chemistry, generating dense and highly correlated spectral features. In contrast, lipid species from biological samples ionised more predictably and contributed largely additive signals, resulting in more stable and interpretable spectral patterns.

Building on this, the feasibility of applying ASAP–MS to formalin-fixed, paraffin-embedded (FFPE) brain tissue samples was evaluated by comparison with frozen brain tissue samples. Focused ultrasonication was shown to be an effective deparaffinisation approach. However, preservation method and fixation time strongly influenced spectral quality. Short-fixed FFPE samples retained more consistent molecular fingerprints than long-fixed samples, yet frozen brain tissue samples consistently outperformed FFPE brain tissue samples for ASAP–MS analysis.

Finally, an optimised protocol for analysing fresh brain tissue samples using ASAP–MS was developed. The resulting mass spectra were used to distinguish brain tumour from normal brain cortex. Robust performance was observed across iterative model updates and predictions on unseen samples. SHAP-based model interpretation indicated that classification was driven primarily by lipid-rich spectral

regions. The best-performing model, a random forest with oversampling, achieved a sensitivity of 0.97, a specificity of 0.95, and a Cohen's kappa of 0.90, indicating excellent agreement beyond chance. These results demonstrate that ASAP–MS, when combined with appropriate protocol optimisation and machine learning, has genuine potential as a rapid support tool for brain tissue classification.

Acknowledgements

Before I began writing this thesis, I had imagined the acknowledgements many times. I thought about the people who have supported me throughout my academic journey, those who helped me when I struggled, and those who, in ways both large and small, changed the direction of my life. Yet when I finally sat down to write this section, I realised how difficult it was to decide where to begin.

So I will begin at the beginning. For as long as I can remember, I have felt to be an 'outlier'. I often believed that I looked different, spoke differently, and behaved differently from those around me since I was a child, and I once wished simply to be "normal". I am deeply grateful to the people who recognised value in that difference and encouraged me to remain myself.

First and foremost, I would like to thank my primary supervisor, Professor Claire Vallance. Thank you for all of your support, from my DPhil application through to the final word of this thesis. During my lowest moments, I once asked whether you regretted choosing me as your student. Your answer truly sustained me. You are one of the most inspiring people I have ever met. You are so multi-talented, but so kind and approachable. You taught me how to be a good scientist, but more importantly, how to be a better person.

I would also like to thank my colleagues in the Vallance Group, especially Dr Annabel Eardley-Brunt. You are a highly capable scientist with amazing organisational and leadership skills. I am sincerely grateful for your patience in teaching me many aspects of ASAP-MS. I learned so much from you and am very glad to have gained a friend as well as a colleague.

My sincere thanks also go to my co-supervisor, Dr Olaf Ansorge. Thank you for recognising and supporting my research ideas. Your ability to generate scientific insight is extraordinary. You encouraged me to think critically and gave me the confidence to explore ideas more broadly and creatively.

I am grateful to all members of the Ansorge Group, especially Dr Jasmine Reese. Your dedication, excellence, and knowledge are appreciated by everyone who works with you. Thank you for teaching me so much about neuroscience and for showing me how to engage with external collaborators. Your curiosity and drive are truly motivating, and I am very thankful for your friendship.

I would also like to thank the neurosurgeons who actively participated in this project, particularly Professor Puneet Plaha and Dr Chris McKinnon. Without

your engagement and support, Chapter Five of this thesis would not have been possible. I will always remember this collaboration with deep gratitude. In addition, I would like to thank Dr Melika Akhbari for your efforts in coordinating sample delivery, connecting neurosurgeons with the mass spectrometry laboratory. Your contributions ensured that this research could proceed smoothly.

Beyond my doctoral work, I would like to express my sincere gratitude to my Master's supervisor, Dr Matthew Grech-Sollars. During my two Master's degrees at Imperial College London, you were the only supervisor who consistently took me seriously and provided genuine guidance and support. I am especially grateful for our weekly online meetings during the COVID pandemic and for your encouragement throughout my PhD application process. Without your support, I would not be where I am today.

我由衷地感谢我的父母在我成长过程中给予我的支持与陪伴。不同于许多传统观念中的东亚家庭，你们始终尊重并允许我为自己的人生作出选择。正是因为你们持续的鼓励和资助、毫不动摇的信任与包容，我才能走到今天。如果没有你们，这段旅程不可能完成。如今，我终于可以坦然且自豪地说：我做到了！

I am also deeply grateful to the NHS, whose care protected my health and saved my life. I hope that through my work, I can contribute scientific research that ultimately benefits patients and returns even a small part of the kindness I received.

To my husband, Dr Michael Platt: meeting you has been one of the greatest gifts of my life. Thank you for showing me the beauty of the world, for bringing so much joy into my life, and for being a constant source of inspiration. Thank you for listening to me rehearse presentations repeatedly, for asking thoughtful questions, and for supporting me with patience and love. You taught me what unconditional love truly means.

I would like to thank the University of Oxford, the Department of Chemistry, the Department of Neuropathology, and the Oxford Brain Bank for providing an exceptional academic and research environment.

I am grateful to everyone who encouraged me during my DPhil. This journey involved significant frustration, and there were times when I felt underappreciated and overlooked. It was the people around me who reminded me that I have something meaningful to contribute.

Finally, I would like to thank all the patients who participated in this research and generously donated precious samples. I sincerely hope that the future research into brain tumours will lead to better understanding, improved treatments, and ultimately, a cure.

Contents

List of Figures

List of Abbreviations

2HG 2-hydroxyglutarate

AAA . . . . . Aminoadipic acid

ANOVA . . . . Analysis of variance

ASAP-MS . . Atmospheric solids analysis probe mass spectrometry.

CSF . . . . . Cerebrospinal fluid.

CUSA . . . . . Cavitron ultrasonic surgical aspirator

CV . . . . . Coefficient of variation

DART . . . . . Direct analysis in real time

DES . . . . . Direct electrical stimulation

DESI . . . . . Desorption electrospray ionisation

EOR . . . . . Extent of resection

FFPE . . . . . Formalin-fixed, paraffin-embedded

GAA . . . . . Guanidinoacetic acid

GC-MS . . . . Gas chromatography-mass spectrometry

H&E . . . . . hematoxylin and eosin

HSD . . . . . Honestly significant difference

HTA . . . . . Human tissue authority

ICA . . . . . Independent component analysis

IHC . . . . . immunohistochemistry

iMRI . . . . . Intra-operative MRI

IONM . . . . . Intra-operative neurophysiological monitoring

IDH . . . . . Isocitrate dehydrogenase

KNN . . . . . k-Nearest Neighbors

kW . . . . . Kruskal–Wallis

LC-MS . . . . Liquid chromatography mass spectrometry

LDA . . . . . Linear discriminant analysis

LR . . . . . Logistic regression

m/z . . . . . Mass to charge ratio

MTBE . . . . . Methyl tert-butyl ether

MEPs . . . . . Motor evoked potentials

MS . . . . . Mass spectrometry

NB . . . . . Naive Bayes classifier

NHS . . . . . National Health Service

OBB . . . . . Oxford Brain Bank

OCT . . . . . Optimal cutting temperature

PCA . . . . . Principal component analysis.

REC . . . . . Research ethics committee

RF . . . . . Random forests

TIC . . . . . Total ion count

SSEPs . . . . . somatosensory evoked potentials

SVM . . . . . Support vector machine

1 Introduction

The complex molecular composition of biological samples reflects both physiological function and pathological changes. Capturing compositional change is therefore essential for understanding pathological progression, distinguishing diseased from normal tissue, and informing diagnostic and therapeutic decisions.

Traditionally, analytical techniques used to determine biological composition prioritise comprehensive molecular identification and/or high analytical resolution. These methods often compromise the acquisition speed and practical deployability. Additionally, such approaches typically rely on large, expensive instrumentation and extensive sample preparation, constraining their use to specialised laboratory environments.

Recent advances in the development of small, lower-cost, and rapid measurement platforms enable the fast and direct capture of molecular changes from complex biological materials. The parallel development of machine learning methods has resulted in data analysis protocols capable of handling the high-dimensional, chemically complex data generated by these measurements.[1] In this context, clinical disciplines such as neurosurgery and neuropathology stand to benefit from new rapid methods that combine fast chemical profiling with machine learning tools.[2]

In the work described in this thesis, a brain tissue classification framework based on rapid mass spectrometric profiling and machine learning is developed and

evaluated, with the aim of exploring its potential to support molecular discrimination of brain tissue types in the contexts of neurosurgery and neuropathology.

1.1 Brain tumour diagnosis and treatment

A brain tumour is defined by the abnormal proliferation of cells within the central nervous system, creating a mass that disrupts healthy brain architecture and function.[3] Brain tumours are broadly classified as either primary or secondary. Primary brain tumours develop within the brain, while secondary brain tumours arise from cancers in other organs that spread to the brain via the bloodstream. Primary brain tumours are further categorised according to their histological features and molecular characteristics. The World Health Organization (WHO) classification system groups tumours based on cell lineage, genetic alterations, and biological behaviour, and assigns grades that reflect tumour aggressiveness and expected clinical outcome.[4] Low-grade tumours generally grow more slowly and may have a better prognosis than high-grade tumours, which are typically more aggressive and associated with poorer survival. Primary brain tumours account for approximately 1.6% of all newly diagnosed cancers worldwide, but they are associated with a disproportionately high burden of mortality and morbidity.[5] Gliomas are the most common primary malignant brain tumours, representing nearly 80% of malignant central nervous system cases.[6] Within this group, glioblastoma (GBM) is the most prevalent and aggressive subtype, characterised by rapid growth and diffuse infiltration into surrounding brain tissue.[7, 8]

Despite advances in neuroimaging, surgery, radiotherapy, and chemotherapy, the management of brain tumours remains challenging. Brain tumours have become the 12th most deadly type of cancer.[9] In particular, the prognosis for high-grade gliomas remains poor, with a five-year survival rate of less than 10% for glioblastoma patients.[10] Current standard-of-care involves maximal safe surgical resection followed by radiotherapy and concomitant chemotherapy.[11] However, tumour heterogeneity, infiltrative growth patterns, and proximity to functionally

critical brain regions limit the effectiveness of treatment and complicate complete surgical resection.[12]

1.1.1 The role of neuropathology

Neuropathology analysis plays a central role in the diagnosis and classification of brain tumours, linking clinical presentation with therapeutic decision-making. The current diagnostic gold standard is the integrated report, formalised in the 2021 WHO Classification of Tumours of the Central Nervous System (WHO CNS5), which combines histological features with molecular information to establish a final diagnosis [4].

Brain tumours include many different types with unique physical and genetic characteristics, and a full review of every type is beyond the scope of this overview. Instead, we focus on gliomas, particularly diffuse gliomas, as a representative tumour group. Gliomas are commonly encountered in neuropathological practice. Figure 1.1 provides a simplified framework showing how morphology, immunohistochemistry, and molecular markers are integrated to reach a final diagnosis.

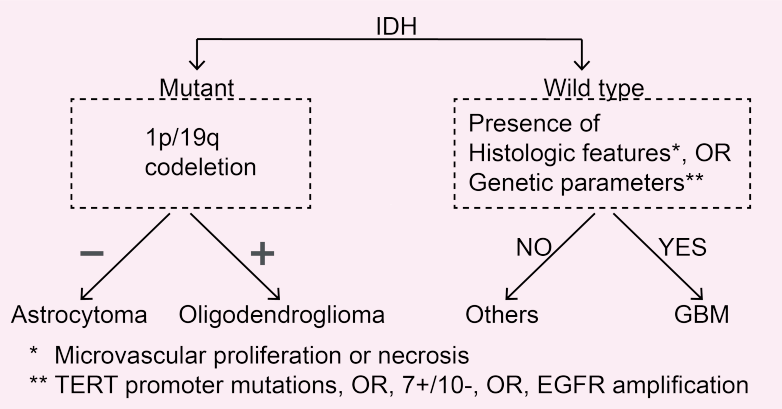

Figure 1.1: Simplified schematic of the adult diffuse glioma classification workflow based on the 2021 World Health Organization (WHO CNS5) guidelines, adapted from [13]. Isocitrate dehydrogenase (IDH) is a metabolic enzyme. Mutations in IDH1/2 define major glioma subtypes and are associated with distinct tumour biology and prognosis.

Histological and immunohistochemical markers describe the microscopic structure of tumour tissue and the expression of specific proteins by tumour cells. These features are assessed using routine hematoxylin and eosin (H&E) staining and

immunohistochemistry (IHC) and provide the foundation for tumour grading and initial classification performed by neuropathologists.[14] Microvascular proliferation refers to abnormal proliferation of blood vessels within the tumour and is a defining feature of high-grade gliomas. Necrosis, characterised by focal areas of cell death, is also associated with aggressive tumour behaviour and contributes to tumour grading. Protein expression assessed by IHC provides additional diagnostic information. Glial fibrillary acidic protein (GFAP) is used to support glial lineage, which denotes tumour cells showing differentiation towards glial cell types[15], while the Ki-67 (MIB-1) proliferation index reflects the proportion of actively dividing cells and provides an estimate of tumour growth activity [16]. Isocitrate dehydrogenase (IDH) is a metabolic enzyme involved in the tricarboxylic acid cycle. Mutations in IDH1 or IDH2 are common in diffuse gliomas and result in the abnormal production of the oncometabolite 2-hydroxyglutarate, which induces widespread epigenetic alterations. As a consequence, IDH mutation status is a central molecular marker in brain tumour classification and is strongly associated with tumour biology, prognosis, and treatment response. In routine practice, mutation-specific immunohistochemistry for IDH1-R132H enables rapid screening for the most common IDH mutation and is widely used as an initial molecular stratification step [17, 18].

Molecular and genetic markers are identified using DNA-based techniques, including sequencing and chromosomal analysis, and provide information on tumour lineage, biological behaviour, and prognosis. These markers are central to contemporary glioma classification frameworks. IDH mutation status, determined by sequencing, represents the primary classification criterion in diffuse gliomas, separating IDH-mutant from IDH-wild-type tumours.[18] Among IDH-mutant gliomas, combined whole-arm 1p/19q codeletion is required for the diagnosis of oligodendroglioma.[19] In IDH-wild-type tumours, additional molecular alterations are used to support tumour classification. These include mutations in the TERT promoter[20], which are associated with telomerase activation, amplification of the EGFR gene, which reflects dysregulated growth signalling[21], and a characteristic chromosomal copy number pattern involving gain of chromosome 7 and loss of

chromosome 10.[22] The presence of these alterations can support a diagnosis of glioblastoma even in the absence of classical histological features.

Together, histological assessment and the biomarker profile enable the neuropathologist to diagnose and classify different types of brain tumours. With a definitive diagnosis, a treatment plan can be made that optimises the outcome and prognosis for the patients. However, the neuropathology diagnostic process is complex, labour intensive, time-consuming, costly, and can be subjective and iterative. Therefore, the development of faster, easier to employ, lower cost, and less subjective methods is an attractive avenue to explore.

1.1.2 The role of neurosurgery

Neurosurgery plays a central role in the management of brain tumours and is often the first therapeutic intervention following diagnosis. The aim of neurosurgery is either to completely remove all the tumour, or simply to reduce pressure in the brain. The stakes in the operating theatre are high: remove too little and the tumour is likely to return; remove too much and vital neurological functions may be lost. The fundamental objective of neurosurgery is maximal safe resection (MSR), defined as removal of the greatest achievable volume of tumour while preserving eloquent cortical regions and critical subcortical white matter tracts, thereby minimising postoperative neurological deficits.[23] Extensive evidence indicates that a greater extent of resection is independently associated with prolonged progression-free survival and overall survival across various glioma grades.[24]

1.1.3 Current intraoperative guidance techniques and limitations

Surgical management of brain tumours has progressed from relying mainly on anatomical landmarks to using a range of intra-operative technologies. Modern neuro-oncological surgery employs multiple tools to address two major challenges: the difficulty of visually distinguishing infiltrative tumour tissue from normal brain, and the movement of brain structures that occurs once the skull is opened.

1.1.3.1 Image-Guidance and visualisation: neuronavigation and the operative microscope

Modern neurosurgery relies heavily on neuronavigation systems. These systems align preoperative MRI or CT images with the patient's anatomy in the operating theatre, providing real-time spatial guidance during surgery. Neuronavigation enables more accurate planning of craniotomies and safer access to deep-seated lesions. However, its major limitation is brain shift: movement of brain tissue caused by gravity, loss of cerebrospinal fluid, and tumour removal. As surgery progresses, this displacement can reduce the accuracy of navigation by up to 10 mm.[25, 26]

The operative microscope remains the primary tool for visualisation in brain tumour surgery.[27] It provides strong illumination and magnification, which are essential for distinguishing tumour from white matter and for preserving small blood vessels.[28] Modern microscopes can display neuronavigation information directly within the surgeon's view using augmented reality. However, the microscope can only visualise exposed tissue and provides no information about tumour below the surgical surface.[29]

1.1.3.2 Metabolic targeting: 5-aminolevulinic acid (5-ALA)

To improve identification of infiltrative tumour margins in high-grade gliomas (HGG), 5-aminolevulinic acid (5-ALA) fluorescence-guided surgery is widely used.[30] Following administration, this prodrug is metabolised to protoporphyrin IX, which preferentially accumulates in malignant cells. When illuminated with blue-violet light ($\approx 400\,nm$), tumour tissue emits a red-pink fluorescence that is not visible under standard white light.

This technique supports maximal safe resection by highlighting tumour regions that may appear macroscopically normal. A phase III randomised trial by Walter Stummer *et al.* showed that 5-ALA increased the rate of gross total resection from 36% to 65%, with an associated improvement in progression-free survival.[31]

However, its effectiveness is largely confined to high-grade gliomas. In low-grade gliomas, reduced metabolic activity results in weak or absent fluorescence.

1.1.3.3 Brain tissue smear

The current gold standard for intraoperative diagnosis is called intraoperative brain tissue smear. A small tissue sample is cut into thin sections, stained, and examined under a microscope by a neuropathologist. This method shows overall accuracy of around 93-95%. [32–34] Despite this performance, the technique has several important limitations. The workflow is time-consuming, with turnaround times commonly between 20 and 40 minutes, which can interrupt surgical decision-making. It is also labour- and resource-intensive, requiring specialised facilities and the immediate availability of expert personnel. Most critically, histological assessment is based solely on morphological features, which provides limited direct biochemical information, limiting its ability to capture tumour heterogeneity or subtle biochemical differences that may be clinically relevant.

1.1.3.4 Awake surgery and neurophysiological monitoring

Awake surgery with direct electrical stimulation (DES) is widely used as the gold standard for diffuse low-grade glioma.[35] By keeping the patient awake during tumour resection, specific cortical and subcortical regions can be tested in real time. Stimulation-induced speech arrest or naming errors indicate functionally critical tissue that must be preserved. Awake surgical techniques have led to substantial improvements in onco-functional outcomes in patients with IDH-mutant grade 2 gliomas. Large series from specialised centres report low rates of permanent neurological deficit (below 2%), high preservation of cognitive function, and return-to-work rates approaching 90%, alongside prolonged survival rates.[36]

Motor pathways are monitored during surgery using intra-operative neurophysiological monitoring (IONM), typically through motor evoked potentials (MEPs) and somatosensory evoked potentials (SSEPs). These techniques provide continuous feedback on the functional integrity of the corticospinal tract during resection, allowing early detection of potentially irreversible injury.

However, IONM is sensitive to depth of anaesthesia and physiological conditions. While it is effective for preserving motor function, it offers limited insight into

higher-order cognitive, behavioural, or emotional processes, which remain difficult to monitor intra-operatively.[37]

1.1.3.5 Intra-operative MRI

To address brain shift and to assess the extent of resection before wound closure, intra-operative MRI (iMRI) was developed. By acquiring updated imaging during surgery, iMRI allows surgeons to reassess tumour boundaries and detect residual disease that may no longer correspond to preoperative navigation data.

Randomised and observational studies have shown that iMRI-guided surgery can increase rates of gross total resection compared with conventional neuronavigation alone.[38]

However, iMRI has substantial practical limitations. The systems are expensive, prolong operative time, and require specialised magnet-compatible instruments and operating theatres. These constraints have limited adoption to a small number of high-resource centres.[39]

Taken together, current intra-operative adjuncts provide either anatomical guidance, surface-level visual contrast, or rapid cytological assessment. What they do not provide is direct, real-time molecular information from tissue at the surgical margin.

1.1.4 Analytical approaches for brain tissue classification

A growing number of analytical techniques aims to classify brain tissue based on its biochemical and metabolic composition. Rather than relying solely on visual or anatomical cues, these approaches use molecular signatures to distinguish tumour from normal brain and to identify different brain tumour subtypes. In principle, this shifts surgical decision-making toward biologically informed resection, particularly at infiltrative margins where visual contrast is poor.

However, the value of an intraoperative analytical method cannot be judged by classification accuracy alone. A clinically useful technique must also provide results within a surgically relevant time window, require minimal sample preparation,

tolerate heterogeneous and contaminated tissue, preserve spatial or anatomical context where possible, and generate outputs that can be interpreted or acted upon by the surgical and neuropathology teams.[40] These requirements create different strengths and limitations for spectroscopy, mass spectrometry, and sequencing-based approaches.

1.1.4.1 Raman Spectroscopy

Vibrational Raman spectroscopy is an optical scattering technique that measures the vibrational modes of molecules within the tissue. By generating a distinct spectral fingerprint based on the chemical composition, Raman spectroscopy can not only distinguish tumour from normal brain with high sensitivity [41], but is also able to distinguish different genetic subtypes of gliomas. [42] However, clinical implementation is often difficult because of weak signal intensities and the complexity of the data. Recent advancements have bridged this gap. Hollon and Orringer (2020) demonstrate that Stimulated Raman Histology (SRH) can now generate rapid, label-free virtual histology images intraoperatively. [43] This is a major translational advantage because it converts spectroscopic contrast into an image format that is familiar to pathologists. However, this strength also defines the limitation of the approach: SRH primarily provides histology-like morphological information with chemical contrast, rather than direct metabolomic or lipidomic profiling. Therefore, Raman-based approaches are powerful for rapid tissue visualisation and tumour infiltration assessment, but are less suited to detailed molecular identification unless combined with additional analytical or computational methods.

1.1.4.2 Mass spectrometry

Mass spectrometry is an analytical technique used to identify and quantify chemical species based on their mass-to-charge ratio. Further technical details will be given in Section 1.2.

Several mass spectrometry techniques have been investigated for use during neurosurgical procedures. Shahi *et al.* have recently provided a comprehensive overview of ambient and rapid mass spectrometry techniques that have been

investigated for brain cancer diagnosis, particularly in intraoperative or near-intraoperative contexts.[44] The review surveys multiple approaches, including desorption electrospray ionisation (DESI-MS), rapid evaporative ionisation mass spectrometry (REIMS/iKnife)[45], laser-based techniques such as SpiderMass[46], and liquid extraction approaches such as the MasSpec Pen.[47] These techniques share the ability to generate molecular information from tissue with minimal or no sample preparation and short analysis times, making them potentially compatible with surgical workflows.

Across multiple studies, the review highlights the fact that most methods rely on lipid- and metabolite-based molecular profiles to distinguish tumour from non-tumour tissue and, in some cases, to infer tumour subtype or molecular status.[44] While many reports demonstrate high classification performance under experimental conditions, the review emphasises that clinical translation remains challenging.[44] Key limitations include tissue heterogeneity, variability in sampling conditions, blood and irrigation fluid contamination, model generalisability, and the need for robust reference data sets.[44]

1.1.4.3 Nanopore DNA sequencing

On the genetic diagnosis side, Oxford Nanopore sequencing has enabled rapid DNA analysis and represents a promising route toward intraoperative molecular profiling. Devices such as the MinION allow for the detection of copy number variations (CNVs) and DNA methylation profiles within hours. Recent studies have demonstrated that low-pass nanopore sequencing can generate a methylation-based classification of brain tumours that aligns with final histopathology[48, 49]. While rapid protocols exist, achieving a comprehensive molecular profile often requires sample preparation, which takes at least 60 minutes.[50] Consequently, reported intraoperative turnaround times can range significantly, with early feasibility studies showing a median time to result of 97 minutes and some workflows requiring up to 24 hours.[50] Consequently, current nanopore-based workflows cannot realistically be completed within the time constraints of an ongoing neurosurgical procedure.

Furthermore, scaling the technology to achieve the high throughput can also put a financial burden on health care systems.

Overall, no single analytical approach satisfies all requirements for intraoperative brain-tissue classification. Raman and SRH provide rapid, label-free optical assessment and can preserve morphological context, but offer limited direct molecular identification. Mass spectrometry provides richer biochemical information, especially from lipid and metabolite profiles, but faces challenges from sampling variability, contamination, matrix effects, and model transfer. Nanopore sequencing provides the most direct genetic and epigenetic information, but is slower and less suited to repeated real-time margin assessment.

Although a mass spectrometry-based molecular profiling approach does not fully satisfy all criteria for an ideal intraoperative tool, it was chosen in this study because it provides rapid access to chemically rich information from brain tissue. Its limitations, including sampling variability, matrix effects, contamination, and incomplete molecular identification, are substantial; however, these are also the central challenges this thesis seeks to address in developing a more practical and reproducible workflow for brain-tissue classification.

1.1.5 Brain tissue sample types

The molecular information accessible from brain tissue is strongly influenced by how the tissue is collected, processed, and preserved. In the context of rapid molecular profiling and data-driven classification, differences between fresh, frozen, and formalin-fixed paraffin-embedded (FFPE) tissue have direct impact on both spectral characteristics and downstream interpretation.

1.1.5.1 Fresh tissue sample

Fresh brain tissue refers to specimens analysed immediately after surgical resection without intentional preservation or stabilisation. Such samples have not undergone freezing, chemical fixation, or dehydration, and therefore most closely reflect the biochemical state of the tissue at the time of removal. However, fresh samples

are chemically unstable, subject to rapid enzymatic activity and degradation, and require immediate handling under controlled conditions.

1.1.5.2 Frozen tissue sample

Frozen brain tissue is obtained by rapidly cooling freshly resected specimens, typically using liquid nitrogen or dry ice, to arrest biological activity and preserve molecular composition. Freezing substantially improves sample stability by inhibiting enzymatic degradation and slowing chemical reactions, while retaining a molecular profile that remains closer to the native tissue state than chemically fixed samples. For this reason, frozen tissue is widely used in molecular pathology and mass spectrometry–based research. Nevertheless, freezing and thawing are not chemically inert processes.

1.1.5.3 FFPE tissue sample

Formalin-fixed paraffin-embedded (FFPE) tissue is produced by chemically fixing tissue in formaldehyde, followed by dehydration and infiltration with paraffin wax for long-term storage. This preparation is the standard format for routine histopathological diagnosis and enables archival storage of tissue samples over many years. However, fixation and embedding fundamentally alter tissue chemistry. Formaldehyde induces extensive cross-linking between biomolecules, while dehydration and paraffin embedding remove or redistribute many low-molecular-weight compounds, including a large fraction of lipids and metabolites. Additionally, deparaffinisation is an essential step before attempting any chemical analysis on FFPE samples, introducing an additional source of variability. As a consequence, the molecular information accessible from FFPE tissue differs substantially from that of fresh or frozen samples and includes fixation-induced modifications and deparaffinisation artefacts.

1.2 Atmospheric solids analysis probe mass spectrometry

1.2.1 Fundamentals of mass spectrometry

Mass spectrometry (MS) is a powerful analytical technique that identifies and quantifies substances by measuring the mass-to-charge ratios m/z of their intact parent and/or fragment ions. It works by ionising a gas-phase sample, separating the resulting ions via their m/z-dependent trajectories in an electric or magnetic field, and then detecting the separated ions as a function of time, field of strength or some other variable to generate a mass spectrum.[51, 52] As shown in Figure 1.2, a mass spectrum is a plot of ion abundance (ion count) vs. m/z. It is used to reveal molecular weight, composition, and structure for unknown or known samples in various fields such as medicine, biology, and chemistry. The origins of MS date back to the late 19th and early 20th centuries.[53, 54]

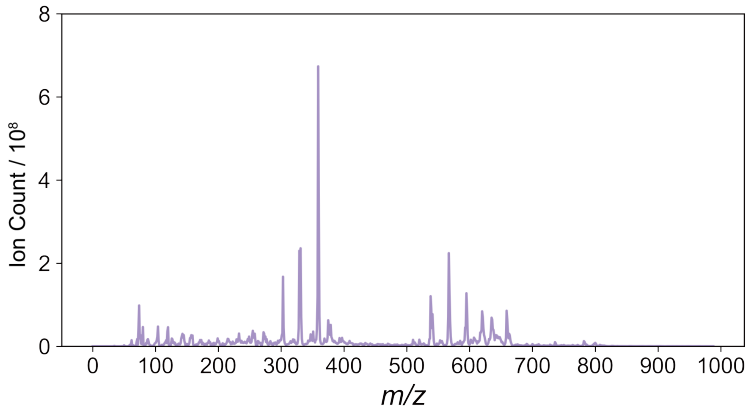

Figure 1.2: An example of a mass spectrum

A mass spectrometer, regardless of its complexity, is composed of four essential modules:

- Sample introduction system: Introduces the sample into the instrument and, where necessary, converts it to the gas phase.
- Ion source: Converts neutral sample molecules into ions.

- Mass analyser: The part of the instrument where ions are separated based on their m/z.
- Detector: Records the arrival of ions in the form of an electrical signal.

For decades, the standard workflow relied heavily on hyphenated techniques in which a chromatographic separation precedes mass spectrometric detection:

- Gas Chromatography–Mass Spectrometry (GC–MS): used for volatile, thermally stable compounds. In GC, analytes are separated in the gas phase based on their interaction with a stationary phase in a capillary column, typically reflecting differences in volatility and column affinity, before sequentially entering the mass spectrometer.[55]
- Liquid Chromatography–Mass Spectrometry (LC–MS): used for polar and non-volatile compounds. In LC, analytes are separated in the liquid phase according to their chemical interactions with the stationary phase (e.g. hydrophobicity in reversed-phase LC), enabling individual components to be introduced into the mass spectrometer at different retention times.[56]

While these methods are robust, they share a common 'bottleneck': sample preparation. Traditional MS requires samples to be purified, extracted, or derivatised, and then introduced into a high-vacuum environment. This process is time-consuming, destructive to the sample matrix, and often requires significant solvent consumption.

1.2.2 Ambient ionisation mass spectrometry

In contrast to traditional ESI, in which analytes are introduced as a prepared solution and ionised through charged-droplet formation from a spray capillary, ambient ionisation refers to a class of ionisation techniques in which ions are generated from untreated or minimally prepared samples under open-air, atmospheric-pressure conditions, typically without prior chromatographic separation.[57–60] The development of electrospray ionisation (ESI)[61] and atmospheric pressure chemical ionisation (APCI)[62] in the 1980s transformed the discipline and paved

the way for practical ambient ionisation. This bridged the gap between liquid-phase chemistry and gas-phase physics. However, even with ESI and APCI, samples still required a delivery system (such as a syringe pump) to reach the ionisation source. In the 2000s, the introduction of desorption electrospray ionisation (DESI) demonstrated that charged droplets could desorb and ionise analytes directly from solid surfaces.[63] Shortly thereafter, direct analysis in real time (DART) showed that gas-phase metastable species could be used to ionise compounds in open air.[64] According to the generally accepted definition[59], an ambient ionisation technique must satisfy three criteria:

- Operation at atmospheric pressure: No vacuum interface is required for the sample itself.
- Direct analysis: The sample is analysed in its natural environment (e.g., a tablet, a leaf, or a piece of fabric).
- Desorption/ionisation in one step: The process of removing the analyte from the surface and ionising it happens near-simultaneously through the interaction of an external agent (gas, liquid, or plasma).

The principal advantage of ambient ionisation methods over chromatographic MS workflows is speed: samples can be analysed directly or with minimal preparation. The corresponding limitation is that, without chromatographic separation, complex tissue matrices enter the ion source simultaneously, increasing the risk of ion suppression, ion enhancement, and overlapping signals. Chromatography does not eliminate matrix effects entirely, but it can substantially reduce them by separating analytes from interfering compounds before ionisation.

Subsequent developments expanded the range of ionisation mechanisms, analyte classes, and sampling geometries. Ambient techniques have since been applied to a wide variety of materials, including biological tissues, biofluids, pharmaceuticals, food products, and forensic samples.[65] Their ability to generate chemically rich information rapidly has driven growing interest in clinical and translational applications in recent years.[66]

Table 1.1: Overview of widely used ambient mass spectrometry techniques

Method	DESI	PS	DART	REIMS	ASAP
Sampling method	Direct surface irradiation by charged solvent spray	Sample deposited onto paper substrate	Sample held in open air within gas stream	Electrosurgical cutting of tissue (iKnife)	Sample touched or dipped with a glass capillary
Desorption mechanism	Liquid extraction	Liquid extraction	Energy transfer from heated, plasma-generated gas species	Joule heating during electrosurgical cutting	Thermal desorption in hot nitrogen stream
Ionisation mechanism	Electrospray ionisation (ESI)	Electrospray ionisation (ESI)	APCI via corona discharge	APCI-like ionisation of aerosolised tissue	APCI via corona discharge
Gas	N_2 (nebulising / drying)	None	He (classical) or N_2	N_2 or air (aerosol transport)	N_2
Destructiveness	Minimal–moderate (surface altered)	Low; chemically altered by solvent extraction	Minimal–moderate	Fully destructive	Minimal–moderate
Typical strengths	Molecular imaging; lipid localisation	Simple hardware; low cost; rapid biofluid analysis	Rapid screening of small molecules	Real-time intraoperative feedback	Rapid bulk sampling; fingerprinting scalability
Structural limitations	Solvent spreading; strong surface and matrix effects	Paper and solvent dominate chemistry	Limited control of sampled volume; low spatial specificity	Severe thermal artefacts; smoke chemistry variability	No spatial resolution; thermal fragmentation
Instrument price	£500k–900k	£10k–50k (source only)	£100k–200k	£350k–450k	£80k–100k

Details of several ambient mass spectrometry techniques are summarised in Table 1.1. For many applications, DESI and paper spray (PS) are currently the two most popular methods, possibly because they are conceptually simple and enable soft ionisation through solvent-mediated extraction[67–69]. REIMS has gained attention for its ability to provide immediate feedback during surgery by directly analysing thermally generated tissue aerosol.[45] DART has been widely adopted for rapid screening because it requires minimal sample preparation and no physical contact with the sample.[70]

Among these methods, ASAP–MS is the least widely adopted ambient MS techniques summarised here[67], but it occupies a distinct and under-explored position. DESI uses charged solvent droplets directed onto a sample surface to desorb and ionise analytes under ambient conditions, whereas REIMS analyses aerosol or vapour generated during rapid thermal evaporation of tissue, such as by electrosurgical or laser-based sampling. In contrast, ASAP–MS directly introduces material on a probe into the ion source, where analytes are thermally desorbed and subsequently ionised by APCI. In contrast to paper spray, ASAP–MS avoids paper- and solvent-driven chemistry that can dominate the measured signal. Although ASAP–MS is less well established in the literature, especially in clinical studies, the combination of low operational cost, direct sampling, and simplicity makes it an attractive method to explore.

1.2.3 ASAP–MS Instrumentation

ASAP–MS works by introducing a sample into the ion source on the tip of a handheld probe, where it is then ionised through an APCI process.[71] The core components of the ion source are a removable glass sampling probe, a heated nitrogen gas stream, and a corona discharge. The source is followed by a hexapole ion guide, a single quadrupole mass analyser, and an electron multiplier detector. The ASAP–MS used in this study is an Advion express**ion**® compact mass spectrometer (CMS-L) equipped with an ASAP–MS source (Advion Ltd., Harlow, United Kingdom).[72]

1.2.3.1 Ionisation

In atmospheric solids analysis probe mass spectrometry (ASAP–MS), ionisation is achieved using atmospheric pressure chemical ionisation (APCI).[72] As shown in Figure 1.3, the sample is thermally desorbed and carried in a heated gas stream to the corona discharge region. In ASAP–MS, the heated nitrogen stream promotes rapid thermal desorption of analytes from the sample-loaded probe into the atmospheric-pressure ionisation region. The source temperature, which may reach approximately 400 °C, should not be interpreted as the equilibrium temperature experienced by all analyte molecules. Instead, analyte transfer occurs over a short residence time and is governed by local probe heating, analyte volatility, and thermal stability. Consequently, ASAP–MS preferentially detects compounds, or fragments of compounds, that can survive desorption and subsequent APCI conditions. Although APCI is generally regarded as a relatively soft ionisation method, thermally labile or strongly non-volatile species may undergo in-source decomposition, neutral loss, or incomplete desorption before ionisation. Therefore, signals observed in the higher-m/z lipid region should be interpreted cautiously as lipid-related molecular ions, adducts, fragments, or thermal degradation products, rather than as unequivocal intact phospholipid molecular ions without further structural confirmation by accurate-mass MS/MS.

During APCI operation, ionisation proceeds via gas-phase chemical reactions initiated by the corona discharge. In positive ion mode, the discharge produces protonated water clusters, predominantly hydronium ions H_3O^+, from the surrounding atmospheric gas. Proton transfer from hydronium to a neutral analyte molecule in the gas phase then generates the protonated analyte:

$$H_3O^+ + M \rightarrow MH^+ + H_2O \tag{1.1}$$

In negative ion mode, ionisation occurs mainly through proton abstraction from the analyte by basic reagent ions, producing deprotonated molecules:

$$B^- + MH \rightarrow M^- + BH \tag{1.2}$$

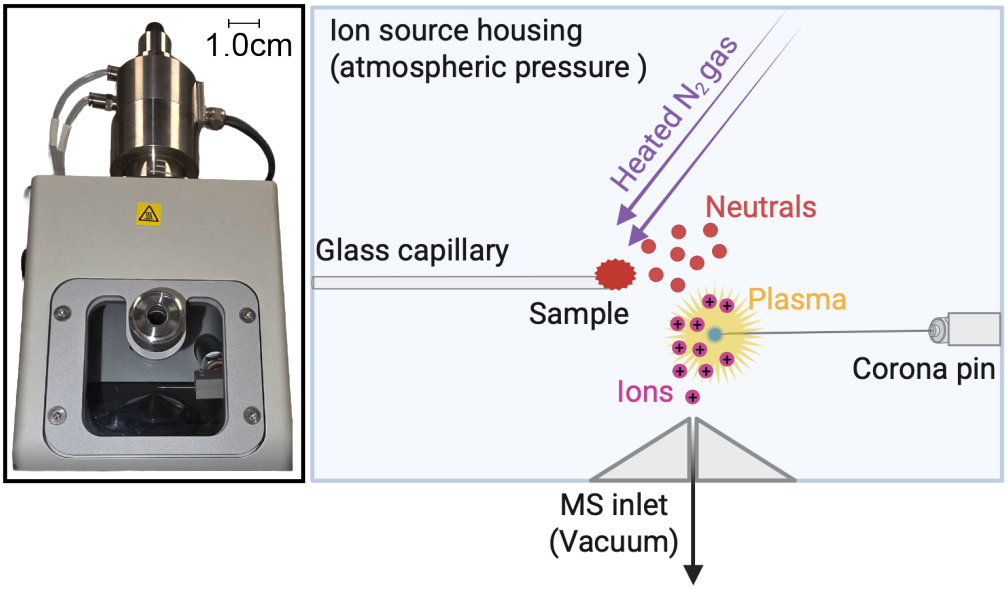

Figure 1.3: ASAP ionisation source housing of the Advion CMS-L instrument with a photograph reference, showing the APCI ionisation region into which thermally desorbed analytes from the probe are introduced.

Following ionisation, the charged analyte ions are guided by electrostatic fields into the mass analyser.

1.2.3.2 Ion guide and mass analyser

Following ionisation, ions are extracted from the source region and enter the hexapole ion guide (Figure 1.4(a)) via the extraction lens.[72] The hexapole region is differentially pumped to an intermediate pressure of approximately 5×10^{-3} mbar. This elevated pressure relative to the analyser region promotes collisional cooling, which reduces the kinetic energy of the ions and dampens their radial motion. As a result, ion trajectories collapse towards the central axis of the ion guide, improving transmission efficiency into the downstream mass analyser.

The hexapole is operated using a radiofrequency voltage scanned dynamically between 0 and 500 V, with an applied direct current bias of approximately $\pm 10V$. This combination of radiofrequency confinement and direct current offset enables efficient ion focusing and transfer from the atmospheric pressure interface into the high-vacuum region of the instrument.

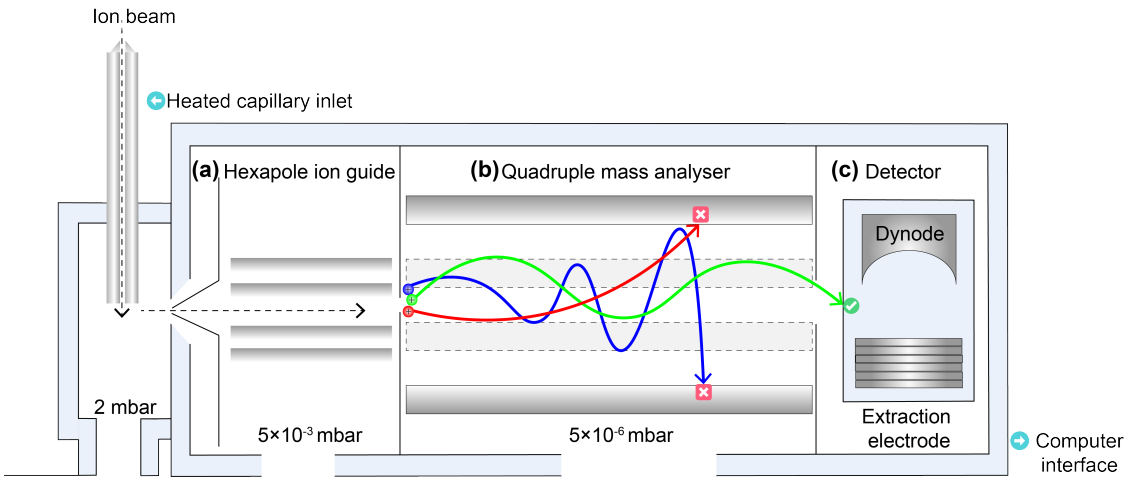

Figure 1.4: Schematic of the hexapole ion guide, quadrupole mass analyser, and detector of the Advion CMS-L instrument. Ions of different m/z experience distinct trajectories within the quadrupole electric fields, such that only ions with stable trajectories (green) are transmitted to the detector. Adapted from [72, 73].

Ions exiting the hexapole region are subsequently introduced into the quadrupole mass analyser (Figure 1.4(b)).[72] The quadrupole consists of four parallel rods to which radiofrequency and direct current voltages are applied. These combined electric fields create conditions under which only ions with a specific m/z follow stable trajectories through the quadrupole at any given time. By scanning the radiofrequency and direct current voltages, ions of different m/z values are sequentially transmitted to the detector, enabling mass spectral acquisition.

Besides a single quadrupole mass analyser, other mass spectrometer types are also compatible with ASAP, as ASAP is primarily an atmospheric-pressure sample introduction and ionisation approach rather than an analyser-specific technique. ASAP has been coupled to triple-quadrupole[74], time-of-flight (ToF)[75], quadrupole time-of-flight (Q-ToF)[76], and high-resolution Orbitrap instruments[77]. Quadrupole-based systems are well suited to rapid, low-cost molecular fingerprinting and targeted monitoring, whereas ToF, Q-ToF, or Orbitrap platforms provide higher mass resolution and mass accuracy, supporting elemental composition assignment, metabolite annotation, and MS/MS-based structural interpretation. Therefore, the choice of mass analyser depends mainly on the analytical objective: nominal-mass profiling and classification can be achieved using compact quadrupole instruments,

while high-resolution analysers are used when accurate-mass information or molecular identification is required.

1.2.3.3 Detector

After mass filtering, ions exiting the quadrupole are detected using an electron multiplier detector equipped with a high-energy conversion dynode (Figure 1.4(c)).[72] The polarity of the conversion dynode is switched depending on the ionisation mode. In positive ion mode, the conversion dynode is biased at approximately −10 kV, whereas in negative ion mode it is biased at approximately +10 kV. When an ion strikes the conversion dynode, secondary electrons are generated. These electrons are directed into the electron multiplier, where successive impact events produce a cascading amplification effect. The resulting amplified electron current is collected at a detection plate, converted from an analogue signal to a digital signal, and recorded as ion intensity by the instrument software to generate the mass spectrum.

The electron multiplier gain is adjustable over a typical voltage range of 0–2000 V, allowing optimisation of sensitivity while maintaining stable detector performance.

1.2.3.4 Instrument settings

The instrument settings for measuring biological samples have been systematically optimised by another group member, Annabel Eardley-Brunt, and subsequently adapted to the present study. These are shown in Table 1.2.[73, 78]

ASAP–MS acquisition was controlled via the Advion Mass Express software (version 6.9.38.1).[79] For biological sample analysis, the ASAP–MS ion source was operated in positive ion mode with the setting "High Temperature, Low Fragmentation".[72] During ASAP–MS analysis, samples were first thermally desorbed from the probe into the APCI region using heated nitrogen gas set to 400°C, a temperature chosen to ensure efficient volatilisation of complex biological compounds, particularly metabolites and lipids, while maintaining stable source operation. Ionisation was achieved in positive ion mode using a corona discharge current of 5 μA, which provides a stable population of reagent ions for proton-transfer reactions without inducing excessive in-source fragmentation. Immediately after ion formation, ions

Table 1.2: ASAP–MS settings

Scan settings	
Starting m/z	10
Finishing m/z	1000
Scan time / ms	900
Scan delay / μs	100
Smoothing radius	7
Smoothing iterations	0
Remove noise	TRUE
Ion source settings	
Polarity	Positive
Capillary temperature / °C	250
Capillary voltage / V	120
Source voltage offset / V	20
Source voltage span / V	0
Source gas temperature / °C	400
Transfer line temperature / °C	100
APCI corona discharge / μA	5

were extracted from the source and directed towards the vacuum interface through a heated inlet capillary held at 250°C. This temperature supports desolvation and prevents condensation of analytes. A capillary voltage of 120 V was applied to draw ions effectively into the ion optics and maintain stable transmission. As ions entered the ion guide region, a source voltage offset of 20 V was applied, increasing ion kinetic energy in the relatively higher-pressure transfer region and promoting controlled collisional activation. This setting was selected to balance ion transmission with limited collision-induced fragmentation, particularly for higher m/z ions. The source voltage span was set to 0 V to maintain a uniform electric field across the source region and avoid unnecessary destabilisation of ion trajectories. The transfer line was maintained at 100°C to minimise ion losses due to condensation prior to mass analysis. Finally, ions were introduced into the single quadrupole mass analyser, where the radiofrequency and direct current fields selectively stabilised ions of a given mass-to-charge ratio for transmission to the detector.

1.2.4 Data pre-processing

After the measurement, raw ASAP–MS data were grouped into unit m/z bins[1] and exported as CSV files using the Advion Data Express data manipulation software (version 6.9.38.1).[79] The raw data were processed using custom scripts written in Python 3.7. Files were imported using Python package `pandas`.

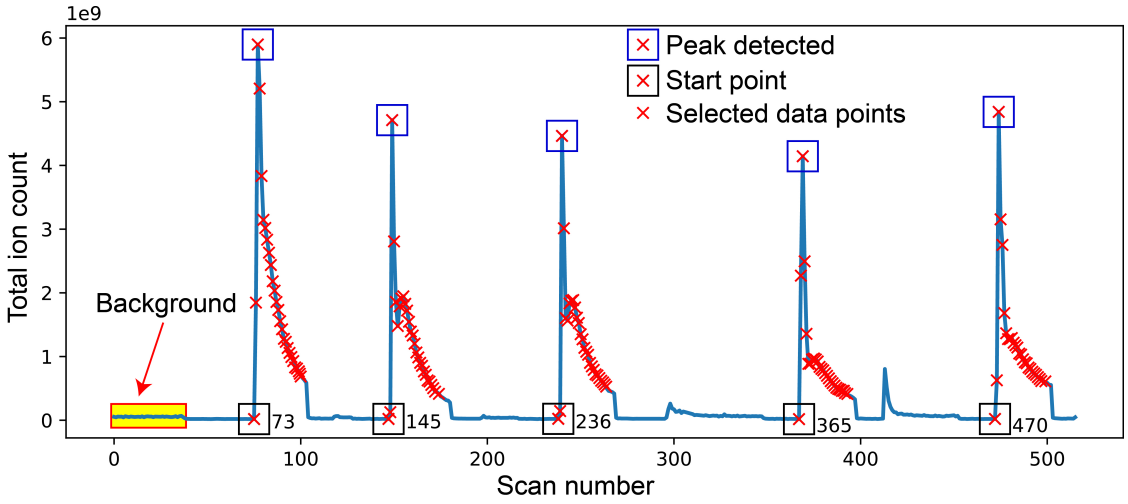

Figure 1.5: Representative chronogram from five measurements of a sample to illustrate how our Python software detects the start point of each measurement.

Figure 1.5 shows a representative chronogram plot of total ion count (TIC) vs time from an acquisition involving five replicate measurements of a sample. The TIC was calculated by summing ion counts across all m/z bins for each scan to monitor signal evolution over scan number and to assist with identifying the onset of sample-related signal. Signal onset was identified through peak detection using the function `find_peaks`, with a user specified a prominence threshold and minimum peak spacing. For each detected peak, the corresponding start point was determined by tracing backwards to the point where the signal gradient changed from positive to non-positive. These start points were used to define the beginning of each measurement.

For each sample measurement i, a continuous series of scans s acquired over the selected measurement period identified by the red crosses in Figure 1.5 was

[1]ASAP-MS relies on a single quadrupole mass analyser with relatively low resolution, so binning is essential to sharpen the data and power the downstream analysis.

averaged to obtain a representative spectrum, denoted as $I_{i,j}^{\text{raw}}$. Background spectra B_j were estimated using an initial set of scans acquired prior to sample introduction, denoted as $((I_{s,j}^{\text{bg}}))$.

For each ion peak j, the background mean $\mu_{\text{bg},j}$ and standard deviation $\sigma_{\text{bg},j}$ were calculated $(\sigma_{\text{bg},j})$ as:

$$\mu_{\text{bg},j} = \frac{1}{|S_{\text{bg}}|} \sum_{s \in S_{\text{bg}}} I_{\text{bg},s,j} \tag{1.3}$$

$$\sigma_{\text{bg},j} = \sqrt{\frac{1}{|S_{\text{bg}}| - 1} \sum_{s \in S_{\text{bg}}} \left(I_{\text{bg},s,j} - \mu_{\text{bg},j}\right)^2} \tag{1.4}$$

A background threshold[78] was then defined for each ion peak as:

$$B_j = \mu_{\text{bg},j} + 3\,\sigma_{\text{bg},j} \tag{1.5}$$

Background subtraction was applied to each raw mass spectrum:

$$I_{i,j} = \max\Big(0,\, I_{i,j}^{\text{raw}} - B_j\Big) \tag{1.6}$$

Normalisation In ASAP–MS, the amount of sample introduced into the ion source cannot be precisely controlled, leading to variability in overall ion counts between measurements. To account for such variability, mass spectrometric data are commonly normalised prior to analysis. Normalisation rescales the ion counts to a common reference, allowing meaningful comparison between measurements by reducing non-biological variability. Our group has systematically evaluated the impact of different normalisation strategies on ASAP–MS mass spectra[80] (See also an additional method described in Appendix D.1), and has demonstrated that total ion count (TIC) normalisation is sufficient for the ASAP–MS analysis.

TIC normalisation scales each m/z peak intensity by the total ion counts in the spectrum and is defined as:

$$N_{i,j}^{\text{TIC}} = \frac{I_{i,j}}{\sum_j I_{i,j}} \tag{1.7}$$

In the following, normalisation refers to TIC normalisation unless otherwise stated.

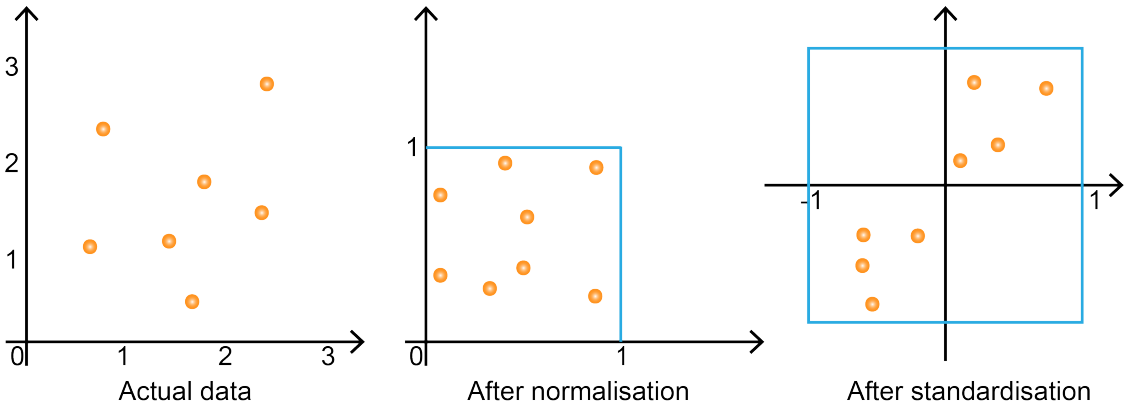

Figure 1.6: Comparison between normalisation and standardisation, adapted from[81]

Standardisation After normalisation, further scaling of the data is often required to ensure that all variables contribute comparably to downstream analysis. A comparison between normalisation and standardisation is shown in Figure 1.6. Standardisation rescales each feature to have zero mean and unit variance across the data set. Unlike normalisation, which corrects for differences in overall signal intensity between samples, standardisation operates at the feature level to account for differences in scale between individual variables. The standardised signal for peak j in the spectrum for sample i is defined as:

$$Z_{i,j} = \frac{I_{i,j} - \mu_j}{\sigma_j} \tag{1.8}$$

where μ_j and σ_j are the mean and standard deviation of feature j across all samples.

After standardisation, for a given sample, m/z peaks with intensities above the average appear as positive values, while those below the average appear as negative values. Dividing by the standard deviation reduces the influence of low-intensity, noise-dominated peaks and prevents a small number of high-intensity features from dominating model training, although it may also reduce the influence of biologically variable peaks.[81] In the present study, standardisation was applied during data pre-processing for supervised machine learning analyses.

1.3 Machine learning

Machine learning (ML) is a branch of artificial intelligence (AI) in which models are trained to learn patterns and relationships from complex data.[82] These ML models are capable of making predictions or decisions without intense programming of pre-defined rules. ML is well suited to mass spectrometry data, where diagnostically relevant information is high-dimensional, complex, and distributed across many spectral features.[83]

1.3.1 Terminology

Throughout this thesis, standard machine-learning terminology is used. For clarity, key terms are defined below.

- Feature: An individual measured variable used as input to a machine-learning model. In this study, features correspond to m/z peak intensities extracted from ASAP–MS spectra.
- Label: The ground-truth category assigned to a sample, derived from the corresponding neuropathological diagnosis and used during supervised model training.
- Class: One of the discrete categories that a model is trained to predict, such as tumour or normal tissue.
- Sample: A single mass spectrum, or an averaged set of replicate spectra, acquired from a specific tissue specimen.
- Hyperparameters: hyperparameters are tunable parameters that form part of the machine learning algorithm itself. These are fixed before the machine learning process begins, and control how the algorithm behaves. Choosing the right hyperparameters is the key to building successful machine learning models.[84]

1.3.2 Unsupervised machine learning

Unsupervised machine learning refers to a class of methods that identify structure, patterns, or relationships in data without reference to predefined class labels.[85] In this study, two unsupervised machine learning methods: Principal component analysis and K-Means, were used.

1.3.2.1 Principal Component Analysis

Principal component analysis (PCA) is a linear dimensionality reduction technique that transforms high-dimensional data into a smaller set of orthogonal components ordered by the amount of variance they explain.[86] Each principal component represents a linear combination of the original variables, enabling correlated spectral features to be summarised into a reduced number of dimensions. An example is shown in Figure 1.7

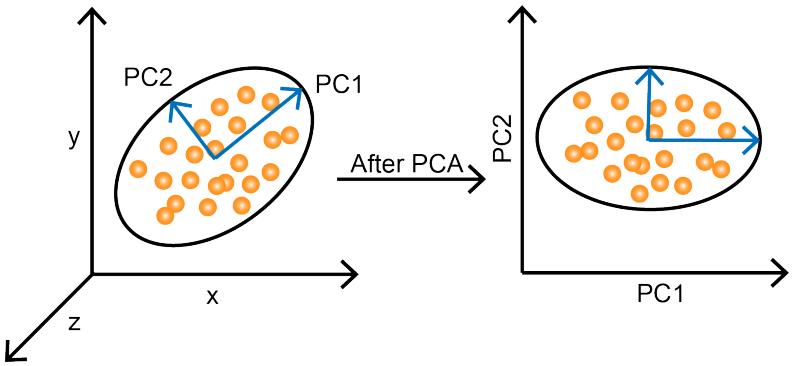

Figure 1.7: Illustration of principal component analysis (PCA). The original high-dimensional feature space (x, y, z) is projected onto principal components (PC1 and PC2).

In mass spectrometry data, PCA is commonly used to visualise global variation, identify dominant trends, and detect batch effects or technical artefacts. Importantly, PCA is unsupervised and variance-driven: the directions of maximal variance do not necessarily correspond to class-discriminative features. As such, PCA is used in this study as an exploratory and diagnostic tool rather than a classifier.

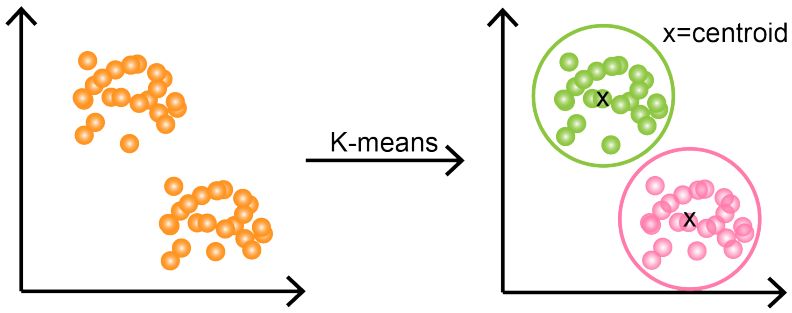

Figure 1.8: Illustration of K-means clustering for $K = 2$. Data points are assigned to two clusters based on the cluster centroids, which are indicated by x.

1.3.2.2 K-means clustering

As illustrated in Figure 1.8, K-means clustering is a method used to find natural groupings in data without prior labelling.[87] It works by randomly picking a user-defined number of starting points K, called centroids, and grouping each data point with its nearest centre. The algorithm then repeatedly adjusts these centres to the 'centre of mass' of their respective groups, and redetermines the data point assignments until the clusters are stable.

1.3.3 Supervised machine learning

Supervised machine learning refers to a class of machine learning methods in which models are trained using labelled data.[88] During training, the model learns a mapping between input features and target labels by minimising a measure of disagreement between its predictions and the true labels, often expressed through a loss or error function. This process enables the model to generalise learned relationships to previously unseen data.

In machine learning, the process of building a model involves splitting data into different groups, training the models as a subset of the data, and fine-tuning the model's hyperparameters to ensure it gives the best possible predictions when presented with data from new, unseen samples.

Training and Test Sets To evaluate a model fairly, the data set is typically split into two parts:

- Training set: This is the portion of the data that the model 'studies' in order to learn patterns.

- Test set: This is a separate portion of data that the model never sees during training. It is used to evaluate the final performance of a trained machine learning model.

Overfitting and Underfitting The goal of machine learning is to discover general patterns within the data set without getting 'distracted' by random noise in the data.

- Overfitting: overfitting occurs when a model is too complex and starts to fit the noise. In this case, the model may fit the training data perfectly, including the random noise, but fails when used to make predictions on the test data.

- Underfitting: underfitting occurs when a model is too simple, and fails to find the underlying patterns in both the training and test sets.

To develop a brain tissue classifier using ASAP–MS mass spectra, supervised learning provides the framework for training models that can distinguish brain tissue types based on complex spectral patterns. In this thesis, six supervised machine learning models based on different algorithmic principles were employed. These are introduced below.

1.3.3.1 k-Nearest Neighbours

The k-Nearest Neighbours (KNN) algorithm classifies a sample based on the labels of the most similar training points.[89]

In KNN, each sample is represented as a vector in $n-$dimensional feature space. For mass spectrometry data, n corresponds to the number of m/z features retained from each spectrum after preprocessing. Similarity between samples is therefore quantified as a distance between these feature vectors. Both methods rely on pairwise distances computed using a defined metric, most commonly the Euclidean distance in this study.

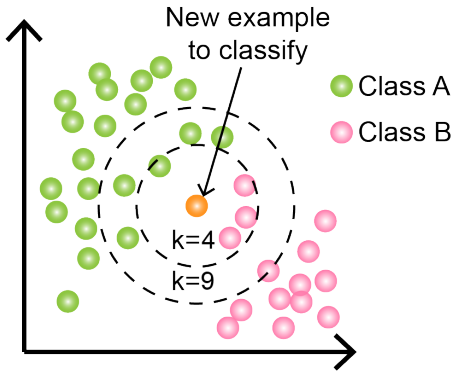

Figure 1.9: Illustration of k-nearest neighbours (KNN) classification. The predicted class for a new sample depends on the choice of k, with majority voting for class B when $k = 4$ and class A when $k = 9$.

Given a new sample x^*, the distance between x^* (represented by an orange circle) and every training sample is computed in feature space. The k data points, with the smallest distance from x^* are identified as the "nearest neighbours". The data point is then assigned to the class to which the majority of these neighbours belong. In the example shown in Figure 1.9, when $k = 4$, x^* is classified as class B, while when $k = 9$, classified as class A.

However, in a high-dimensional space (too many features), the distance between all points starts to look almost the same. If every sample is far from every other sample, the concept of a nearest neighbour becomes rather meaningless.[90] PCA is then used to reduce the dimensionality down so that real similarities can be measured again.

Hyperparameters The behaviour of KNN depends on several hyperparameter choices:

- k (number of neighbours): too small can cause highly variable, noise-driven predictions; too large can cause class boundary blurring.
- PCA components: defines the dimensionality of the data point vectors after dimension reduction using PCA

- Weighting scheme: uniform (assigns equal weights to all points.) vs. distance-weighted voting (assigns higher weights to closer points and lower weights to further points.)

Advantages and limitations KNN is conceptually simple, requires no assumptions about the distribution of the data points in k-space, and can estimate complex decision boundaries given sufficient data. However, KNN is an unsuitable choice when variables are numerous, or sparsity is high. As introduced above, PCA is often used to reduce the dimensionality of the data. However, this can reduce the explainability of a KNN model.

1.3.3.2 Logistic Regression

Logistic regression (LR) is a parametric probabilistic classifier that models the relationship between input features and the log-odds of class membership. For binary classification, the model estimates the probability that a given sample belongs to the positive class (class 1) as a function of its feature vector.

Given an input sample represented by a feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where each x_i denotes a measured feature (for example, an intensity value at a given m/z feature in a mass spectrum), logistic regression first computes a linear predictor

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n, \tag{1.9}$$

where β_0 is the intercept and β_i are model coefficients learned from the training data.

This linear predictor is related to the log-odds of class membership, defined as

$$\log\left(\frac{P(y = 1 \mid \mathbf{x})}{1 - P(y = 1 \mid \mathbf{x})}\right) = z, \tag{1.10}$$

where $P(y = 1 \mid \mathbf{x})$ denotes the predicted probability that the sample $\mathbf{x}$ belongs to class 1. Applying the inverse logit (sigmoid) function transforms the log-odds into a probability between 0 and 1:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-z}}. \tag{1.11}$$

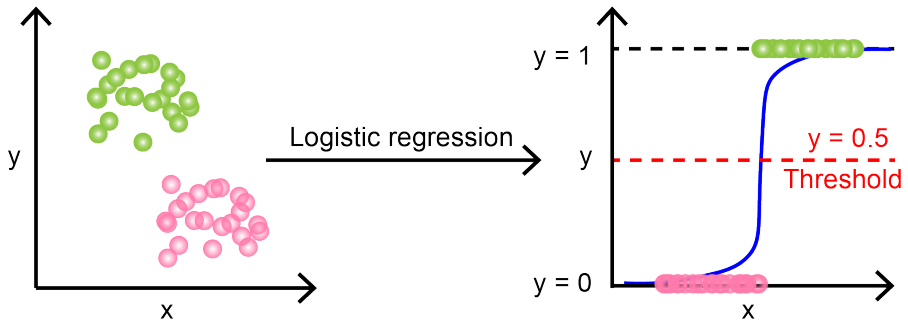

Figure 1.10: Illustration of logistic regression classification, where predicted probabilities are mapped to class labels using a decision threshold (e.g. values greater than 0.5 are assigned to class 1, and values less than 0.5 to class 0).

As shown in Figure 1.10, for binary classification, logistic regression using the sigmoid function, which transforms the linear combination of predictors z into a value between 0 and 1.The decision boundary is obtained by thresholding this probability at 0.5.

Hyperparameters The behaviour of LR depends on several hyperparameter choices:

- Regularisation penalty: Logistic regression estimates a coefficient for each input feature, where each coefficient quantifies the contribution of that feature to the log-odds of the outcome. Regularisation adds a penalty term to the optimisation objective to discourage overly large coefficients, thereby reducing variance and limiting overfitting.
  - L1 regularisation (lasso) applies an absolute-value penalty to coefficients. This encourages sparsity by driving some coefficients exactly to zero, effectively performing feature selection.
  - L2 regularisation (ridge) applies a squared-magnitude penalty. This shrinks coefficients towards zero but typically does not eliminate them entirely, which stabilises estimates in the presence of correlated or noisy features.

  - Elastic net combines L1 and L2 penalties, allowing simultaneous coefficient shrinkage and sparsity. This can be advantageous when groups of correlated features are expected to contribute jointly.

- Regularisation strength C: smaller C means stronger penalty; too large invites overfitting; too small overshrinks signal.

Advantages and limitations Logistic regression is mathematically transparent. Coefficients have a clear interpretation, which makes it one of the few classifiers that genuinely supports statistical inference (confidence intervals, hypothesis tests, effect directions). It is fast, works well with modest sample sizes relative to feature count when regularised appropriately, and is surprisingly competitive when class separation can be achieved by a simple weighted sum of the input features.[91] However, it is a poor choice when the data structure is highly nonlinear, when interactions are numerous but unknown, or when users expect the model to "discover" complex structure automatically.

1.3.3.3 Linear Discriminant Analysis

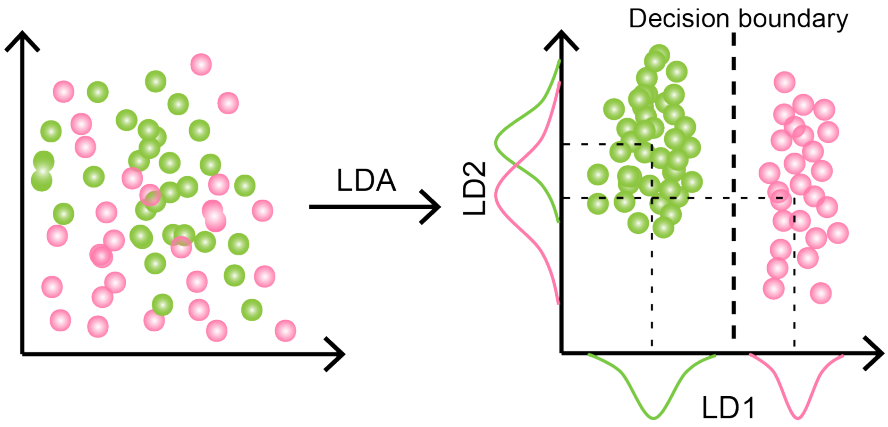

Figure 1.11: Illustration of Linear Discriminant Analysis (LDA) classification.

Linear Discriminant Analysis (LDA) classifies data by modelling the distinct distribution of features within each class.[92] LDA assumes that each class follows a multivariate Gaussian distribution, and that all classes share the same covariance matrix but have different means. Under these assumptions, the log-posterior

probabilities simplify to linear functions of the features. The classifier therefore assigns a sample to the class with the largest discriminant score, which corresponds geometrically to separating classes with linear boundaries. As shown in Figure 1.11 LDA finds projections of the data that maximise between-class variance while minimising within-class variance. These projections also provide dimensionality reduction. How LDA works is sometimes considered as a supervised version of PCA.

Hyperparameters In the present study when feature number is much larger than the sample number, reducing the dimensionality is the priority. This is usually achieved using PCA, in which case the number of principal components used in the training data becomes a hyperparameter

Advantages and limitations When its assumptions are roughly met, LDA is efficient. It produces linear decision boundaries with strong statistical foundations, and performs well even with small samples provided the dimensionality is controlled. However, real-world data often violate LDA's assumptions. In high-dimensional settings where features outnumber samples, the sample covariance matrix becomes nearly singular and naïve LDA collapses; any apparently good performance in this scenario is typically overfitting disguised as success.[93]

1.3.3.4 Naïve Bayes classifier

Naïve Bayes classifier(NB) is a family of simple generative classifiers based on Bayes' theorem coupled with an unrealistic assumption: features are independent.[94]

As shown in Figure 1.12, NB models the joint distribution of features $X = x_1....x_p$ and class label y. Bayes' theorem is defined as:

$$P(y|X) = \frac{P(X|y) \cdot \ P(y)}{P(X)} \tag{1.12}$$

where:

- $P(y|X)$ represents the probability that a sample belongs to class y;

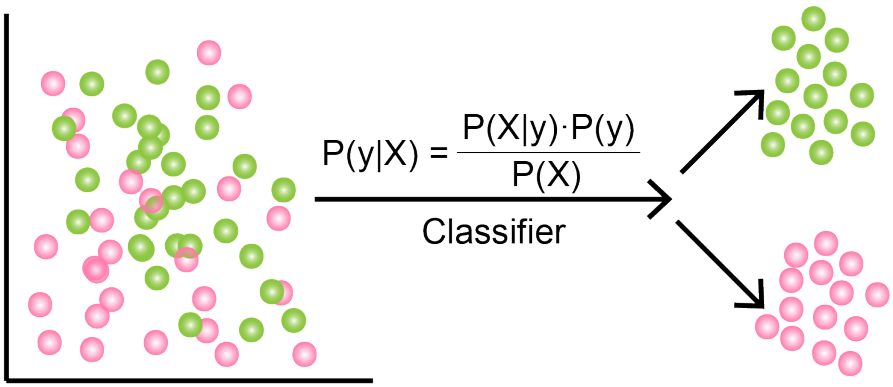

Figure 1.12: Illustration of Naïve Bayes classification.

- $P(X|y)$ represents the probability of seeing these specific features X if we already know that the sample belongs to class y.
- $P(y)$ represents the baseline probability of a class before looking at any new evidence.
- $P(X)$ represents the total probability of observing the features X across all possible classes.

By combining these individual probabilities, the model can classify high-dimensional data without requiring a large number of training samples.

Hyperparameters

- Smoothing parameter: Adds a small constant to all feature probability estimates so that features not observed in the training data for a given class do not receive zero probability. This prevents a single missing feature from forcing the overall class probability to zero and improves numerical stability and generalisation.
- Selected feature k: Specifies the number of features retained during feature selection prior to model training. Limiting k reduces dimensionality and redundancy in the input space, which can improve robustness and reduce overfitting, particularly for high-dimensional mass spectrometry data.

Advantages and limitations Naïve Bayes is fast to train, fast to predict, and scalable to extremely high-dimensional spaces. However, the core independence assumption is fundamentally wrong in almost any real scientific data set. When features are correlated, which they usually are, the model double-counts evidence, leading to overconfident and sometimes systematically biased results.[95]

1.3.3.5 Random Forest

Random Forests (RF) are ensemble classifiers that aggregate the predictions of many decision trees built on random variations of the data and features.[96] A decision tree is a flowchart-like model that maps out decisions and their potential consequences, using a tree-like structure with nodes, branches, and leaves to visualise choices and outcomes, helping in classification, regression, and strategic planning for complex problems.[97] As shown in Figure 1.13, a RF grows many decision trees, Each tree

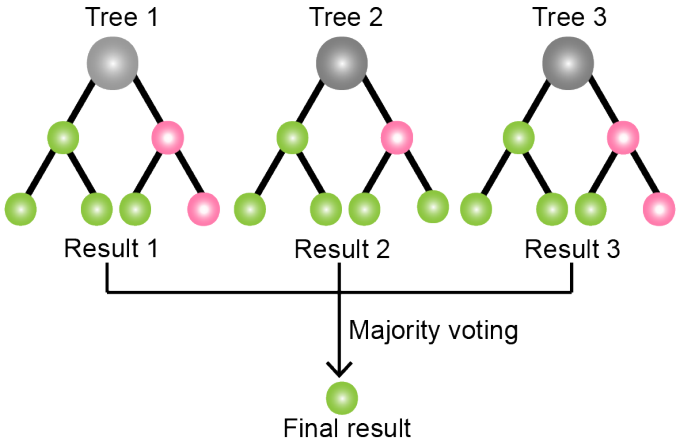

Figure 1.13: Illustration of random forest (RF) classification, where an ensemble of decision trees contributes to the final prediction through majority voting.

in the forest is trained on a bootstrap sample of the training data, and at each split only a random subset of spectral features is considered. Consequently, different trees learn different decision boundaries, even when trained on the same data set. For a given tissue sample, each tree produces an independent class prediction based on its learned rules. The final RF prediction is obtained by majority voting across all trees in the ensemble, thereby reducing the influence of noise or outlier features from any single tree and improving overall classification robustness. At every split

within a tree, only a random subset of features is considered. Each tree votes for a class label, and the forest prediction for classification is based on the majority vote.

Hyperparameters

- Number of trees (`n_estimators`): more trees reduce variance but increase computation. After a certain point, the accuracy of the model plateaus.
- Maximum tree depth (`max_depth`): the maximum number of successive splits from the root node to a leaf node in a decision tree. Increasing the tree depth allows the model to represent more complex decision boundaries, but excessively deep trees may fit noise in the training data and lead to overfitting.
- Number of features per split (`max_features`): controls how many different variables each individual tree is allowed to look at when making a decision. If the number is too high, every tree in the forest will look at the same best features. This makes the trees too similar to each other and makes the voting pointless; If the number is too low, each tree is only allowed to see one or two features. This might not be enough information to make a good decision.
- Minimum samples per split/leaf: regularises trees, prevents splits on tiny sample subsets.

Advantages and limitations RFs are strong general-purpose learners with minimal tuning. They are excellent default choices when feature relationships are nonlinear, interactions matter, and predictive performance is the priority over interpretability. However, RF can overfit badly in small data sets, especially when trees are deep and signal-to-noise is low. While powerful, RFs are often seen as 'black boxes' because they are difficult to interpret. Unlike simpler models, their inner workings are not naturally transparent. Instead, we have to rely on secondary tools like SHAP, which will be introduced in Section 1.3.6. Furthermore, they can be slow and resource-heavy when dealing with large data sets.[98, 99]

1.3.3.6 Support Vector Machines

Support Vector Machines (SVM) are margin-based classifiers that construct a decision boundary by maximising the separation between classes in feature space.[100] SVM focus strictly on the geometry of classification, finding the hyperplane that separates classes with the largest possible margin while penalising misclassified or ambiguous points.

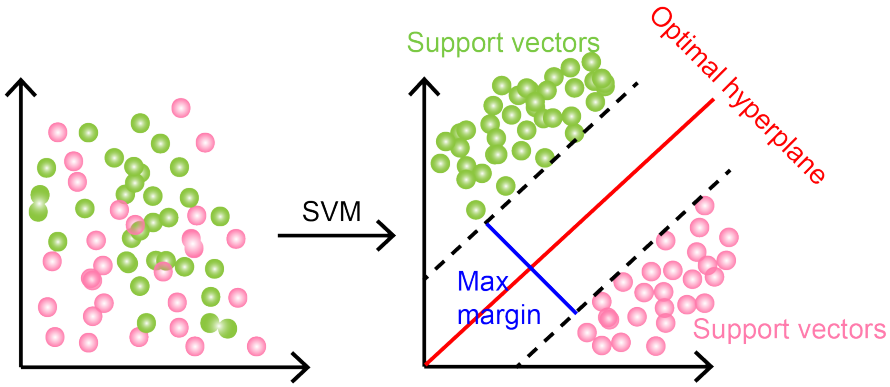

Figure 1.14: Illustration of a linear case of using Support Vector Machines (SVM).

As shown in Figure 1.14, SVM finds an optimal hyperplane defined by:

$$\mathbf{w}^\top \cdot \mathbf{x} - b = 0 \tag{1.13}$$

Where:

- $\mathbf{w}^\top$ is the normal vector to the hyperplane.
- $\mathbf{x}$ is the feature vector.
- b is the bias term that determines the offset of the hyperplane from the origin.

A max-margin is the widest gap between two perfectly separable classes. It defines two parallel boundaries: one that marks the start of Class A (where $\mathbf{w}^\top \cdot \mathbf{x} - b = 1$) and another that marks the start of Class B (where $\mathbf{w}^\top \cdot \mathbf{x} - b = -1$). To ensure the best performance, the algorithm maximises the distance between these boundaries. The final decision is made using the hyperplane; any sample falling on or above the top boundary is classified as one group, while anything on or below the bottom boundary belongs to the other.

Hyperparameters

- Kernel choice: linear, radial basis function (RBF). This defines the geometry of the feature space.

- Regularisation parameter C: high C punishes misclassification harshly (risk of overfitting), while low C tolerates errors to gain margin (risk of underfitting).

- Kernel parameters (γ in RBF): control the locality of influence; overly large γ memorises noise, overly small γ washes out structure.

- Feature scaling: Support Vector Machines are highly sensitive to the scale of input features because the optimisation of the separating hyperplane depends on distances in feature space. If features are on different scales, those with larger numerical ranges dominate the margin calculation, regardless of their true discriminative importance. Scaling ensures that all features contribute comparably to the optimisation process, which is essential for obtaining a meaningful decision boundary.

- Selected feature k: as with NB, limiting the number of selected features reduces the effective dimensionality of the input space, which helps prevent the SVM from overfitting to noise.

Advantages and limitations SVMs handle high-dimensional data extremely well and are robust to overfitting when appropriately tuned. The main limitations of SVMs are their sensitivity to tuning and poor scalability. An SVM model that is not perfectly calibrated will probably perform poorly on new data.[101] Besides, SVMs are difficult to interpret; the logic the model uses to separate classes is hidden within a complex mathematical transformation that does not relate directly to the original data.[102]

1.3.4 Cross-validation

Cross-validation[103] is a statistical method used to evaluate how well the performance of a machine learning model generalises given different, but equivalent training and test data. Instead of relying on a single split of the complete set of data into training and testing sets, the data set is partitioned into multiple subsets, or folds. In a standard $k-$fold approach, the model is trained on $k-1$ folds and validated on the remaining fold, with this process repeating k times so that every data point serves as validation data exactly once. This technique provides a more reliable estimate of how the model could perform on unseen samples and reduces the risk of performance metrics being skewed by an unusually 'easy' or 'difficult' train-test split.

Several variants of cross-validation have been developed in order to accommodate specific experimental needs:

- `Stratified cross-validation` preserves class proportions in each fold, essential under class imbalance.
- `Grouped cross-validation` ensures that correlated samples (e.g. repeated measurements from the same subject) never appear in both training and validation sets.

A well-recognised limitation of cross-validation is that it does not inherently prevent overfitting during optimisation. Even when each fold is held out correctly, repeatedly tuning hyperparameters or other pipeline decisions with the same data set can lead to 'overhyping', where noise in the data influences those choices and produces optimistically biased performance estimates that will not generalise to truly independent data.[104, 105]

1.3.5 Model evaluation

In medical classification tasks, where class imbalance and asymmetric error are common, reliance on a single metric to evaluate the model performance can lead to misleading conclusions.[106] This section therefore introduces the variety of metrics that are used in model evaluation within the present work.

1.3.5.1 Confusion matrix

The confusion matrix provides a straightforward way to summarise the correct and the incorrect predictions of a model. As an example, consider the task of classifying mass spectra as belonging to tumour vs normal tissue. For this binary classification problem, each prediction falls into one of four categories:

- True Positive (TP): tumour tissue correctly classified as tumour
- False Positive (FP): normal tissue incorrectly classified as tumour
- True Negative (TN): normal tissue correctly classified as normal
- False Negative (FN): tumour tissue incorrectly classified as normal

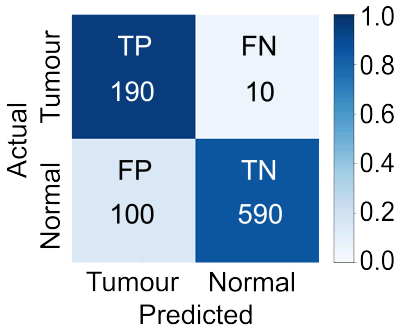

Figure 1.15: Example confusion matrix for a binary classification task. The optional heatmap colour intensity represents the proportion of samples within each true class.

The confusion matrix shown in Figure 1.15 provides a compact representation of classification outcomes by summarising the counts of TP, FP, TN, and FN.

1.3.5.2 Performance metrics

All performance metrics[107] were calculated from the confusion matrix components defined above and calculated as follows.

Accuracy Accuracy measures overall correctness but does not distinguish between error types, and therefore can provide misleading results in imbalanced data sets.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1.14}$$

For example, in a data set with a 9:1 class imbalance, a classifier that predicts all the test data as the majority class would report 90% accuracy while providing no discriminatory power for the minority class. This demonstrates that accuracy alone is insufficient for evaluating model performance on imbalanced data sets.

Sensitivity Sensitivity quantifies the ability of the model to correctly identify positive cases.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1.15}$$

Specificity Specificity quantifies the ability of the model to correctly identify negative cases.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{1.16}$$

Precision Precision reflects how reliable a positive prediction is once the model has made it.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1.17}$$

Negative Predictive Value(NPV) Precision reflects how reliable a negative prediction is once the model has made it.

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{1.18}$$

F1 score The F1 score provides a single summary measure that balances sensitivity and precision. It prevents a model from appearing to be successful just because it labels everything as the dominant class.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{1.19}$$

Cohen's kappa Cohen's kappa (kappa score) corrects for chance agreement and is therefore particularly informative in imbalanced classification problems, where accuracy may be inflated by dominant class predictions.[108] The interpretation of kappa score is shown in Table 1.3

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1.20}$$

where the observed agreement p_o is

$$p_o = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1.21}$$

and the expected agreement by chance p_e is

$$p_e = \frac{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN}) + (\mathrm{FN} + \mathrm{TN})(\mathrm{FP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})^2} \tag{1.22}$$

Table 1.3: Kappa score interpretation

Kappa Score	Agreement
<0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Excellent

1.3.5.3 Brier score

The Brier score provides a complementary measure by quantifying the accuracy of predicted probabilities.[109] For a binary classification problem, it is defined as the mean squared difference between the predicted probability of the positive class and the true outcome. Lower Brier scores indicate better probabilistic calibration and overall predictive performance.

$$BS = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - y_i)^2 \tag{1.23}$$

where:

- N is the number of samples,
- $\hat{p}_i$ is the predicted probability of the class for sample i
- $y \in \{0, 1\}$ is the true class label.

Unlike accuracy or the F1 score, the Brier score penalises both incorrect predictions and poorly calibrated confidence, making it sensitive to overconfident errors. This metric is especially important in medical classification tasks. For example, a model that assigns predicted probabilities clustered around the decision threshold (e.g. 0.49 and 0.51 for a threshold of 0.5) may achieve acceptable classification accuracy, yet provide little meaningful confidence about its predictions. A perfectly calibrated and perfectly accurate classifier achieves a Brier score of 0, whereas larger values indicate increasing uncertainty between predicted probabilities and true outcomes.

1.3.6 Model interpretability - SHAP

SHAP (SHapley Additive exPlanations) is an interpretability framework based on cooperative game theory that assigns each feature (in our case, each m/z intensity) an importance value, for a specific prediction.[110] It calculates Shapley values, or 'SHAP values' by comparing a model's prediction with and without a specific feature across all possible combinations of features. The SHAP value for a feature i is defined as:

$$\phi_i = \sum_{S \subseteq \{x_1,\ldots,x_n\} \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [f(S \cup \{i\}) - f(S)] \tag{1.24}$$

where:

- n is the total number of features;
- S is a subset of features excluding feature i;
- $f(S)$ is the model's prediction for that subset;
- ϕ_i is the SHAP value for feature i;

- $\sum_{S \subseteq \{x_1,\ldots,x_n\} \setminus \{i\}}$ is the sum over all possible subsets S of features that do not include feature i;
- $\frac{|S|!(n-|S|-1)!}{n!}$ is a weighting factor based on the number of ways a subset can be formed;
- $[f(S \cup \{i\}) - f(S)]$ is the difference in the model's prediction when feature i is added to the subset S.

This approach ensures that the total prediction is fairly distributed among the input variables. By providing a clear measure of how much each feature increases or decreases the probability of a classification, SHAP can identify the specific features that drive the model's decision-making process. In our case, the SHAP values reveal to what extent each m/z peak in the mass spectrum influences the classification of the tissue as tumour or normal.

SHAP has gained popularity in medical research.[111, 112] This is because it reveals both the global importance of the set of input features across the whole data set, and local interpretability, which explains exactly why the model made a specific decision for a single sample. By providing a clear link between the model's logic and the original data, SHAP helps clinicians to trust and validate automated diagnostic tools.

1.4 Aims and scope of this thesis

The aim of this study was to evaluate whether atmospheric solids analysis probe mass spectrometry (ASAP–MS), when combined with appropriately designed machine-learning workflows, can provide robust and interpretable molecular discrimination of brain tumour tissue from neuropathologically normal brain in a clinically relevant setting. The study has been reviewed and approved by HRA and Health and Care Research Wales (HCRW) (Reference: 24/WS/0013) and Oxford University Hospitals NHS Foundation Trust's Research and Development Office (Reference: PID17637). The University of Oxford is acting as the study sponsor.

Experimentally, we investigated the factors that affect data quality, including ion source cleanliness, probe temperature, cleaning protocols, consumables, inter-operator variability, and batch effects. Optimisation strategies are proposed within this defined instrumental framework.

Biologically, the primary focus is on human brain tissue. Fresh and frozen samples are used for model development and evaluation, while formalin-fixed, paraffin-embedded (FFPE) brain tissue is examined to assess feasibility and limitations.

Chemically, the study concentrates on understanding how different molecular classes behave under ASAP ionisation. In particular, the contrasting behaviour of small polar metabolites and lipids is examined to define constraints on spectral interpretation and downstream data analysis.

From a computational perspective, the work focuses on classification tasks, primarily tumour versus normal brain tissue. Model selection and evaluation are tailored to small-to-moderate data sets, with emphasis on controlling class imbalance, avoiding data leakage, accounting for batch structure, and ensuring interpretability through model-explanation techniques. The study does not aim to exhaustively optimise all possible machine-learning architectures, nor to perform molecular biomarker discovery.

Finally, the clinical scope of this work is exploratory. The study does not seek to replace established histopathological assessment or to deliver a deployable diagnostic tool. Instead, it aims to establish a robust experimental and analytical foundation upon which future, larger-scale, and clinically integrated studies may be built.

2

Optimising Clinical Data Acquisition with ASAP–MS

Contents

This chapter presents a series of approaches to optimising the use of ASAP–MS. The work has been published in *Analyst* (DOI: 10.1039/D5AN00166H) and is adapted here with permission from the Royal Society of Chemistry.

2.1 Introduction

ASAP–MS shows strong potential for clinical use. However, obtaining high-quality and reliable datasets remains challenging. This is often due to poorly standardised sample handling and measurement procedures. Considering the issue more broadly, in a survey involving 1576 scientists,[113] over 70% disclosed difficulties in reproducing others' experiments, and more than half faced challenges in repeating their own experiments. Almost 90% of chemists participating admitted to experiencing failures in experiment replication. While not directly related to ASAP–MS, this study highlights the critical need for standardisation and careful experimental design.

In the context of ASAP–MS this requires careful investigation and optimisation of a number of key experimental parameters, in order to enable the generation of trustworthy and reproducible clinical data sets, thereby enhancing the reliability of ASAP–MS for clinical applications. In a previous study, we addressed standardisation issues in the measurement of human plasma using ASAP–MS.[78] The present study builds on this earlier work to further optimise the methodology for generating meaningful and trustworthy clinical data sets, in this case on samples of human brain and cerebrospinal fluid. We investigate the influence on the mass spectra of calibration and post-calibration procedures, including the presence of residual calibration mix within the ion source;[114] cooling of the probe tip following recording of background spectra; probe cleaning procedures; contamination from consumables such as lens tissue and sample storage containers;[115] and variation between different instrument users. Lack of standardisation in any of these can lead to variations in the data in the form of contaminant peaks, changes in ionisation probability for some or all sample components, or sample degradation. If these variations are interpreted as significant features in classification models the result is significant skewing of results.

As part of the study, we also consider batch effects in some detail. As the name suggests, batch effects arise when samples are measured in different 'batches', resulting in systematic differences between subsets of data within a large data

set.[116–118] These have become more important in the age of big data.[119] Some of the inter-batch differences may be reasonably easy to explain (*e.g.* batches of data from different labs or recorded using different instruments or different experimental protocols, or subjected to different data processing protocols). However, even if all experimental parameters are standardised as comprehensively as possible, some batch effects will usually remain. If these are not considered carefully and corrected for as far as possible, they can mask, or worse, mimic biological variation, leading to highly misleading results. The conclusions of more than one study have been found to be invalid due to improper treatment of such effects; for example, Mertens *et al.*[118] highlighted a number of clinical mass spectrometry studies in which poorly designed experiments resulted in the perfect confounding impacts of batch variation and biological variation. Although these experiments produced high accuracy rates in machine learning training, the affected data sets and conclusions were ultimately considered useless and abandoned.

In the present study, alongside standardising the experimental protocol as far as possible, we evaluate several different methods for batch effect correction, and are able to make a series of recommendations in relation to optimised use of ASAP–MS for clinical data set generation.

2.2 Methods

2.2.1 Samples and experimental

The frozen post-mortem brain and cerebrospinal fluid (CSF) samples used in this study were provided by the Oxford Brain Bank; a research ethics committee (REC) approved and Human Tissue Authority (HTA)-regulated research tissue bank (REC reference 15/SC/0639, issued by the NHS Health Research Authority 'South-Central – Oxford C'). Upon collection of whole post-mortem brains by the Oxford Brain Bank, brains were dissected fresh. If the dura was present, it was removed from the cerebrum, and an incision was made along the longitudinal fissure to separate the cerebral hemispheres. The cerebellum was detached from the cerebrum at the level of the fourth ventricle and the brain stem were removed. The cerebellum was

divided along the posterior Cerebellar notch, and one of the Cerebellar hemispheres was dissected into 1 cm thick slabs. The brain was then sectioned into 1 cm thick coronal sections starting at the level of the mammillary bodies. Whole hemisphere slabs were snap-frozen in liquid nitrogen vapour and stored at −80 °C until further dissection on dry ice, whereby the anterior cingulate gyrus was dissected at the level of the genu of the corpus callosum.

Samples were prepared for analysis as follows:

1. Brain samples were equilibrated to −20,° C and mounted onto a cryostat block using optimal cutting temperature (OCT) compound, a water-soluble cryo-embedding medium containing polyethylene glycol, polyvinyl alcohol, and resin components. To avoid OCT contamination, the compound was applied only to a small area on the opposite side of the tissue from the cutting surface. When tissue blocks were re-used, sections were cut only from the OCT-free face, and any region in direct contact with OCT was excluded from analysis. Three 10 µm sections were obtained from each sample, and transferred to a polypropylene sample tube (see later for details of the various tubes employed). To create a suitable medium for homogenisation, 100 µL of LC-MS grade water (Fisher) was added to the tube. A bead homogeniser (OMNI International bead ruptor elite) was then employed to thoroughly homogenise the sample (settings are shown in Table 2.1).

2. CSF samples were thawed from −80 °C and centrifuged at 12,000g for 15 minutes at 4 °C to separate the cellular components and particulate matter from the liquid fraction. The supernatant was then pipetted into a new tube, taking care not to disturb the pellet.

All measurements were made on an Advion Expression version L compact quadrupole mass spectrometer equipped with an ASAP ion source and controlled by Advion Mass Express data acquisition software (version 6.9.38.1). The ion source was run in 'high temperature, low fragmentation' positive ion mode (Table 1.2). Prior to measurements, the glass capillaries that comprise the tip of the ASAP

Table 2.1: OMNI Homogenizer settings

Parameter	Value
Tube volume / mL	1.5
Speed / m/s	4.00
Cycles	2
Time / s	30
Dwell / s	15

probe were baked in an oven at 250 °C for 30 minutes to remove any surface contaminants as far as possible.

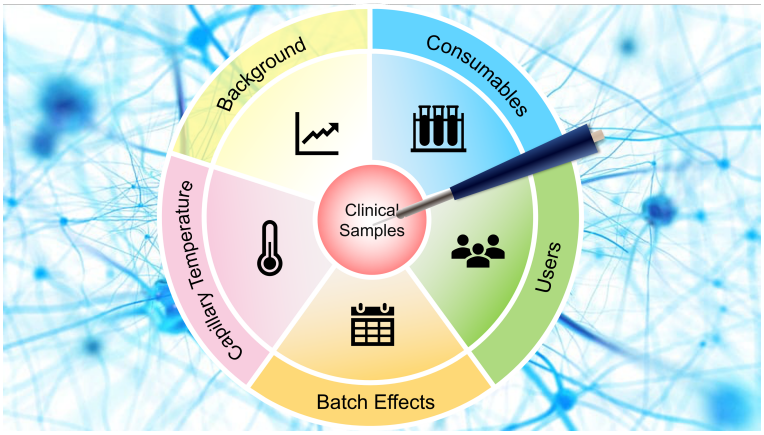

Figure 2.1: The effects of the experimental factors investigated in this study

To make a measurement, the probe was fitted with a clean glass capillary and inserted into the ion source for 30 s to record a background mass spectrum. The probe was then removed from the ion source and allowed to cool before being brought into contact with a small amount of sample and reinserted into the ion source for a measurement time of 25 s. Due to the high sensitivity of the instrument, in general the smallest possible amount of sample should be transferred to the probe tip for measurement. When the probe is inserted into the ion source. This should result in an almost instantaneous rise in the total ion signal, followed by a rapid decay over the next 20–30 s. Too much sample leads to signal saturation, characterised by high signals that do not decay or strange time-dependent behaviour

of the total ion signal.[78] The obtained raw data were processed using the method explained in Section 1.2.4

2.2.2 The effects of experimental factors

The effects of a number of experimental factors (Figure 2.1) were investigated using the following methods:

2.2.2.1 Calibrant and background effects

The mass spectrometer is calibrated daily with Advion APCI calibration tuning mix in order to ensure repeatable peak positions and widths in the mass spectra. A considerable amount of tuning mix is injected through a capillary tube inlet into the ion source during the calibration process, and any residual mix can have a significant effect on both the 'background' and 'sample' mass spectra recorded subsequently. Figure 2.2 shows examples of background spectra recorded in the presence and absence of residual tuning mix.

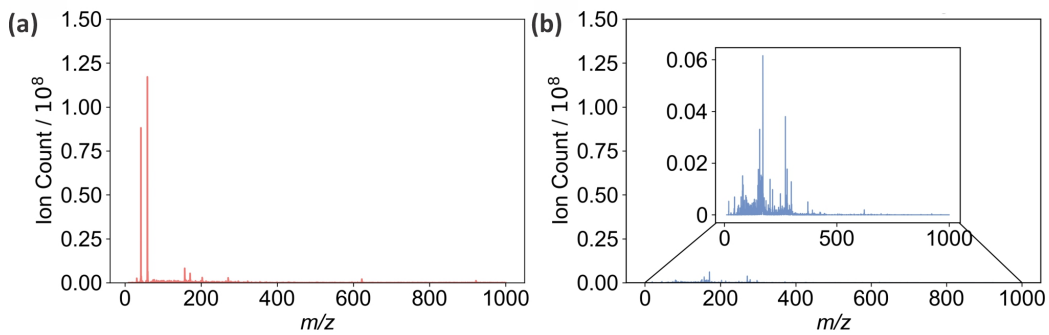

Figure 2.2: Examples of background mass spectra taken under different conditions in ASAP–MS: (a) a background spectrum taken when residual calibration tuning mix was present in the ion source; (b) an example of a clean background spectrum obtained after cleaning the ion source with a 50:50 mixture of LC-MS ethanol and LC-MS water. The inset to (b) shows the low-mass peaks on a magnified scale.

To assess the impact of residual tuning mix on clinical sample data sets, a CSF sample prepared as detailed above was split into two. Each of the two sub-samples were subjected to 25 repeat ASAP–MS measurements, with the measurements on

one sub-sample performed immediately after calibration when residual tuning mix was present, and those on the second sub-sample performed after allowing the instrument to run without sample introduction until a clean and stable background was observed. After analysing the repeatability of the two groups of measurements, two methods for removal of residual tuning mix from the ion source were investigated. These involved flowing either (i) air or (ii) a 1 : 1 mix of LC-MS water and ethanol through the capillary inlet for 1.5 minutes, then running the instrument for 8.5 minutes to clear the ion source and allow the background signal to stabilise prior to making any measurements.

2.2.2.2 Temperature of ASAP probe tip

After each background measurement, during which the glass capillary that forms the tip of the ASAP probe is exposed to nitrogen gas at 400 °C, a period of cooling is required prior to sample loading and measurement. Cooling curves were measured using a thermal imaging camera (FLIR C3-X Compact Thermal Camera) immediately after removing the probe from the ion source. The experimental setup is shown in Figure 2.3 and camera settings are shown in Table 2.2). The highest temperature region of the image was identified by the thermal camera's 'hot spot' function, and was used to define the measurement region corresponding to the probe tip. The temperature in the measurement region was recorded every second during imaging of the probe, and the results from ten repeat measurements were used to generate a cooling curve. The room temperature was 23 °C on the day of the measurement.

Table 2.2: FLIR C3-X Compact Thermal Camera settings

Parameter	**Value**
Emissivity	0.85
Distance / m	1.00
Atmospheric temperature / °C	23
Relative humidity / %	50
Image scale maximum / °C	85
Image scale minimum / °C	25

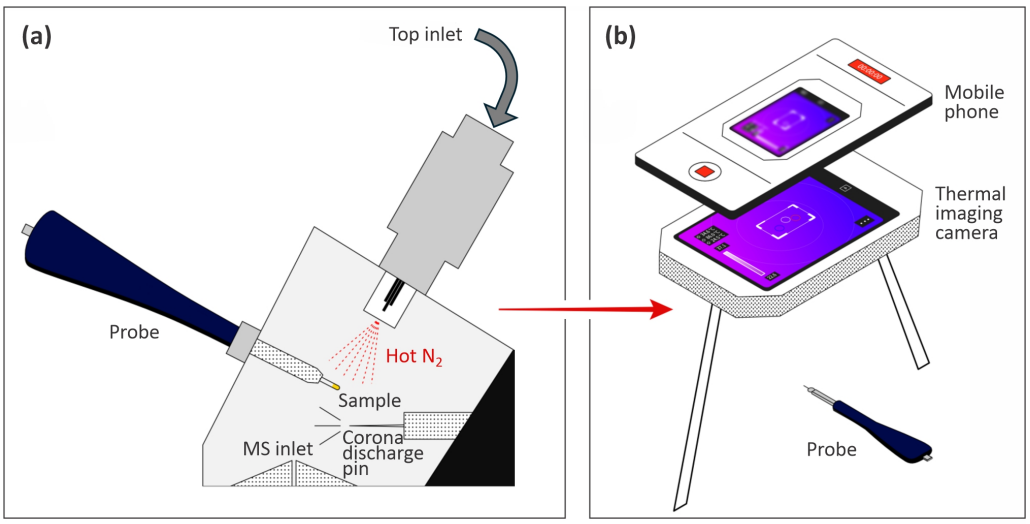

Figure 2.3: (a) Schematic of the ASAP–MS ion source, showing the inserted probe with glass capillary tip, hot N_2 gas flow, corona discharge, and inlet to the ion optics and mass analyser; (b) Setup used for thermal imaging of the ASAP probe tip. A region of interest in the image is defined, corresponding to the tip. Additional time-resolved imaging was achieved by using a mobile phone to capture video.

2.2.2.3 Glass capillary cleaning/reuse

In the interests of cost reduction and sustainability, we have established in previous work [78] that with appropriate cleaning the glass capillary forming the tip of the ASAP probe can be reused for up to five measurements on the same sample without any negative impact on the measured mass spectra. Between measurements, the capillary is cleaned by rinsing with deionised water, followed by gentle wiping and drying with lens tissue. In the present work, we investigated whether the performance could be improved further by inserting the cleaned capillary into the ASAP source in order to expose it to the hot nitrogen gas flow, followed by removal and cooling for 20 s. The comparison was performed by analysing one frozen brain sample 25 times with the two different cleaning protocols, recording five repeats with five different capillaries for each protocol.

2.2.2.4 Consumables

To investigate the impact of different consumables on the mass spectra, we selected three different brands of polypropylene tubes (1.5 mL polypropylene tubes (Tube

Brand 1, PCR clean, manufactured without slip agents, plasticisers, and biocides) were purchased from Eppendorf; 2 mL polypropylene tubes (Tube Brand 2, DNase- and RNase-free) were obtained from OMNI; 10 mL optically clear polypropylene tubes (Tube Brand 3, DNase-, RNase-, Endotoxin-, and Pyrogen-free) were sourced from Appleton) and three different brands of lens tissues (MC-5 Lens Tissues (Lens Tissue Brand 1) were purchased from THORLABS; additional lens tissues were obtained from KimTec (Lens Tissue Brand 2) and Fisher, UK (Lens Tissue Brand 3)).

To investigate contaminant peaks in the mass spectra arising from the polypropylene tubes, we first used moderate force to swab the inner wall of the empty tubes with the glass capillary tip of the ASAP probe, and recorded mass spectra. To investigate any diffusion of materials from the tube to the solvent, we added LC-MS water to the tubes and left them overnight, before recording mass spectra of the resulting solutions. To investigate contaminant peaks arising from the lens tissues, we mimicked a standard cleaning procedure by gently wiping the glass capillary ASAP tip with each brand of lens tissue and then recorded mass spectra. To ensure consistent results, the measurements on each consumable were repeated five times.

2.2.2.5 Measurement repeatability between users

In order to investigate variations between different instrument operators in mass spectra recorded using a standardised protocol, we selected four representative users with varying scientific backgrounds and different levels of experience with ASAP–MS measurements. The four users made independent measurements on the same human cerebellum sample on the same day in a random order. We assessed both the repeatability within each individual user's measurements and the reproducibility across all four users.

2.2.3 Data analysis

Principal component analysis (PCA) was used to investigate the repeatability and reproducibility of measurements made under the various different conditions investigated. For each data set, the individual measurements were reduced to

their first five principal components. To visualise the data, the first two principal components (PC1 and PC2), which together account for around 85% of the variance in the mass spectra were plotted, together with a 'confidence ellipse' of one standard deviation (i.e. the ellipse defines a region within which 68% of the data points are expected to lie, assuming the data follow an approximately normal distribution). Prior to this, we performed a Shapiro–Wilk test to establish the normality of the data[120] (see Table A.1 for details). The resulting plots enable a comparative analysis between groups of measurements. The centroid or mean of each group was calculated by averaging each principal component across all measurements within the group, while the variability within each group was quantified by finding the standard deviation of the Euclidean distances between each measurement and the centroid. The significance of any similarities and differences between the measurement groups was evaluated using various statistical methods. Since several of the data sets did not follow strictly normal distributions, for single-pair comparisons, we employed the Mann–Whitney U test[121] within the SciPy (version 1.5.4) Python package[122] (this package was also used for the Shapiro–Wilk test mentioned above). For multiple group comparisons, we used a two-step approach: first a one-way Analysis of Variance (ANOVA) to determine overall differences among groups, followed by Tukey's Honestly Significant Difference (HSD) test for pairwise comparisons.[123] These tests were implemented using the Python statmodels package (version 0.13.5).[124] Significance levels were set at $p < 0.05$, $p < 0.01$, and $p < 0.001$ for all tests.

2.2.4 Batch effects

Batch effects were investigated using frozen brain samples from the cerebellum and anterior cingulate cortex regions of six patients, *i.e.* 12 samples in total. Each sample was split into two, and used to generate two data sets as follows:

1. All 12 samples were measured within a single day, generating a data set that serves as a control without batch effects;

2. The 12 samples were measured in two batches on two different days, generating a data set with the potential presence of batch effects. The sample distribution between the two batches was randomised in order to mitigate any bias, with each batch including samples from both brain regions of three different patients.

3. To examine the batch effects, PCA plots were generated for all the data points in each batch, with confidence circles centred at the centroid and extending to the furthest data point in each batch. The function Kruskal from SciPy (version 1.5.4) was used to perform a Kruskal–Wallis (kW) H-test in order to evaluate the statistical significance of correlations between the principal components (PCs) and the batches.[122, 125] This dual approach enabled us to rigorously investigate any potential batch-related discrepancies, ensuring the reliability of our findings. Having identified batch effects, we explored the efficacy of correcting for these using two different batch-effect correction tools available in Python, namely ComBat and Independent Component Analysis (ICA).[116, 126, 127]

2.3 Results and Discussion

2.3.1 Calibrant and background effects

As noted in Section 2.1, residual tuning mix in the ion source following spectrometer calibration can have a significant effect on both the background and sample mass spectra recorded subsequently. This is shown quantitatively in Figure 2.4. Panel (a) shows PCA plots for the two sets of 25 mass spectra recorded for a sample of CSF immediately after calibration, and after running the spectrometer until no tuning mix remained in the ion source, respectively. The corresponding mass spectra are shown in Figure A.1. We see from Figure 2.4 that the data recorded under 'clean' conditions, shown in blue, are tightly clustered and highly reproducible, while those recorded with residual tuning mix present, shown in red, are scattered over a large area of PCA space, showing a much higher degree of variation. A Mann–Whitney U test performed on the Euclidean distances of the principal

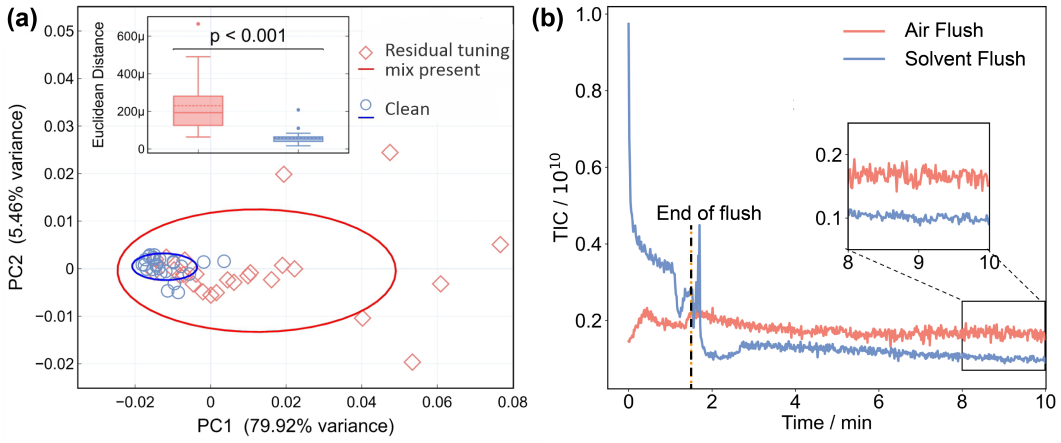

Figure 2.4: The effect of residual calibration tuning mix in the ion source on the repeatability of ASAP–MS measurements: (a) PCA plot (see Section 2.1 for detailed description) of 25 repeat measurements on the same CSF sample with (red diamonds) and without (blue circles) residual tuning mix present in the ion source. The inset shows the results of a Euclidean distance Mann–Whitney U test. comparison between the two conditions, demonstrating a significant difference; (b) comparison of the total ion signals (total ion count, TIC) recorded as a function of time when the ion source is flushed with air (red line) or a 1:1 mixture of LC-MS water and ethanol ASAP–MS. The ion source is flushed for the first 1.5 minutes, before the inlet valve is closed and the signal allowed to stabilise over the following 8.5 minutes.

component vectors from the centroid for each data set shows a significant difference between the two sets of measurements, with $p < 0.001$. The variability in the mass spectra recorded in the presence of residual tuning mix is most probably caused by a combination of time-varying contributions to the mass spectra from the tuning mix 'contaminant' peaks and ion suppression effects that alter the contributions from sample peaks. Unsurprisingly, we can conclude that ensuring a clean background and removing contaminant m/z peaks (from any source) as far as possible yields improved measurement repeatability. Before moving on to consider active methods for eradicating signals from residual tuning mix prior, we note that frequent cleaning of the ASAP ion source is an important measure that should be taken in order to minimise background interference and enhance data quality. The required cleaning frequency is sample-dependent, but as an example, when processing large numbers of plasma samples in our own laboratory we clean the source after around 150 measurements.

As explained in Section 2.2, we investigated two different flushing methods for removing residual tuning mix from the ASAP ion source, employing air or a 50:50 mix of ethanol and LC-MS water, respectively. The results are shown in Figure 2.4(b) in the form of plots of total ion count (i.e. the integrated signal across the entire mass spectrum) as a function of time. Air (red line) or solvent (blue line) is flushed through the ion source for the first 90 seconds, before the inlet valve is closed and signal is recorded for a further 8.5 minutes. Compared with the air flush, flushing with solvent results in much higher signal levels during the flush, but considerably lower and more stable signal levels (see inset to figure) once equilibrium is reached following the flush. As well as reducing the overall ion count more effectively, the solvent flush was also more effective at removing the specific m/z peaks arising from the tuning mix, and was the flushing method of choice for all subsequent measurements.

2.3.2 Temperature of ASAP probe tip

Figure 2.5 shows the results of the thermal imaging experiments used to record cooling curves for the ASAP probe after removal from the ion source. The probe is simply left to stand under ambient conditions in the laboratory during the cooling period. Panel (a) shows thermal images of the probe at four different time points, while panel (b) shows the temperature as a function of time. The cooling curve shows the expected exponential decay, with rapid cooling to a little over 100°C occurring during the few seconds taken to remove the probe from the ion source and position it in front of the camera. The tip reaches a temperature of ∼37 °C after around 20 s, after which it is cool enough for loading of clinical tissue samples. After 30 s the tip has cooled to 26°C, just a little over room temperature. The threshold of ∼37 °C was selected because the method was being developed for potential intraoperative use. Cooling the probe to approximately body temperature provided a practical criterion for minimising heat-induced changes during sample loading. For other sample types or applications, the acceptable loading temperature should be adjusted according to the thermal stability and biological relevance of the material being analysed.

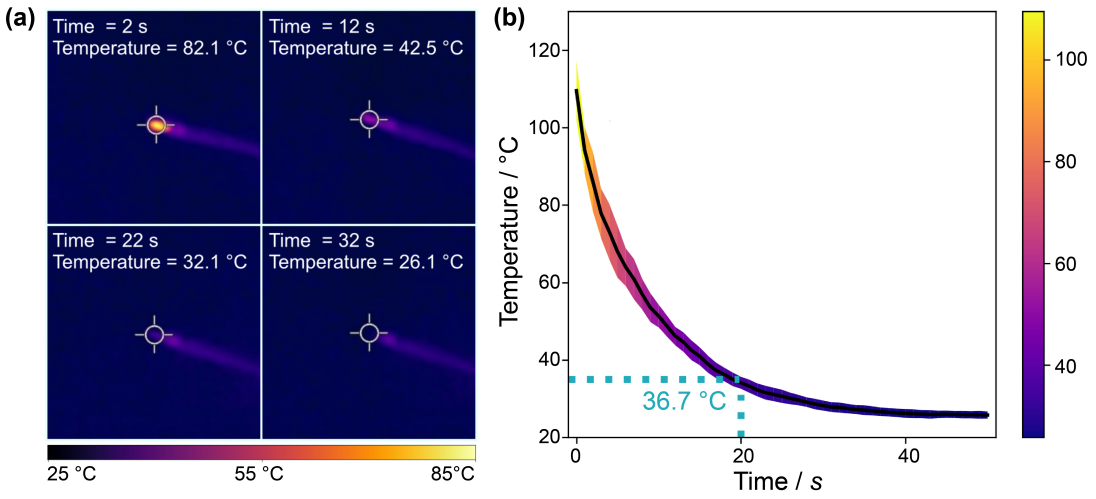

Figure 2.5: (a) Thermal images of the ASAP probe tip recorded 2 s, 12 s, 22 s, and 32 s after removal of the probe from the mass spectrometer ion source; (b) cooling curve for the probe tip following removal from the ion source, averaged over ten measurements. The solid line is the mean, with the shaded area indicating the standard deviation of the ten measurements.

2.3.3 Glass capillary cleaning/reuse

As noted in Section 2.2, we have established previously that with appropriate cleaning between measurements, the glass capillary tip of the ASAP probe can be reused up to five times before it should be replaced with a clean capillary.[78] In the previous work, the tip was cleaned with deionised water and lens tissue between uses ('Method 1'). In the present work, we investigated whether performance can be improved further by including an additional cleaning step after the manual cleaning, in which the clean probe is inserted into the ion source and exposed to the hot nitrogen flow for 25 s before being removed, allowed to cool for 20 s, and used for the next measurement ('Method 2'). Figure 2.6(a) shows the total ion count recorded as a function of time during a sample measurement, during the period when the probe is reinserted after manual cleaning and exposed to the hot nitrogen flow, and when the probe is reinserted again to check the effectiveness of the cleaning protocol. The mass spectra recorded for the last two ('cleaning' and 'check') measurements are shown in Figure 2.6(b). We see that there is still a significant ion signal during the time when the probe is inserted into the ion

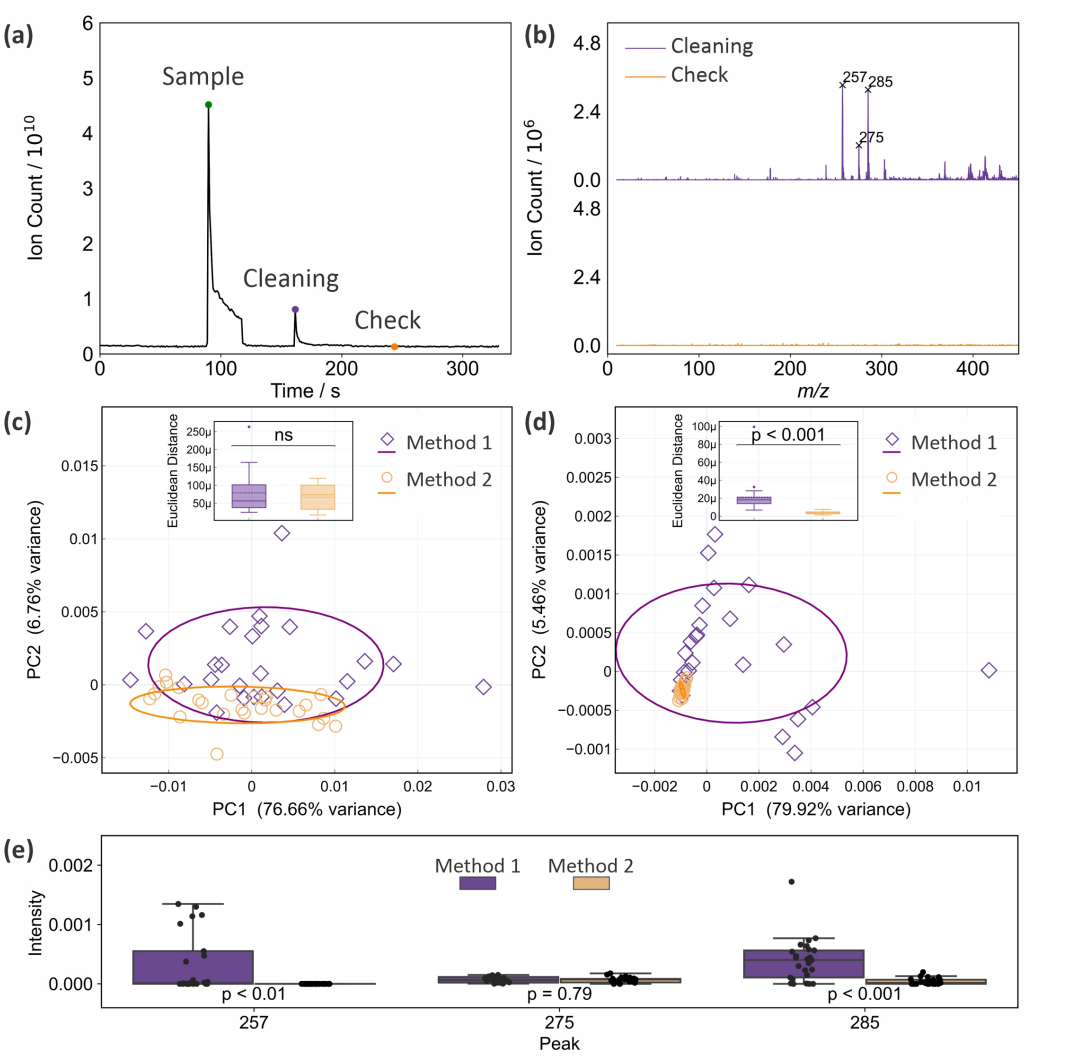

Figure 2.6: Evaluation of two different cleaning protocols for the glass capillary tip of the ASAP probe,(a) total ion count recorded as a function of time during a sample measurement, cleaning step, and 'check' measurement for cleaning Method 2; (b) mass spectra recorded during the 'cleaning' and 'check' insertions of the ASAP probe; (c) PCA plots for 25 mass spectra (full m/z 10–1000 range) of a frozen brain sample recorded using cleaning methods 1 and 2; (d) as for (c), but considering only the mass range from m/z 200–300; (e) Box and whisker plots comparing the intensities of three mass peaks arising from contaminants associated with the lens tissue used for cleaning at m/z 257, 275, and 285 when cleaning methods 1 and 2 are employed. p values are from a Mann–Whitney U test.

source for cleaning, with significant peaks that can be attributed to residue from the lens tissue used in the manual cleaning step (see Section 2.3.4). This residual signal disappears following the two cleaning steps.

Figure 2.6(c) and (d) show PCA plots for the two sets of 25 mass spectra

recorded for a frozen brain sample using cleaning methods 1 and 2 (the complete set of mass spectra used in this analysis can be found in Figure A.2 and Figure A.3). Panel (c) shows principal components for the full m/z 10–1000 mass range of the mass spectrometer, while panel (d) shows the results for a truncated mass range of m/z 200–300, chosen to isolate the contributions from mass peaks associated with the lens tissue used in the manual cleaning step. In both cases, the spread in the data is reduced considerably when using cleaning method 2. Despite the clear reduction of spread in the PCA plot, a Mann–Whitney U test shows that this reduction is not statistically significant when the complete mass spectra are considered; however, it is highly significant ($p < 0.001$) over the m/z range in which the lens tissue peaks appear. These results suggest that while it can be difficult to detect significant influences of contaminants on the complete mass spectra, explaining why such contamination can sometimes be accidentally and subtly introduced into clinical data sets, local influences can be highly significant and should be minimised wherever possible.

The peaks at m/z 257 and 285 are tentatively assigned to protonated palmitic acid and stearic acid, respectively. This assignment is chemically plausible because C16-C18 fatty acids and fatty-acid-derived compounds are widely associated with paper and papermaking chemistry, including hydrophobic sizing agents and alkylketene dimer precursors.[128] The peak at m/z 275 could not be confidently assigned from the available information and may represent a related fragment ion, a degradation product, or another lens-tissue-derived contaminant. However, no direct chemical analysis of the specific lens tissue used in this study was performed. Therefore, these assignments should be regarded as putative, and the peaks are described as lens-tissue-associated contaminant ions rather than definitively identified compounds.

Figure 2.6(d) shows box and whisker plots of the intensities recorded using both cleaning methods for the three most intense peaks arising from lens tissue residue. Significant reductions are seen in the intensities of the peaks at 257 and 285, while the peak at m/z 275 is sufficiently low in intensity under all conditions that the second step does not lead to a statistically significant reduction.

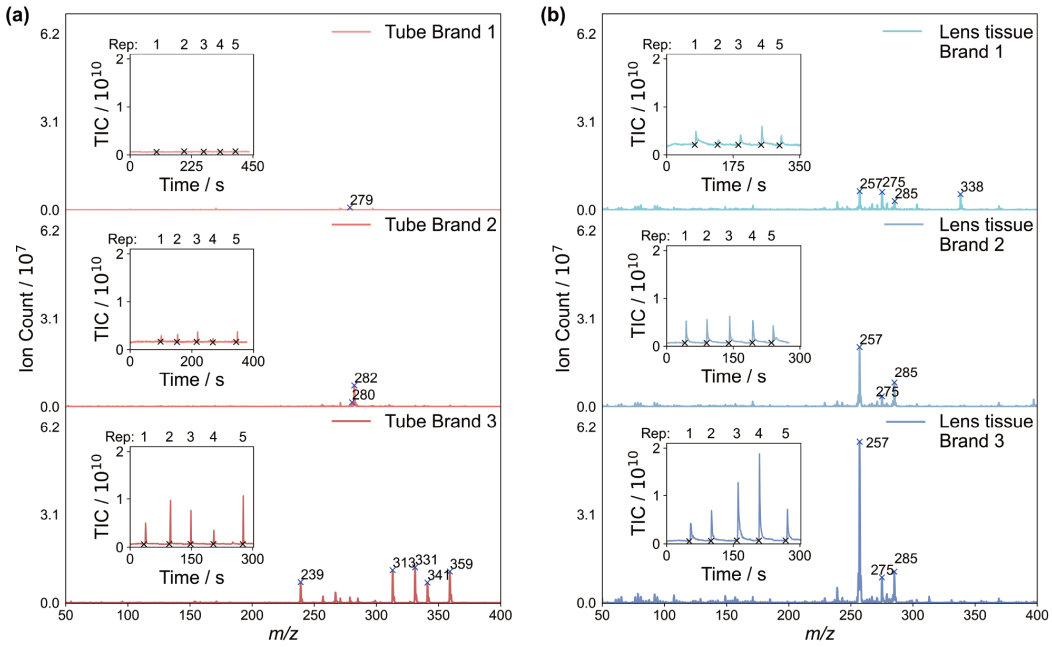

Figure 2.7: Mass spectra recorded for (a) swabs from the inner wall of three different brands of polypropylene tubes, and (b) three brands of lens tissue (see Section 2.2 for brand details), averaged over five measurements. The total ion count recorded during the five measurements is shown in the inset to each mass spectrum. Note the difference in intensity scales used to plot the two sets of mass spectra.

We can conclude that cleaning method 2 is superior to cleaning method 1, and this method was employed in all subsequent measurements. This includes all measurements reported in this study, unless stated otherwise.

2.3.4 Consumables

Consumables can add a variety of different contaminant peaks to the mass spectra, which we investigated by performing ASAP–MS measurements on three different types of sample storage tubes, and three different types of lens tissue. Details of the products used can be found in Section 2.2. Mass spectra recorded by swabbing the sample tubes and lens tissues with the ASAP probe, averaged over five repeat measurements, are shown in Figure 2.7(a) and (b), respectively. The total ion counts recorded during the five measurements are shown as insets to each mass spectrum. For both the storage tubes and lens tissue, we see large differences in the mass spectra recorded for the three different brands. Considering first the sample

tubes, brand 1 yields consistently low total ion signals and a small number of very low intensity peaks in the mass spectrum, in line with previous work by Canez *et al.*[115] In contrast, brands 2 and 3 yield significantly higher total ion signals and particularly in the case of brand 3, with a large number of high-intensity peaks in their mass spectra. This suggests that the polypropylene material may be unstable when being swabbed. A second set of measurements, made on LC-MS water left overnight in each brand of tube (see Figure A.4), yielded very low signal for all three brands of tube. This suggests that all three tubes are generally suitable for use with ASAP–MS analysis as long as direct contact of the ASAP probe with the tube walls is avoided. Considering next the three brands of lens tissue, we see large total ion signals and numerous peaks in the mass spectra for all three brands. In all three cases, three of the most intense peaks appear at m/z 257, 275, and 285, with brand 1 yielding another peak of high intensity at m/z 338. While these signals have the potential to contaminate the sample mass spectra, as shown in Section 2.3.3, residue from lens tissue used to clean the ASAP probe tip can be removed effectively simply by inserting the probe into the hot N_2 gas flow inside the mass spectrometer ion source for a short period of time prior to making measurements on samples. We can conclude that careful and consistent selection of consumables, and clear and consistent handling protocols, are likely to be key in order to avoid the potential for brand-specific contamination of the sample mass spectra. New consumables should be evaluated carefully in order to establish their potential to contaminate mass spectra, either directly or *via* ion suppression effects. Direct contact of the sample or probe with the consumables should be avoided where possible, and rigorous cleaning techniques should be employed when reusing glass capillaries in the ASAP probe. Caution should also be exercised in situations where m/z peaks arising from consumables overlap with 'biological' peaks arising from the sample. Ideally, such peaks should not be used for classification, and in cases where this is not possible, interferences should be carefully considered and quantified.

2.3.5 Measurement reproducibility between users

Figure 2.8 shows PCA plots and the corresponding reproducibility analysis for ten repeat measurements on a frozen cerebellum sample made by four different users. Based on the results shown in Figure 2.8(b), there are no significant differences in the repeatability achieved by each of the four users, with mean deviations of each users measurements from their own centroid of between 0.002 and 0.0035; each user is able to measure the sample consistently, with minimal variability within their own data sets. However, when the results obtained by different users are compared with each other, we see more significant differences, with mean deviations from the centroid of other users of up to 0.006. These results suggest that despite consistent performance of individual users, variations between users occur even when the same protocol is followed rigorously. In the context of a metabolomics or other clinical study, it is therefore important to evaluate user reproducibility carefully and quantitatively to ensure that any variability in the data introduced by multiple users is (significantly) smaller than the biological variation under study.

2.3.6 Batch effects

As noted in Section 2.1, when acquiring large clinical data sets, variations in the mass spectra recorded for different batches of samples can be very significant in comparison with the biological variations under study. The reasons for these variations are not well understood, but any study of this type needs to take batch effects into account and implement measures to mitigate or remedy them. The problem is illustrated quite clearly even using the small number of measurements employed in the present study. As explained in Section 2.2, we recorded spectra from 12 brain tissue samples from six patients, with cerebellum and anterior cingulate cortex samples for each patient. The patients were randomised into two different groups, with their samples comprising 'batch 1' and 'batch 2', respectively. To generate Data Set 1, which we can use as a control, measurements were made on all 12 samples (both batches) on the same day, while for Data Set 2 the two batches were analysed on different days. Figure 2.9(a) and (b) show PCA plots (PC1 vs.

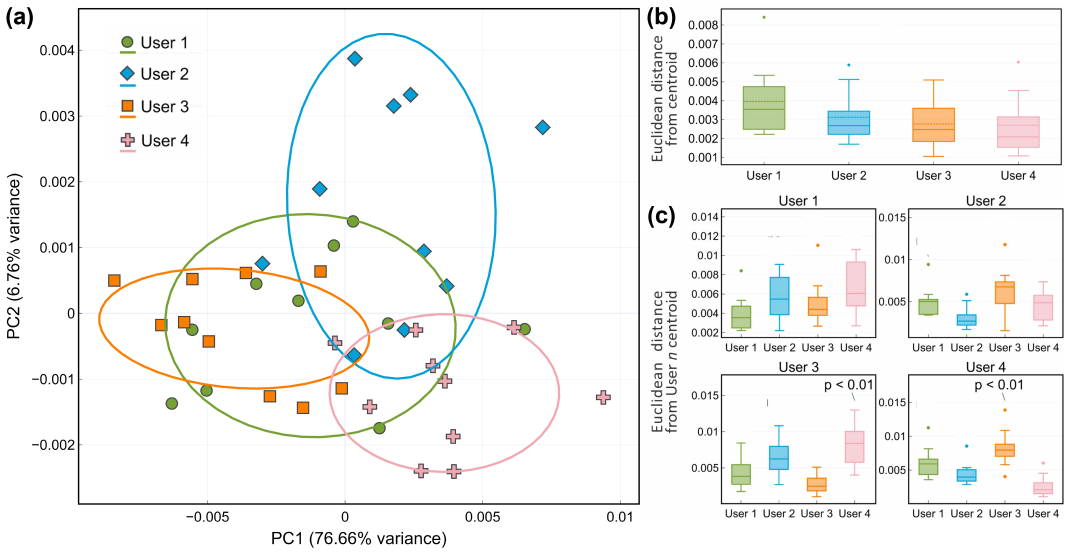

Figure 2.8: Evaluation of measurement reproducibility for a single sample of cerebellum between four different users: (a) PCA plot, with confidence ellipses representing one standard deviation for each user; (b) Box and whisker plot of the repeatability achieved by each user, determined from the Euclidean distances of each user's measurements from their own measurement centroid; (c) As for (b), but comparing each user's measurements with the centroids of other users. Statistically significant deviations are labelled with their p value, determined from ANOVA and Tukey's HSD test.

PC2 and PC2 vs. PC3 in each case) for the two data sets, with the centroid rings superimposed as shaded regions in each case. For Data Set 1 (Figure 2.9(a)), we see good overlap between the measurements made on the two batches, implying that they were recorded under consistent measurement conditions. For Data Set 2 (Figure 2.9(b)) we see significant separation between the two sets of measurements, particularly in the first principal component, and more overlap between data points for the two different tissue types. To the best of our ability, the measurements were all made under identical conditions, so this is a clear example of a batch effect, in line with those observed by other authors.[117] If not corrected or appropriately accounted for, the batch effect could easily confound any effects arising from true biological variability, and lead to the drawing of highly misleading conclusions. In Figure 2.9(c) and (d), we show the results of performing a batch correction on Data Set 2, using the ComBat and independent component analysis (ICA) methods, respectively. The ComBat algorithm[116, 126] assumes that batch effects

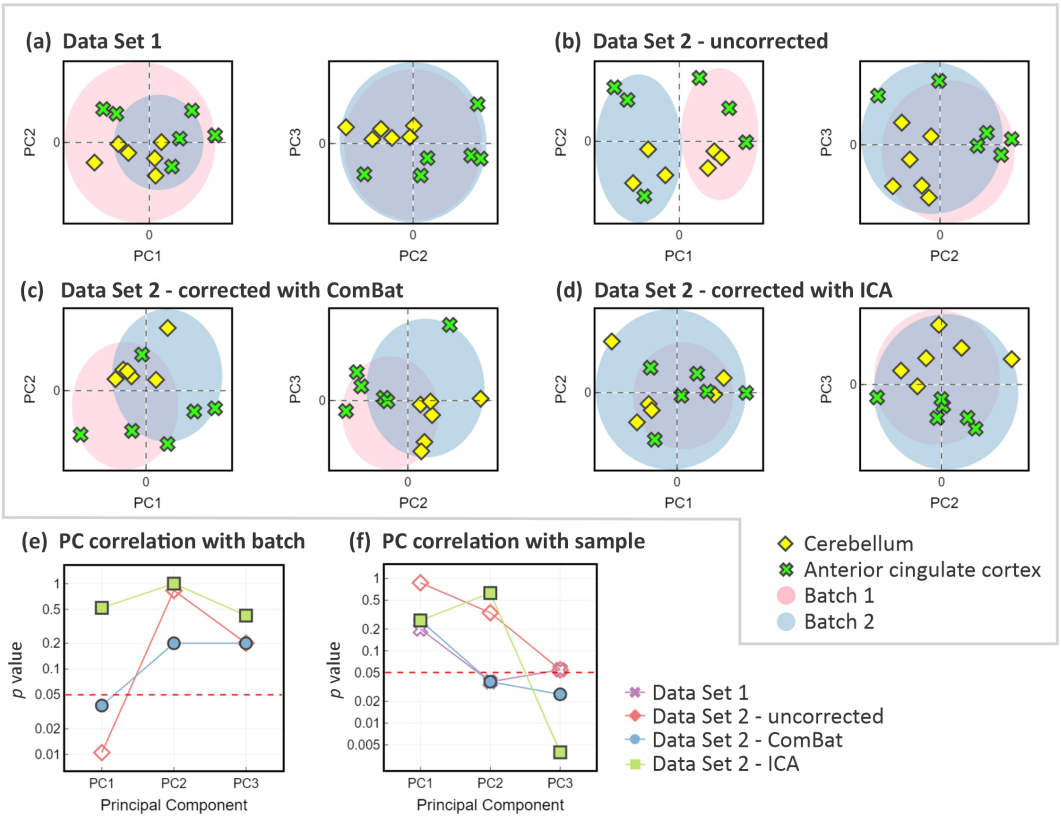

Figure 2.9: Batch effects and mitigation through the use of batch effect correction methods: Panels (a) and (b) show PCA plots (PC1 vs. PC2 and PC2 vs. PC3) with one-standard-deviation confidence ellipses (shaded regions) for Data Set 1 (recorded in a single batch) and Data Set 2 (recorded in two batches on separate days), respectively. Panels (c) and (d) show the analogous PCA plots for Data Set 2 after correction using the ComBat and ICA methods, respectively. Correlations (in the form of p values) between the first three principal components of the mass spectra and (e) batch and (f) sample type, for Data Set 1 and Data Set 2 before and after batch correction. The red dashed line indicates the upper threshold for the region of statistical significance, with $p < 0.05$ (Kruskal–Wallis (kW) H-test).

affect many m/z peaks in similar ways and uses an empirical Bayesian approach to adjust for these effects. In the ICA method,[127] the complete data set is factorised into components using one of a number of matrix factorisation methods, and the components that show significant correlation with the individual batches are removed. Inspecting the data in Figure 2.9(c) and (d), we see that both methods yield a reduction in the batch-related separation along PC1 and improve the clustering according to sample type (cerebellum vs anterior cingulate cortex). However, the data corrected using the ComBat method still show separation by

batch, while the separation is almost completely removed by the ICA method. The latter results in almost complete overlap of the centroid rings, similar to that observed in Data Set 1, which was recorded in a single batch. In the present example, the ICA method appears to perform better than the ComBat method, but this would need to be tested much more extensively with larger data sets before drawing any firm conclusions on this point.

Figure 2.9(e) and (f) show the results of a more in-depth statistical analysis of the batch effect correction, in the form of a correlation analysis. Panel (e) shows p values for the correlation between batch and the first three principal components of Data Set 2 before and after batch correction using the ComBat and ICA methods. Before batch correction (red diamonds), the data show a statistically significant correlation with batch (in the plot they lie below the red dashed line indicating $p < 0.05$). Batch correction with the ComBat method (blue circles) reduces the correlation, but p still indicates a significant correlation, while the ICA method reduces the correlation substantially ($p \sim 0.5$). Panel (f) explores the correlation between the principal components of the data sets and the sample type for Data Set 1 and for Data Set 2 before and after batch correction. For Data Set 1, recorded in a single batch, PC2 (and perhaps PC3) reveals a statistically significant correlation with sample type. For Data Set 2, PC2 shows no significant correlation with sample type before batch correction. The correlation becomes significant following correction with the ComBat algorithm, but (interestingly given the apparent overall better performance of this algorithm) not with the ICA algorithm. Correlation between sample type and PC3 becomes statistically significant for Data Set 2 following correction with either approach. These findings suggest that while batch effects can obscure true biological variation, employing batch effect correction methods can recover these biological differences to varying degrees. However, as noted by other authors,[129] it is often not possible to remove batch effects completely. In order to identify the features responsible for the separation caused by batch effects and observed along the first principal component (PC1) in Figure 2.9, we examined the PC1 loadings, which represent the contribution of each original

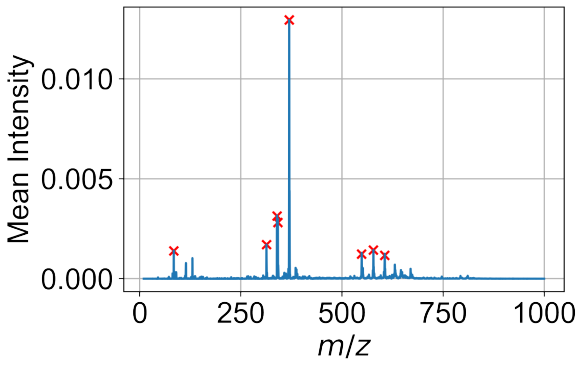

Figure 2.10: The eight peaks with the highest loadings for PC1 in the analysis of batch effects, marked with red crosses. The spectrum shown in blue is the average across all samples.

variable to the principal component. Features (m/z peaks) with higher absolute loading values were interpreted as having a greater impact on the direction of PC1. The eight most significant mass peaks according to their PC1 loadings are shown in Figure 2.10, with their intensity distributions within each batch shown in the form of box and whisker plots in Figure 2.11. There is a strong

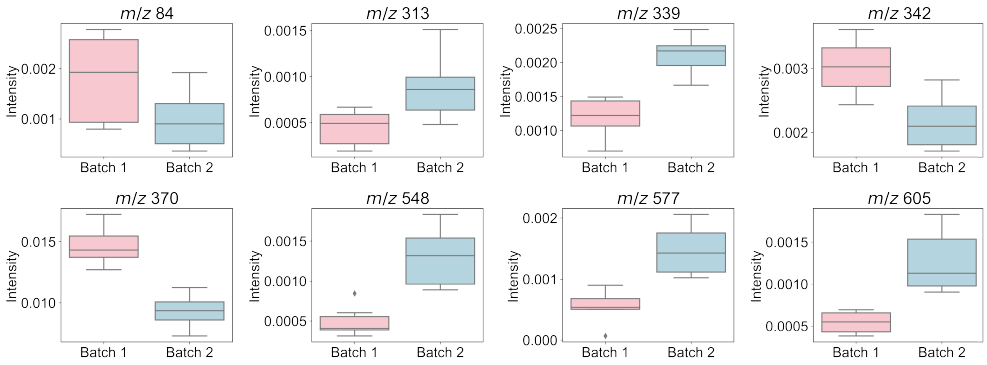

Figure 2.11: Box and whisker plots of the intensity distributions within each batch for each of the peaks identified in Figure 2.10.

correlation between the absolute intensity of the m/z and their contribution to batch effects, with the most intense features in the mass spectra playing the largest role. Unsurprisingly, any small technical differences between batches can cause noticeable changes in the most intense peaks. Intensity changes in less intense peaks are then induced either directly via the technical differences or as a secondary effect of the intense peak variation, via the normalisation procedure employed as

part of the data pre-processing. The more intense peaks therefore make the most significant contribution to the observed batch effects.

Based on our small proof-of-concept demonstration, we can conclude cautiously that batch effects are significant in ASAP–MS measurements, but that they can be mitigated by the use of batch effect correction methods. However, further investigation with much larger data sets is needed before drawing any firm or quantitative conclusions on the most effective correction methods or the extent to which biological variations can be recovered.

2.4 Conclusion

We have investigated a number of factors that can degrade the quality of clinical data sets based on ASAP–MS measurements of tissue or fluid samples and have identified ways to mitigate these effects. Based on the current investigation, and our previous work on plasma,[78] we recommend:

(i) Ensuring utmost cleanliness of the ion source, which should be free from all contaminants, including residual calibration tuning mix;

(ii) Ensuring sufficient cooling of the ASAP probe tip before loading with biological samples;

(iii) Developing an appropriate cleaning protocol for the probe between measurements. Standard cleaning approaches can be enhanced by an additional step in which the probe tip is exposed to the hot N_2 gas flow inside the ASAP ion source for a short period of time;

(iv) Carefully evaluating consumables for their potential to introduce contaminant peaks into the mass spectra, and selecting those that minimise such contamination;

(v) Standardising all measurement procedures as far as possible and documenting them clearly and comprehensively so that different users can follow them closely;

(vi) Recognising that even with strict adherence to well-designed standard operating procedures, batch effects—arising from both known (e.g., different users) and unknown sources—cannot be entirely avoided. These should be carefully considered and corrected where possible. In the present study, we have shown using a small set of sample data that both the ComBat and ICA methods for batch correction can mitigate these effects to varying degrees while preserving true biological variation, opening the way to further, more quantitative studies with larger data sets.

As a relatively low-cost, rapid, and straightforward measurement technique, ASAP–MS holds considerable appeal for clinical applications. However, it is still a relatively new technique in the clinical arena, and measurement protocols must be carefully optimised in order to maximise its potential. The work presented in this study, along with our previous findings on plasma measurements,[78] lays the foundation for achieving these goals.

3

From Molecules to Matrices: Interpretation of ASAP–MS Mass Spectral Behaviour

Contents

3.1 Introduction

As introduced in previous chapters, Atmospheric Solids Analysis Probe Mass Spectrometry (ASAP–MS) has the potential to be developed as a metabolic fingerprinting tool because it offers rapid, direct analysis with little or no sample

preparation. Furthermore, the instrumentation is relatively affordable compared to other high-end mass spectrometry platforms, making it a more accessible option for routine use in healthcare and research settings. In the previous chapter, we described the development of multifaceted approaches to optimising the protocol for complex biological sample measurement.

Previous studies within the group have demonstrated that ASAP–MS, combined with machine-learning analysis, can successfully distinguish clinically relevant plasma phenotypes, despite limited molecular specificity in the resulting spectra.[130] These findings highlight the potential of ASAP–MS for rapid biological classification. However, like other metabolic fingerprinting techniques, ASAP–MS faces important limitations, particularly in data interpretation.[131] These challenges arise from adducts, overlapping signals, and the absence of chromatographic separation, which make it difficult to distinguish and identify individual compounds. Furthermore, biomedical samples contain a mix of lipids, metabolites, salts, and other components. ASAP–MS applies atmospheric pressure chemical ionisation (APCI) to ionise the molecules. As introduced in Section 1.2, during the measurement time, the sample is heated and desorbed from the probe tip by a flow of hot nitrogen. It is generally believed that during the measurement time, more volatile molecules desorb at early times and less volatile molecules at later times. However, a commonly overlooked aspect of ASAP–MS is the possibility of chemical reactions occurring during the measurement. Under the combined influence of heat and energetic ion–molecule interactions in APCI, components of complex biological mixtures may react with one another. For example, amino acids and carbohydrates—both abundant in biological samples—can undergo Maillard-like reactions[132], generating reaction-derived ions not originally present in the sample. This is especially relevant in untargeted metabolic fingerprinting studies, which aim to capture broad chemical profiles without knowing in advance what compounds are present. Without careful spectral interpretation, these studies could suffer from false positives, where peaks related to noise or by-product peaks are wrongly interpreted as meaningful biochemical features.[133, 134] This risk increases when in-source transformations and matrix

effects are not well understood. While several methods have been developed to filter out false signals,[135–138] these typically rely on retention time or MS^2 information, which cannot be adapted to ASAP–MS. One early study used the consistent appearance of peaks across multiple spectra to distinguish true signals. This is an idea that can be adapted for ASAP–MS;[139] however, the interpretation still remains a challenge. Although high-resolution platforms such as ASAP-QTOF facilitate interpretation for biological samples, access to such instruments is limited. As a result, methods that depend on high-resolution data are not universally applicable, underlining the need for statistical interpretation strategies.

To address some of the challenges outlined above, previous work within the group examined the behaviour of individual small molecules and simple mixtures under different ion source conditions. These studies demonstrated that, under gentler ASAP–MS settings, molecular ions and associated supramolecular species such as dimers could be preserved, enabling insight into aggregation and association behaviour in relatively simple chemical systems.[140]

Building on this foundation, the present study adopted a complementary strategy by investigating selected molecules and simple mixtures under the conditions used for complex biological sample measurements. In addition, intact human brain tissue and plasma, together with their corresponding lipid extracts, were analysed to provide mechanistic insight into how ASAP–MS mass spectra for biological samples evolve in time and how this behaviour depends on molecular class and sample composition. This perspective provides an initial but mechanistically grounded understanding of ASAP–MS performance in complex biological contexts. Understanding these effects is important because time-dependent desorption, fragmentation, and potential in-source reactions directly shape the spectral features used for downstream analysis.

3.2 Materials and Methods

3.2.1 Single-molecule and binary-mixture measurements

Individual reference compounds were analysed by ASAP-MS to assess their ionisation and fragmentation behaviour under the high-temperature, relatively low-

fragmentation conditions used in this study. The analytes included glycine ($\geq$ 99% purity, HPLC grade; Sigma-Aldrich, Merck KGaA, Darmstadt, Germany), glucose ($\geq$ 99.5% purity, GC grade; Sigma-Aldrich, Merck KGaA, Darmstadt, Germany), triolein ($\geq$ 95% purity; MP Biomedicals, Germany), and 1-stearoyl-2-arachidonoyl-*sn*-glycero-3-phosphocholine (SAPC; $\geq$ 98% purity; Cayman Chemical, USA). Additional brain-relevant small molecules and lipids were also analysed, including dopamine (98% purity, TLC grade; Sigma-Aldrich, Merck KGaA, Darmstadt, Germany), epinephrine (HPLC grade; Sigma-Aldrich, Merck KGaA, Darmstadt, Germany), glutamate ($\geq$ 99% purity, FG; Sigma-Aldrich, Merck KGaA, Darmstadt, Germany), and cholesterol ($\geq$ 99% purity; Sigma-Aldrich, Merck KGaA, Darmstadt, Germany). D-2-hydroxyglutarate (D-2-HG; $\geq$ 98% purity, GC grade; Sigma-Aldrich, Merck KGaA, Darmstadt, Germany), an oncometabolite associated with IDH-mutant brain tumours,[17] was examined separately to assess whether this clinically relevant metabolite could be detected directly under the ASAP-MS conditions used.

A 1:1 mixture of glycine and glucose, as well as a 1:1 mixture of triolein and SAPC, were prepared separately to investigate potential molecular interactions during ASAP–MS analysis. For each metabolite or metabolite mixture, samples were prepared in LC-MS-grade water at a concentration of 1 mM. For each lipid or lipid mixture, samples were prepared in LC-MS-grade ethanol at a concentration of 1 mM, unless otherwise stated.

3.2.2 Complex samples and sample preparation

Pooled human plasma and frozen human cerebellum were analysed in this study. The pooled plasma sample was purchased from Bioivt and stored at $-20\,^{\circ}$C until use. Before processing, the plasma was thawed on ice and vortexed for 30 seconds to ensure it was well mixed. Then, 50 μL was pipetted into each of six tubes (Eppendorf, Germany).

The cerebellum sample was obtained from the Oxford Brain Bank (Sample code: NP10/18). Six cryosections were placed into a single tube, followed by 800 μL of water. The tissue was homogenised using a bead homogeniser (OMNI), followed

by vortexing for 30 seconds. After mixing, the homogenate was divided into six equal portions and transformed to 1.5 mL tubes (Eppendorf, Germany).

For both plasma and brain samples, three portions were used for lipid extraction. The other three were measured directly using ASAP–MS. This allowed comparison between lipid-extracted and intact samples for each biological matrix.

3.2.3 Lipid extraction

Lipid extraction was carried out for both brain and plasma samples using a modified MTBE (Methyl tert-butyl ether)-based protocol.[141] All solvents were LC-MS grade and kept ice-cold before use. All steps were performed on ice unless stated otherwise.

For brain samples, we pipetted 130 μL of sample from each aliquot of homogenized cerebellum tissues and placed this into a glass vial. To this was added 620 μL of MTBE (Supelco, USA) and 200 μL of methanol (Supelco, USA). The mixtures were vortexed for 30 seconds and kept on ice for 30 minutes. Then, 30 μL of water was added, followed by another vortexing step (30 s) and centrifugation at 1000 $\times$ g for 10 minutes at 4 °C.

For plasma samples, 30 μL of plasma was pipetted into each glass vial and 100 μL of water was added to match the volume with the brain samples. The same extraction steps as above were followed.

After centrifugation, the upper (organic) phase was carefully collected into a new glass vial. A second extraction was performed by adding 0.3 times the volume of a solvent mixture (62:20:13 ratio of MTBE, methanol and water) to the remaining lower phase. This was vortexed and centrifuged again under the same conditions. The supernatants from both extractions were pooled and dried overnight in a SpeedVac (Thermo Fisher Scientific, UK). Dried lipid extracts were resuspended the following day. All steps were performed at room temperature using LC-MS grade solvents.

To each dried sample, 46.5 μL of lipid resuspension buffer (8:23:69 ratio of butanol, isopropanol, and water) was added. The samples were mixed on a vertical mixer for 10 minutes at 800 RPM. After mixing, 103 μL of 0.1% formic acid

was added. The samples were vortexed for 30 seconds and left to stand at room temperature for 30 minutes. Sonication was then performed for 10 minutes.

Finally, samples were centrifuged at 5000 $\times$ g for 10 minutes. 140 μL of clear supernatant was transferred to fresh glass vials for ASAP–MS analysis.

ASAP–MS mass spectra for each sample above were measured according to the optimised protocol described in Chapter 2.

3.2.4 Data analysis

All data processing and statistical analyses were performed using Jupyter Notebook with Python (v3.7) with standard libraries including NumPy, pandas, scikit-learn, and matplotlib.

ASAP–MS data were acquired as time-resolved spectra, with one mass spectrum recorded approximately every 900 milliseconds, resulting in about 27 spectra over a 25-second run for each sample. Spectra were binned to 1 m/z intervals due to the low-resolution nature of the instrument.

Mass spectra for each individual molecule were normalised using max normalisation (described in Section D.1) and plotted directly in order to examine the observed ionisation patterns, and the possible ionisation processes were determined and compared with previously reported observations.

For complex samples, a max normalised intensity matrix was generated across all time points. Besides plotting the mass spectra, principal component analysis (PCA) was used to visualise the variation and temporal evolution of spectral fingerprints across intact and lipid-extracted samples. PCA was applied to time-averaged spectra to examine sample separation and dynamic trends.

K-means clustering was employed to group the profiles of individual m/z features based on their temporal behaviour during the ionisation process. Clusters were interpreted based on signal onset and trajectory.

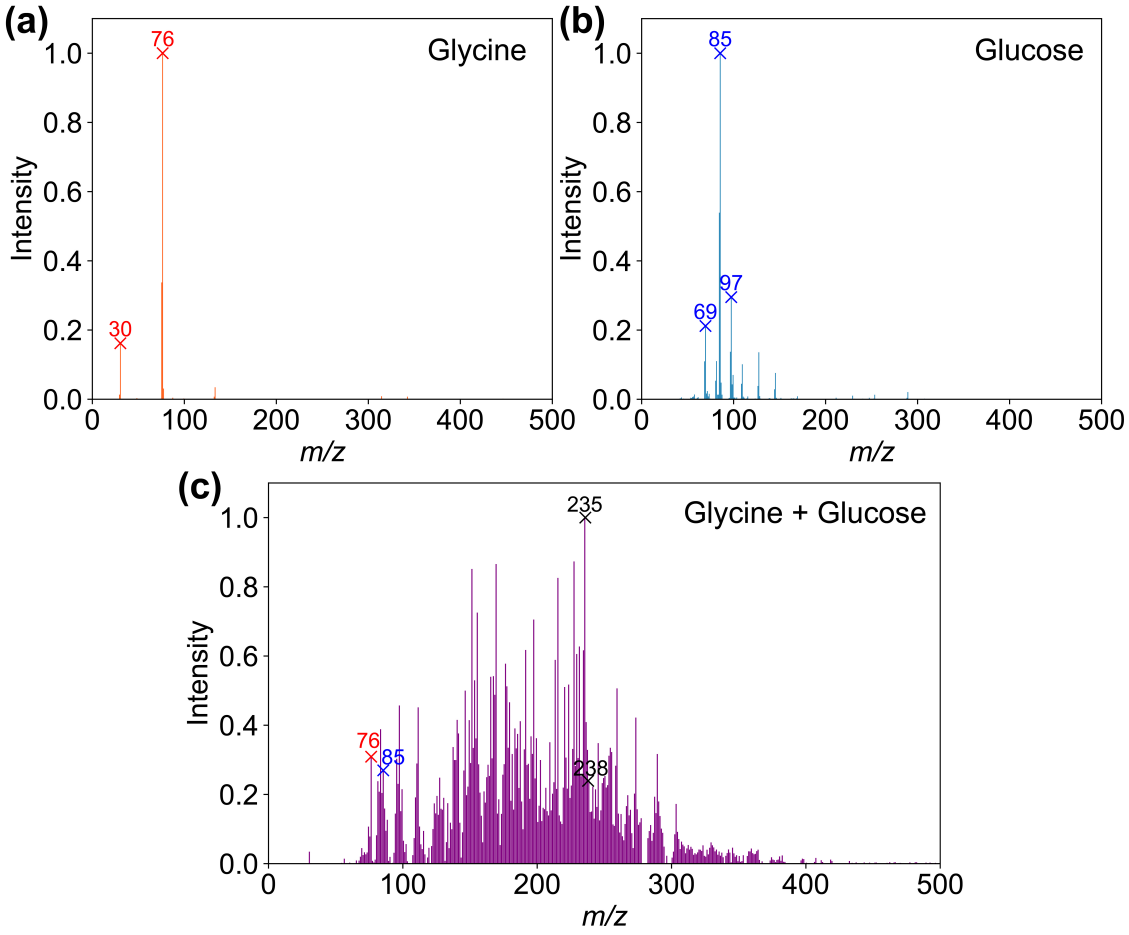

Figure 3.1: Average positive-ion ASAP–MS mass spectra acquired from five replicate measurements of (a) glycine, (b) glucose, and (c) a 1:1 glycine–glucose mixture.

3.3 Results and Discussion

3.3.1 Single molecules and molecular mixture

Several small metabolites, including dopamine, glutamate, epinephrine and cholesterol, produced relatively simple and readily interpretable ASAP–MS mass spectra, with dominant ions corresponding to the intact molecule or characteristic low-energy fragments. In contrast, D-2-hydroxyglutarate (D-2-HG) did not yield a detectable signal under the same conditions.

The ASAP mass spectrum of glycine (Figure 3.1(a)) was dominated by the protonated molecular ion at $m/z = 76$ ($[\mathrm{M+H}]^+$). In contrast, glucose (Figure 3.1(b)) produced a slightly more complex spectrum in which the dominant peak arose

at $m/z = 85$ rather than at the protonated molecular ion mass of $m/z = 181$. This peak is consistent with a multiply dehydrated and decarbonylated hexose fragment ($[\mathrm{M+H} - 2\mathrm{CH_2O} - 2\mathrm{H_2O}]^+$).[142] Other fragments observed for glucose are also consistent with a previous study.[142]

In contrast to the relatively simple spectra of the individual compounds, the glycine–glucose mixture produced a much more complicated spectrum (Figure 3.1(c)). The most intense peak was observed at $m/z = 235$. This peak is interpreted as a mixture-derived reaction or fragmentation product. One possible explanation is a product containing two glycine units and an 84 Da glucose-derived fragment, giving 2×75 Da + 84 Da + $\mathrm{H}^+ = 235$. The spectrum also contained a peak at $m/z = 238$, which is consistent with protonated monoglycated glycine, $[\mathrm{Gly + Glc - H_2O + H}]^+$.[143]

However, the assignment above remains tentative. MS^1 data alone cannot resolve the exact identities of the ions, and the high density of overlapping signals in this mass region likely reflects the simultaneous formation of multiple species arising from in-source dehydration, fragmentation, Maillard-type chain reactions, and ion-molecule association. Such spectral congestion is typical for amino acid-sugar mixtures analysed by ASAP or other APCI-like techniques, where the high-temperature corona-discharge environment promotes extensive in-source chemistry.[144]

The lipid spectra exhibit simpler behaviour. Triolein (Figure 3.2(a)) and SAPC (Figure 3.2(b)) each produce well-defined high-mass ions with limited fragmentation. The mass spectrum for triolein is dominated by the peak at $m/z = 603$, which can be assigned to fragments formed by loss of a single fatty acid RCOOH from the protonated molecule, i.e. $[\mathrm{M+H-RCOOH}]^+$.[145]. The mass spectrum for SAPC is dominated by the peak at $m/z = 627$, which is formed by loss of a phosphocholine, i.e. $[\mathrm{M+H-phosphocholine}]^+$. When the two lipids are measured as a mixture (Figure 3.2(c)), the spectrum is essentially a sum of the individual profiles: the dominant SAPC ions at $m/z = 627$ and the triolein ions at $m/z = 603$ are retained with only minor changes in relative intensity. No new mixed-lipid adducts or complex interaction-dependent species are observed.

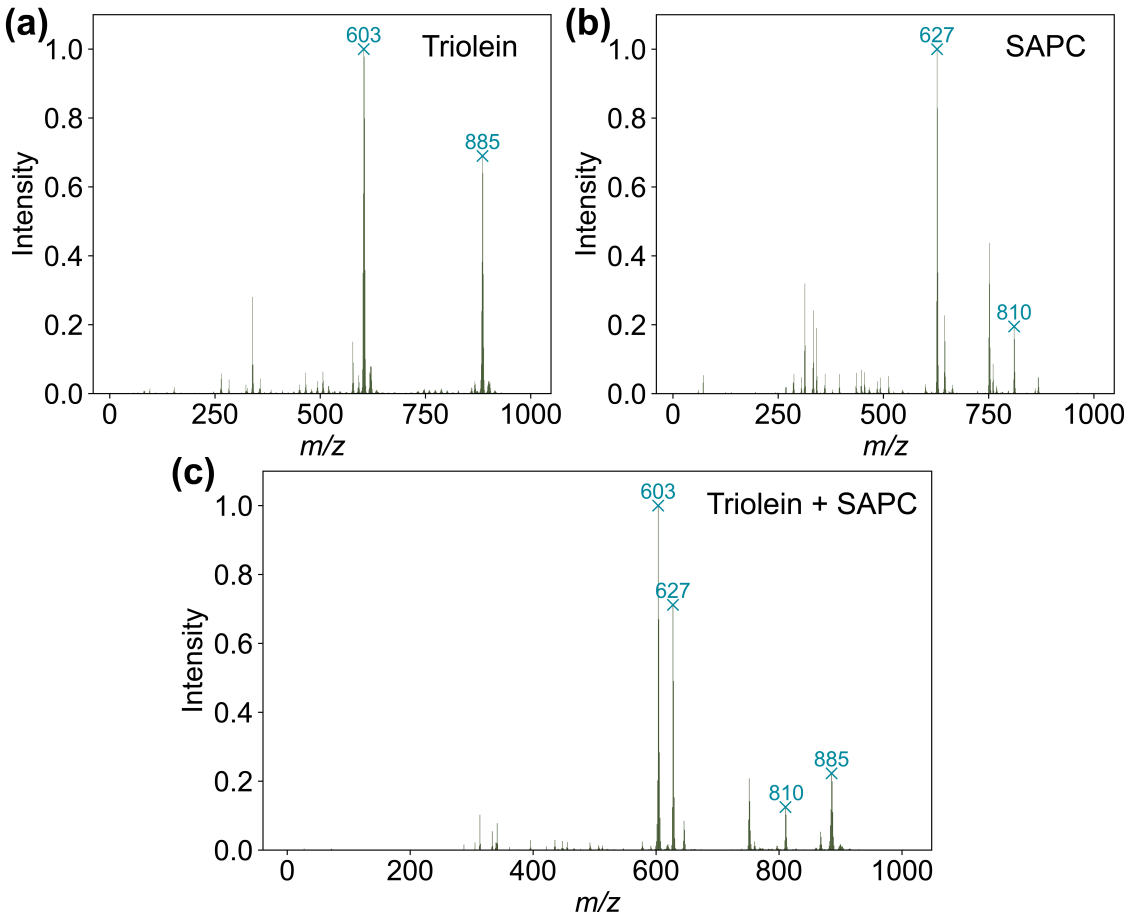

Figure 3.2: Average positive-ion ASAP–MS mass spectra acquired from five replicate measurements of (a) triolein, (b) SAPC, and (c) a 1:1 triolein-SAPC mixture

A clear difference was observed between metabolite- and lipid-type analytes under ASAP–MS conditions. The glycine-glucose mixture produced a highly complex spectrum dominated by overlapping ions arising from extensive in-source chemistry. In contrast, the triolein-SAPC mixture was essentially the linear sum of the individual profiles, with no evidence for extensive in-source chemistry. These data indicate that, under the present conditions, complex samples rich in lipids are more likely to produce interpretable ASAP–MS mass spectra than metabolite-rich samples dominated by small, highly reactive polar molecules.

Besides the differences described above, another important observation is the distribution of discriminative m/z features across the mass range. Many low-m/z signals below m/z 500 are likely to arise from small metabolites, amino acids, fragments, or other low-molecular-weight compounds, whereas features above m/z

500 are more consistent with lipid-related ions.

However, the assignments in this study remain putative. All the discriminative features were treated primarily as statistical variables rather than confirmed chemical identities, because ASAP–MS on a single quadrupole instrument does not provide sufficient structural information for confident molecular annotation. Future work should therefore focus on the chemical identification of these discriminative features. This could include accurate-mass analysis using high-resolution mass spectrometry, tandem MS fragmentation, and comparison with interested molecule standards.

3.3.2 Time-resolved spectral differentiation of intact and lipid-extracted samples

Figure 3.3 explores the progression of spectral features during ASAP–MS analysis of intact and lipid-extracted plasma and brain samples. Figure 3.3 (a) captures the evolving spectral signature during measurement and shows how it diverges between intact and extracted samples. The trends observed in Figure 3.3 (a) are quantified in Figure 3.3 (b), which shows the Euclidean distance between the mass spectra of the intact and extracted samples over time. The intact and lipid-extracted plasma samples initially follow a similar trajectory in PCA space, but begin to separate significantly after 5 seconds. This early similarity likely arises due to the volatility-driven mechanism of ASAP–MS. Hydrophobic lipid species, possessing higher vapour pressures, require less thermal energy to desorb and are therefore detected in the early, low-temperature phase of the scan.[71] Furthermore, these lipids likely induce charge competition in the APCI source, effectively suppressing the ionisation of less volatile polar metabolites. As the measurement progresses and the probe temperature increases, intact plasma samples diverge sharply from their lipid-extracted counterparts, as quantified by the increasing Euclidean distance in PCA space(Figure 3.3(b)). This divergence likely reflects a reduction in lipid-driven ion suppression, enabling the delayed emergence of low-mass, polar metabolites that are abundant in plasma but largely removed by lipid extraction. Such observation is also supported by the mass spectra in Figure 3.3(c). All cumulative average

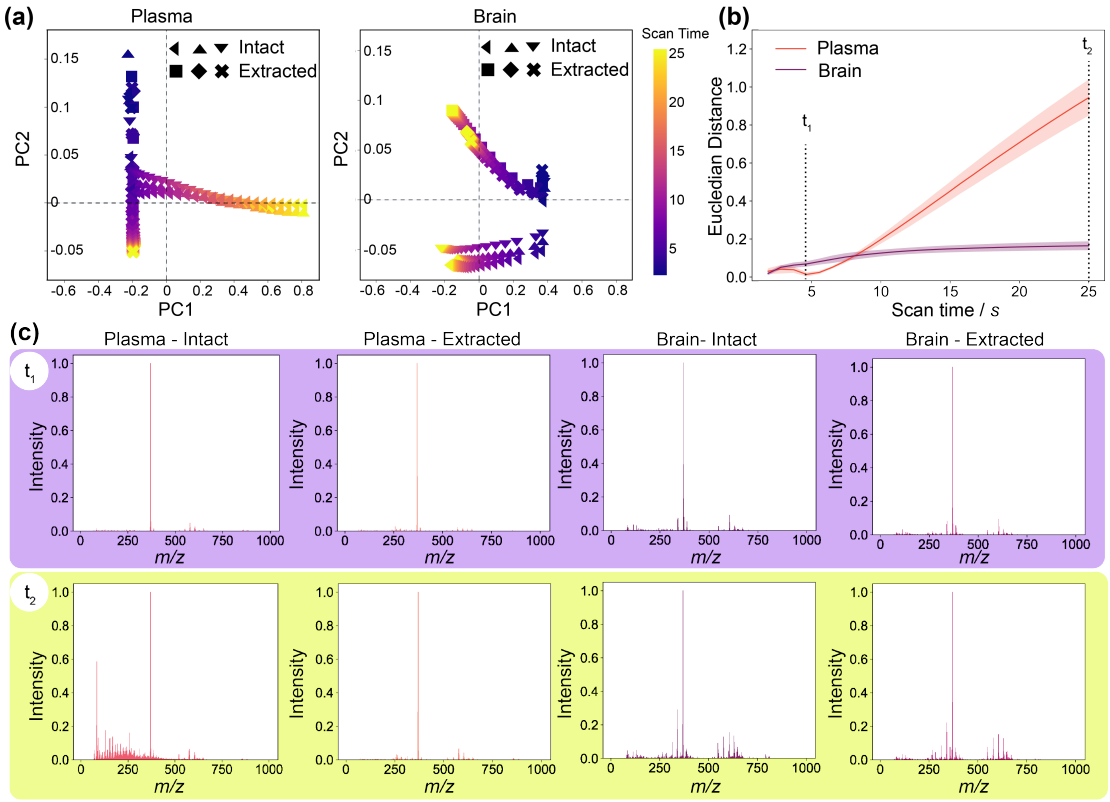

Figure 3.3: Comparison of mass spectra obtained from intact and lipid-extracted biomedical samples using ASAP–MS. (a) Principal component analysis (PCA) of plasma and brain samples, where each point represents the cumulative average mass spectrum from 0 s to the given time point. Data points are colored by scan time over the 25 s acquisition. (b) Euclidean distance between the intact and lipid-extracted samples over time. Shaded areas represent standard deviation. t_1 marks the time point with the smallest Euclidean distance between intact and extracted plasma sample; t_2 marks the final time point of measurement. (c) Mass spectra of intact and extracted plasma and brain samples at t_1 and t_2.

ASAP–MS mass spectra at different time points for the samples mentioned above are provided in Section B.

Brain samples show far less separation between intact and extracted forms than plasma samples over the course of the 25 s scan. The intact and extracted samples cluster closely in PCA space (sharing the same PC1 and only diverge in PC2), indicating that lipid extraction has a smaller influence on the dominant ion population detected by ASAP–MS. This observation is consistent with the brain's high lipid content (50–60% dry weight) and extremely low abundance of free polar metabolites ($<$ 1%).[146]

The comparative behaviour of plasma and brain samples (Table 3.1), together

Table 3.1: Comparative ASAP–MS behaviour of plasma and brain samples

Feature	**Plasma**	**Brain**
Dominant biochemical class	Metabolites+lipids	Mainly lipids
Spectral complexity	High	Moderate
Interpretability of ASAP–MS data	Lower	Higher
Temporal spectral evolution	Strong	Minimal
Effect of lipid extraction	Significant	Low

with the molecular analysis show that sample composition strongly governs the complexity of ASAP–MS mass spectra, with metabolite-rich samples producing more complex spectra than lipid-rich samples. Previous work from our group has demonstrated that ASAP–MS can generate reproducible and biologically informative mass spectra from plasma samples[130], thereby establishing the feasibility of applying ASAP–MS to complex biological samples. The study in this chapter extends our previous understanding by providing a mechanistic context, showing that ASAP–MS mass spectra for plasma are shaped by intense in-source chemistry that evolve over the time of acquisition. However, because spectra are typically averaged over the acquisition period and yield reproducible mass spectral patterns, this temporal evolution does not preclude the use of ASAP–MS for plasma analysis. By contrast, brain tissue shows simpler temporal evolution because of its lipid-rich composition and relatively lower abundance of polar metabolites. These findings indicate that ASAP–MS does not uniformly favour all biological matrices and should be interpreted as providing a reproducible spectral fingerprint of sample chemistry rather than a direct or quantitative representation of molecular composition.

Our single- and mixed-molecule experiments demonstrate that individual analytes can contribute multiple correlated spectral features. Additionally, metabolite-rich mixtures can introduce additional non-linear interactions that further distort simple feature-analyte relationships. This "one-molecule-many-peaks" behaviour introduces significant multicollinearity, where spectral features are highly correlated rather than independent. Consequently, ASAP–MS datasets for biological samples violate several assumptions implicit in many commonly applied classification models; these models often assume feature independence, linear separability, and a uniform

Table 3.2: Machine-learning models and hyper-parameters considered for ASAP–MS data

Model	Assumption	Hyper-parameters	Rationale
LR	Linear separability	Regularisation type (L1/L2), regularisation strength (C)	Controls overfitting arising from correlated multi-peak contributions
NB	Feature independence	Smoothing parameter	Mitigates numerical instability but cannot remove peak dependence
KNN	Uniform local density in feature space	Number of neighbours (k), distance metric	Balances noise sensitivity against neighbourhood structure
SVM	Fixed decision margin	Kernel type, margin parameter (C), kernel width (γ)	Enables modelling of non-linear interactions between features
LDA	Shared covariance structure	Shrinkage, dimensionality control	Stabilises covariance estimation in heterogeneous spectra
RF	Unbiased feature sampling	Tree depth, number of trees, features per split	Limits dominance of high-intensity lipid-derived ions

variance structure. On this basis, hyper-parameter optimisation is treated here as a chemistry-driven methodological requirement rather than a purely computational refinement. As summarised in Table 3.2, each model considered in this study is sensitive to some aspects of the behaviour of ASAP–MS mass spectra.

3.3.3 Temporal analysis of ASAP–MS features using *K*-means clustering

As shown in Figure 3.3, ASAP–MS mass spectra evolve over the time of acquisition, with dominant signal contributions shifting from lipid-associated ions at early time points to metabolite-related features at later times. Motivated by this temporal structure, K-means clustering was applied to the time-resolved ion count profiles for each binned m/z feature. Each feature was represented by its ion count as a function of time, and clustering was performed with K=3 to capture broad classes

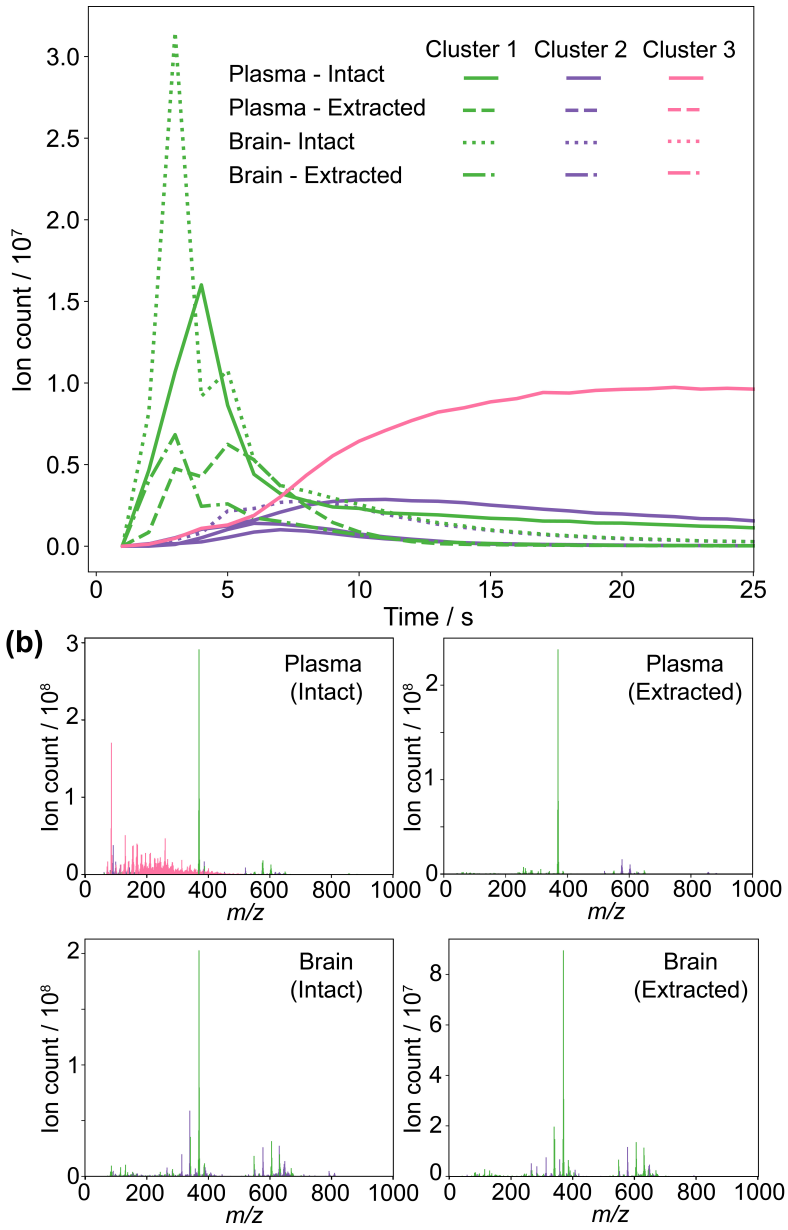

Figure 3.4: Temporal K-means clustering of time-resolved ASAP–MS features. (a) Average ion count profiles over time for three clusters identified by K-means clustering; (b) Corresponding mass spectra with features coloured according to cluster assignment for intact and lipid-extracted plasma and brain samples.

of temporal behaviour. The resulting clusters, shown in Figure 3.4(a), correspond to three temporal trends: Cluster 1: early-time high-intensity peaks; Cluster 2: gradually rising and falling peaks; and Cluster 3: continuously increasing peaks. Notably, Cluster 3 is observed only in intact plasma samples.

Because Cluster 3 consists of ions whose count keeps increasing during mea-

surement, we attribute this behaviour to reaction products formed on the glass capillary tip under the ASAP ionisation environment. When cluster assignments are projected back onto the mass spectra (Figure 3.4(b)), features in Cluster 3 are localised to the dense low-m/z region, which is typically associated with metabolites and their reaction by-products.

This pattern suggests two points: first, that ongoing chemical reactions may occur during the analysis of intact plasma; and second, that time-resolved clustering can highlight metabolite-rich regions of the spectrum. On this basis, K-means clustering provides a useful framework for the effect of lipid extraction on the molecular composition and mass spectra of biofluids. However, the approach has clear limitations. Approximately 67% of features present only in intact plasma, when compared with lipid-extracted samples, overlapped with Cluster 3. This observation indicated the limited discriminatory power of this clustering approach as an independent method to perform such a task. This limitation most probably reflects the fact that not all metabolites participate in continuous or detectable reactions during ASAP–MS analysis. Moreover, the unsupervised nature of K-means clustering and the imposed choice of K constrain the robustness of cluster assignments. Consequently, while K-means clustering offers insight into the temporal organisation of ASAP–MS features, it is not sufficient on its own to define a reliable lipid extraction metric.

However, such information is still valuable. In future work, researchers could exclude the features that show continuously increasing signals during sample measurement before data analysis and machine learning training, as these are likely to reflect time-dependent reaction artefacts rather than native sample composition.

3.4 Conclusions and Future Work

The work described in this chapter shows that the behaviour of ASAP–MS mass spectra is governed by both sample complexity and the ionisation characteristics of different molecular classes. Experiments with single compounds and simple mixtures demonstrate that small polar metabolites undergo extensive in-source

chemistry, producing dense and highly correlated spectral features, whereas lipids ionise more predictably and largely additively. These differences carry through to biological samples: plasma, which is rich in metabolites, yields complex and strongly time-dependent spectra, while lipid-dominated brain tissue produces more stable and easier-to-interpret profiles. Together, these findings define the constraints that direct how ASAP–MS data should be analysed. In particular, they motivate the careful choice and tuning of machine learning models that can integrate weak, distributed signals while remaining robust to correlated features. This will guide the machine learning model training in Chapter 5.

However, the experimental scope was intentionally restricted, both in terms of the range of molecules examined and the number of biological matrices considered. While this focused approach enabled clear interpretation of temporal trends and ionisation behaviour, it does not capture the full chemical and biological diversity encountered in real-world applications.

All measurements in this chapter were performed under a fixed set of ion source conditions. In practical analyses, however, variations in sample concentration and molecular composition are likely to alter ionisation efficiency, ion–molecule reactions, and suppression effects between molecular classes. These effects were not systematically explored here. Similarly, exploratory analyses such as PCA and K-means clustering were used to identify dominant patterns in the data, but these methods are inherently descriptive, sensitive to parameter choices, and not designed to provide definitive separation or mechanistic attribution.

Future work should adopt a controlled, incremental strategy, beginning with single-molecule systems and progressively increasing molecular complexity by adding additional compounds. By recording how mass spectral features change as additional molecules are introduced, the molecular origin of individual peaks can be assigned. Combining this information with the temporal behaviour of each peak then allows a machine-learning model to be trained to distinguish peaks arising from metabolites, or lipids in biological samples. Such models could then be validated using different

biological sample types, comparing intact tissues with their corresponding lipid-extractions, to assess classification robustness and biological relevance.

Another possible direction would be to investigate whether the surface chemistry of the ASAP probe contributes to the formation of some observed ions. However, the first step would therefore be to determine where these reactions are taking place. This is because the relevant chemical processes could occur in the bulk sample, at the capillary surface, during thermal desorption, or in the gas phase. This could be examined by comparing spectra obtained using probes with different surface properties, different loading times, different heating profiles, and different sample-preparation conditions. If specific ions were found to arise from surface-catalysed reactions, then probe surface modification could in principle be used to suppress or modify these processes. For example, alternative glass treatments, inert coatings, or different capillary materials could be tested empirically to determine whether the abundance of reaction-derived ions changes. A more mechanistic approach would involve identifying the reaction pathway and selecting surface chemistries designed to interfere with key steps in that process. However, such work would require dedicated chemical-identification and mechanistic studies.

4

ASAP–MS Brain Region Classifiers using FFPE and Frozen Neuropathological Samples

Contents

4.1 Introduction

As introduced in Section 1.1.1, neuropathology plays a crucial role in diagnosing diseases of the central nervous system (CNS). Traditionally, neuropathological diagnosis includes histological examination, immunohistochemistry, and expert interpretation under a microscope. These steps are often labour-intensive, time-

consuming, and highly dependent on specialised expertise. The shortage of workforce in pathology is a global issue.[147] Thus, there is a growing demand for faster, more automated, and objective diagnostic tools. As introduced, ASAP–MS is an attractive option for neuropathological applications.[71]

Formalin-fixed, paraffin-embedded (FFPE) brain samples are heavily used for research and diagnostics in neuropathology. These FFPE samples serve as a valuable resource. With millions of FFPE samples stored in biobanks worldwide, these samples hold a high potential for accelerating biomarker discovery using ASAP–MS. However, the limitations of FFPE samples have been discussed in several studies. An LC/MS-MS–based metabolomics study on colorectal tissues comparing matched FFPE and fresh frozen samples detected 200 metabolites in FFPE samples versus 536 in fresh frozen tissues, with only 15 metabolites commonly identified in both sample types. [148] Besides this, a variety of studies have demonstrated that the FFPE process significantly impacts the metabolic profile of tissues, including causing washout of polar metabolites, lipid oxidation, and the formation of fixative-related reaction products. [149] Despite this, a recent untargeted brain tumour metabolomics study employing FFPE samples identified 2-hydroxyglutarate (2HG), associated with isocitrate dehydrogenase (IDH) mutation, and aminoadipic acid (AAA) and guanidinoacetic acid (GAA), both used in diffuse glioma grading, with models developed from frozen samples showing moderate diagnostic performance when validated on FFPE tissues.[150]

One important limitation of FFPE samples is variability in the fixation protocol. In particular, fixation duration can influence the molecular composition preserved in the tissue. FFPE samples may therefore be broadly distinguished as either short-fixed or long-fixed, with short-fixed samples referring to tissues fixed for less than 48 h and long-fixed samples referring to those fixed for more than 48 h. The differences in fixation duration may contribute to variation in the molecular profiles detected by mass spectrometry.

Another important limitation of FFPE samples in the context of mass spectrometric measurements is the presence of paraffin. In ASAP–MS measurements,

excessive residual paraffin (Figure 4.1) can cause ion suppression, reduce analyte detection, and increase background noise, often resulting in mass spectra that bear little relationship to the original tissue samples. Deparaffinisation, therefore, is

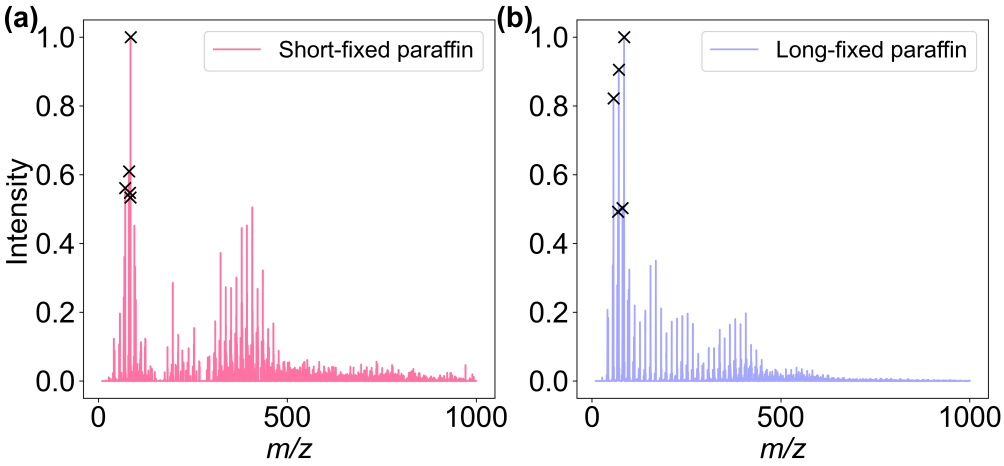

Figure 4.1: Paraffin mass spectra: (a) paraffin from a short-fixed sample, and (b) paraffin from a long-fixed sample. The top five highest m/z peaks are 85, 71, 57, 81, and 83.

crucial for obtaining reliable mass spectra using ASAP–MS. Several metabolic studies on FFPE samples using traditional xylene-ethanol deparaffinisation have effectively removed paraffin; however, this method is limited by the potential loss of biological molecules during the process and the inherent toxicity of xylene. Our research group evaluated multiple FFPE tissue samples deparaffinised using the xylene–ethanol protocol, and found that approach consistently yielded low-quality mass spectral data when measured by ASAP–MS.

The Covaris ultrasonicator is a device widely used for sample processing in molecular biology and analytical workflows. It uses adaptive focused acoustics (AFA) technology to deliver controlled ultrasonic energy, which allows for precise and reproducible disruption, emulsification, or mixing of samples. Unlike traditional sonicators, the Covaris system provides better temperature control and consistency, reducing the risk of sample degradation. This method is appealing and shows strong potential for effective paraffin removal in ASAP–MS analysis.

Thus, the aim of this study is to explore whether the Covaris ultrasonicator offers a quicker and safe deparaffinisation method that preserves a reasonable range of

biological molecules, and to evaluate the feasibility of using ASAP–MS for analysing FFPE brain samples by comparing with frozen brain samples.

4.2 Methods

4.2.1 Protocol optimisation

The measurement protocol was established in Chapter 2, Section 2.2. The deparaffinisation protocol was developed by Covaris for general sample preparation and DNA/RNA extraction. We adapted it to make it more suitable for ASAP–MS analysis. In the original protocol, the tissue section was placed in a microTUBE-130 AFA Fiber Screw-Cap (Covaris), and 100 μL of aqueous buffer was added. The sample was sonicated for 5 minutes using the following settings: peak incident power (PIP) 350 W, duty factor (DF) 20%, cycles per burst (CPB) 200, water level 15, and temperature 20°C. This step helped emulsify the paraffin. After 2 minutes of centrifugation, the paraffin floats to the top of the mixture and the supernatant was collected.

In the context of MS measurements, the buffer was found to introduce strong background signals. Such strong signals interfered with ASAP–MS analysis. To address this, we explored alternative solvents to replace the buffer. We tested water, methanol, and a 1 : 1 mixture of water and methanol. The ideal solvent should enable clear separation of the sample pellet from the paraffin layer.

As shown in Figure 4.2. Methanol produced the clearest separation between solvent and sample pellet. Water was less effective than methanol, and the mixed solvent showed no separation. Therefore, methanol was selected for further comparison with the xylene–ethanol method.

4.2.2 Comparison of deparaffinisation methods

A formalin-fixed, paraffin-embedded (FFPE) cerebellum tissue block from Oxford Brain Bank (OBB) with a specimen code NP10/18 was sectioned, with one section placed into each of two tubes. A parallel experiment was conducted using two different deparaffinisation methods.

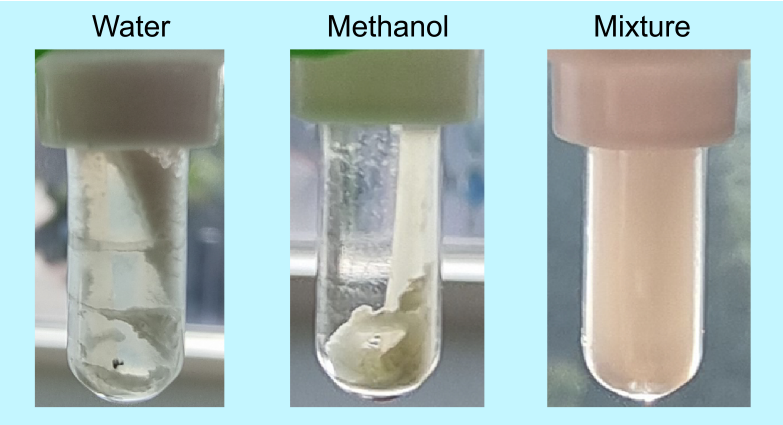

Figure 4.2: Visual comparison of solvent selection for the Covaris deparaffinisation method. Three solvents were tested: water, methanol, and a 1:1 mixture of water and methanol. The images show the samples after Covaris deparaffinisation treatment using the corresponding solvents.

1. Method 1 followed the traditional xylene-ethanol approach. The section was treated with xylene and centrifuged at 14.8 kRPM for 2 minutes. The supernatant was removed, and the sample was treated with ethanol and centrifuged at 14.8 kRPM for 2 minutes, followed by removal of the supernatant. The sample was then placed on a thermo mixer (Eppendorf ThermoMixer F1.5) to vaporise any remaining ethanol. Subsequently, 100 µL of LC-MS grade water (Supelco) was added, and the sample was homogenised using the OMNI bead homogeniser.

2. Method 2 utilised the protocol described above in Section 4.2-1, employing the Covaris E220 Demo ultra-sonicator.

Samples processed using both methods were subsequently analysed with ASAP–MS. Five repeat measurements were performed for each sample. The total ion count (TIC) was used to quantify the residual molecular content after deparaffinisation. The intensities of the five most intense residual paraffin peaks within the two sets of spectra were analysed to evaluate and compare the effectiveness of these two methods.

4.2.3 Comparison between short- and long-fixed FFPE samples

The Oxford Brain Bank (OBB) provided cerebellum samples from the same five patients, collected from the same brain region but different hemispheres, and fixed using both short- and long-fixation methods. The fixation time for the Long-fixed samples is shown in Table 4.1. Method 2 was used to deparaffinise the samples, followed by analysis using ASAP–MS. All samples were measured in a single batch.

Spectral overlays were generated to visually compare molecular fingerprints within each group. Principal Component Analysis (PCA) was used to visualise global differences between the paired samples. Pearson correlation coefficients were calculated to assess intra-group and inter-group spectral similarity. Total ion counts were extracted for each spectrum and compared across conditions to assess signal intensity.

Table 4.1: Fixation time for Long-fixed Sample

OBB ID	Fixation Time / days
NP10/18	3
NP11/18	3
NP30/18	3
NP92/18	5
NP95/18	4

4.2.4 Comparison between frozen and FFPE samples

Two sets of samples stored using different preservation methods (frozen and FFPE) were measured in this study. A total of 19 frozen cerebellum and 19 frozen motor cortex samples were obtained from 19 individual patients, with each patient providing both tissue types. Similar to frozen samples, 19 FFPE cerebellum and 19 FFPE motor cortex samples were collected from a separate cohort of 19 patients, with each patient providing both cerebellum and motor cortex tissues (details in Table C.1). Frozen samples were prepared and measured using a previously established protocol(see in Section 2.2). FFPE samples were prepared and measured

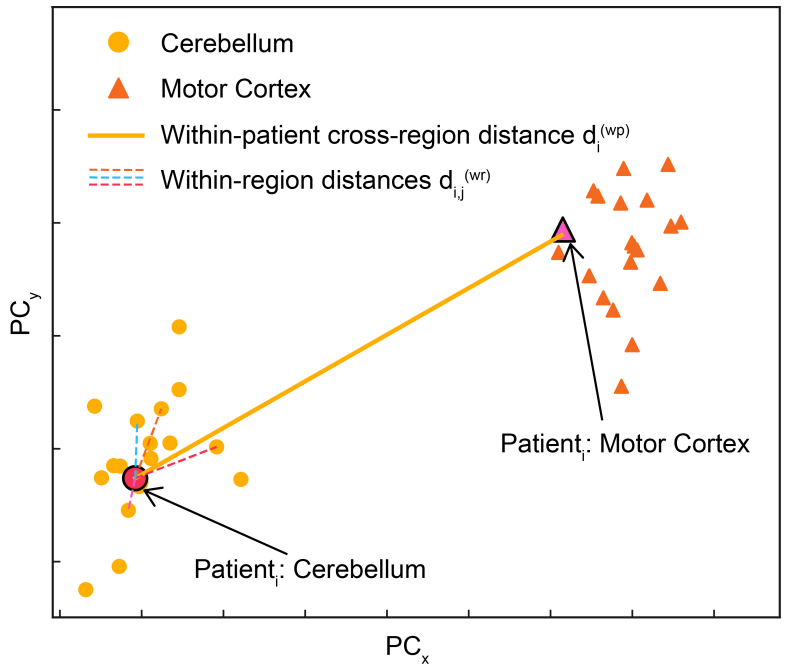

Figure 4.3: Illustration of the distance comparison method used in this study. Each point represents a sample in PCA space. For a given patient, the Euclidean distance between their two samples from different brain regions is calculated as the within-patient cross-region distance ($d^{(\mathrm{wp})}$, shown as a solid line). The average Euclidean distance between a sample and other samples from the same region (excluding itself) is computed by averaging the within-region distances ($d^{(\mathrm{wr})}$, shown as dashed lines).

using Method 2. Each data set was measured in two batches and ComBat was used to perform batch effect correction.[116, 126] The correction reduced batch-associated variation while preserving the main sample-group differences, so that the observed trends were not solely driven by measurement batch.

For each dataset, PCA was performed to examine the overall variation among samples. The Kruskal–Wallis (H) test was then used to assess correlations between principal components from the PCA and brain region.[122, 125] Finally (see Figure 4.3), we compared (1) Euclidean distances between samples from different regions within the same patient, and (2) average Euclidean distances between each sample and other samples from the same brain region, in order to assess the impact of the sample preservation method. The distance functions are defined as follows:

Let $i = 1, \ldots, N$ denote the individual patients, and $r \in \{C, M\}$ denote the two brain regions (C = cerebellum and M = motor cortex). The quantity $\mathbf{x}_{i,r} \in \mathbb{R}^p$ denotes the PCA score vector (e.g., $(\mathrm{PC1}, \mathrm{PC2}, \ldots)$) for patient i and brain region r.

The within-patient cross-region distance is defined by:

$$d_i^{(\mathrm{wp})} \;=\; \|\mathbf{x}_{i,C} - \mathbf{x}_{i,M}\|_2 \,. \tag{4.1}$$

Where "$\|\quad\|_2$" denotes the Euclidean distance.

For a given sample i, the average Euclidean distance to all other samples from the same region, denoted the "with-in region average distance" is

$$\bar{d}_i^{(\mathrm{wr})} = \frac{1}{|N-1|} \sum_{j \neq i} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \,, \tag{4.2}$$

After evaluating the distance metrics of interest, the set $\{d_k^{(\mathrm{wp})}\}$ is compared with the set $\{\bar{d}_i^{(\mathrm{wr})}\}$ using a two-sided Mann–Whitney U test to evaluate whether within-region distances are systematically smaller than within-patient cross-region distances.

4.3 Results and Discussion

4.3.1 Comparison of deparaffinisation methods

Figure 4.4(a) shows ASAP–MS data for samples deparaffinised with two methods discussed in Section 4.2.2. The FFPE samples deparaffinised via Method 2 (focused-ultrasonication) yielded about 4.5 times higher total ion count across all scan times compared to those treated with the xylene–ethanol method (Method 1). The green trace (Method 2) exhibits sharp and well-defined ion peaks, consistent with the optimal peak characteristics reported in our prior work.[78] In contrast, the purple trace (Method 1) shows reduced and flat peak profiles, indicating compromised ionisation efficiency and molecular loss during the deparaffinisation process.

The ion counts for the five most intense paraffin-associated peaks (marked $\times$) in both mass spectra are similar (Figure 4.4 (b)). This suggests that both deparaffinisation methods appear to perform similarly in terms of paraffin removal. However, samples treated with Method 2 show an obviously richer and more complex spectral profile across the full m/z range. This also suggests that the higher total ion count observed in the samples treated using Method 2 is not merely due to residual

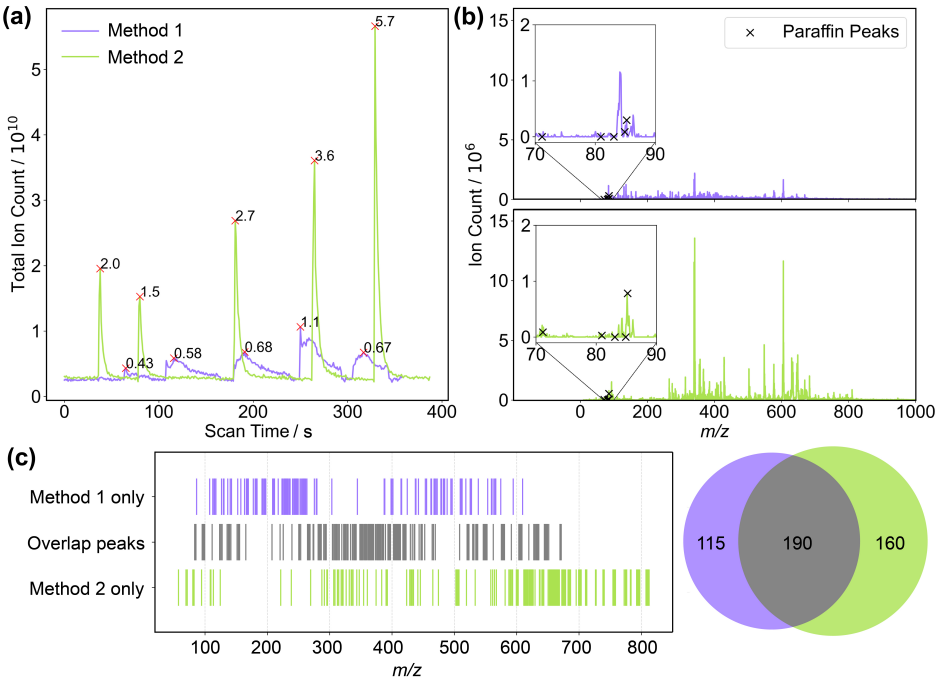

Figure 4.4: Comparison of deparaffinisation methods for the same FFPE brain tissue sample measured by ASAP–MS. Method 1: xylene-ethanol; Method 2: focused-ultrasonication. Quantitative comparison between two deparaffinisation methods: (a) Total ion current (TIC) traces of FFPE samples deparaffinised using Method 1 and Method 2. Each sample was measured in five replicates; (b) Averaged mass spectra from each method, illustrating spectral complexity, richness and paraffin-related background (Top five paraffin related peaks are marked as '×', see Figure 4.1); (c) Unique and overlapping peaks between the two deparaffinisation methods: Purple bars indicate peaks unique to the Method 1; Green bars indicate peaks unique to the Method 2; Grey bars represent peaks common to both methods. Venn diagram quantifies the overlap between the two extraction methods in terms of the total number of detected m/z peaks: Method 1 yielded 115 unique peaks. Method 2 yielded 160 unique peaks. 190 peaks were common to both methods.

paraffin but instead reflects a true increase in signal intensity from biologically relevant species. The comparable levels of paraffin-related ions across both methods supports the conclusion that the enhanced spectral complexity and total ion yield are driven by improved extraction method.

Figure 4.4(c) qualitatively compares the distribution of detected mass-to-charge (m/z) peaks across Method 1 and Method 2. The visual pattern shows that Method 2 produces a broader range of unique peaks across the m/z spectrum, particularly in the mid-to-high m/z range. In contrast, Method 1 appears to yield fewer unique peaks, and these are more concentrated in the low-to-mid m/z range. The

overlapping region (grey) indicates a substantial number of shared features, though differences between the methods are clearly visible. Furthermore, the Venn Diagram in Figure 4.4(c) shows that while there is a substantial core set of common features between both methods, Method 2 provided more unique peaks, suggesting it may provide richer molecular coverage. The data support the conclusion that Method 2 is more effective at liberating a broader set of molecular species from FFPE tissue. Based on these results, we selected Method 2 (Focused-ultrasonication) as our deparaffinisation approach for subsequent analyses.

4.3.2 Comparison between short- and long-fixed FFPE samples

As explained in Section 4.2.3, to assess the impact of formalin fixation duration on the quality and consistency of mass spectrometric molecular profiling, five matched short-fixed and long-fixed FFPE cerebellum samples were used in this study. After deparaffinisation, these samples were measured using ASAP–MS. Overlaying the mass spectra for each sample group (Figure 4.5(a)) shows that the short-fixed samples have strong spectral alignment, with clear and consistent peak positions and intensities across the five samples. In contrast, the mass spectra for long-fixed samples show less overlap. In line with these observations, the PCA plot shows a clear separation between short-fixed and long-fixed samples along the first principal component (PC1), which explains 78.09% of the total variance (Figure 4.5(b)). Furthermore, short-fixed samples show much tighter clustering on the PCA plot than long-fixed samples.

To further examine intra-group consistency, Pearson correlation heatmaps were generated for each fixation group (see Figure 4.5(c). Short-fixed samples show very high inter-sample correlations. In contrast, long-fixed samples show considerable variability. A direct comparison of total ion counts (TIC) between mass spectra recorded for matched short- and long-fixed samples (Figure 4.5(d)) suggests a general trend of reduced signal intensity in long-fixed tissues. Despite some shared features, pairwise Pearson correlations between matched short- and long-fixed

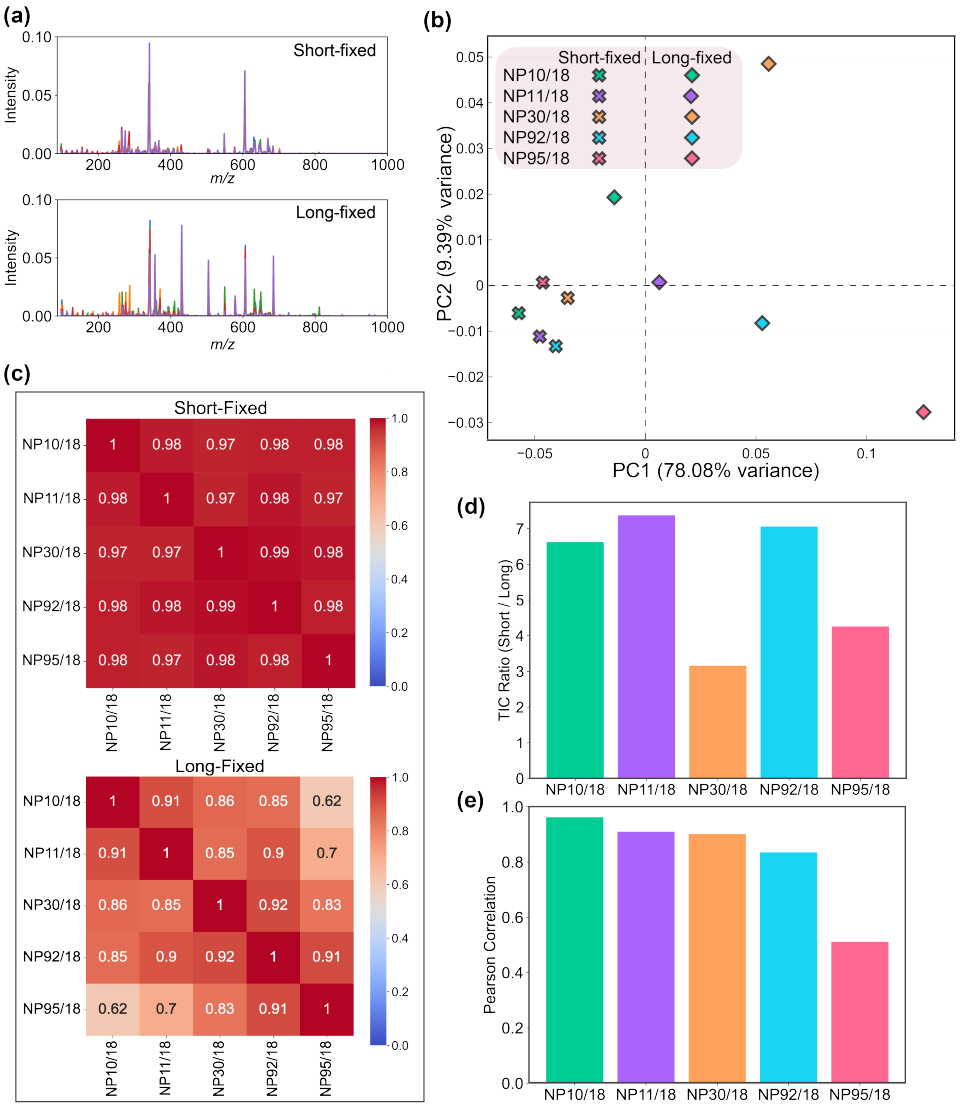

Figure 4.5: Comparison between mass spectra for short-fixed and long-fixed FFPE samples after deparaffinisation: (a) Overlaid mass spectra of the five short-fixed (top) and long-fixed (bottom) samples; (b) PCA plot of five matched short-fixed and long-fixed FFPE cerebellum samples; (c) Intra-group Pearson correlation heatmaps for short-fixed (top) and long-fixed (bottom) sample mass spectra; (d) Comparison of total ion counts from ASAP mass spectra between matched short-fixed and long-fixed samples, revealing a decrease in overall signal intensity in long-fixed tissues; (e) Pearson correlation of ASAP mass spectra between matched short- and long-fixed samples, reflecting the degree of molecular preservation with prolonged fixation.

samples (Figure 4.5(e)) are generally lower than the short-fixed sample intra-group correlations ($0.97 - 0.99$). Interestingly, even among samples fixed for the same duration, variability was observed. For example, NP10/18, NP11/18, and NP30/18 were all fixed for 3 days, yet their correlations with the corresponding short-fixed

samples varied: 0.96, 0.90, and 0.90, respectively. This suggests that factors beyond just the fixation time may contribute to molecular variability. Moreover, sample NP95/18, which was fixed for 4 days, shows a much lower matched short-fixed sample correlation of 0.51, substantially lower than NP92/18, which was fixed for 5 days with a correlation of 0.83. This inconsistency highlights the factor that prolonged fixation does not correspond linearly to degradation, but rather introduces unpredictable effects on molecular profiles.

To summarise, the spectra from short-fixed samples are more reproducible and overlap better across replicates. Long-fixed samples, on the other hand, show weaker signals, and greater differences between replicates. The results above indicate that short fixation helps to preserve consistent and information-rich mass spectral profiles in FFPE tissues. In contrast, long fixation introduces more variability and leads to a clear and unpredictable loss of signal. Our findings are consistent with previous studies reporting that prolonged fixation times of three days or more are detrimental to sample quality. [151, 152] One mass spectrometry imaging study reported that formalin fixation can cause washout of polar metabolites, and that slower fixation increases lipid hydrolysis in the tissue core, which may help to explain our findings to some extent.[149] These findings address the importance of using standardised, short fixation protocols in tissue processing, especially when introducing new technologies such as molecular profiling or mass spectrometry-based studies.

4.3.3 Comparison between frozen and FFPE samples

As explained in Section 4.2.4, we collected ASAP–MS data from both frozen and FFPE brain samples in order to assess whether ASAP mass spectra could distinguish two brain regions: cerebellum and motor cortex. We employed PCA to determine whether the natural sample variations could separate these brain regions. As illustrated in Figure 4.6(a), the PCA plot for the frozen-sample dataset shows two clearly separated clusters, with slightly overlapped confidence ellipses. In contrast, the FFPE-sample PCA plot (Figure 4.6(b)) shows two ellipses that overlap perfectly, suggesting almost no separation between the spectra recorded for

different brain regions. These findings indicate that the spectral variation in frozen samples is sufficient to distinguish cerebellum from motor cortex. However, this feature appears to be lost in FFPE samples. This observation is further confirmed by the Kruskal–Wallis H test results shown in Figure 4.6(c), which examine the correlation between brain region and principal components from the PCA analysis above. In the frozen-sample dataset, both PC1 and PC4 demonstrate significant associations with brain region. No such correlations are observed for any component in the FFPE data, although PC4 has a p-value only slightly higher than 0.05, which may suggest a weak trend. Such results suggest that frozen samples preserve considerably more metabolic information than FFPE samples.

Similar trends have also been reported in other untargeted metabolomics studies comparing FFPE and frozen tissue. For example, Cacciatore *et al.* demonstrated that, while FFPE samples can yield reproducible metabolic profiles, the number and intensity of detected metabolite signals are generally lower than in matched frozen samples, with certain metabolite classes, particularly polar compounds, being disproportionately reduced [153]. Other studies have reached similar conclusions, showing that fresh-frozen tissue better preserves the native metabolic state, whereas FFPE preparation can lead to selective loss or alteration of metabolites [148, 154, 155]. However, some of these studies have also reported that FFPE samples can still support meaningful classification or diagnostic modelling, sometimes approaching the performance of frozen tissue when workflows are carefully optimized [148, 155]. Our data did not show such an effect. This could be because our FFPE samples were long-fixed (fixation day from two days to three years, details in Table C.1.), which likely reduced total ion count signal significantly based on the results from Section 4.3.2. In addition, between-patient batch effects could also mask the biological differences.

To further understand between-patient batch effects, we compared: (1) the Euclidean distances between samples from different brain regions within the same patient; and (2) the average Euclidean distances between each sample and other samples from the same brain region, for both the frozen and FFPE sample PCA plots (Figure 4.6(d) and (e)). We used the Mann-Whitney U test to perform significance

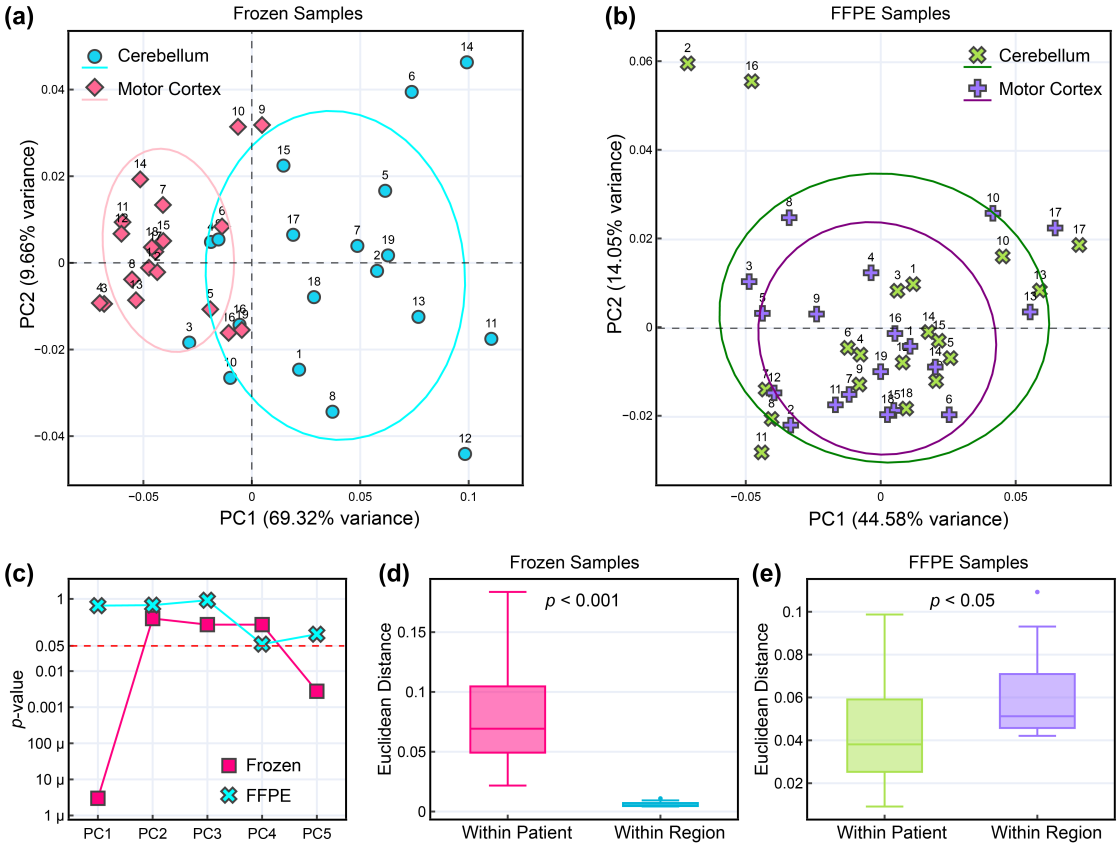

Figure 4.6: Comparison between frozen and FFPE samples: (a) PCA plot of matched frozen cerebellum and motor cortex samples from 19 patients, with one-standard-deviation confidence ellipses. (b) PCA plot of matched FFPE cerebellum and motor cortex samples from the same 19 patients, with one-standard-deviation confidence ellipses. For both data sets, each point is labelled with a unique number representing an individual patient. (c) Kruskal–Wallis (H) test showing p values for the correlation between the first five principal components of the mass spectra and brain regions in both data sets. The red dashed line indicates the upper threshold for statistical significance. (d) Box and whisker plots for the frozen data set comparing: (1) Euclidean distances between samples from different regions within the same patient, and (2) average Euclidean distances between each sample and other samples from the same region. p values are from a Mann–Whitney U test. (e) Same comparison as in panel (d), applied to the FFPE data set.

analysis (details in Table C.2). In the frozen-sample dataset, the distances between data points corresponding to the samples from the same brain region are significantly smaller than the distances between data points from different brain regions within the same patient. This implies that frozen brain samples are more similar to other samples from the same region than to the other sample from the same patient, and that freezing brain samples could introduce few to no batch effects. In contrast, the FFPE data set shows the opposite pattern. In FFPE samples, the distances

between data points from the same patient are smaller than those between data points from the same brain region. This implies that FFPE samples are more similar within the same patient than within the same brain region, suggesting that the fixation process introduced strong batch effects in the FFPE samples.

As mentioned above, the FFPE samples used in this study are all long-fixed samples due to limitations associated with sample access to short-fixed samples with a standardised fixation protocol. Unsurprisingly, the information we can extract from these long-fixed samples is very limited.

4.4 Conclusion and Future Work

We have developed a rapid deparaffinisation protocol designed to preserve as much molecular information in FFPE tissue samples as possible for ASAP–MS measurements. Through a comparison of short- and long-fixed FFPE tissues, we found that short-fixed samples retain more consistent molecular fingerprinting features, and are therefore more suitable for ASAP–MS measurements. When comparing preservation methods, we found that frozen samples contain more spectral information than FFPE samples. It is important to note that the FFPE samples analysed in the brain region data set were all long-fixed, which may have contributed to the lower molecular information.

These findings demonstrate the strong influence of fixation time and preservation method on the quality of ASAP–MS data. In particular, fixation in FFPE processing appears to introduce unpredictable batch effects and diminish spectral diversity. On this basis, frozen tissue should be preferred for ASAP–MS analysis whenever feasible. When FFPE material is unavoidable, a rigorously standardised short-fixation protocol applied uniformly across samples is essential. Consequently, meaningful retrospective ASAP–MS analysis using FFPE samples prepared without such control is, in practice, unlikely.

For future work, it will be important to evaluate short-fixed FFPE samples using strict fixation protocol under the same ASAP–MS workflow used here, to determine whether the reduced fixation time and strict fixation protocol can improve data

quality and reduce batch effects. Additionally, expanding the study to include a wider range of brain regions, larger sample sizes, and other tissue types could provide a more comprehensive understanding of how preservation methods affect spectral profiles. Ultimately, these improvements could help make FFPE tissue a more viable source for large-scale ASAP–MS-based studies.

5

Development of a Brain Tumour Classifier Using ASAP–MS

Contents

5.1 Introduction

The practical challenge for an intraoperative method is to obtain chemically informative data rapidly and at low consumable cost. As discussed in Chapters 1 and 3, ASAP–MS, despite being a simple technique with certain limitations, provides a viable approach that is worthy of further investigation.

In this study, we developed an optimised protocol for the application of ASAP–MS to the analysis of fresh brain tissue samples. Subsequently, we assessed whether

the resulting mass spectra contained sufficient information to enable machine learning–based classification of tumour versus normal brain tissue. The trained models were then applied to classify newly acquired brain tissue sample mass spectra before the availability of the official neuropathological diagnosis. Finally we explored the potential of using the mass spectral information to distinguish subtypes of brain tumour. This work represents an initial step towards establishing ASAP–MS–based brain tissue analysis as a potential alternative or complementary diagnostic tool.

5.2 Optimisation: Data Acquisition and Model Training

5.2.1 Introduction

Brain tissue samples were transferred directly from the operating theatre. Consequently, the protocol for measuring fresh brain tissue samples cannot be adapted from Section 2.2, which describes the protocol for measuring frozen brain tissue samples, because fresh and frozen tissues exhibit distinct physical properties. Furthermore, certain fresh brain tissue samples exhibit residual blood contamination, which may compromise the accuracy and reliability of ASAP–MS signal acquisition. In light of these considerations, it is necessary to establish an optimised ASAP–MS measurement protocol specifically tailored for fresh brain tissue samples. A preliminary investigation was conducted using brain tissue samples from 50 patients to determine the optimal sampling strategy, including measurement methodology and the required number of repetitions. We also systematically tuned machine learning model hyperparameters to assess which configurations perform best on ASAP–MS spectral data.

5.2.2 Methods

Figure 5.1 shows the experimental design of the establishment of an optimal protocol.

Biopsy samples were split, with half of each sample delivered directly from the operating theatre to the mass spectrometry laboratory and the other half sent to the neuropathology laboratory for diagnosis. Two methods were compared to

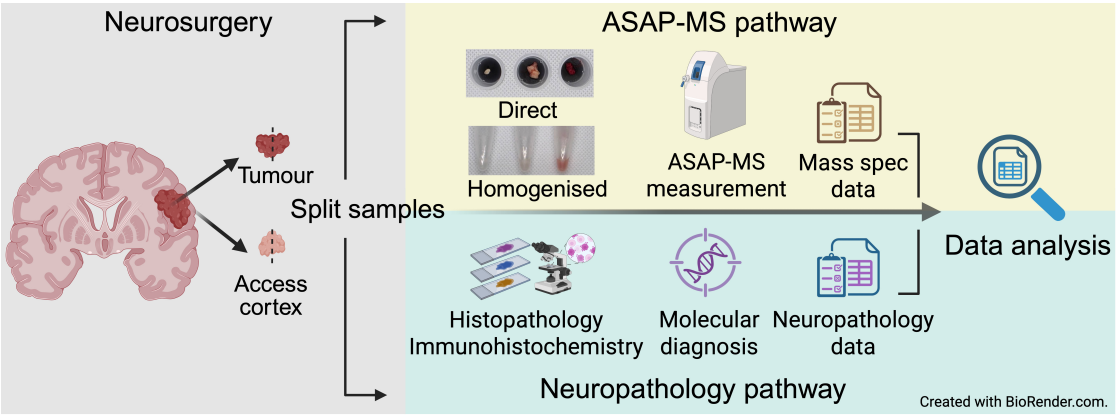

Figure 5.1: Overview of protocol development workflow. ASAP–MS pathway: brain tissue samples were prepared and measured using the ASAP–MS with either direct or homogenised sampling methods, with mass spec data reported. Neuropathology pathway: Histological and molecular diagnosis of the samples, with neuropathology data reported. Data analysis: Merging mass spec data and neuropathology data using specimen code and performing data analysis.

introduce samples to the mass spectrometer, in order to evaluate which approach offered higher repeatability and diagnostic relevance when combined with machine learning (ML) analysis.

1. Direct sampling: Fresh brain tissue samples delivered to the lab from the theatre were gently touched using a glass capillary mounted on the ASAP, with repeat measurement taken from different point on the sample surface.

2. Homogenised sampling: A small portion of brain tissue was excised, placed in a 1.5 mL microcentrifuge tube, and weighed. For homogenization, 100 μL of LC–MS-grade water was added to each sample. Samples were then homogenised using a bead homogeniser with the following settings: speed = 4 m/s, cycle time = 10 s, number of cycles = 6, dwell time between cycles = 10 s. After homogenisation, the samples were vortexed to ensure uniform mixing. The glass capillary from ASAP probe tip was dipped into the homogenised liquid.

Mass spectra for each type of samples were measured, exported and pre-processed using the method introduced in Section 2.2.

To compare signal quality and repeatability between the two sampling methods, we analysed two brain tissue samples within the same batch: one with visible blood contamination and one without. Blood and surgical irrigation fluid can contaminate the samples and confound the mass spectra. To understand the impact of such contamination on the two sampling methods, three samples obtained via CUSA (Cavitron Ultrasonic Surgical Aspirator), consisting primarily of blood and irrigation fluid, were analysed. These samples were centrifuged to remove tissue fragments, and the remaining fluid was analysed directly by ASAP–MS. The evaluation of the two sampling methods focused on the following aspects:

- Data quality and repeatability:

  The chronograms of the total ion count (TIC) between the two sampling methods were compared. Then, principal component analysis (PCA) was performed on the mass spectra to evaluate and visualise the repeatability of the two sampling methods.

- Blood and surgical irrigation fluid contamination:

  For each measurement, a sequence of 27 spectra was acquired over 25 s. After averaging 10 technical replicates and completing the standard preprocessing pipeline, PCA was performed on spectra from brain tissue measured using the two sampling methods and from the three CUSA liquid samples. Scores for PC1 and PC2 were plotted to assess clustering by sampling method and the influence of contamination on spectral variation.

To further assess the clinical relevance of the two protocols described above, measurements were made using both protocols for samples from the first 50 patients, and 10 replicate measurements were made for each sample. To avoid bias, all samples were labelled using anonymised specimen codes. Mass spectrometry data collection and data preprocessing were completed before diagnostic information was released and matched with the mass spectrometry results via specimen codes.

To ensure a fair comparison, the analysis was restricted to samples that were measured using both protocols.[1]

5.2.2.1 Batch effect correction

To enable statistically valid batch correction, only batches containing both tumour and normal samples were retained for `neuroComBat` batch effect correction[156], as these balanced batches allow estimation of technical variation independently from biological class effects. Details of these batches are given in Table 5.1. *Batch 11* was defined as the reference batch against which all other batches were harmonised. Only features (m/z peaks) present in all of the samples were retained for downstream analysis.[139] Linear regression was used to evaluate the proportion of variance explained by batch before and after batch effect correction.[157] After batch effect correction, PCA was employed to get an initial view of how the samples vary in each data set.

Table 5.1: Tumour and normal sample counts per batch, indicating ComBat-eligible batches.

Batch	Normal	Tumour	Eligible	Batch	Normal	Tumour	Eligible
1	2	7	TRUE	11	3	7	TRUE
2	1	4	TRUE	12	0	4	FALSE
3	0	3	FALSE	13	2	3	TRUE
4	2	1	TRUE	14	4	1	TRUE
5	0	3	FALSE	15	3	7	TRUE
6	0	3	FALSE	16	2	2	TRUE
7	1	4	TRUE	17	0	2	FALSE
8	1	2	TRUE	18	0	0	FALSE
9	2	3	TRUE	19	0	4	FALSE
10	0	2	FALSE	20	0	3	FALSE

5.2.2.2 Hyperparameter optimisation, and model training and validation

To evaluate classification potential, a set of common supervised machine learning models were employed (introduced in Section 1.3): Logistic Regression(LR)[158],

[1]This selection was necessary due to the exclusion of specific samples that were either lost during direct measurement due to insufficient quantity, or only measured using the homogenised method due to time constraints.

Naïve Bayes (NB)[159], Support Vector Machine (SVM)[100], k-Nearest Neighbors (KNN)[89], Linear Discriminant Analysis (LDA)[92], and Random Forest (RF)[96]. Initial benchmarking for the ML analysis was performed using these models with their hyperparameters set to the default values. A range of algorithmic approaches and hyper-parameters were chosen to optimise and fine-tune the models based on the findings from Chapter 3. The relevant hyperparameters are listed in Table 5.2. The hyperparameter grids were designed to explore the bias–variance trade-off under our conditions of high-dimensional, strongly correlated spectral features. As introduced in Section 1.3, for regularised linear models and SVMs, the penalty parameter C was varied over several orders of magnitude to control model complexity, while the number of selected features k determined the effective dimensionality of the classifier. For tree-based models, depth and minimum node size parameters were tuned to limit variance amplification due to correlated predictors. For KNN and LDA, dimensionality reduction via PCA was optimised, reflecting the need for well-conditioned covariance or distance metrics. For Naïve Bayes, both univariate feature selection and variance smoothing were tuned to mitigate violations of the conditional independence assumption. Overall, the search spaces prioritised statistical stability and generalisability rather than maximal model flexibility.

The data set was balanced using two data augmentation strategies (see Table 5.3):

- Random undersampling

  The larger class was reduced by randomly removing samples until both classes contained the same number of samples.

- Random oversampling

  The smaller class was increased by randomly repeating samples until both classes contained the same number of samples.

Features (m/z values) were standardised using `StandardScaler` from `scikit-learn` (v1.7.2) before model fitting. Partitioning of the data set into training and validation data was performed at the level of pseudo-anonymised patient IDs. Data

Table 5.2: Hyperparameter search spaces explored for each model.

Model	Hyperparameter search space
LR	$C \in \{0.01, 0.1, 1, 10\}$; l_1-ratio $\in \{0.1, 0.5, 0.9\}$
SVM (linear kernel)	$C \in \{0.01, 0.1, 1, 10, 100\}$; Selected feature $k \in \{10, 20, 50, 100, 200\}$
SVM (RBF kernel)	$C \in \{0.01, 0.1, 1, 10, 100\}$; $\gamma \in \{\text{scale}, 10^{-3}, 10^{-2}, 10^{-1}\}$; Selected feature $k \in \{10, 20, 50, 100, 200\}$
RF	Maximum depth $\in \{\text{None}, 5, 10, 20, 40\}$; Minimum samples split $\in \{2, 5, 10\}$; Minimum samples leaf $\in \{1, 2, 5, 10\}$; Maximum features $\in \{\sqrt{p}, \log_2 p\}$
KNN	PCA components $\in \{1, 2, 3, 5, 10, 20, 50\}$; Number of nearest neighbors $k \in \{1, 2, 3, 5, 7, 9\}$; weights $\in \{\text{uniform}, \text{distance}\}$
LDA	PCA components $\in \{5, 10, 20, 50\}$
NB	Selected feature $k \in \{10, 20, 50, 100, 200\}$; Variance smoothing $\in \{10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}\}$

Table 5.3: Class distribution of samples across the data set before and after sampling.

Sample	N	N(Undersampling)	N(Oversampling)
Normal	23	23	41
Tumour	41	23	41

augmentation was applied only after this split, with synthetic samples inheriting the corresponding patient identifier to prevent any leakage from validation to training sets. Model performance was evaluated using 5-fold `GroupKFold`. For each model, Cohen's kappa was computed, and the mean and standard deviation across folds were reported.

5.2.2.3 Minimum repetition estimation

To determine the minimum number of replicates required for reliable model training, we created averaged spectra using between 2 and 10 replicates per sample with mass spectra added to the average in order of acquisition. To quantify the effect of replicate averaging on data quality, we adapted the feature-accuracy metric proposed by Mohammed *et al.*[160]. In their formulation, feature accuracy $A(c)$

for a variable c is defined as

$$A(c_i) = 1 - \frac{\bar{d}(c_i)}{\bar{g}(c_i)} \tag{5.1}$$

$$\bar{d}(c_i) = \frac{1}{n}\sum_{j=1}^{n} |g_{i,j} - v_{i,j}| \tag{5.2}$$

represents the mean absolute deviation between the observed feature values $v_{i,j}$ and the corresponding ground-truth values $g_{i,j}$ across n samples. Overall feature accuracy is then computed as

$$\text{FeatureAccuracy} = \frac{1}{n}\sum_{i=1}^{n} A(c_i) \tag{5.3}$$

In the original study, "ground truth" values were obtained from high-quality reference measurements. In our setting, no external ground truth exists; therefore, we defined the 10-repeat averaged spectrum for each sample to be a reasonable approximation of the ground truth spectrum for that sample. For each repetition level $k = 2, ..., 9$, spectra were averaged across the k technical repeats per sample, and these averaged values were compared with the corresponding 10-repeat averages. Specifically, for each feature c_i and repetition level k:

- $g_{i,j}$ denotes the intensity of feature i in the 10-repeat average of sample j
- $v_{i,j}^{(k)}$ denotes the intensity of the feature i in the k-repeat average of sample j

The mean absolute deviation $\bar{d}(c_i)$ was then computed across samples for each k , and substituted into the equations above to derive $A(c_i)$ and the overall Feature accuracy for a given number of averaged measurement k. Higher values indicate closer agreement between the k-repeat data and the 10-repeat reference, i.e. higher effective feature accuracy.

The model performance was then evaluated at each replicate level using the same optimised machine-learning pipeline. The effect of replicate number on classification quality was quantified using (i) Cohen's kappa to assess classification reliability and (ii) the Brier score to assess the accuracy of probabilistic predictions. This analysis

was used to determine whether increasing the number of technical repetitions yielded meaningful gains in predictive performance, and to identify a practical balance between measurement throughput and model accuracy.

5.2.3 Results and Discussion

5.2.3.1 Comparison between measurements on intact and homogenised samples

We compared the mass spectral quality obtained from a brain tissue sample without any obvious blood contamination (Sample 1) and a brain tissue sample with blood contamination (Sample 2) using the two different approaches described in Section 5.2.2: direct sampling and homogenised sampling. The results of this comparison are summarised in Figure 5.2, which illustrates the ion chronograms and PCA results used to evaluate the quality and repeatability of these measurements.

Figure 5.2(a) shows the chronograms for five repeat measurements for both samples using the two sampling methods. For Sample 1, both direct and homogenised methods produced consistent, sharp peaks in the chronogram, reflecting stable ionisation and minimal interference. In contrast, Sample 2 measured using the direct sampling method showed considerable variability in TIC, with distorted peak shapes. Notably, one cleaning-related peak[2] was wide and stubborn, requiring repetition of the cleaning step. By comparison, homogenised sampling of Sample 2 produced more intense and stable peaks, with only minor cleaning peaks between each measurement. These results suggest that direct sampling becomes unreliable in the presence of the blood contamination, whereas homogenisation reduces the interference and preserves spectral quality.

The PCA results further support this observation. Figure 5.2(b) illustrates the clustering of replicate measurements: homogenised samples form tight clusters, showing strong repeatability, whereas direct sampling results in a wider spread. Such variability is likely to be a consequence of the inherent heterogeneity of brain

[2]The optimised protocol used a single capillary for five replicate measurements. Between measurements, the capillary was rinsed with LC–MS-grade water and wiped with lens tissue, and residual material was removed by source heating.

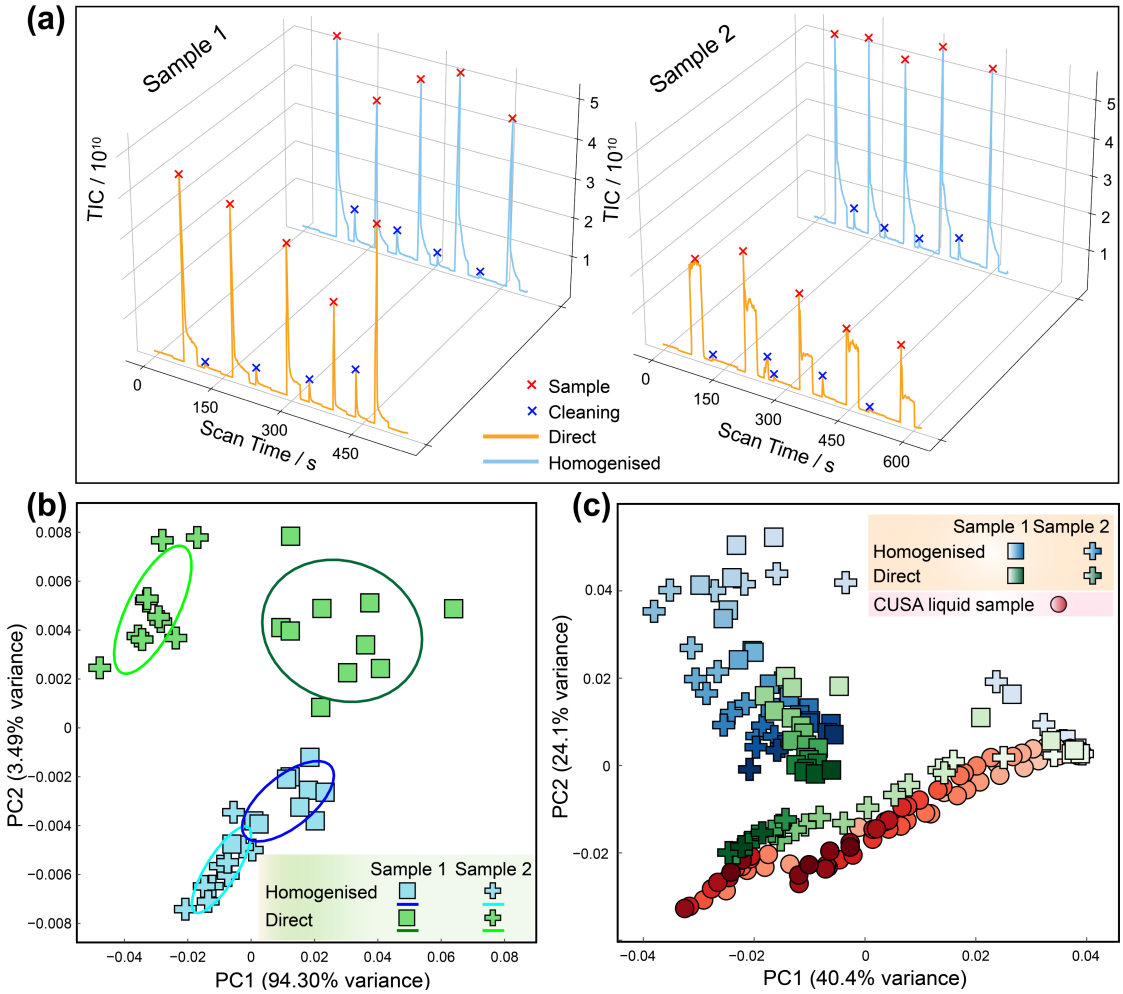

Figure 5.2: Comparison between direct and homogenised sampling methods for ASAP–MS analysis. (a) Chronograms of fresh brain tissue samples measured directly or after homogenisation by ASAP–MS, showing sample measurement peaks (red '×') and cleaning-related peaks (blue '×') between each measurement. Sample 1 is a brain tissue sample without clear blood contamination, Sample 2 is a brain tissue sample with blood contamination. (b) PCA plot showing the measurement repeatability of Sample 1 and Sample 2 under both direct and homogenised sampling conditions. (c) PCA plot showing all the mass spectra obtained during measurement (when the ASAP probe and sample were inserted in the mass spectrometer) to examine the impact of blood and surgical irrigation fluid contamination in the brain tissue sample. Each data point was coloured by time using group-specific light-to-dark gradients with earlier spectra appearing lighter and later spectra darker.

tissue, especially in pathological regions such as tumours, which often exhibit diverse cellular composition and density.[161]

The influence of contamination is more clearly visualised in Figure 5.2(c). Here, spectra from Sample 2 measured using direct sampling overlap closely with those

of fluid samples, suggesting that the mass spectral signature is dominated by the contamination. In contrast, both homogenised Sample 2 and uncontaminated Sample 1 (whether measured using direct or homogenised sampling) cluster together, indicating that homogenisation substantially reduces the confounding effect of blood and surgical irrigation fluid contamination. This can be explained by the sampling mechanism: direct sampling probes only a small region of the tissue surface, where residual blood and surface fluid may disproportionately contribute to the ionised material, thus biasing the measurement towards a liquid sample-like profile. Homogenisation, by mixing the tissue into a liquid matrix, dilutes the contribution of surface liquid, thereby providing a more representative molecular profile of the underlying brain tissue.

The above findings suggest that homogenised sampling provides superior robustness and repeatability for ASAP–MS analysis of brain tissues, especially in the presence of contamination. This observation aligns with previous studies reporting that homogenisation improves analyte extraction efficiency and reduces sample-to-sample variability in mass spectrometry-based tissue analysis [162, 163]. At this stage, homogenised sampling can be considered a more reliable protocol for ensuring spectral quality and minimising the impact of contamination caused by the presence of blood or surgical irrigation fluid.

To further assess the clinical relevance of these two methods, we analysed samples from 50 patients using both direct and homogenised sampling. After batch effect corrections (result shown in Table D.1), the resulting data sets were examined by PCA. Figure 5.3(a) shows that in the case of direct sampling, there is overlap in the PCA plots between tumour and normal brain tissue samples, indicating limited discriminatory power. By contrast, the PCA plot of homogenised data, shown in Figure 5.3(b), reveals a clearer separation between tumour and normal tissue groups, with only minor overlap at the margins of the confidence ellipses.

This comparison reinforces our earlier observation that homogenised sampling yields higher quality spectra when using ASAP–MS to measure fresh brain tissue samples. The improved separation in PCA space suggests that homogenisation

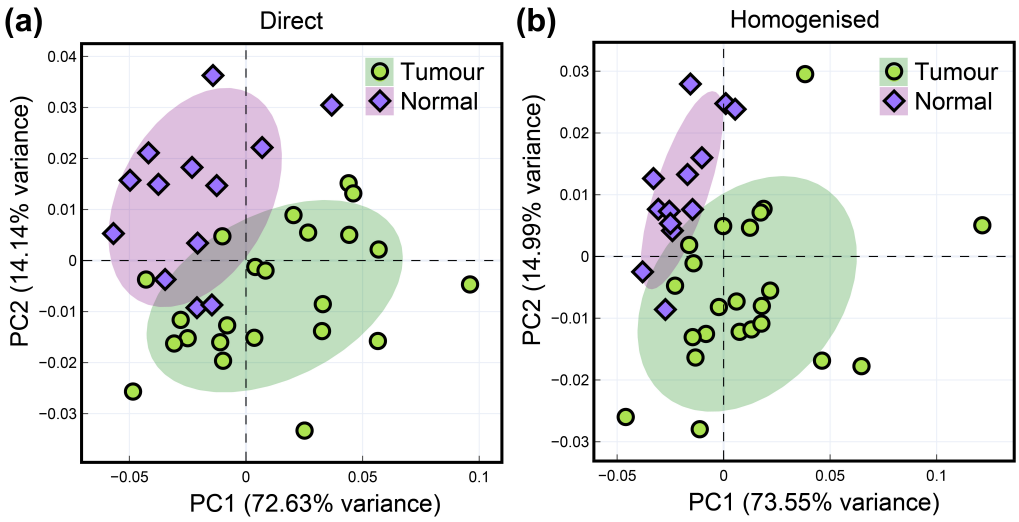

Figure 5.3: Comparison of data sets obtained by direct and homogenised sampling methods. The label 'Tumour' refers to brain tumour biopsies, whereas the label 'Normal' refers to access cortex biopsies that were neuropathologically normal. (a) PCA plot of the data set generated using the direct sampling method. (b) PCA plot of the data set generated using the homogenised sampling method. The confidence ellipse corresponds to two standard deviations.

enhances the capture of both normal- and tumour-specific molecular features, while reducing variability arising from local heterogeneity and contamination. However, a possible confounding factor is the mass of sample used for homogenisation. The mass excised for homogenisation is hard to control accurately and larger masses yield higher concentration in the homogenised sample solution, and higher ion counts in mass spectra. To examine this effect, we compared sample weights for tumour and normal tissues (Figure 5.4). The difference was not significant, suggesting weight is unlikely to drive the observed separation.

Overall, these results indicate that homogenisation not only improves measurement quality but also enhances the extraction of clinically relevant information. On this basis, we selected the homogenised data set for training our diagnostic models, since it offered superior spectral quality and a more distinct separation between tumour and normal tissue compared with direct sampling. Although we proceeded using only the homogenisation sampling method in this study, direct sampling still holds potential for future diagnostic applications. The PCA results have already

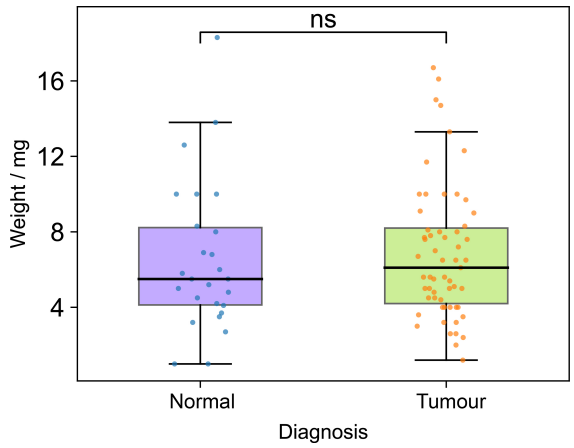

Figure 5.4: Box plot of sample weight for tumour and normal groups. Jittered points show individual samples. The line in each box is the median. The box spans the interquartile range.

demonstrated that tumour and normal samples retain a degree of separation even without homogenisation. However, to make direct sampling clinically feasible, further optimisation would be necessary. One possible method for improvement is surface washing to minimise blood contamination. A key challenge, however, is identifying an appropriate washing solution that is both compatible with the mass spectrometer ion source and safe for patient use. Besides, bioengineering solutions could also improve this method. For example, reducing the size of the probe and glass capillary could allow sampling from deeper tissue layers than can be achieved by superficial swabbing, while enabling more precise targeting of regions of interest. This could be particularly advantageous in heterogeneous tumour environments. Such an improvement would also limit the amount of surface liquid collected during sampling, thereby decreasing contamination and improving the specificity of the tissue signal.

5.2.3.2 Hyperparameter optimisation and model performance

Given the clear separation observed in the PCA plot for the data set generated using the homogenised samples (see Figure 5.5(a) for the complete data set), we trained and optimised supervised machine learning models on this data set. The model performance is shown in Figure 5.5(b). All six models demonstrated robust performance under both undersampling and oversampling conditions used

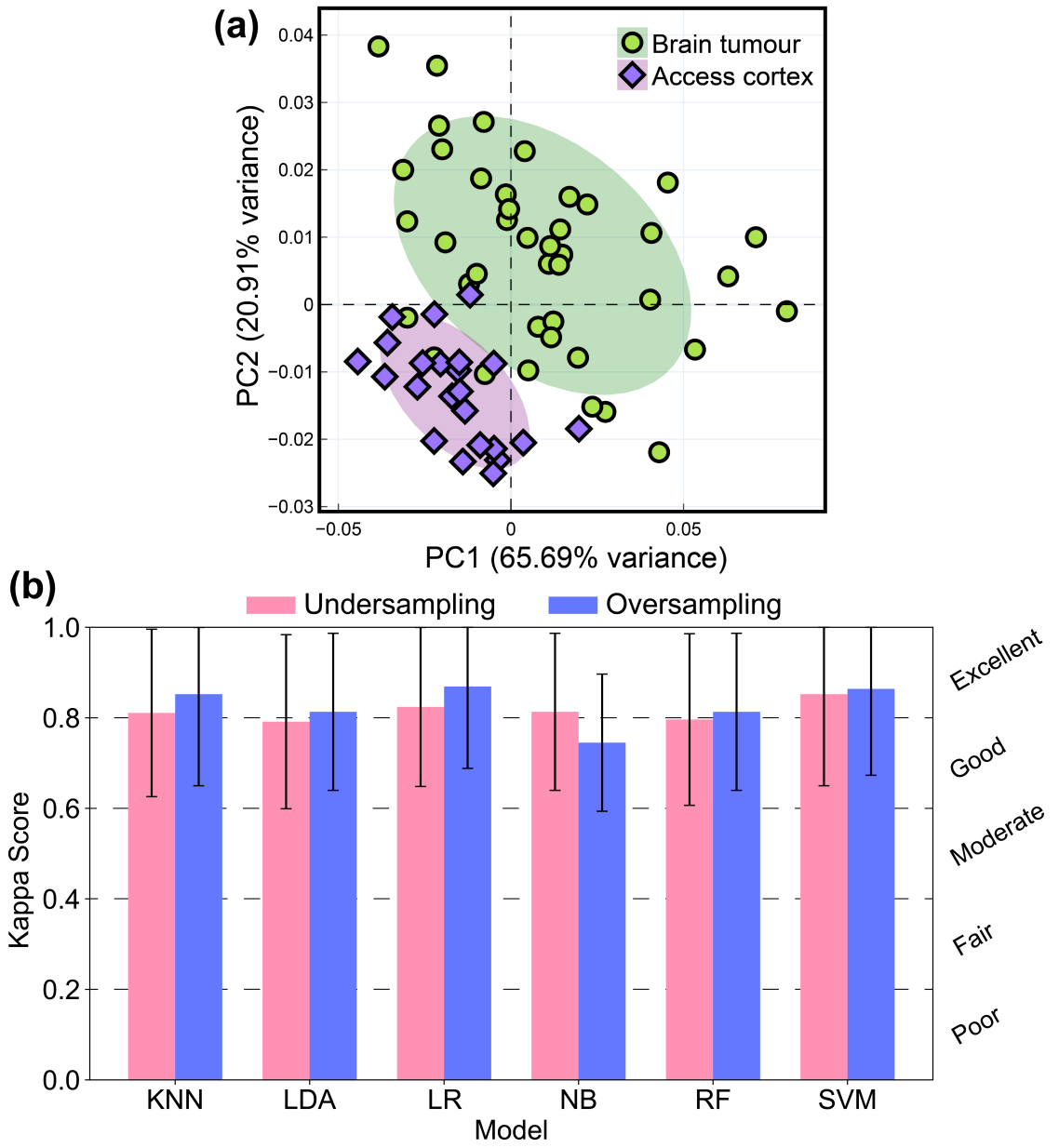

Figure 5.5: Model performance: (a) PCA plot of the data set of the first 50 patients. The confidence ellipse corresponds to two standard deviations; (b) Cohen's kappa scores obtained during cross-validation for all classifier–sampling combinations (undersampling and oversampling). Each bar shows the mean kappa score across folds, with error bars indicating the standard deviation. The dashed background colour bands represent the accepted interpretation thresholds for kappa (0–0.20 = poor, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, >0.80 = almost perfect agreement.

to balance the data set. The mean Cohen's kappa score was approximately 0.8, which is generally interpreted as indicating excellent predictive capability. However, the standard deviation in kappa score was large across folds, which is likely due to the combination of limited sample size, high feature dimensionality, and inter-patient heterogeneity [164]. Besides, `GroupKFold` assigns all samples from each patient to a single fold, different folds contain different patient subsets, resulting in variable classification difficulty.

Table 5.4: Optimised hyperparameters for each classifier under undersampling and oversampling strategies.

	Model	Optimised hyperparameters
Undersampling	KNN	PCA components = 20, neighbours = 3, uniform weights
	LDA	PCA components = 5
	LR	$C = 10$, l_1-ratio = 0.9
	NB	selected features $k = 20$, variance smoothing = 10^{-12}
	RF	max depth = None, max features = $\sqrt{p}$
		min samples leaf = 1, min samples split = 2
	SVM	RBF kernel, $C = 1$, γ = scale, selected features $k = 100$
Oversampling	KNN	PCA components = 10, neighbours = 5, uniform weights
	LDA	PCA components = 10
	LR	$C = 0.1$, l_1-ratio = 0.9
	NB	selected features $k = 20$, variance smoothing = 10^{-12}
	RF	max depth = None, max features = $\sqrt{p}$
		min samples leaf = 1, min samples split = 2
	SVM	linear kernel, $C = 0.1$, selected features $k = 200$

The optimised hyperparameters are shown in Table 5.4. A comparison between default and optimised model performance evaluated by kappa score is shown in Figure 5.6. Overall, hyperparameter optimisation improved performance primarily by tailoring regularisation strength, which controls model complexity, and feature dimensionality, which limits the number of variables used, to suit the high-dimensional, small-sample number nature of the data. The following sections discuss the theoretical justification for the hyperparameter optimisations[3] performed.

[3]These hyperparameters are introduced in Section 1.3

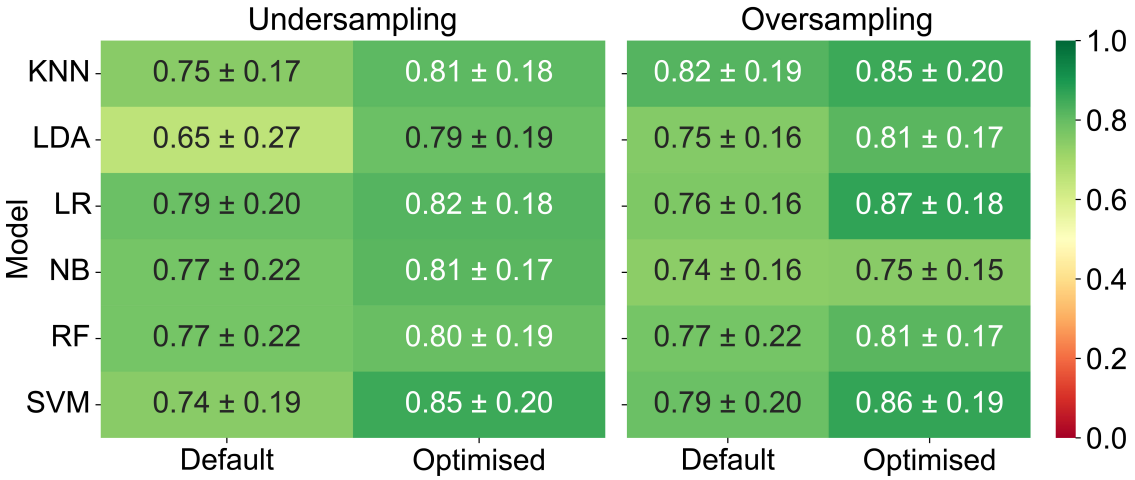

Figure 5.6: Comparison of supervised model training performance (Cohen's kappa) with and without optimisation using data balancing by undersampling and oversampling.

- **k-Nearest Neighbors (KNN)**: KNN performs poorly in very high-dimensional spaces because distances between samples become less informative.[90] To rectify this, we employed PCA to reduce the dimensionality of the data set before applying KNN. This preprocessing step removes redundant, highly correlated features and focuses the distance calculation to considering only the directions of maximum variance. The selection of $k = 3$ and $k = 5$ suggests that the local class structure is relatively tight; a larger k would have introduced "neighbourhood blurring" from the majority class due to the high-dimensional overlap.

- **Linear Discriminant Analysis (LDA)**: LDA requires inverting the within-class covariance matrix, $\sum$. When the number of features is larger than the number of samples ($p > n$), $\sum$ becomes singular (non-invertible), so LDA becomes numerically unstable or impossible to fit.[93, 165] In this study, we therefore used PCA before LDA not as a minor optimisation step but as a mathematical requirement. By reducing the original 270 features to a lower-dimensional space (5 principal components for undersampled data sets and 10 principal components for oversampled data sets), the covariance matrix became invertible and well-behaved, which in turn allowed LDA to estimate reliable discriminant boundaries.

- **Logistic Regression (LR)**: The baseline LR model could overfit easily because the maximum likelihood estimator could fit noise in the many correlated m/z features. Adding an elastic-net penalty addressed this by using the sparsity of LASSO (L_1) with the shrinkage of Ridge ($L2$) regularisation.[166] The optimised $l_1 - ratio = 0.9$ indicates that the model strongly favoured a sparse solution, which is appropriate for mass-spectrometry data where biological signal is concentrated in a limited set of peaks rather than spread uniformly across the spectrum. For undersampled data, the optimal logistic regression parameter was $C = 10$, indicating weaker regularisation. This is consistent with the fact that undersampling reduces the effective sample size and increases estimator variance, so a more flexible model is required. In contrast, for oversampled data the optimal $C = 0.1$ reflects stronger regularisation: oversampling inflates the apparent sample size with highly correlated observations, increasing the risk of overfitting unless the coefficients are heavily penalised.

- **Naive Bayes (NB)**: NB operates on the strong assumption of conditional independence between features.[94, 159] In spectral data, this assumption is fundamentally violated as different m/z bins are often physically and biologically correlated. By restricting the model to the top 20 features, we reduced the "dependency noise" that typically causes NB class probabilities to collapse. Furthermore, the application of variance smoothing stops Naïve Bayes from becoming overconfident because of tiny variances estimated from too few data points. The optimised variance-smoothing parameter was smaller than the `scikit-learn` default. Usually a smaller smoothing parameter allows Naïve Bayes to trust the empirical variances more, rather than flattening them artificially. This indicates that the default level of smoothing was overly conservative for our data set. This could be because filtering to m/z peaks present in all samples removes many zero-variance and unstable peaks so that the model benefits from less variance inflation.

- **Random Forests (RFs)**: RFs are generally robust. For the Random Forest models, the optimised configuration used unrestricted tree depth (`max_depth=None`) but constrained the way trees were grown. Setting `max_features`=$\sqrt{p}$ forced each split to consider only a random subset of the available features, which decorrelates the trees and prevents any single high-intensity peak from dominating the ensemble. The values `min_samples_split`=2 and `min_samples_leaf`=1 allow individual trees to remain flexible, but the randomness introduced through feature subsampling provides the main regularising effect. This combination increases diversity between trees while limiting overfitting, which is particularly important in high-dimensional, small-sample number mass-spectrometry data.

- **Support Vector Machine (SVM)**: The key finding from the SVM analysis was that the optimal kernel depended on the sampling strategy. With random oversampling, the best-performing model used a linear kernel with a small C. Random oversampling increases the weight of the minority class by duplicating existing samples rather than generating new ones; this amplifies existing noise and class imbalance artefacts without adding new geometric structure. In this setting, a linear separator with a soft margin generalises better because it discourages the model from fitting duplicated noisy points too closely.

  Under undersampling, the optimal model used an RBF kernel. Undersampling removes many samples and leaves only the core structure of the classes, which requires a more flexible non-linear boundary to separate them effectively. Thus, sampling strategy and kernel choice interacted strongly: random oversampling favoured simpler linear decision boundaries with stronger regularisation, whereas undersampling favoured more flexible RBF kernels.

Because our data are characterised by many interdependent features and a small number of samples, we deliberately avoided automated hyperparameter optimisation tools such as `Optuna`. This is because these tools offer extensive hyperparameter search, which can lead to overtuning, i.e. the hyperparameter

configuration overfits the validation procedure rather than improving genuine generalisation performance. This phenomenon has been documented in recent large-scale studies of hyperparameter optimisation, which show that overtuning is more frequent than previously assumed and particularly problematic for small data sets.[167] We therefore constrained optimisation to a small, theoretically motivated set of parameters controlling model capacity and effective dimensionality, rather than performing a wide, unconstrained search.

5.2.3.3 Minimum repetition estimation

The ML models explored in the previous section exhibited satisfactory cross-validated performance with the optimised hyperparameters. Following hyperparameter optimisation, we next examined whether the number of replicate measurements could be reduced without compromising the model performance. As shown in Figure 5.7, averaging additional repetitions progressively improved feature accuracy (Equation 5.3), with the mean feature accuracy increasing from 0.93 (two repeats) to 0.99 (nine repeats) when compared with the ten-repeat reference.

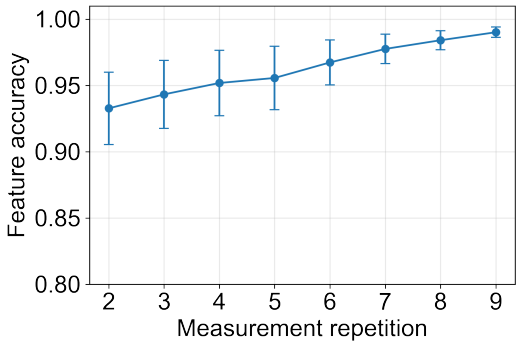

Figure 5.7: Evaluation of the impact of sample measurement repetition on feature accuracy.

Using the same optimised pipeline and 5-fold stratified group cross-validation, models were retrained on data sets generated from different numbers of averaged replicate measurements. The results are shown in Figure 5.8. Across all classifiers, increasing the number of technical measurements per sample from 2 to 10 produced no systematic improvement in Cohen's kappa score, and the associated variability

intervals largely overlapped between repetition levels. We further investigated the model performance by computing the Brier score (Figure 5.9), which similarly remained effectively unchanged with increasing numbers of repeats. Thus, although replicate averaging improved feature accuracy, this did not translate into measurable gains in model performance.

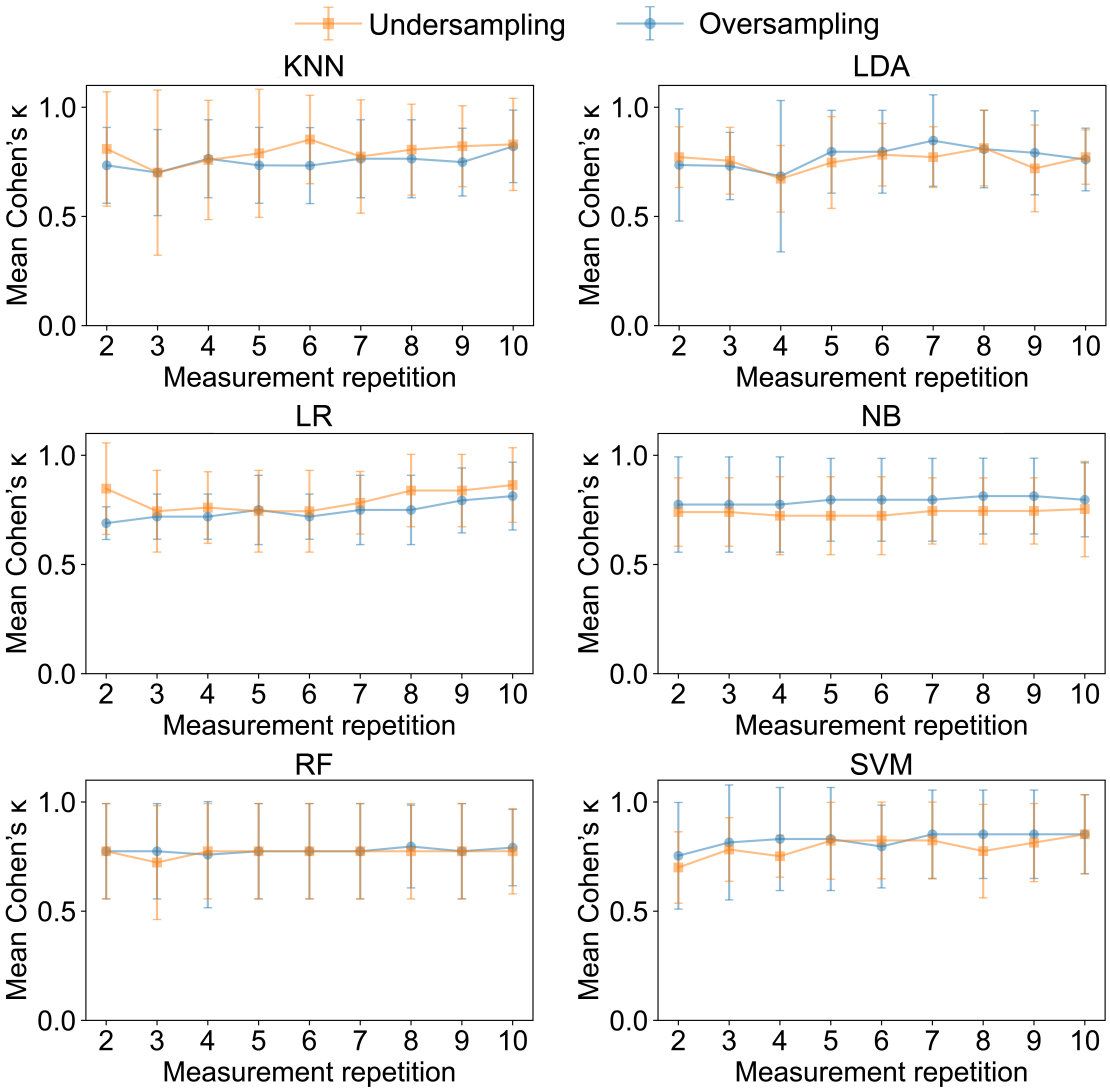

Figure 5.8: Evaluation of the impact of sample measurement repetition using Cohen's kappa score across all models.

Many studies have discussed the importance of data quality on machine learning model training.[168–170] In our previous study[78], the mean coefficient of variation (CV) in m/z peak intensities of 40% indicates substantial measurement variation. Besides, Classical measurement theory states that repeated measurements reduce

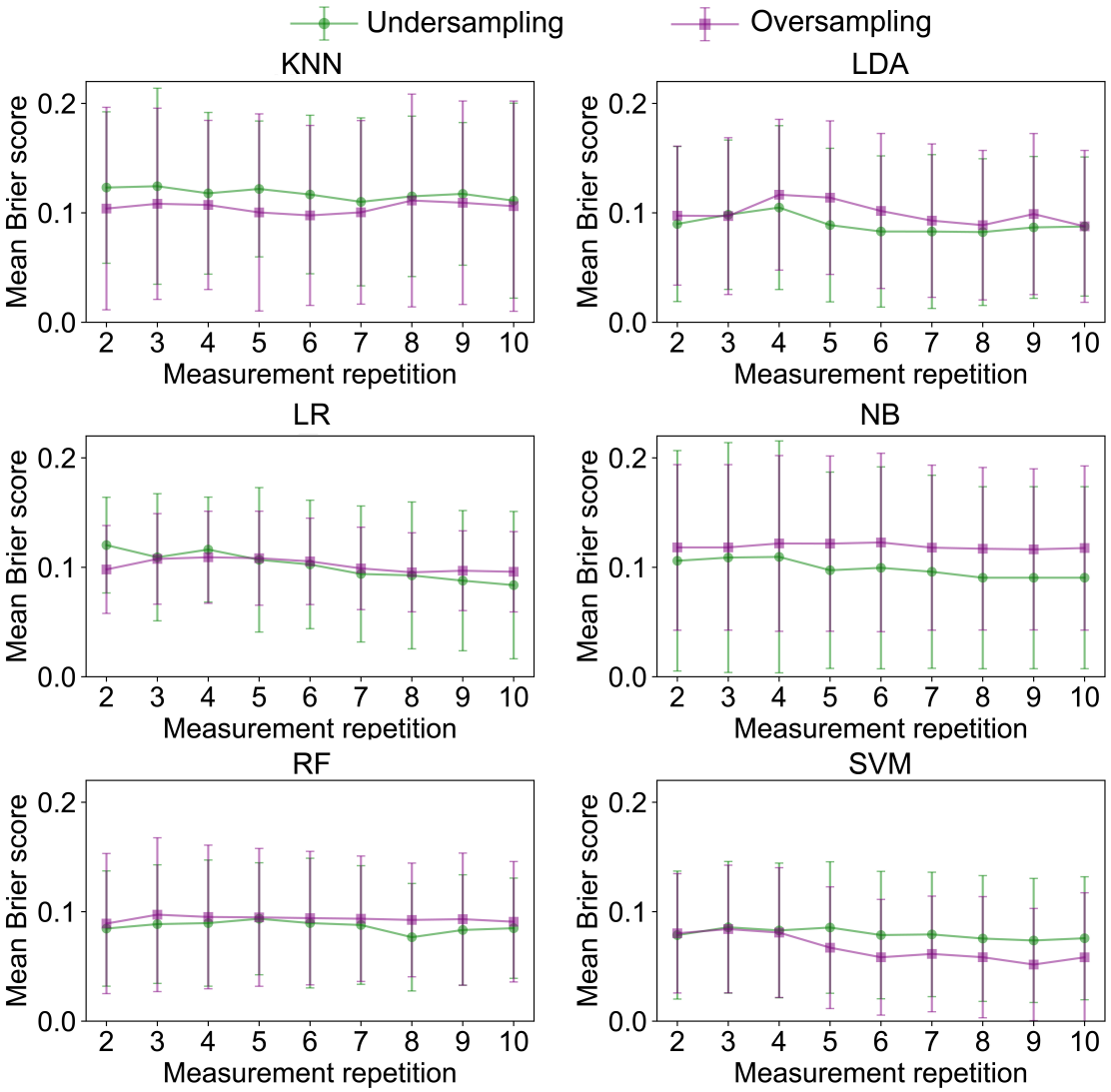

Figure 5.9: Evaluation of the impact of sample measurement repetition using Brier score across all models.

random error and improve precision of the data quantity.[171] This is also the core assumption motivating our use of repeated measurements. The goal was to reduce measurement variance and increase feature reliability so more accurate and stable classification models could be obtained. Our previous study has also found that averaging ten repeat measurements per sample provides a suitable balance between improving spectral quality while maintaining a rapid acquisition protocol when using ASAP–MS.[78] However, the observed minimal impact of increasing repeats on machine learning performance (Figure 5.8 and Figure 5.9) contradicts the intuitive expectation that additional averaging should improve

machine learning model training, yet it is consistent with the findings of the OXAMI and OXAAA ASAP–MS studies by Eardley-Brunt.[73] and with broader evidence that many machine-learning models are comparatively robust to moderate degradation in feature accuracy.[172]

Large-scale empirical work has shown that substantial decreases in model performance occur primarily when feature accuracy falls below approximately 0.8.[160] In our data, even two repeats yielded a mean feature accuracy of 0.93, well above this regime, which likely explains the limited effect of additional averaging. In practice, we selected five technical repetitions for the default acquisition scheme. Choosing five repeats achieves a pragmatic balance: it substantially reduces measurement variance relative to two or three repeats, while avoiding the unnecessary doubling of experimental effort required to acquire ten repeats, for no observable gain in predictive performance.

Notably, recent studies have demonstrated that the impact of noisy features depends more strongly on model complexity than on baseline accuracy: highly accurate models may nonetheless generalise poorly if they rely on a small number of unstable predictors.[170] This highlights the need to evaluate the robustness of the models above when used to make predictions based on independent unseen data, which is addressed in the next section.

To conclude this section, preliminary data suggest that ASAP–MS can rapidly distinguish brain tumour tissue from surrounding brain parenchyma when combined with properly optimised supervised machine learning models. Based on the investigation above, the recommended measurement protocol for fresh brain tissue samples is as follows: a small portion of brain tissue is excised in the operating theatre, to which 100 μL of LC–MS-grade water is added. Following thorough homogenisation, the resulting preparation is analysed by ASAP–MS, performing five replicate measurements for each sample.

5.3 Model Generalisability: Evaluation and Interpretation

5.3.1 Introduction

In the previous section, we trained and cross-validated machine-learning models using six supervised machine learning algorithms, and obtained excellent results. However, several studies have shown that models with strong internal performance can exhibit marked degradation when evaluated on independent cohorts, in some cases approaching random-chance performance (≈ 0.50 accuracy).[173, 174] This well-recognised failure of model generalisation motivated the development of a prospective evaluation pipeline. Specifically, the trained models were used to classify mass spectra from newly measured, previously unseen samples obtained *before* release of the official neuropathology report, in order to assess robustness on genuinely independent data.

5.3.2 Methods

5.3.2.1 Model generalisability and iterative learning

To evaluate model robustness and generalisation, an iterative learning cycle pipeline was designed (see Figure 5.10). As described in Section 5.2, using data from the first 50 patients, 12 initial models were trained, combining six machine learning algorithms with two class-balancing strategies. Each model was saved with the package `joblib` (1.3.2). Following the standardised measurement protocol detailed in Section 5.2, newly acquired samples underwent batch effect correction using *Batch 11* as the reference batch. The resulting spectra were presented to the 12 models, which classified them as either brain tumour or normal access cortex. The predictions were then submitted to a third party researcher with access to the official neuropathological diagnoses in order to determine whether or not they were correct. Model performance was assessed through this comparison of predictions with the neuropathological diagnoses, recording the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). After each

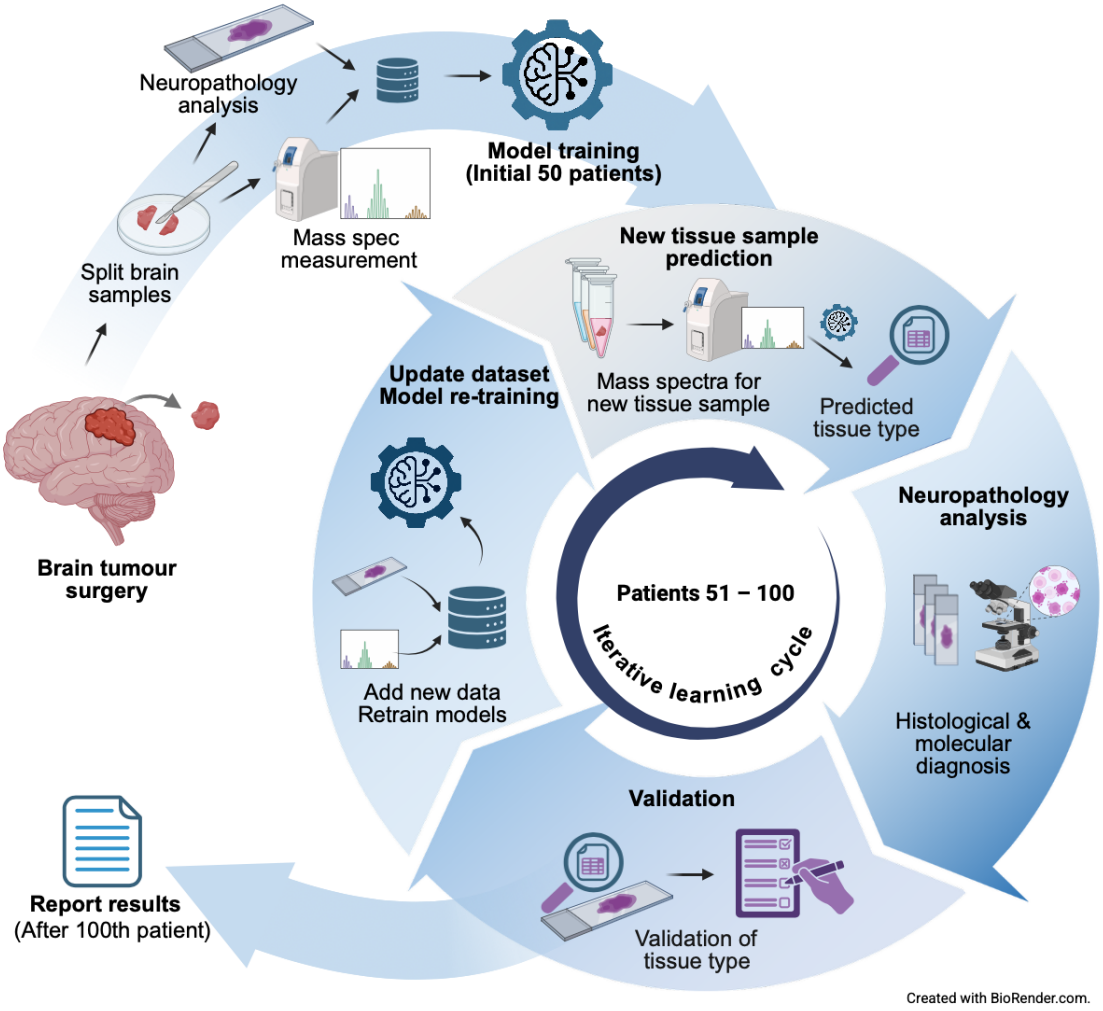

Figure 5.10: Model application and iterative learning workflow.

evaluation cycle, the newly acquired data were appended to the training data, and all 12 models were retrained using the hyperparameters in Table 5.4. This iterative learning cycle pipeline was followed for patients numbered 51 to 100, with a total of 89 samples (20 normal and 69 tumour) in 13 batches. Based on the cumulative totals of TP, FP, TN, and FN recorded throughout the iterative learning cycle, confusion matrices were determined for each model. Performance metrics (introduced in Section 1.3), including sensitivity, specificity, precision, negative predictive value (NPV), F1 score and Cohen's kappa, were also calculated in order to evaluate the model performance.

5.3.2.2 Model explanation

Feature contributions were explored using SHAP (SHapley Additive exPlanations) to explore correlations between m/z peak intensities and classification into each of two tissue classes (See details in Section 1.3).[110] For the best-performing model in this study (Random Forest with oversampling), SHAP values were computed using `TreeExplainer` on samples from all 100 patients. Global importance was summarised as the mean absolute SHAP value per feature, and visualised using summary plots to identify the most influential spectral features.

To visualise how individual features drove the model's predictions on new data, cumulative SHAP decision paths were generated for the final, unseen batch of samples.[4] Using the `TreeExplainer`, the SHAP values for each feature were calculated and summed progressively. The resulting decision path plot illustrated the journey of a prediction: starting from a base value, each feature's contribution pulls the trajectory either right (toward a tumour diagnosis) or left (toward a normal tissue diagnosis).

To understand which region of the mass spectrum contributed to the classification, the SHAP contributions were mapped back to their corresponding m/z values. These m/z values were partitioned into equal-width bins across the entire mass range. This resulted in a summary histogram that shows which region of the spectrum provided the most useful information for distinguishing between tumour and normal brain tissues.

5.3.2.3 Learning-curve analysis using the final data set

To understand how much data the model needs to reach its full potential, we tested its performance across the entire cohort of samples from 100 patients. Learning curves were constructed using stratified group k-fold cross-validation, with data split at the patient level into training and test data. Such data split method prevented information leakage.

[4]This model was the last updated model using the second last batch of data, and this model was used to predict the last batch of data

Model performance was summarised using two metrics: Cohen's kappa score (discrimination) and the Brier score (probability calibration). For the kappa score, higher values indicate better agreement beyond chance, whereas for the Brier score lower values indicate better calibrated probabilistic predictions. Learning curves were plotted as kappa and Brier scores against the number of training samples.

5.3.3 Results and Discussion

5.3.3.1 Model prediction

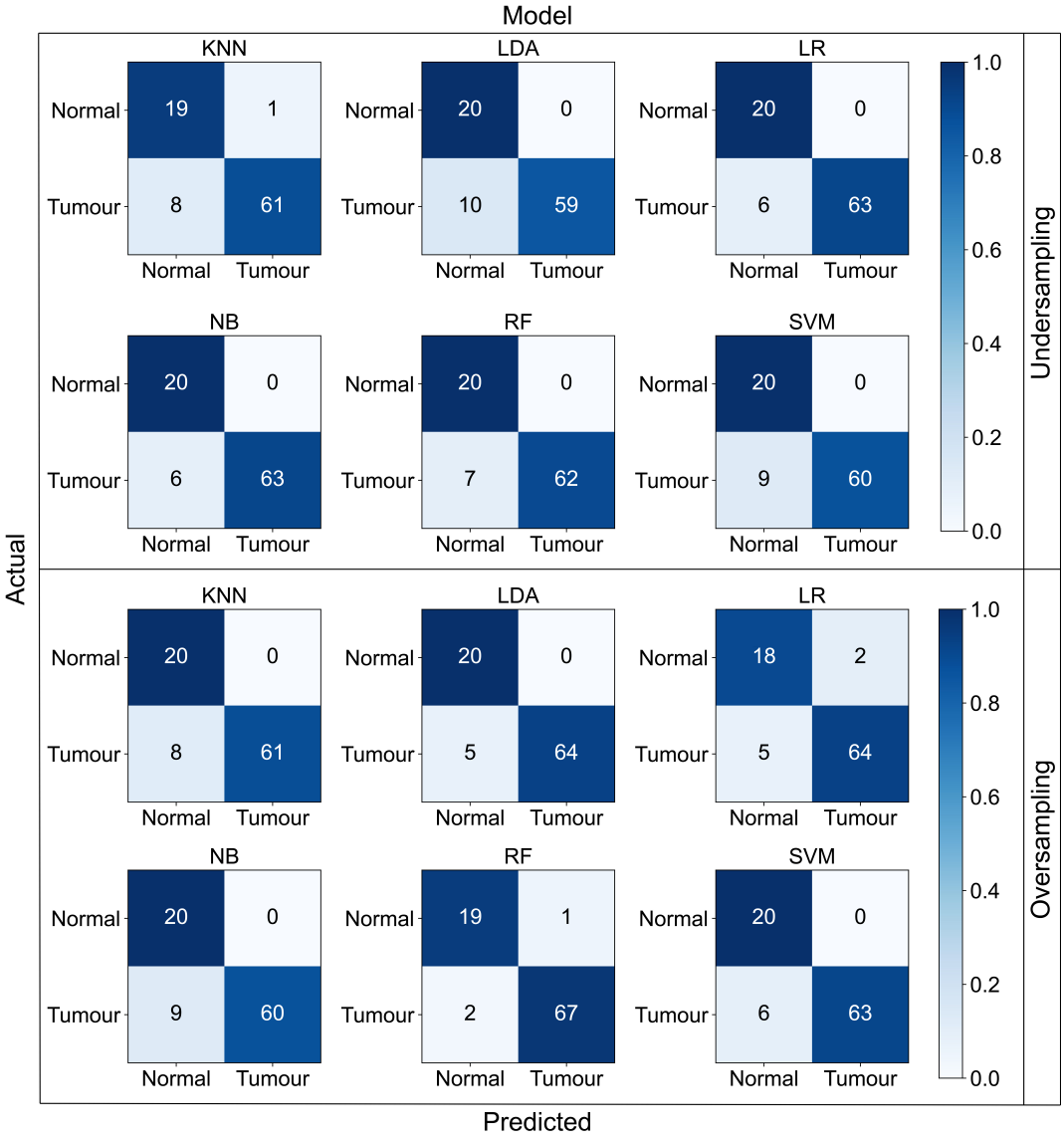

Figure 5.11: Confusion matrices of classification outcomes for 12 models trained in the iterative learning cycle.

The confusion matrices resulting from the analysis described above are shown in Figure 5.11 with the performance of the various models summarised further in Figure 5.12. The details for each round of prediction are shown in Table D.15−Table D.27

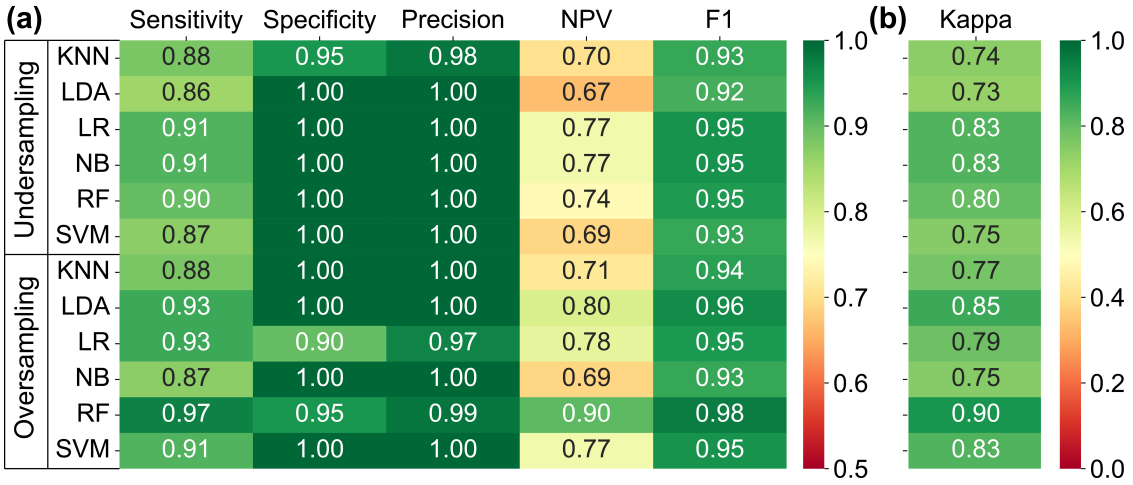

Figure 5.12: Performance metrics of classification outcomes for 12 models trained in the iterative learning cycle.

Performance was found to be uniformly high across all models, with F1-scores generally between 0.92 and 0.98 for both undersampling and oversampling data sets. Specificity was consistently excellent and frequently reached 1.00. However, sensitivity showed greater variation across classifiers (approximately 0.86–0.97), and negative predictive value was the least stable metric, indicating that some models tend to predict some tumour samples as normal. The number of normal samples is relatively small ($n = 20$). Several models misclassified between five and ten tumour samples as normal. This combination of a small normal class and a number of false negatives leads to large fluctuations in the NPV. As shown in Figure 5.12(b), Cohen's kappa score indicated substantial to almost-perfect agreement between predicted and true tissue type, ranging from 0.73 to 0.90. Overall, oversampling did not systematically outperform undersampling; instead, performance differences were model dependent rather than sampling-strategy dependent.

The best model within this analysis was found to be random forest with oversampling, which achieved a sensitivity of 97% and a specificity of 95% ,

corresponding to a Cohen's kappa score of 0.90. These results indicate strong tumour-normal discrimination within the present data set. These values are comparable to those reported for MS-based methods.

For example, rapid evaporative ionisation mass spectrometry (REIMS) has been used to distinguish normal brain tissue from multiple brain tumour types [175]. In a reanalysis of the published data, we grouped all tumour types into a single tumour class, resulting in an accuracy of 0.91 and a kappa score of 0.72. Jarmusch *et al.*[176] used desorption electrospray ionisation-mass spectrometry (DESI-MS) profiling of lipids and metabolites to distinguish brain parenchyma from gliomas, reporting 97.4% sensitivity and 98.5% specificity in 58 patients. By contrast, some DESI-MS studies have reported overall accuracies above 90%. However, when their published confusion matrices are re-expressed in terms of Cohen's kappa scores, the resulting agreement is only moderate to fair (0.51 in one study, and 0.28 in the other), demonstrating that raw accuracy alone overestimated diagnostic performance in the presence of skewed class distributions [177, 178] This indicates the importance of kappa score as a stricter performance metric that is more sensitive to imbalanced data than raw accuracy.

Raman spectroscopy has also shown strong performance for intraoperative discrimination between brain tumour and normal tissue. A meta-analysis reported pooled sensitivity and specificity of 0.96 and 0.99 for glioma, and 0.98 and 1.00 for meningioma, respectively.[179] In a more clinically realistic handheld "optical biopsy needle" study, high-wavenumber Raman spectroscopy achieved 84% accuracy, with 80% sensitivity and 90% specificity for detecting dense tumour (>60% neoplastic cells) versus tissue that did not contain sufficient tumour cells for diagnosis.[180]

Most of these studies relied on cross-validation within a fixed data set. Therefore, the strength of this work is that we used an iterative learning design in which models were evaluated on genuinely unseen samples and updated over new batches. In addition, five specimens were labelled as "cortex or tumour" by neurosurgeons, reflecting intraoperative uncertainty. For all five cases, the random forest model with oversampling produced predictions consistent with the final neuropathological

diagnosis (three tumour and two normal cortex; Table D.29). The results suggest that our ASAP–MS–based classifiers perform accurately, but with the added confidence that its predictions are truly generalisable to new cases.

A further limitation of this comparison is that the classification tasks are not equivalent in difficulty. The present study focused primarily on binary discrimination between tumour and non-tumour brain tissue, whereas several published studies attempted more complex tasks, including separation of multiple tumour entities[175, 176, 179], molecular subtype prediction[176], or detection of infiltrative tumour with variable tumour-cell content[180]. These tasks are more challenging than binary tumour-normal classification because class boundaries are less distinct. Therefore, the performance metrics reported here should not be interpreted as demonstrating the superiority of ASAP–MS over other analytical methods. More complex diagnostic tasks using ASAP–MS will require further validation.

5.3.3.2 Analysis of misclassification

Samples misclassified by the majority of models (>6/12) are summarised in Table 5.5. All cases in this group were tumour samples predicted as normal (false negatives). One case (Specimen 4071777517) was a brain metastasis from lung carcinoma, which represents an out-of-distribution case rather than a simple classification error. The remaining six cases were primary brain tumours. For each of these, each of the five individual replicates were run through the models to investigate whether individual measurements were the cause of the misclassification.

Table 5.5: Summary of samples misclassified by the majority of models, N is the number of models misclassifying the sample.

Specimen code	True class	Error type	N	Tumour subtype
4071777414	Tumour	FN	12	IDH mutant astrocytoma
4071777200	Tumour	FN	9	GBM
4071777357	Tumour	FN	8	GBM
4071777517	Tumour	FN	8	Metastasis(lung carcinoma)
4071777534	Tumour	FN	12	IDH-mutant astrocytoma
4071777551	Tumour	FN	8	GBM
4071777586	Tumour	FN	7	GBM

For samples 4071777534 and 4071777586, none of the five replicate spectra was correctly classified by the models that misclassified them during the iterative learning cycle, indicating genuine model failure rather than instability between repeated measurements. In contrast, samples 4071777414, 4071777200, 4071777357 and 4071777551 were mostly classified correctly when individual replicates were analysed. At the replicate level, the models were often able to recognise these samples as tumour, yet the final predictions shifted to normal only after the data had undergone batch effect correction using `neuroComBat` with other data in the same batch. This inconsistency prompted the examination of the batch effect correction step.

5.3.3.3 The impact of batch effect correction

For comparison, we replicated the iterative learning framework described above without applying batch effect correction. The detailed classification outcomes for this uncorrected approach are presented in Figure D.2. Figure 5.13 shows the change in model performance, measured by Cohen's kappa score, before and after batch effect correction. Batch effect correction resulted in improved kappa scores for nine models, indicating that batch effect correction is generally associated with improved overall model performance.

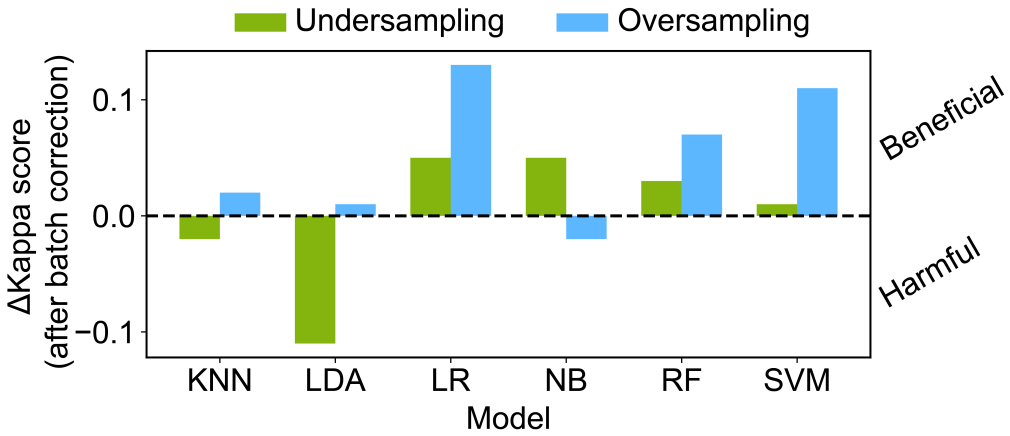

Figure 5.13: Change in Cohen's kappa score after batch effect correction across models. Positive values indicate that batch effect correction is beneficial to the overall model performance, whereas negative values indicate that batch effect is harmful to the overall model performance.

Samples with inconsistent prediction results before and after batch effect correction are summarised in Table D.30. In total, predictions changed for seven samples after batch effect correction. Three normal samples became correctly classified after batch effect correction, whereas four tumour samples changed from correct to incorrect after batch effect correction. Such pattern indicates that the effects of batch correction within this pipeline are not always neutral to prediction of brain tissue types. Batch correction can push samples towards the "normal" region of the feature space. PCA of the training data (Figure 5.5 (a)) supports this observation: normal samples cluster closely together, whereas tumour samples are much more spread out. `NeuroComBat` corrects batch effect towards the reference distribution.[181] When tumour biology is partially confounded with batch, ComBat could remove tumour-specific signal as if it were technical variation.[182, 183]

This outcome is clinically important. False-positive predictions cause damage to patients, while false-negative predictions risk under-treatment and recurrence. To overcome the limitation of the ComBat method for batch effect correction, in future work, each batch should include repeated measurements of a shared quality-control (QC) material, such as a standard plasma sample. These QC measurements can be used to monitor and correct technical variation more robustly. This is because QC samples provide a direct, biology-independent measure of technical variation, allowing correction of batch effect without relying on assumptions about the distribution of biological classes across batches.

5.3.3.4 SHAP analysis

The SHAP beeswarm plot, shown in Figure 5.14(a), summarises the global contribution of the most influential 20 m/z peaks to the model output. Features are ranked by mean absolute SHAP value. 17 out of 20 of these peaks are from the higher mass range ($m/z > 500$). Based on the results from Chapter 3, these features are likely to relate to the lipid-dominated part of the spectrum. These results are biologically reasonable given the metabolic shifts in lipids observed in brain tumours, especially glioblastoma. [184–188]

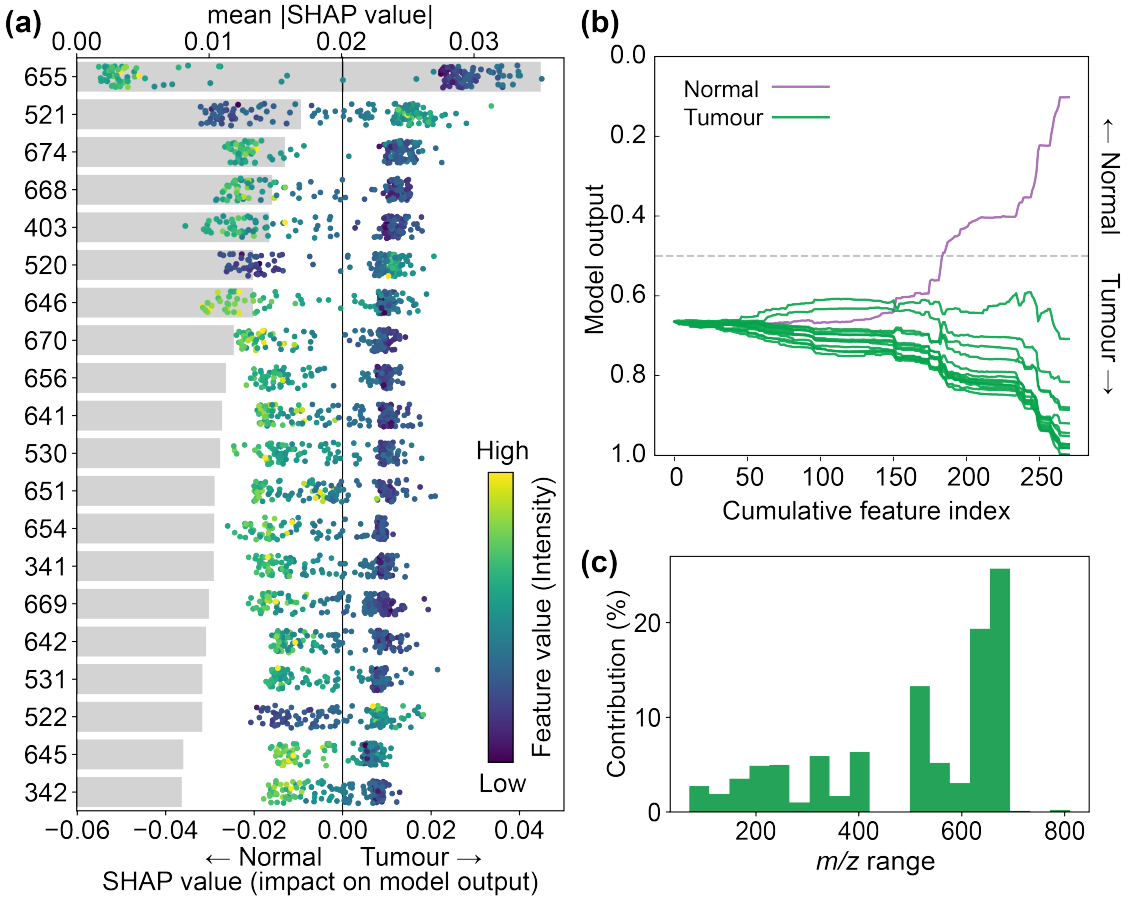

Figure 5.14: (a) Global SHAP importance and beeswarm plot showing the top features ranked by mean absolute SHAP value; point colour indicates raw feature value and horizontal position reflects contribution to the tumour versus normal decision. (b) SHAP decision paths for the final model applied to the last batch of patients, showing the cumulative effect of the features on the model output; (c) Percentage contribution of SHAP values aggregated across m/z bins.

A SHAP decision plot (Figure 5.14(b)) was generated for the final model applied to the last batch of samples, which had not been used for model training. The normal sample correctly trended toward the normal class, while all 12 tumour samples showed clear cumulative contributions driving the decision toward the tumour class. The clear separation of trajectories indicates that, for these unseen cases, the model aligned with the true labels instead of random guessing at the decision border. Additionally, the model output remains close to the base value until a large number of features have been included. This indicates that no single feature dominates the prediction. Instead, the decision emerges from the accumulation of many weak but consistently signed contributions. This finding aligns with the results presented in

Chapter 3, which showed that ASAP–MS rarely yields a single decisive biomarker: one molecule can generate several peaks through adduct formation, isotopologues and in-source fragmentation, and inter-molecular reactions. Although ASAP–MS spectra of brain tissue are simpler than those of plasma, genuine biological differences are still spread across groups of correlated peaks rather than concentrated in several single features. Consequently, classification performance is driven by the biochemical pattern across many features rather than by a few isolated dominant peaks.

To relate model behaviour to the underlying spectral structure, SHAP contributions were calculated across the m/z range. Figure 5.14(c) shows that predictive information is not spread evenly across the spectrum. Instead, some of m/z ranges account for a large proportion of the total contribution, while some other regions contribute very little. The high contributing regions lie mainly at higher m/z and are consistent with lipid-rich parts of the spectrum. Although the most influential features are located at higher $m/z(500 - 700)$, the lower m/z region $(100 - 400)$ still contributes to classification. As discussed in Chapter 3, brain tissue spectra changed little before and after lipid extraction, so it is highly likely that some of the low m/z features could also be lipid fragments.

5.3.3.5 Learning-curve analysis using the final data set

Across all models, the kappa learning curves (Figure 5.15(a)) shows a consistent pattern: performance reached a plateau rapidly. Once approximately 50–100 samples were included, no substantial further improvement was observed. While adding more samples reduced the variability of the kappa score, it did not significantly boost the model's overall accuracy.

The Brier score (Figure 5.15(b)) continued to decrease slightly with increasing sample size, whereas the kappa score had already plateaued. This suggests that more data helped the model to become more confident and better calibrated in its predictions. However, further modest increases in sample size are unlikely to increase the kappa score. These findings suggest that future performance gains

will depend on optimising feature representation or model architecture rather than purely enlarging the size of the patient cohort.

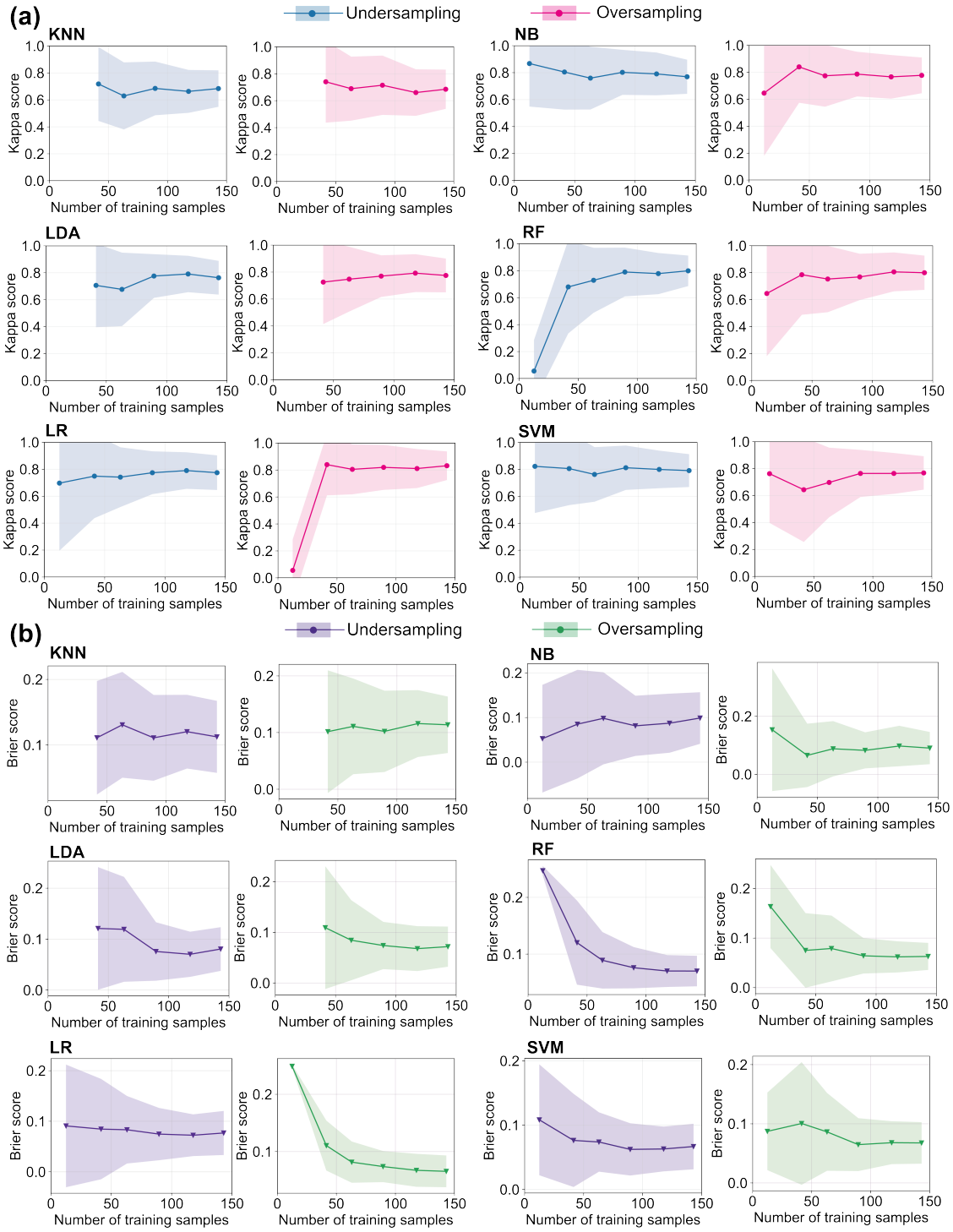

Figure 5.15: Cohen's kappa (a) and Brier score (b) learning curves as a function of training sample size for all classification models. Note: the KNN and LDA models were evaluated only from 50 samples onward because the dimensionality-reduction PCA step requires the number of samples N to exceed the number of retained principal components.

5.4 Conclusion and Future Work

This study applied an iterative learning framework to classify brain tumour versus neuropathologically normal cortex using ASAP–MS data. Across model updates and predictions on unseen samples, the method demonstrated robust performance, and SHAP-based model explanations confirmed that predictions were mostly driven by lipid-rich m/z regions of the spectrum. While the data set is modest in size, the results indicate that ASAP–MS, a comparatively simple and low-cost mass spectrometric technique, has genuine potential as a support tool for fast brain tissue classification.

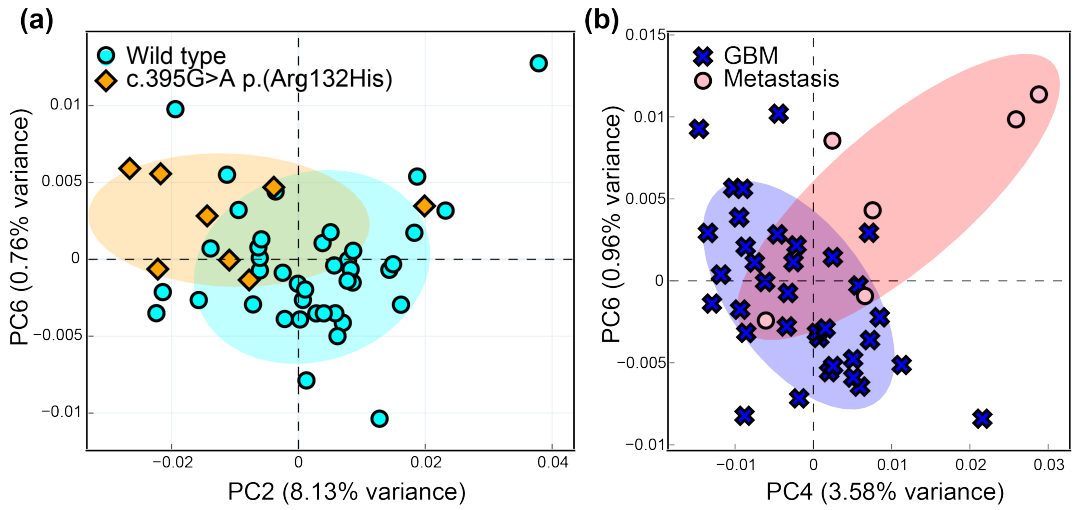

Figure 5.16: (a) PCA plot for IDH-mutant and IDH-wildtype gliomas; (b) PCA plot for glioblastoma and brain metastases.

To explore the broader utility of the data set, we also evaluated whether unsupervised analysis could separate clinically relevant tumour subtypes. We selected one tumour sample per batch to avoid batch-related clustering in PCA. Under this design, as shown in Figure 5.16, PCA shows only limited separation between (a) IDH-mutant and IDH-wildtype gliomas and (b) between glioblastoma and brain metastases. The weak separation is most likely to reflect both real biological overlap and methodological limitations: PCA is a linear technique, the data set is small, and batch effects were not explicitly modelled. Future studies

should employ quality-control samples and explicit batch-correction strategies to enable more reliable investigation of tumour subtypes.

Another important direction for future work is the investigation of tumour infiltration zones, where tumour and normal brain tissue coexist and discrete class boundaries become ill-defined. Rather than framing this problem solely as a classification task, future ASAP–MS studies could explore regression-based approaches to estimate tumour burden on a continuous scale (e.g. 0–100%). Such an approach would better reflect the underlying biological heterogeneity of infiltrative tumours and align more closely with the clinical reality of margin assessment during neurosurgical resection.

Finally, moving from statistical discrimination to biochemical interpretation requires molecular identification of key spectral features. ASAP–MS cannot directly identify compounds, so complementary techniques such as tandem MS or total-correlation mass spectrometry (TOC-MS) will be required in order to link discriminative m/z features to specific molecular species. This is a natural next step toward understanding the biology of different tissues rather than purely performing predictive classification.

6
Conclusion and Future work

Contents

6.1 Conclusion

This thesis evaluated whether atmospheric solids analysis probe mass spectrometry (ASAP–MS), when combined with carefully designed machine-learning workflows, can provide reliable molecular discrimination between brain tissue types in clinically relevant settings. Taken together, the results demonstrate that this goal is achievable, but requires careful experimental and analytical control.

At the methodological level, Chapter 2 establishes that data quality is the dominant limiting factor for clinical ASAP–MS applications. Multiple sources of technical variability, including ion source contamination, probe temperature, probe cleaning procedures, consumables, inter-user differences, and batch effects, were shown to introduce systematic artefacts that can overwhelm biological signals if left unaddressed. The optimisation strategies proposed here, together with batch-correction approaches such as ComBat and ICA, demonstrate that these issues

can be mitigated. This imposes a clear constraint: robust clinical deployment of ASAP–MS requires disciplined standardisation, comprehensive documentation, and explicit consideration of batch structure in downstream analysis.

Chapter 3 provides mechanistic insight into how molecular class and sample composition shape ASAP–MS spectra. The contrast between metabolite-rich and lipid-rich samples is clear. Small polar metabolites undergo extensive in-source chemistry, generating dense, time-dependent, and highly correlated spectral features, whereas lipids ionise more predictably and contribute largely additive signals. This distinction defines how the data must be analysed. Models that assume feature independence or rely on sparse, dominant peaks are poorly matched to the general behaviour of ASAP–MS data, which is characterised by complex ionisation dynamics, overlapping fragment patterns, and sensitivity to changes in sample composition and instrument conditions. Understanding these general properties of ASAP–MS is essential for developing robust analytical and computational strategies that can extract meaningful biological information from its spectra. Importantly, these findings justify the modelling choices adopted later in the thesis.

Chapter 4 demonstrates that sample preservation has a profound influence on ASAP–MS data quality. Although a rapid ultrasonication-based deparaffinisation protocol was developed, fixation time proved to be a key factor affecting the quality of the spectra. Short-fixed FFPE samples retained more consistent molecular fingerprints than long-fixed samples, but frozen tissue consistently outperformed FFPE tissue. Moreover, FFPE processing introduced unpredictable batch effects and reduced spectral diversity. In practice, fresh or frozen tissue is the better option for ASAP–MS when it is available, while reliable analysis of FFPE samples is generally difficult to achieve.

Finally, Chapter 5 integrates these experimental and analytical insights into an iterative learning framework for brain tumour classification. Despite a modest data set size, the optimised workflow achieved robust performance across model updates and predictions on unseen samples. A random forest model with oversampling achieved high sensitivity (0.97) and specificity (0.95), with a Cohen's kappa (0.90)

indicating excellent agreement beyond chance. Model interpretation using SHAP analysis indicated that classification was driven by lipid-rich regions of the spectrum, in direct agreement with the mechanistic findings of Chapter 3.

In summary, this thesis demonstrates that ASAP–MS can support rapid brain tissue classification when its limitations are clearly recognised, carefully controlled, and systematically addressed. The findings offer a practical basis for further investigation and outline considerations for potential application in clinical research.

6.2 Future work

While this study demonstrates the feasibility of combining ASAP–MS with machine learning for rapid brain tissue classification, several paths remain essential for strengthening both robustness and scope.

Although this work demonstrated that the inclusion of one or more quality-control (QC) samples could improve data reliability, these were not incorporated at the start of the study due to unavailability during the initial experimental phase. From patient 60 onwards, a pooled plasma sample was introduced for QC purposes within the ASAP–MS workflow. To maintain consistency with the initial measurements and to avoid introducing retrospective bias[1], no QC-based batch correction was applied in the current study. Future studies should therefore formalise the use of QC samples across all batches and apply appropriate batch-correction strategies. As the data set grows, QC-driven batch effect correction strategies can be evaluated more rigorously and benchmarked against ComBat, the batch effect modelling approach used in this study.

In parallel, the clinical data set should be expanded in both size and biological diversity. In the work described within this thesis, the classification task was limited to identification of tumour versus neuropathologically normal cortex. Future work should include a broader range of brain tumour subtypes, including low-grade glioma, high-grade glioma, metastatic tumours, and specific genetic subtypes such

[1]Retrospective bias refers to a systematic distortion introduced when information or methods that were not available at the time of data generation are applied after the fact, in a way that alters the interpretation of earlier results.

as IDH-mutant and IDH-wildtype gliomas. This would allow direct assessment of whether ASAP–MS can resolve clinically meaningful molecular distinctions beyond gross tissue type. At the same time, the inclusion of additional neuropathologically normal samples should be prioritised to improve class balance and reduce reliance on resampling strategies

As the data set grows, more advanced machine learning models can be explored. In this thesis, random forest (RF) provided a strong balance between performance and robustness in a limited-data regime. However, ensemble methods such as extreme gradient boosting (XGBoost) are known to perform well on structured, high-dimensional data when sufficient training samples are available. Future work should, therefore, explore these methods against the models used here, potentially improving performance.

A further direction for this work is the identification of the molecular features that drive classification performance. Model interpretation indicated that tumour–normal discrimination is dominated by lipid-rich regions of the ASAP–MS spectrum. However, the mass resolving power of the ASAP–MS instrument used in this study is insufficient to enable definitive molecular assignment of these features. As a result, while discriminative spectral regions can be identified (explained in Section 3.3.1), unambiguous molecular identification would require follow-up experiments using higher-resolution mass spectrometry platforms. Such measurements would be essential for biomarker discovery and biological interpretation. Total correlation mass spectrometry (TOC–MS) is a novel technique that combines ultraviolet photodissociation (UVPD) with mass spectrometry (MS) to offer an innovative approach for detailed lipid characterisation from lipid extraction of biological samples.[189] This method can link fragments to precursor molecules without requiring extensive derivatisation or chromatographic separation, which could be used to confirm molecular identity and relate predictive ASAP–MS signals to specific lipid structures.

Beyond data and modelling, there is substantial scope for optimisation at the hardware and sampling level. Our study compared sampling from both intact

and homogenised brain tissue finding that homogenisation improves mass spectra quality. However, there is scope for optimisation of other experimental parameters. Systematic optimisation of glass capillary geometry and tip diameter could enhance ionisation efficiency, repeatability, and signal stability. Because the local electric field and reagent-ion environment are primarily defined by the corona discharge, capillary design should be co-optimised with discharge conditions (needle position/current and gas flow) to maintain a stable discharge and consistent ion–molecule reaction zone. Treating the probe as an engineered component rather than a fixed consumable also opens the possibility of developing small, sharp glass capillaries for direct measurement on brain samples. Such designs could enable minimally invasive, spatially resolved sampling of intact tissue, reduce preparation time, and bridge the gap between *ex vivo* analysis and real-time intraoperative assessment, further demonstrating the potential of ASAP–MS as a rapid diagnostic tool.

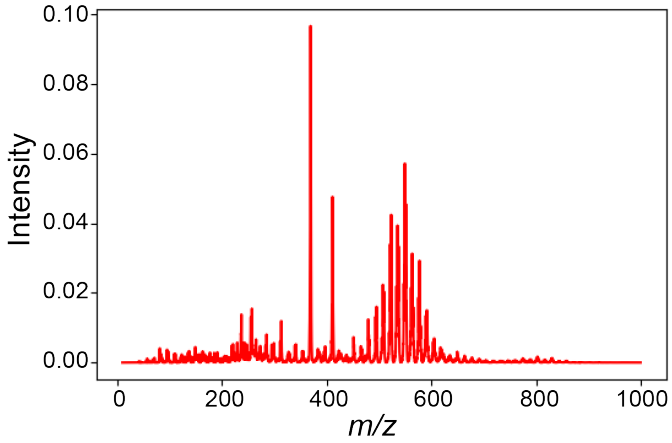

Figure 6.1: Representative ASAP–MS mass spectrum acquired from direct skin contact with the probe. The spectrum is dominated by lipid-associated m/z features, highlighting the suitability for ASAP–MS analysis.

Finally, there is plenty of scope to explore clinical application of ASAP–MS beyond the classification of brain tissue types. Lipid-rich tissues are particularly promising targets, as demonstrated by the mechanistic findings of this work. In Parkinson's disease, skin manifestations are common and include both disease-related abnormalities and therapy-associated effects.[190] This offers a compelling next step because skin samples are accessible, clinical relevant, and high in lipid

content. An example mass spectrum is shown in Figure 6.1. Extending the workflow to additional tissue types would allow assessment of how generalisable the methods are, and whether similar classification performance can be achieved in less controlled sampling environments.

References

[1] Albina Jetybayeva et al. “A review on recent machine learning applications for imaging mass spectrometry studies”. In: *Journal of Applied Physics* 133.2 (2023).

[2] Javier E Villanueva-Meyer et al. “Artificial Intelligence for Response Assessment in Neuro Oncology (AI-RANO), part 1: review of current advancements”. In: *The Lancet Oncology* 25.11 (2024), e581–e588.

[3] Manimekalai Pichaivel et al. *An overview of brain tumor*. IntechOpen, 2022.

[4] David N Louis et al. “The 2021 WHO classification of tumors of the central nervous system: a summary”. In: *Neuro-oncology* 23.8 (2021), pp. 1231–1251.

[5] Quinn T Ostrom et al. “CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019”. In: *Neuro-oncology* 24.Supplement_5 (2022), pp. v1–v95.

[6] Kimberly D Miller et al. “Brain and other central nervous system tumor statistics, 2021”. In: *CA: a cancer journal for clinicians* 71.5 (2021), pp. 381–406.

[7] Roger Stupp et al. “Effect of tumor-treating fields plus maintenance temozolomide vs maintenance temozolomide alone on survival in patients with glioblastoma: a randomized clinical trial”. In: *Jama* 318.23 (2017), pp. 2306–2316.

[8] Aaron C Tan et al. “Management of glioblastoma: State of the art and future directions”. In: *CA: a cancer journal for clinicians* 70.4 (2020), pp. 299–312.

[9] Irena Ilic and Milena Ilic. “International patterns and trends in the brain cancer incidence and mortality: An observational study based on the global burden of disease”. In: *Heliyon* 9.7 (2023).

[10] Quinn T Ostrom et al. “National-level overall survival patterns for molecularly-defined diffuse glioma types in the United States”. In: *Neuro-oncology* 25.4 (2023), pp. 799–807.

[11] Roger Stupp et al. “Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma”. In: *New England journal of medicine* 352.10 (2005), pp. 987–996.

[12] Kenneth Aldape et al. “Challenges to curing primary brain tumours”. In: *Nature reviews Clinical oncology* 16.8 (2019), pp. 509–520.

[13] Mónica Leiria de Mendonça et al. “Updating TCGA glioma classification through integration of molecular data following the latest WHO guidelines”. In: *Scientific Data* 12.1 (2025), p. 935.

[14] Guido Reifenberger et al. “Advances in the molecular genetics of gliomas—implications for classification and therapy”. In: *Nature reviews Clinical oncology* 14.7 (2017), pp. 434–452.

[15] Zhihui Yang and Kevin KW Wang. "Glial fibrillary acidic protein: from intermediate filament assembly and gliosis to neurobiomarker". In: *Trends in neurosciences* 38.6 (2015), pp. 364–374.

[16] Paolo Tini et al. "Low expression of Ki-67/MIB-1 labeling index in IDH wild type glioblastoma predicts prolonged survival independently by MGMT methylation status". In: *Journal of Neuro-oncology* 163.2 (2023), pp. 339–344.

[17] Lenny Dang et al. "Cancer-associated IDH1 mutations produce 2-hydroxyglutarate". In: *Nature* 462.7274 (2009), pp. 739–744.

[18] Sue Han et al. "IDH mutation in glioma: molecular mechanisms and potential therapeutic targets". In: *British journal of cancer* 122.11 (2020), pp. 1580–1589.

[19] Gregory Cairncross and Robert Jenkins. "Gliomas with 1p/19q codeletion: aka oligodendroglioma". In: *The Cancer Journal* 14.6 (2008), pp. 352–357.

[20] Branka Powter et al. "Human TERT promoter mutations as a prognostic biomarker in glioma". In: *Journal of cancer research and clinical oncology* 147.4 (2021), pp. 1007–1017.

[21] Fadi S Saadeh, Rami Mahfouz, and Hazem I Assi. "EGFR as a clinical marker in glioblastomas and other gliomas". In: *The International journal of biological markers* 33.1 (2018), pp. 22–32.

[22] Damian Stichel et al. "Distribution of EGFR amplification, combined chromosome 7 gain and chromosome 10 loss, and TERT promoter mutation in brain tumors and their potential for the reclassification of IDH wt astrocytoma to glioblastoma". In: *Acta neuropathologica* 136.5 (2018), pp. 793–803.

[23] Nader Sanai and Mitchel S Berger. "Surgical oncology for gliomas: the state of the art". In: *Nature Reviews Clinical Oncology* 15.2 (2018), pp. 112–125.

[24] Annette M Molinaro et al. "Association of maximal extent of resection of contrast-enhanced and non–contrast-enhanced tumor with survival within molecular subgroups of patients with newly diagnosed glioblastoma". In: *JAMA oncology* 6.4 (2020), pp. 495–503.

[25] Ian J Gerard. *Improving the accuracy of image-guided neurosurgical tools for brain tumour resections*. McGill University (Canada), 2018.

[26] Ian J Gerard et al. "Brain shift in neuronavigation of brain tumors: an updated review of intra-operative ultrasound applications". In: *Frontiers in oncology* 10 (2021), p. 618837.

[27] Alexander F Haddad, Manish K Aghi, and Nicholas Butowski. "Novel intraoperative strategies for enhancing tumor control: Future directions". In: *Neuro-oncology* 24.Supplement_6 (2022), S25–S32.

[28] Nader Sanai et al. "Intraoperative confocal microscopy for brain tumors: a feasibility analysis in humans". In: *Operative Neurosurgery* 68 (2011), ons282–ons290.

[29] Martina Piloni et al. "Resection of intracranial tumors with a robotic-assisted digital microscope: a preliminary experience with robotic scope". In: *World Neurosurgery* 152 (2021), e205–e211.

[30] Herbert Stepp and Walter Stummer. "5-ALA in the management of malignant glioma". In: *Lasers in surgery and medicine* 50.5 (2018), pp. 399–419.

[31] Walter Stummer et al. “Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase III trial”. In: *The lancet oncology* 7.5 (2006), pp. 392–401.

[32] Karl Roessler, Wolfgang Dietrich, and Klaus Kitz. “High diagnostic accuracy of cytologic smears of central nervous system tumors: A 15-year experience based on 4,172 patients”. In: *Acta cytologica* 46.4 (2002), pp. 667–674.

[33] Muhammad Shakir et al. “Unveiling the potential application of intraoperative brain smear for brain tumor diagnosis in low-middle-income countries: A comprehensive systematic review”. In: *Surgical Neurology International* 14 (2023), p. 325.

[34] Muhammad Shakir et al. “Diagnostic Accuracy of Intraoperative Brain Smear: A Meta-Analysis of Studies from Resource-Limited Settings”. In: *World Neurosurgery* 185 (2024), pp. 493–502.

[35] Fabien Almairac and Hugues Duffau. “Awake surgery with direct electrical stimulation mapping and real-time cognitive monitoring for functionally guided tumor resection: how we do it”. In: *Acta Neurochirurgica* 167.1 (2025), p. 239.

[36] Sam Ng et al. “Long-term autonomy, professional activities, cognition, and overall survival after awake functional-based surgery in patients with IDH-mutant grade 2 gliomas: a retrospective cohort study”. In: *The Lancet Regional Health–Europe* 46 (2024).

[37] Georg Neuloh et al. “Continuous motor monitoring enhances functional preservation and seizure-free outcome in surgery for intractable focal epilepsy”. In: *Acta neurochirurgica* 152.8 (2010), pp. 1307–1314.

[38] Christian Senft et al. “Intraoperative MRI guidance and extent of resection in glioma surgery: a randomised, controlled trial”. In: *The lancet oncology* 12.11 (2011), pp. 997–1003.

[39] Peter Abraham et al. “Cost-effectiveness of intraoperative MRI for treatment of high-grade gliomas”. In: *Radiology* 291.3 (2019), pp. 689–697.

[40] Zoe Z Zhang et al. “The art of intraoperative glioma identification”. In: *Frontiers in oncology* 5 (2015), p. 175.

[41] Michael Jermyn et al. “Intraoperative brain cancer detection with Raman spectroscopy in humans”. In: *Science translational medicine* 7.274 (2015), 274ra19–274ra19.

[42] Laurent James Livermore et al. “Rapid intraoperative molecular genetic classification of gliomas using Raman spectroscopy”. In: *Neuro-Oncology Advances* 1.1 (2019), vdz008.

[43] Todd Hollon and Daniel A Orringer. “Label-free brain tumor imaging using Raman-based methods”. In: *Journal of neuro-oncology* 151.3 (2021), pp. 393–402.

[44] Mahdiyeh Shahi and R Graham Cooks. “Ambient ionization mass spectrometry in brain cancer diagnosis”. In: *Journal of Mass Spectrometry and Advances in the Clinical Lab* (2025).

[45] Angus RJ Barber et al. “Rapid evaporative ionization mass spectrometry in surgery: a systematic review”. In: *British Journal of Surgery* 112.11 (2025), znaf228.

[46] Yanis Zirem et al. “Real-time glioblastoma tumor microenvironment assessment by SpiderMass for improved patient management”. In: *Cell Reports Medicine* 5.4 (2024).

[47] Prajwal Gowda. “Molecular Diagnosis of Glial Cell Brain Cancers Using Handheld Mass Spectrometry Device”. PhD thesis. The University of Texas at Austin, 2021.

[48] Philipp Euskirchen et al. “Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing”. In: *Acta neuropathologica* 134.5 (2017), pp. 691–703.

[49] Nik Sol et al. “Glioblastoma, IDH-wildtype with primarily leptomeningeal localization diagnosed by nanopore sequencing of cell-free DNA from cerebrospinal fluid”. In: *Acta Neuropathologica* 148.1 (2024), p. 35.

[50] Simon Deacon et al. “Nanopore-based brain tumour classification: the harbinger of near-patient, ultra-rapid tumour sequencing”. In: *Diagnostic Histopathology* 30.12 (2024), pp. 691–698.

[51] Jürgen H Gross. *Mass spectrometry: a textbook*. Springer Science & Business Media, 2006.

[52] Edmond De Hoffmann and Vincent Stroobant. *Mass spectrometry: principles and applications*. John Wiley & Sons, 2007.

[53] Joseph John Thomson. “Bakerian Lecture:—Rays of positive electricity”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 89.607 (1913), pp. 1–20.

[54] Frederick W Aston. “The constitution of the elements”. In: *Nature* 104.2616 (1919), pp. 393–393.

[55] Francis W Karasek and Ray E Clement. *Basic gas chromatography-mass spectrometry: principles and techniques*. Elsevier, 2012.

[56] Wilfried MA Niessen. *Liquid chromatography-mass spectrometry*. CRC press, 2006.

[57] Marek Domin and Robert Cody. *Ambient ionization mass spectrometry*. Royal Society of Chemistry, 2014.

[58] R Graham Cooks et al. “Ambient mass spectrometry”. In: *Science* 311.5767 (2006), pp. 1566–1570.

[59] Maria Eugenia Monge et al. “Mass spectrometry: recent advances in direct open air surface sampling/ionization”. In: *Chemical reviews* 113.4 (2013), pp. 2269–2308.

[60] Min-Zong Huang et al. “Ambient ionization mass spectrometry”. In: *Annual review of analytical chemistry* 3 (2010), pp. 43–65.

[61] Masamichi Yamashita and John B Fenn. “Electrospray ion source. Another variation on the free-jet theme”. In: *The Journal of physical chemistry* 88.20 (1984), pp. 4451–4459.

[62] DI Carroll et al. “Atmospheric pressure ionization mass spectrometry. Corona discharge ion source for use in a liquid chromatograph-mass spectrometer-computer analytical system”. In: *Analytical Chemistry* 47.14 (1975), pp. 2369–2373.

[63] Zoltán Takáts et al. “Mass spectrometry sampling under ambient conditions with desorption electrospray ionization”. In: *Science* 306.5695 (2004), pp. 471–473.

[64] Robert B Cody et al. “Direct analysis in real time (DART) mass spectrometry”. In: *JEOL news* 40.1 (2005), pp. 8–12.

[65] Xuelu Ding and Yixiang Duan. “Plasma-based ambient mass spectrometry techniques: The current status and future prospective”. In: *Mass spectrometry reviews* 34.4 (2015), pp. 449–473.

[66] Forough Doustkhahvajari and Stephanie Rankin-Turner. “Adopting ambient ionization mass spectrometry into bioanalytical laboratories”. In: *Bioanalysis* 17.7 (2025), pp. 439–443.

[67] Stephanie Rankin-Turner, Patrick Sears, and Liam M Heaney. “Applications of ambient ionization mass spectrometry in 2022: An annual review”. In: *Analytical Science Advances* 4.5-6 (2023), pp. 133–153.

[68] Yan Wang et al. “Desorption Electrospray Ionization Mass Spectrometry Imaging: Principles, Advancements, and Multidisciplinary Applications”. In: *Journal of Mass Spectrometry* 61.1 (2026), e70004.

[69] Nicholas E Manicke, Brandon J Bills, and Chengsen Zhang. “Analysis of biofluids by paper spray MS: advances and challenges”. In: *Bioanalysis* 8.6 (2016), pp. 589–606.

[70] Jürgen H Gross. “Direct analysis in real time—a critical review on DART-MS”. In: *Analytical and bioanalytical chemistry* 406.1 (2014), pp. 63–80.

[71] Charles N McEwen, Richard G McKay, and Barbara S Larsen. “Analysis of solids, liquids, and biological tissues using solids probe introduction at atmospheric pressure on commercial LC/MS instruments”. In: *Analytical Chemistry* 77.23 (2005), pp. 7826–7831.

[72] *CMS Product Family User’s Manual*. Instrument user manual for the expression® CMS. Advion Ltd. Harlow, United Kingdom, 2023.

[73] Eardley-Brunt Annabel S. J. “Rapid mass spectrometry coupled with machine learning to predict clinically relevant variables in cardiovascular disease”. PhD thesis. Oxford, United Kingdom: University of Oxford, 2025.

[74] David Fabregat-Safont et al. “Direct and fast screening of new psychoactive substances using medical swabs and atmospheric solids analysis probe triple quadrupole with data-dependent acquisition”. In: *Journal of the American Society for Mass Spectrometry* 31.7 (2020), pp. 1610–1614.

[75] Diane Lebeau and Muriel Ferry. “Direct characterization of polyurethanes and additives by atmospheric solid analysis probe with time-of-flight mass spectrometry (ASAP-TOF-MS).” In: *Analytical & Bioanalytical Chemistry* 407.23 (2015), p. 7175.

[76] Daniel Carrizo et al. “Direct screening of tobacco indicators in urine and saliva by Atmospheric Pressure Solid Analysis Probe coupled to quadrupole-time of flight mass spectrometry (ASAP-MS-Q-TOF-)”. In: *Journal of Pharmaceutical and Biomedical Analysis* 124 (2016), pp. 149–156.

[77] Gabriel Gaiffe et al. “Characterization of fluorinated polymers by atmospheric-solid-analysis-probe high-resolution mass spectrometry (ASAP/HRMS) combined with Kendrick-mass-defect analysis”. In: *Analytical chemistry* 90.10 (2018), pp. 6035–6042.

[78] Annabel SJ Eardley-Brunt et al. “Development of an optimised method for the analysis of human blood plasma samples by atmospheric solids analysis probe mass spectrometry”. In: *International Journal of Mass Spectrometry* 508 (2025), p. 117386.

[79] Advion Ltd. *Mass Express Software Manual for Advion’s Compact Mass Spectrometer (CMS)*. Software user manual for Mass Express and Advion CMS instruments. Advion Ltd. Harlow, United Kingdom, 2023.

[80] Annabel SJ Eardley-Brunt et al. “Optimising the choice of normalisation method for use in machine-learning classification of human blood plasma ambient ionisation mass spectra”. In: *International Journal of Mass Spectrometry* 520 (2026), p. 117553.

[81] Tushar Babbar. *Standardization: The Secret to Better Data Science*. May 2023. URL: `https://blog.alliedoffsets.com/standardization-the-secret-to-better-data-science`.

[82] Issam El Naqa and Martin J Murphy. “What is machine learning?” In: *Machine learning in radiation oncology: theory and applications*. Springer, 2015, pp. 3–11.

[83] Armen G Beck et al. “Recent developments in machine learning for mass spectrometry”. In: *ACS Measurement Science Au* 4.3 (2024), pp. 233–246.

[84] Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. “Tunability: Importance of hyperparameters of machine learning algorithms”. In: *Journal of Machine Learning Research* 20.53 (2019), pp. 1–32.

[85] Zoubin Ghahramani. “Unsupervised learning”. In: *Summer school on machine learning*. Springer, 2003, pp. 72–112.

[86] Ian Jolliffe. “Principal component analysis”. In: *Encyclopedia of statistics in behavioral science* (2005).

[87] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. “The global k-means clustering algorithm”. In: *Pattern recognition* 36.2 (2003), pp. 451–461.

[88] Amanpreet Singh, Narina Thakur, and Aakanksha Sharma. “A review of supervised machine learning algorithms”. In: *2016 3rd international conference on computing for sustainable global development (INDIACom)*. Ieee. 2016, pp. 1310–1315.

[89] Leif E Peterson. “K-nearest neighbor”. In: *Scholarpedia* 4.2 (2009), p. 1883.

[90] Nikolaos Kouiroukidis and Georgios Evangelidis. “The effects of dimensionality curse in high dimensional knn search”. In: *2011 15th panhellenic conference on informatics*. IEEE. 2011, pp. 41–45.

[91] Jill C Stoltzfus. “Logistic regression: a brief primer”. In: *Academic emergency medicine* 18.10 (2011), pp. 1099–1104.

[92] Suresh Balakrishnama and Aravind Ganapathiraju. “Linear discriminant analysis-a brief tutorial”. In: *Institute for Signal and information Processing* 18.1998 (1998), pp. 1–8.

[93] Ping Xu, Guy N Brock, and Rudolph S Parrish. "Modified linear discriminant analysis approaches for classification of high-dimensional microarray data". In: *Computational Statistics & Data Analysis* 53.5 (2009), pp. 1674–1687.

[94] Irina Rish et al. "An empirical study of the naive Bayes classifier". In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22. Seattle, USA. 2001, pp. 41–46.

[95] PJ Beslin Pajila et al. "A comprehensive survey on naive bayes algorithm: Advantages, limitations and applications". In: *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE. 2023, pp. 1228–1234.

[96] Steven J Rigatti. "Random forest". In: *Journal of insurance medicine* 47.1 (2017), pp. 31–39.

[97] Anthony J Myles et al. "An introduction to decision tree modeling". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 18.6 (2004), pp. 275–285.

[98] Qiong Ren, Hui Cheng, and Hai Han. "Research on machine learning framework based on random forest algorithm". In: *AIP conference proceedings*. Vol. 1820. 1. AIP Publishing LLC. 2017, p. 080020.

[99] Thirupathi Kandadi. "DRAWBACKS OF RANDOM FOREST ALGORITHM TO EXAMINE EXTENSIVE DATASETS". In: *Available at SSRN 5236759* (2025).

[100] Marti A. Hearst et al. "Support vector machines". In: *IEEE Intelligent Systems and their applications* 13.4 (1998), pp. 18–28.

[101] Sascha Klement, Amir Madany Mamlouk, and Thomas Martinetz. "Reliability of cross-validation for SVMs in high-dimensional, low sample size scenarios". In: *International Conference on Artificial Neural Networks*. Springer. 2008, pp. 41–50.

[102] Jair Cervantes et al. "A comprehensive survey on support vector machine classification: Applications, challenges and trends". In: *Neurocomputing* 408 (2020), pp. 189–215.

[103] Mervyn Stone. "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the royal statistical society: Series B (Methodological)* 36.2 (1974), pp. 111–133.

[104] Mahan Hosseini et al. "I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data". In: *Neuroscience & Biobehavioral Reviews* 119 (2020), pp. 456–467.

[105] Michael A Lones. "Avoiding common machine learning pitfalls". In: *Patterns* 5.10 (2024).

[106] Manahel Altalhan, Abdulmohsen Algarni, and Monia Turki-Hadj Alouane. "Imbalanced data problem in machine learning: A review". In: *IEEE Access* (2025).

[107] Haibo He and Edwardo A Garcia. "Learning from imbalanced data". In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[108] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[109] W Brier Glenn et al. "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1 (1950), pp. 1–3.

[110] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[111] Viswan Vimbi, Noushath Shaffi, and Mufti Mahmud. "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection". In: *Brain Informatics* 11.1 (2024), p. 10.

[112] Johannes Allgaier et al. "How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare". In: *Artificial Intelligence in Medicine* 143 (2023), p. 102616.

[113] Monya Baker. "Reproducibility crisis". In: *nature* 533.26 (2016), pp. 353–66.

[114] Shou-Ze Wang et al. "Evaluation of atmospheric solids analysis probe mass spectrometry for the analysis of coal-related model compounds". In: *Fuel* 117 (2014), pp. 556–563.

[115] Carlos R Canez and Liang Li. "Studies of Labware Contamination during Lipid Extraction in Mass Spectrometry-Based Lipidome Analysis". In: *Analytical Chemistry* 96.8 (2024), pp. 3544–3552.

[116] W Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127.

[117] Jeffrey T Leek et al. "Tackling the widespread and critical impact of batch effects in high-throughput data". In: *Nature Reviews Genetics* 11.10 (2010), pp. 733–739.

[118] Bart JA Mertens. "Transformation, normalization, and batch effect in the analysis of mass spectrometry data for omics studies". In: *Statistical analysis of proteomics, metabolomics, and lipidomics data using mass spectrometry* (2017), pp. 1–21.

[119] Wilson Wen Bin Goh, Chern Han Yong, and Limsoon Wong. "Are batch effects still relevant in the age of big data?" In: *Trends in Biotechnology* 40.9 (2022), pp. 1029–1040.

[120] Samuel Sanford Shapiro and Martin B Wilk. "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3-4 (1965), pp. 591–611.

[121] Henry B Mann and Donald R Whitney. "On a test of whether one of two random variables is stochastically larger than the other". In: *The annals of mathematical statistics* (1947), pp. 50–60.

[122] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272.

[123] John W Tukey. "Comparing individual means in the analysis of variance". In: *Biometrics* (1949), pp. 99–114.

[124] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[125] Wilson Wen Bin Goh et al. "Can peripheral blood-derived gene expressions characterize individuals at ultra-high risk for psychosis?" In: *Computational Psychiatry (Cambridge, Mass.)* 1 (2017), p. 168.

[126] Abdelkader Behdenna et al. “pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods”. In: *BMC bioinformatics* 24.1 (2023), p. 459.

[127] Aapo Hyvärinen and Erkki Oja. “Independent component analysis: algorithms and applications”. In: *Neural networks* 13.4-5 (2000), pp. 411–430.

[128] Martin A Hubbe et al. “Self-assembly of alkyl chains of fatty acids in papermaking systems: A review of related pitch issues, hydrophobic sizing, and pH effects”. In: *BioResources* 15.2 (2020), p. 4591.

[129] Jason P Ross et al. “Batch-effect detection, correction and characterisation in Illumina HumanMethylation450 and MethylationEPIC BeadChip array data”. In: *Clinical Epigenetics* 14.1 (2022), p. 58.

[130] Annabel SJ Eardley-Brunt et al. “Prediction of clinical outcomes of ST-elevated myocardial infarction patients using atmospheric solids analysis probe mass spectrometry and machine learning”. In: *Analyst* 150.22 (2025), pp. 4982–4996.

[131] Alyssa K Kosmides et al. “Metabolomic fingerprinting: challenges and opportunities”. In: *Critical Reviews™ in Biomedical Engineering* 41.3 (2013).

[132] Ashbala Shakoor et al. “Maillard reaction chemistry in formation of critical intermediates and flavour compounds and their antioxidant properties”. In: *Food Chemistry* 393 (2022), p. 133416.

[133] Wenyun Lu et al. “Metabolite measurement: pitfalls to avoid and practices to follow”. In: *Annual review of biochemistry* 86.1 (2017), pp. 277–304.

[134] Pengwei Zhang et al. “Susceptibility to false discovery in biomarker research using liquid chromatography–high resolution mass spectrometry based untargeted metabolomics profiling”. In: *Clinical and Translational Medicine* 11.6 (2021), e469.

[135] Ran Ju et al. “Removal of false positive features to generate authentic peak table for high-resolution mass spectrometry-based metabolomics study”. In: *Analytica Chimica Acta* 1067 (2019), pp. 79–87.

[136] Wei Chen et al. “A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics”. In: *Molecular plant* 6.6 (2013), pp. 1769–1780.

[137] Tianwei Yu et al. “Hybrid feature detection and information accumulation using high-resolution LC–MS metabolomics data”. In: *Journal of proteome research* 12.3 (2013), pp. 1419–1427.

[138] Kelsey Chetnik, Lauren Petrick, and Gaurav Pandey. “MetaClean: a machine learning-based classifier for reduced false positive peak detection in untargeted LC–MS metabolomics data”. In: *Metabolomics* 16 (2020), pp. 1–13.

[139] Weichaun Yu et al. “Improving mass spectrometry peak detection using multiple peak alignment results”. In: *Journal of proteome research* 7.01 (2008), pp. 123–129.

[140] Harish Mantika. “Investigating the Behaviour of Biologically Relevant Small Molecules within Ambient Ionisation Mass Spectrometry”. MA thesis. Oxford, United Kingdom: University of Oxford, 2025.

[141] Vitali Matyash et al. “Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics”. In: *Journal of lipid research* 49.5 (2008), pp. 1137–1146.

[142] Vivien F Taylor et al. “A mass spectrometric study of glucose, sucrose, and fructose using an inductively coupled plasma and electrospray ionization”. In: *International Journal of Mass Spectrometry* 243.1 (2005), pp. 71–84.

[143] Haoran Xing and Varoujan Yaylayan. “Insight into isomeric diversity of glycated amino acids in Maillard reaction mixtures”. In: *International Journal of Molecular Sciences* 23.7 (2022), p. 3430.

[144] Chiara Salvitti et al. “Kinetic Study of the Maillard Reaction in Thin Film Generated by Microdroplets Deposition”. In: *Molecules* 27.18 (2022), p. 5747.

[145] Francesco Saliu. “Soft Ionization mass spectrometry of lipid residues in archaeological findings: ESI vs APCI”. In: *Journal of Physics: Conference Series*. Vol. 2204. 1. IOP Publishing. 2022, p. 012044.

[146] John S O’Brien and E Lois Sampson. “Lipid composition of the normal human brain: gray matter, white matter, and myelin”. In: *Journal of lipid research* 6.4 (1965), pp. 537–544.

[147] Elizabeth Walsh and Nicolas M Orsi. “The current troubled state of the global pathology workforce: a concise review”. In: *Diagnostic Pathology* 19.1 (2024), p. 163.

[148] Kota Arima et al. “Metabolic profiling of formalin-fixed paraffin-embedded tissues discriminates normal colon from colorectal cancer”. In: *Molecular Cancer Research* 18.6 (2020), pp. 883–890.

[149] Andreas Dannhorn et al. “Evaluation of formalin-fixed and FFPE tissues for spatially resolved metabolomics and drug distribution studies”. In: *Pharmaceuticals* 15.11 (2022), p. 1307.

[150] David Chardin et al. “Identification of metabolomic markers in frozen or formalin-fixed and paraffin-embedded samples of diffuse glioma from adults”. In: *International Journal of Molecular Sciences* 24.23 (2023), p. 16697.

[151] Reinhard von Wasielewski et al. “Tissue array technology for testing interlaboratory and interobserver reproducibility of immunohistochemical estrogen receptor analysis in a large multicenter trial”. In: *American journal of clinical pathology* 118.5 (2002), pp. 675–682.

[152] Anthony SY Leong and Peter N Gilham. “The effects of progressive formaldehyde fixation on the preservation of tissue antigens”. In: *Pathology* 21.4 (1989), pp. 266–268.

[153] Stefano Cacciatore et al. “Metabolic profiling in formalin-fixed and paraffin-embedded prostate cancer tissues”. In: *Molecular cancer research* 15.4 (2017), pp. 439–447.

[154] Anna Wojakowska et al. “An optimized method of metabolite extraction from formalin-fixed paraffin-embedded tissue for GC/MS analysis”. In: *PloS one* 10.9 (2015), e0136902.

[155] Di Feng et al. "UPLC-MS/MS-based metabolomic characterization and comparison of pancreatic adenocarcinoma tissues using formalin-fixed, paraffin-embedded and optimal cutting temperature-embedded materials". In: *International journal of oncology* 55.6 (2019), pp. 1249–1260.

[156] Jean-Philippe Fortin et al. "Harmonization of cortical thickness measurements across scanners and sites". In: *Neuroimage* 167 (2018), pp. 104–120.

[157] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[158] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[159] Nir Friedman, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers". In: *Machine learning* 29.2 (1997), pp. 131–163.

[160] Sedir Mohammed et al. "The effects of data quality on machine learning performance on tabular data". In: *Information Systems* 132 (2025), p. 102549.

[161] Maria-del-Mar Inda, Rudy Bonavia, and Joan Seoane. "Glioblastoma multiforme: a look inside its heterogeneous nature". In: *Cancers* 6.1 (2014), pp. 226–239.

[162] Paul D Piehowski et al. "Sources of technical variability in quantitative LC–MS proteomics: human brain tissue sample analysis". In: *Journal of proteome research* 12.5 (2013), pp. 2128–2137.

[163] Pei Li and Michael G Bartlett. "A review of sample preparation methods for quantitation of small-molecule analytes in brain tissue by liquid chromatography tandem mass spectrometry (LC-MS/MS)". In: *Analytical Methods* 6.16 (2014), pp. 6183–6207.

[164] Gaël Varoquaux. "Cross-validation failure: Small sample sizes lead to large error bars". In: *Neuroimage* 180 (2018), pp. 68–77.

[165] Pasquale J Di Pillo. "The application of bias to discriminant analysis". In: *Communications in Statistics-Theory and Methods* 5.9 (1976), pp. 843–854.

[166] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.

[167] Lennart Schneider, Bernd Bischl, and Matthias Feurer. "Overtuning in Hyperparameter Optimization". In: *arXiv preprint arXiv:2506.19540* (2025).

[168] Anders Haug, Frederik Zachariassen, and Dennis Van Liempd. "The costs of poor data quality". In: *Journal of Industrial Engineering and Management (JIEM)* 4.2 (2011), pp. 168–193.

[169] Mary E McNamara et al. "Not just "big" data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions". In: *Behaviour research and therapy* 153 (2022), p. 104086.

[170] Yingbin ZHANG. "The Effect of Feature Reliability on the Generalization of Machine Learning Models in Educational Data". In: *International Conference on Computers in Education*. 2024.

[171] John R Taylor. *An introduction to error analysis: the study of uncertainties in physical measurements*. MIT Press, 2022.

[172] Daniele Foroni, Matteo Lissandrini, and Yannis Velegrakis. "Estimating the extent of the effects of Data Quality through Observations". In: *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE. 2021, pp. 1913–1918.

[173] Martin P Paulus and Wesley K Thompson. "Computational approaches and machine learning for individual-level treatment predictions". In: *Psychopharmacology* 238.5 (2021), pp. 1231–1239.

[174] Pasa Sukson et al. "Towards Robust Risk-Based Screening of Early-Stage Diabetes: Machine Learning Models with Union Features Selection and External Validation". In: *Diabetology* 7.1 (2025), p. 2.

[175] Laura Van Hese et al. "Towards real-time intraoperative tissue interrogation for REIMS-guided glioma surgery". In: *Journal of Mass Spectrometry and Advances in the Clinical lab* 24 (2022), pp. 80–89.

[176] Alan K Jarmusch et al. "Lipid and metabolite profiles of human brain tumors by desorption electrospray ionization-MS". In: *Proceedings of the National Academy of Sciences* 113.6 (2016), pp. 1486–1491.

[177] Ana L Seidinger et al. "Tumor-Promoted Changes in Pediatric Brain Histology Can Be Distinguished from Normal Parenchyma by Desorption Electrospray Ionization Mass Spectrometry Imaging". In: *Biomedicines* 12.11 (2024), p. 2593.

[178] Thais Maria Santos Bezerra et al. "Deep learning outperforms classical machine learning methods in pediatric brain tumor classification through mass spectra". In: *Intelligence-Based Medicine* 10 (2024), p. 100178.

[179] Jing Zhang et al. "Accuracy of Raman spectroscopy in differentiating brain tumor from normal brain tissue". In: *Oncotarget* 8.22 (2017), p. 36824.

[180] Joannie Desroches et al. "A new method using Raman spectroscopy for in vivo targeted brain cancer tissue biopsy". In: *Scientific reports* 8.1 (2018), p. 1792.

[181] Fanny Orlhac et al. "A guide to ComBat harmonization of imaging biomarkers in multicenter studies". In: *Journal of Nuclear Medicine* 63.2 (2022), pp. 172–179.

[182] Andrew E Jaffe et al. "Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis". In: *BMC bioinformatics* 16.1 (2015), p. 372.

[183] Harvard Wai Hann Hui, Weijia Kong, and Wilson Wen Bin Goh. "Thinking points for effective batch correction on biomedical data". In: *Briefings in Bioinformatics* 25.6 (2024), bbae515.

[184] Kirstie S Opstad et al. "An investigation of human brain tumour lipids by high-resolution magic angle spinning 1H MRS and histological analysis". In: *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* 21.7 (2008), pp. 677–685.

[185] Trang Thi Thu Nguyen et al. "HDAC inhibitors elicit metabolic reprogramming by targeting super-enhancers in glioblastoma models". In: *The Journal of clinical investigation* 130.7 (2020), pp. 3699–3716.

[186] Corina Tamas et al. “Metabolic contrasts: fatty acid oxidation and ketone bodies in healthy brains vs. glioblastoma multiforme”. In: *International Journal of Molecular Sciences* 25.10 (2024), p. 5482.

[187] Magdalena Gaca-Tabaszewska, Joanna Bogusiewicz, and Barbara Bojko. “Metabolomic and lipidomic profiling of gliomas—a new direction in personalized therapies”. In: *Cancers* 14.20 (2022), p. 5041.

[188] Lu Lu et al. “Lipid metabolism: the potential therapeutic targets in glioblastoma”. In: *Cell Death Discovery* 11.1 (2025), p. 107.

[189] Jack Rice, Benjamin Jenkins, and Albert Koulman. *Lipid feature characterisation using Total Correlation Mass Spectrometry and ultraviolet photodissociation.* LIP001. Application Note. Application note. Verdel Instruments. Verdel Instruments, Jan. 2023. URL: `https://verdelinstruments.boost.gdn/wp-content/uploads/2023/08/Application-Note-LIP001_Final.pdf`.

[190] Matej Skorvanek and Kailash P Bhatia. “The skin and Parkinson’s disease: review of clinical, diagnostic, and therapeutic issues”. In: *Movement Disorders Clinical Practice* 4.1 (2017), pp. 21–31.

[191] M José Blanca Mena et al. “Non-normal data: Is ANOVA still a valid option?” In: *Psicothema, 2017, vol. 29, num. 4, p. 552-557* (2017).

Appendices

A

Optimising Clinical Data Acquisition with ASAP-MS

Contents

A.1 Additional information for optimising the acquisition with ASAP-MS

Table A.1: Data distribution analysis and statistical significance test method choice

Dataset	Shapiro-Wilk test			Statistical significance test
	Statistic	df	p-value	Method
Background Effects				
Contaminated background	0.8573	25	0.002430	Mann-Whitney U test
Clean background	0.9620	25	1.03E-05	
Glass Capillary Cleaning				
Method 1 (Whole spectra)	0.9557	25	0.335960	Mann-Whitney U test
Method 2 (Whole spectra)	0.8718	25	0.004693	
Method 1 (m/z 200-300)	0.8689	25	0.004108	Mann-Whitney U test
Method 2 (m/z 200-300)	0.9617	25	0.448487	
Lens Tissue Peak Test				
Peak258 (Method 1)	0.6513	25	1.72E-06	Mann-Whitney U test
Peak258 (Method 2)	1.0000	25	1.000000	
Peak275 (Method 1)	0.9260	25	0.070427	Mann-Whitney U test
Peak275 (Method 2)	0.9396	25	0.145646	
Peak285 (Method 1)	0.8301	25	0.000750	Mann-Whitney U test
Peak285 (Method 2)	0.7676	25	6.86E-05	
User Repeatability				
User 1 to User 1 Centroid	0.8454	10	0.051221	ANOVA and Tukey's HSD
User 2 to User 2 Centroid	0.8543	10	0.065274	
User 3 to User 3 Centroid	0.9409	10	0.563233	
User 4 to User 4 Centroid	0.8734	10	0.109352	
User Reproducibility				
User 1 to User 1 Centroid	0.8454	10	0.051221	ANOVA and Tukey's HSD
User 1 to User 2 Centroid	0.9571	10	0.752081	
User 1 to User 3 Centroid	0.8136	10	0.021196	
User 1 to User 4 Centroid	0.9374	10	0.524356	
User 2 to User 1 Centroid	0.7920	10	0.011601	ANOVA and Tukey's HSD
User 2 to User 2 Centroid	0.8543	10	0.065274	
User 2 to User 3 Centroid	0.9482	10	0.647420	
User 2 to User 4 Centroid	0.9598	10	0.784061	
User 3 to User 1 Centroid	0.9430	10	0.586380	ANOVA and Tukey's HSD
User 3 to User 2 Centroid	0.9739	10	0.924502	
User 3 to User 3 Centroid	0.9409	10	0.563233	
User 3 to User 4 Centroid	0.9756	10	0.937322	
User 4 to User 1 Centroid	0.8941	10	0.188340	ANOVA and Tukey's HSD
User 4 to User 2 Centroid	0.8084	10	0.018355	
User 4 to User 3 Centroid	0.9418	10	0.573641	
User 4 to User 4 Centroid	0.8734	10	0.109352	

$*p$-value < 0.001: data deviation is likely severe; $0.001 < p$-value ≤ 0.01: data deviation is significant; $0.01 < p$-value ≤ 0.05: deviation is moderate; p-value > 0.05: data is normally distributed. ANOVA and Tukey's HSD are relatively robust to moderate deviations from normality [191], so this method was applied for user reproducibility analysis.

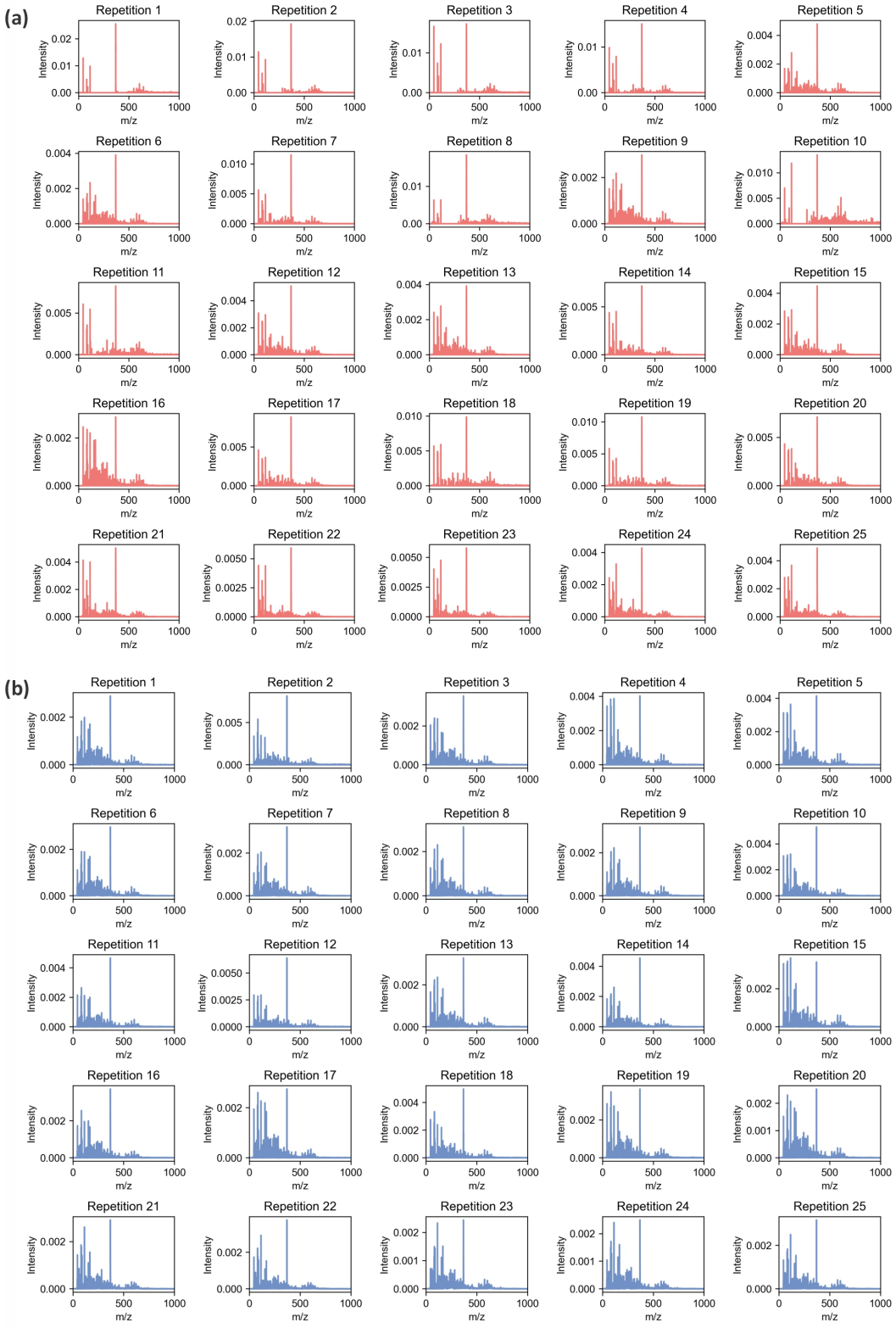

Figure A.1: All mass spectra recorded for a CSF sample in the evaluation of the effect of residual background mass peaks on the repeatability of the measurements (see main text for details): (a) CSF sampled shortly after calibration, with residual tuning mix peaks present; (b) CSF sampled after running the spectrometer until no tuning mix remained in the source, *i.e.* clean background.

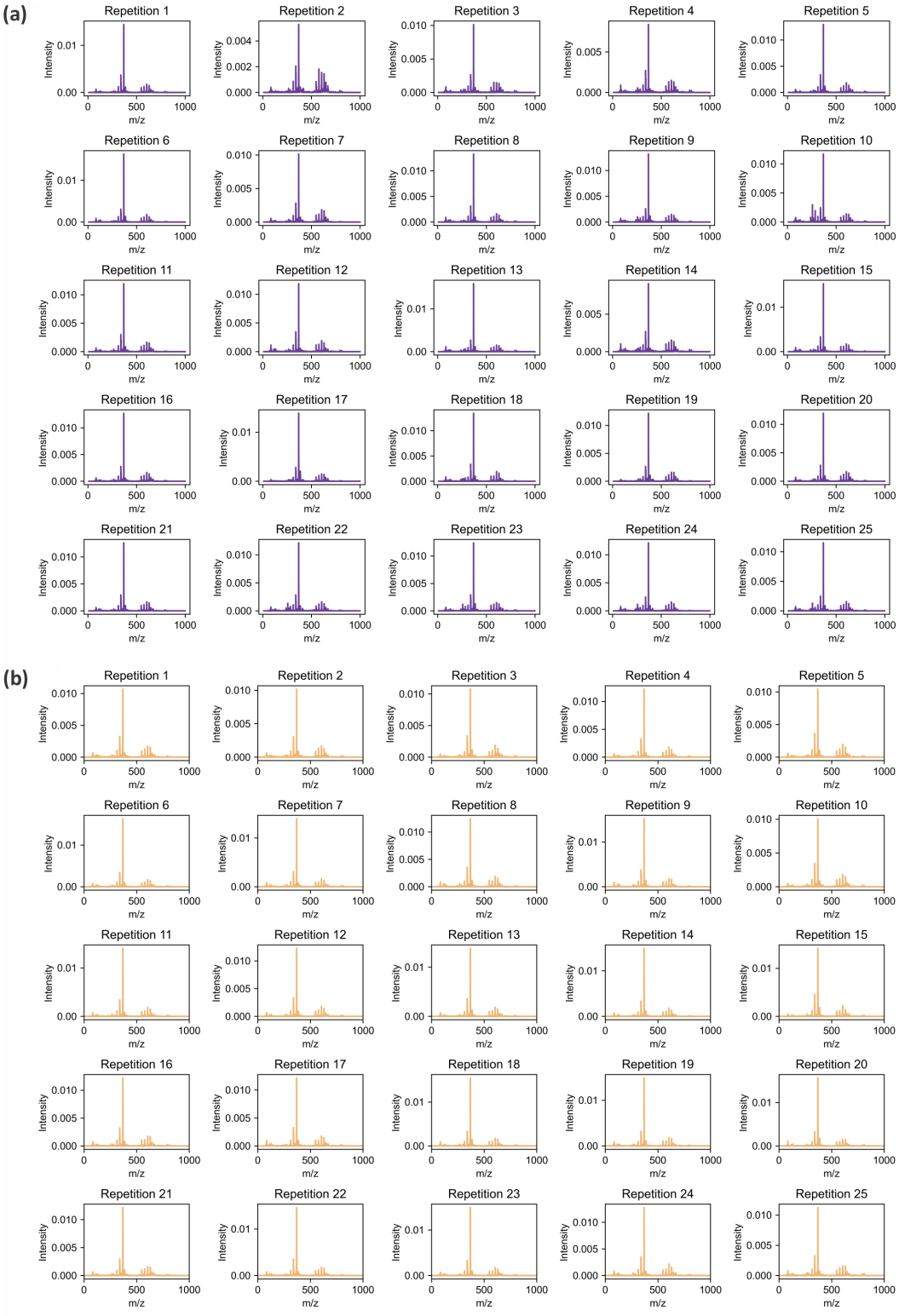

Figure A.2: Mass spectra recorded for a frozen brain sample using: (a) cleaning method 1 (without a hot nitrogen cleaning process between each repetition); and (b) cleaning method 2 (with a hot nitrogen cleaning process between each repetition).

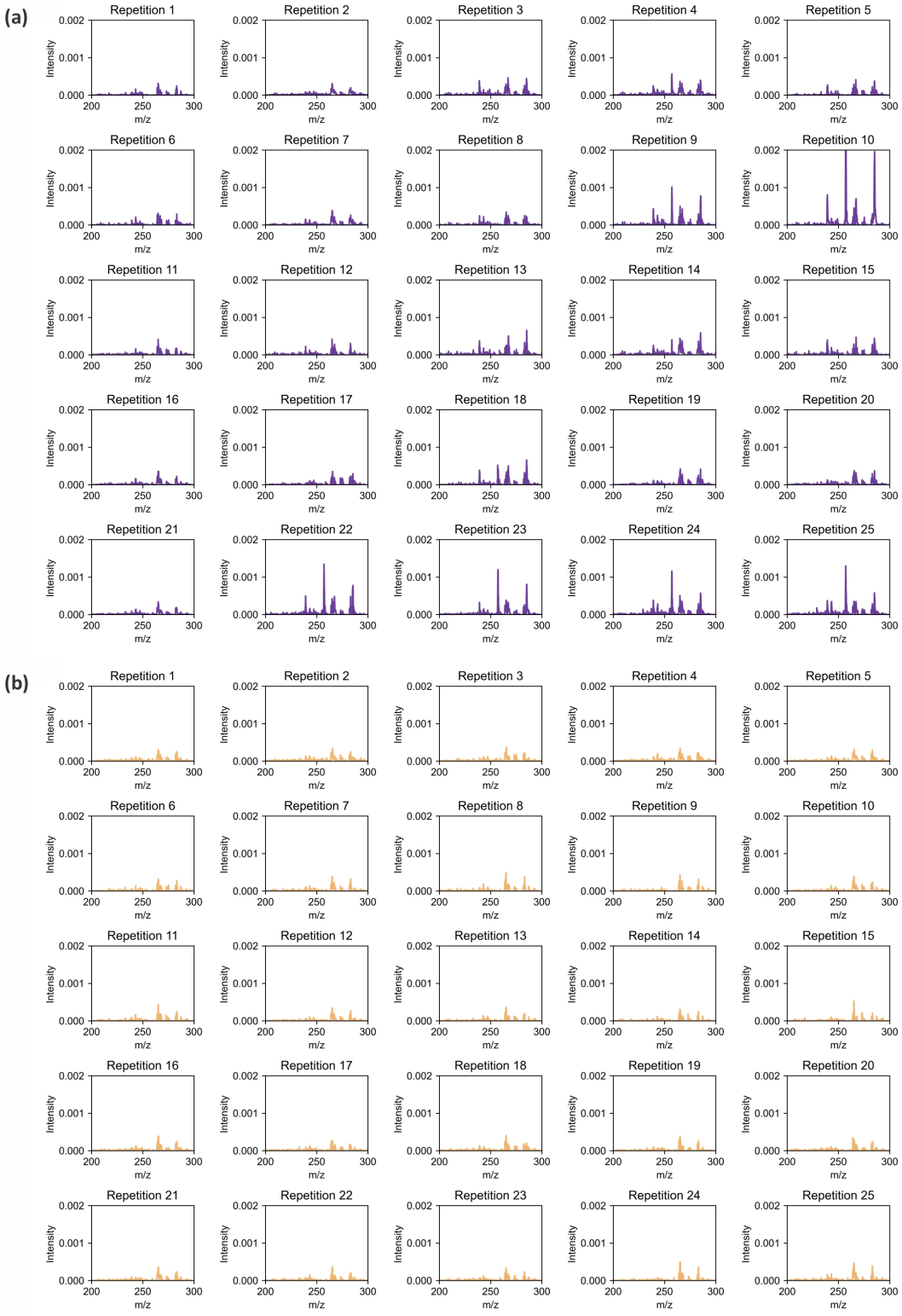

Figure A.3: As for Figure A.2, but showing only the low mass range between m/z 200 and 300.

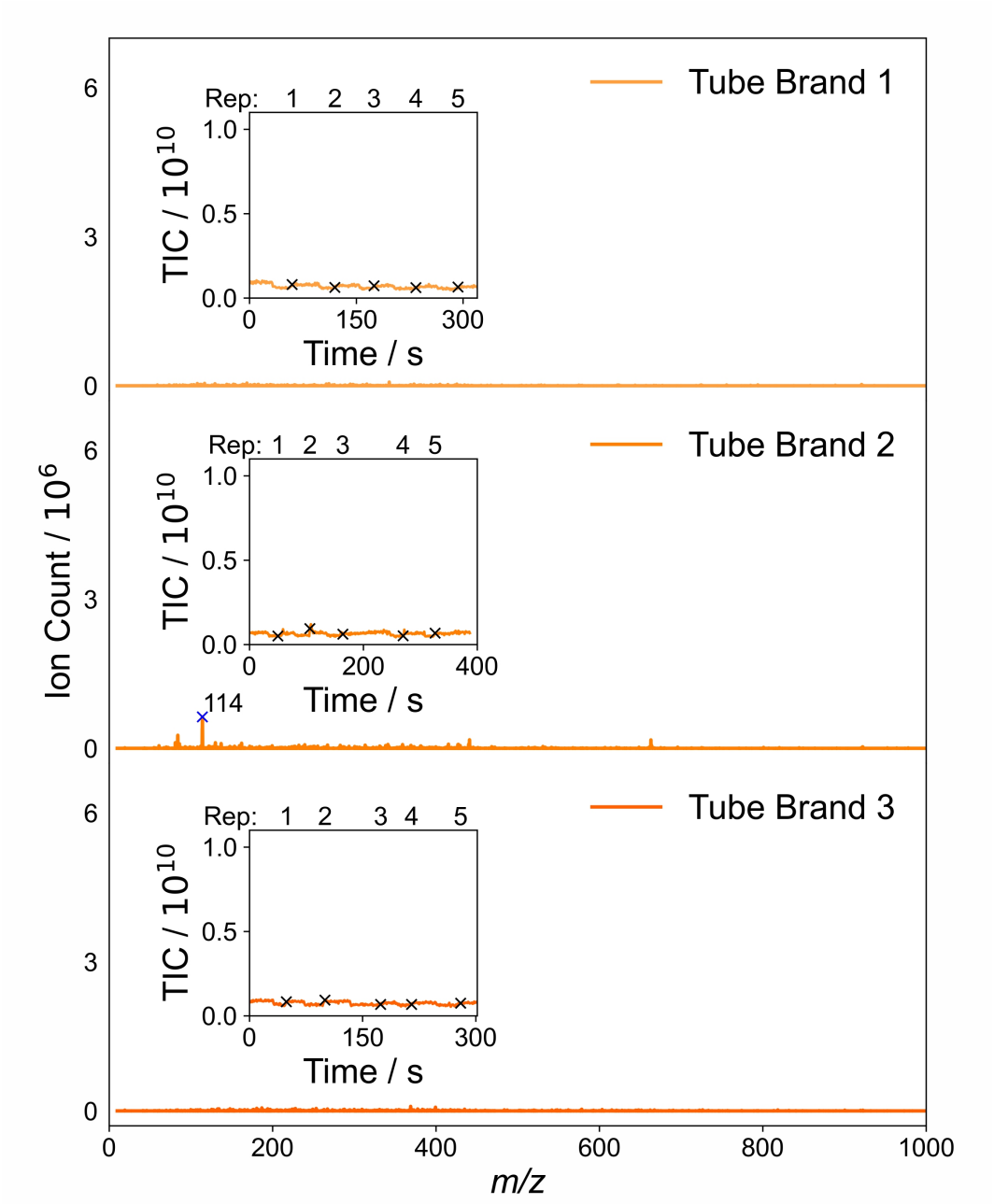

Figure A.4: Mass spectra recorded for samples of LC-MS water left overnight in three different brands of sample tube, with insets showing chromatograms for the total ion count (TIC) across five repetitions (Reps). While substantial signal was observed when sampling directly from the walls of the tubes (see main article), there appears to be no signficant leaching of material into solution, with no obvious peaks in the chromatograms and very low signal in the mass spectra.

B

Cumulative Average ASAP–MS Mass Spectra for Different Sample Types

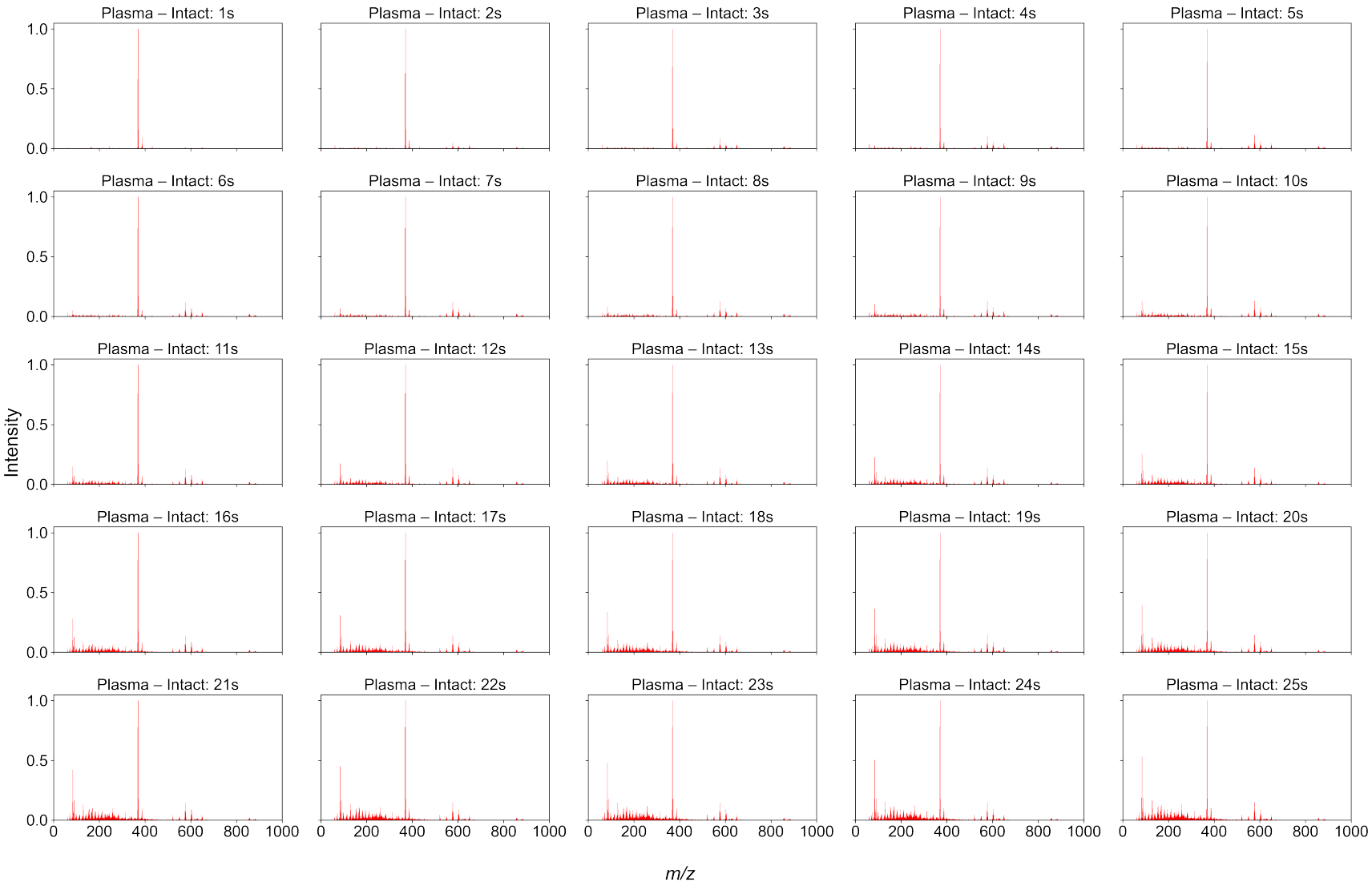

Figure B.1: Cumulative average ASAP–MS mass spectrum of an intact plasma sample from 0 s to the given time point.

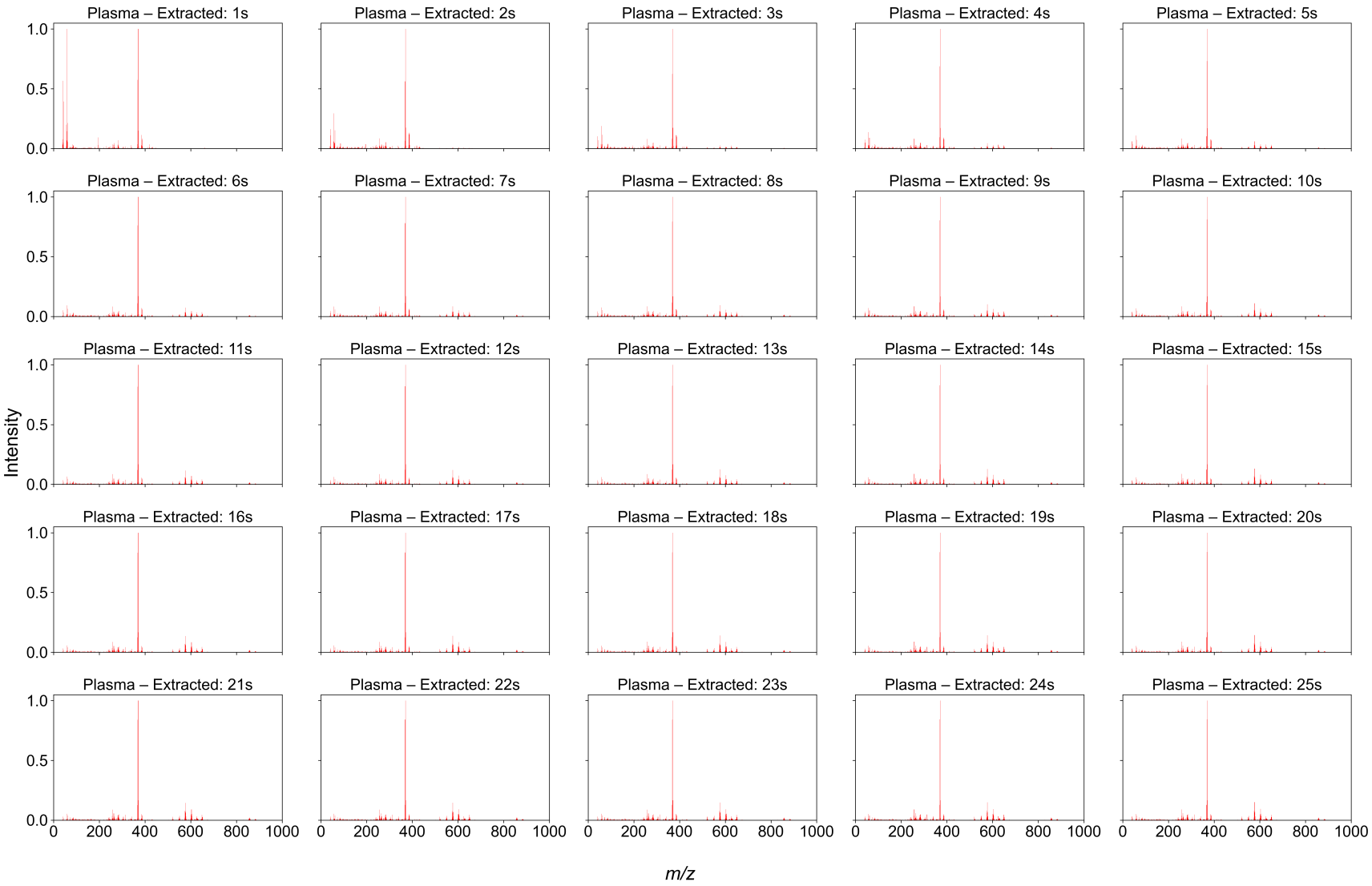

Figure B.2: Cumulative average ASAP–MS mass spectrum of a plasma lipid-extracted sample from 0 s to the given time point.

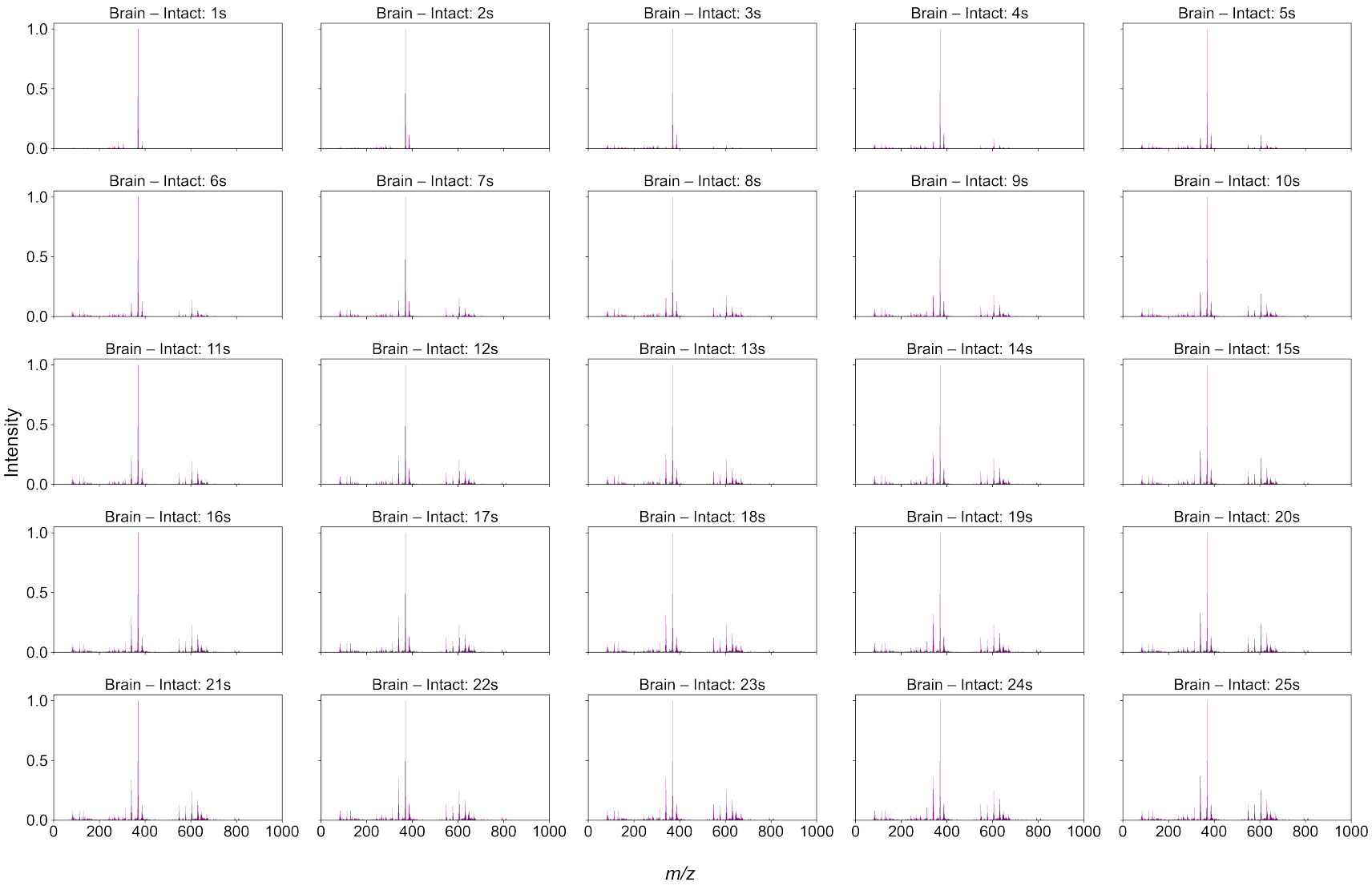

Figure B.3: Cumulative average ASAP–MS mass spectrum of an intact brain sample from 0 s to the given time point.

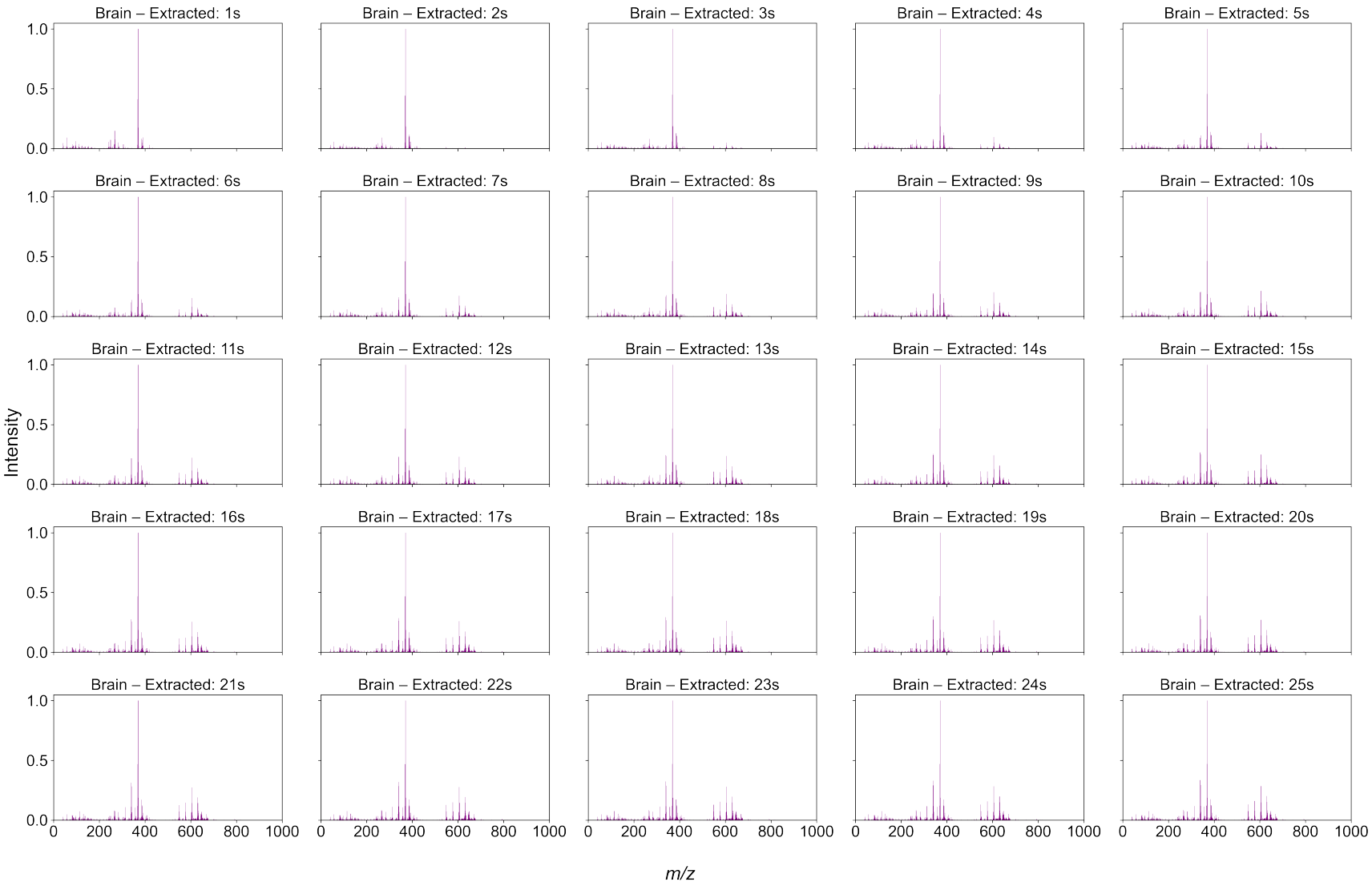

Figure B.4: Cumulative average ASAP–MS mass spectrum of a brain lipid-extracted sample from 0 s to the given time point.

C

ASAP-MS Brain Region Classifiers using FFPE and Frozen Neuropathological Samples

Contents

C.1 Fixation time for FFPE sample data set

Table C.1: Fixation time for FFPE sample data set

OBB ID	Fixation time / days
NP189/09	30
NP10/10	unknown
NP79/10	60
NP42/13	unknown
NP80/15	240
NP132/15	120
NP22/16	30
Np32/16	150
NP45/16	90
NP146/16	unknown
NP394/16	1095
NP428/16	180
NP24/17	120
NP92/17	4
NP100/17	4
NP103/17	4
NP13/18	4
NP14/18	2
NP23/18	3

C.2 Shapiro-Wilk test results

Table C.2: Data distribution analysis using Shapiro-Wilk test and statistical significance test

Data set	Within Patient	Within Region	Method	p-value
Frozen	0.346302569	0.000186589	Mann-Whitney U test	5.22E-10
FFPE	0.318671495	0.000186589	Mann-Whitney U test	0.002056831

D

ASAP-MS Brain Tumour Classifier Using Fresh Brain Tissue Samples

Contents

D.1 Normalisation

A modified normalisation strategy was also evaluated alongside two commonly used approaches: TIC normalisation and maximum intensity (Max) normalisation. To reduce the influence of unstable or highly variable features on the scaling factor, the coefficient of variation (CV) of each peak was calculated across all samples. Peaks with CV>0.5 were excluded from the normalisation denominator. The remaining, more stable peaks were summed to define a modified total ion count (MTIC), and each peak intensity was normalised by this value. The performance of this MTIC-based normalisation was then compared directly with standard TIC and Max normalisation.

- Maximum intensity (Max) normalisation Each peak intensity is scaled by the maximum peak ion count in the spectrum:

$$N_{i,j}^{\text{Max}} = \frac{I_{i,j}}{\max_k(I_{i,k})} \tag{D.1}$$

- Modified total ion count (MTIC) normalisation Only peaks with acceptable

stability ($CV_k \leq 0.5$) are included in the denominator:

$$N_{i,j}^{\text{MTIC}} = \frac{I_{i,j}}{\sum_{k \in S} I_{i,k}}, \tag{D.2}$$

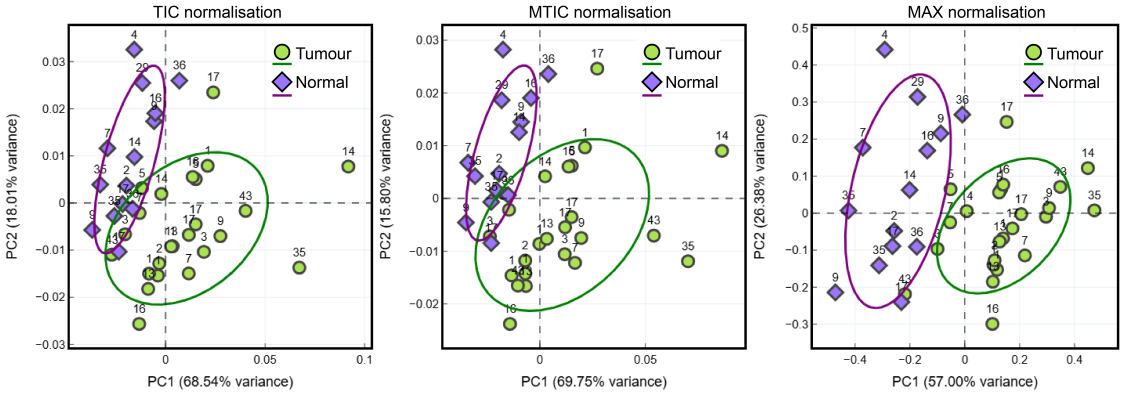

Figure D.1: PCA plot of the data set of the first 50 patients. The confidence ellipse corresponds to two standard deviations.

The results are shown in Figure D.1. As no clear differences were observed among the three normalisation methods, TIC normalisation was used for subsequent machine learning analyses.

D.2 Batch effect correction

Batch effects were corrected using the ComBat method. To evaluate the effectiveness of batch correction, a linear regression model was applied to assess the association between mass spectra and batch labels before and after correction.

Table D.1: Batch effect correction evaluation

Data set	R^2 before ComBat	R^2 after ComBat
Direct	0.514	0.077
Homogenised	0.453	0.086

D.3 Summary of patient demographics and sample composition.

Table D.2: Summary of patient demographics and sample composition.

Age	(Mean ± SD)	56 ± 13	
Category	**Group**	**Patient** n	**Sample** n
Sex	Female	38	–
	Male	36	–
Sample type	GBM	36	76
	Astrocytoma	12	21
	Ependymoma	2	4
	Metastasis	6	13
	Meningioma	3	6
	Others	4	8
	Normal cortex	37	44

D.4 Training sample counts updates

Table D.3: Training sample counts after Update 1

Sample	**N**	**N(Undersampling)**	**N(Oversampling)**
Normal	23	23	43
Tumour	43	23	43

Table D.4: Training sample counts after Update 2

Sample	**N**	**N(Undersampling)**	**N(Oversampling)**
Normal	23	23	46
Tumour	46	23	46

Table D.5: Training sample counts after Update 3

Sample	**N**	**N(Undersampling)**	**N(Oversampling)**
Normal	23	23	49
Tumour	49	23	49

Table D.6: Training sample counts after Update 4

Sample	N	N(Undersampling)	N(Oversampling)
Normal	24	24	58
Tumour	58	24	58

Table D.7: Training sample counts after Update 5

Sample	N	N(Undersampling)	N(Oversampling)
Normal	26	26	62
Tumour	62	26	62

Table D.8: Training sample counts after Update 6

Sample	N	N(Undersampling)	N(Oversampling)
Normal	29	29	70
Tumour	70	29	70

Table D.9: Training sample counts after Update 7

Sample	N	N(Undersampling)	N(Oversampling)
Normal	32	32	74
Tumour	74	32	74

Table D.10: Training sample counts after Update 8

Sample	N	N(Undersampling)	N(Oversampling)
Normal	35	35	80
Tumour	80	35	80

Table D.11: Training sample counts after Update 9

Sample	N	N(Undersampling)	N(Oversampling)
Normal	36	36	85
Tumour	85	36	85

Table D.12: Training sample counts after Update 10

Sample	N	N(Undersampling)	N(Oversampling)
Normal	39	39	94
Tumour	94	39	94

Table D.13: Training sample counts after Update 11

Sample	**N**	**N(Undersampling)**	**N(Oversampling)**
Normal	42	42	96
Tumour	96	42	96

Table D.14: Training sample counts after Update 12

Sample	**N**	**N(Undersampling)**	**N(Oversampling)**
Normal	43	43	109
Tumour	109	43	109

D.5 Iterative learning cycle results

Table D.15: Confusion matrix for unseen brain tissue samples (Batch 21)

Model	TP	FP	TN	FN
NB_oversampling	2	0	0	0
NB_undersampling	2	0	0	0
KNN_oversampling	2	0	0	0
KNN_undersampling	2	0	0	0
LDA_oversampling	2	0	0	0
LDA_undersampling	1	0	0	1
LR_oversampling	2	0	0	0
LR_undersampling	2	0	0	0
RF_oversampling	2	0	0	0
RF_undersampling	2	0	0	0
SVM_oversampling	2	0	0	0
SVM_undersampling	2	0	0	0

Table D.16: Confusion matrix for unseen brain tissue samples (Batch 22)

Model	TP	FP	TN	FN
NB_oversampling	2	0	0	1
NB_undersampling	2	0	0	1
KNN_oversampling	2	0	0	1
KNN_undersampling	2	0	0	1
LDA_oversampling	2	0	0	1
LDA_undersampling	2	0	0	1
LR_oversampling	2	0	0	1
LR_undersampling	2	0	0	1
RF_oversampling	2	0	0	1
RF_undersampling	2	0	0	1
SVM_oversampling	2	0	0	1
SVM_undersampling	2	0	0	1

Table D.17: Confusion matrix for unseen brain tissue samples (Batch 23)

Model	TP	FP	TN	FN
NB_oversampling	2	0	0	1
NB_undersampling	3	0	0	0
KNN_oversampling	2	0	0	1
KNN_undersampling	2	0	0	1
LDA_oversampling	3	0	0	0
LDA_undersampling	3	0	0	0
LR_oversampling	3	0	0	0
LR_undersampling	3	0	0	0
RF_oversampling	3	0	0	0
RF_undersampling	2	0	0	1
SVM_oversampling	3	0	0	0
SVM_undersampling	2	0	0	1

Table D.18: Confusion matrix for unseen brain tissue samples (Batch 24)

Model	TP	FP	TN	FN
NB_oversampling	1	0	0	0
NB_undersampling	1	0	0	0
KNN_oversampling	1	0	0	0
KNN_undersampling	1	0	0	0
LDA_oversampling	1	0	0	0
LDA_undersampling	1	0	0	0
LR_oversampling	1	0	0	0
LR_undersampling	1	0	0	0
RF_oversampling	1	0	0	0
RF_undersampling	1	0	0	0
SVM_oversampling	1	0	0	0
SVM_undersampling	1	0	0	0

Table D.19: Confusion matrix for unseen brain tissue samples (Batch 25)

Model	TP	FP	TN	FN
NB_oversampling	6	0	1	3
NB_undersampling	7	0	1	2
KNN_oversampling	6	0	1	3
KNN_undersampling	5	0	1	4
LDA_oversampling	8	0	1	1
LDA_undersampling	5	0	1	4
LR_oversampling	9	0	1	0
LR_undersampling	7	0	1	2
RF_oversampling	9	0	1	0
RF_undersampling	8	0	1	1
SVM_oversampling	8	0	1	1
SVM_undersampling	6	0	1	3

Table D.20: Confusion matrix for unseen brain tissue samples (Batch 26)

Model	TP	FP	TN	FN
NB_oversampling	4	0	2	0
NB_undersampling	4	0	2	0
KNN_oversampling	4	0	2	0
KNN_undersampling	4	1	1	0
LDA_oversampling	4	0	2	0
LDA_undersampling	4	0	2	0
LR_oversampling	4	0	2	0
LR_undersampling	4	0	2	0
RF_oversampling	4	0	2	0
RF_undersampling	4	0	2	0
SVM_oversampling	4	0	2	0
SVM_undersampling	4	0	2	0

Table D.21: Confusion matrix for unseen brain tissue samples (Batch 27)

Model	TP	FP	TN	FN
NB_oversampling	8	0	3	0
NB_undersampling	8	0	3	0
KNN_oversampling	8	0	3	0
KNN_undersampling	8	0	3	0
LDA_oversampling	8	0	3	0
LDA_undersampling	8	0	3	0
LR_oversampling	7	0	3	1
LR_undersampling	8	0	3	0
RF_oversampling	8	0	3	0
RF_undersampling	8	0	3	0
SVM_oversampling	8	0	3	0
SVM_undersampling	8	0	3	0

Table D.22: Confusion matrix for unseen brain tissue samples (Batch 28)

Model	TP	FP	TN	FN
NB_oversampling	4	0	3	0
NB_undersampling	4	0	3	0
KNN_oversampling	4	0	3	0
KNN_undersampling	4	0	3	0
LDA_oversampling	4	0	3	0
LDA_undersampling	4	0	3	0
LR_oversampling	4	0	3	0
LR_undersampling	4	0	3	0
RF_oversampling	4	0	3	0
RF_undersampling	4	0	3	0
SVM_oversampling	4	0	3	0
SVM_undersampling	4	0	3	0

Table D.23: Confusion matrix for unseen brain tissue samples (Batch 29)

Model	TP	FP	TN	FN
NB_oversampling	5	0	3	1
NB_undersampling	5	0	3	1
KNN_oversampling	6	0	3	0
KNN_undersampling	6	0	3	0
LDA_oversampling	5	0	3	1
LDA_undersampling	5	0	3	1
LR_oversampling	6	0	3	0
LR_undersampling	5	0	3	1
RF_oversampling	6	0	3	0
RF_undersampling	5	0	3	1
SVM_oversampling	5	0	3	1
SVM_undersampling	5	0	3	1

Table D.24: Confusion matrix for unseen brain tissue samples (Batch 30)

Model	TP	FP	TN	FN
NB_oversampling	4	0	1	1
NB_undersampling	4	0	1	1
KNN_oversampling	4	0	1	1
KNN_undersampling	4	0	1	1
LDA_oversampling	4	0	1	1
LDA_undersampling	4	0	1	1
LR_oversampling	4	0	1	1
LR_undersampling	4	0	1	1
RF_oversampling	4	0	1	1
RF_undersampling	4	0	1	1
SVM_oversampling	4	0	1	1
SVM_undersampling	4	0	1	1

Table D.25: Confusion matrix for unseen brain tissue samples (Batch 31)

Model	TP	FP	TN	FN
NB_oversampling	8	0	3	1
NB_undersampling	8	0	3	1
KNN_oversampling	8	0	3	1
KNN_undersampling	9	0	3	0
LDA_oversampling	8	0	3	1
LDA_undersampling	9	0	3	0
LR_oversampling	8	2	1	1
LR_undersampling	9	0	3	0
RF_oversampling	9	1	2	0
RF_undersampling	8	0	3	1
SVM_oversampling	8	0	3	1
SVM_undersampling	8	0	3	1

Table D.26: Confusion matrix for unseen brain tissue samples (Batch 32)

Model	TP	FP	TN	FN
NB_oversampling	1	0	3	1
NB_undersampling	2	0	3	0
KNN_oversampling	2	0	3	0
KNN_undersampling	2	0	3	0
LDA_oversampling	1	0	3	1
LDA_undersampling	1	0	3	1
LR_oversampling	1	0	3	1
LR_undersampling	1	0	3	1
RF_oversampling	2	0	3	0
RF_undersampling	1	0	3	1
SVM_oversampling	1	0	3	1
SVM_undersampling	1	0	3	1

Table D.27: Confusion matrix for unseen brain tissue samples (Batch 33)

Model	TP	FP	TN	FN
NB_oversampling	13	0	1	0
NB_undersampling	13	0	1	0
KNN_oversampling	13	0	1	0
KNN_undersampling	13	0	1	0
LDA_oversampling	13	0	1	0
LDA_undersampling	11	0	1	2
LR_oversampling	13	0	1	0
LR_undersampling	12	0	1	1
RF_oversampling	13	0	1	0
RF_undersampling	13	0	1	0
SVM_oversampling	13	0	1	0
SVM_undersampling	13	0	1	0

Table D.28: Summary of all misclassified samples

Specimen code	Wrong models	$N_{\text{wrong models}}$
4071777200	NB_over, NB_under, KNN_over, KNN_under, LDA_over, LDA_under, LR_under, SVM_over, SVM_under	9
4071777210	LR_over	1
4071777272	NB_over, KNN_over, KNN_under, RF_under, SVM_under	5
4071777312	NB_over, KNN_over, KNN_under, LDA_under, SVM_under	5
4071777334	KNN_under, LDA_under	2
4071777357	NB_over, NB_under, KNN_over, KNN_under, LDA_under, LR_under, RF_under, SVM_under	8
4071777414	NB_over, NB_under, KNN_over, KNN_under, LDA_over, LDA_under, LR_over, LR_under, RF_over, RF_under, SVM_over, SVM_under	12
4071777432	KNN_under	1
4071777517	NB_over, NB_under, LDA_over, LDA_under, LR_under, RF_under, SVM_over, SVM_under	8
4071777534	NB_over, NB_under, KNN_over, KNN_under, LDA_over, LDA_under, LR_over, LR_under, RF_over, RF_under, SVM_over, SVM_under	12
4071777542	LR_over	1
4071777543	LR_over,RF_over	2
4071777550	LR_over	1
4071777551	NB_over, NB_under, KNN_over, LDA_over, LDA_under, RF_under, SVM_over, SVM_under	8
4071777586	NB_over, KNN_over, KNN_under, LR_over, RF_under, SVM_over, SVM_under	7
4071777601	LDA_under	1
4071777602	LDA_under, LR_under	2

Table D.29: Comparison of neurosurgeon intra-operative assessment, final neuropathological diagnosis, and machine-learning (ML) model predictions for specimens labelled as ambiguous by neurosurgeons.

Specimen code	**Neurosurgeon**	**Neuropathologist**	**ML model**
4071777384	Cortex or Tumour	Tumour	Tumour
4071777405	Cortex or Tumour	Tumour	Tumour
4071777357	Cortex or Tumour	Tumour	Tumour
4071777432	Cortex or Tumour	Cortex	Normal
4071777255	Cortex or Tumour	Cortex	Normal

Table D.30: Samples with prediction changes after neuroComBat batch effect correction, N is the number of affected models

Specimen code	True class	Predicted class (not corrected)	Predicted class (corrected)	Impact	N
4071777414	Tumour	Tumour	Normal	Harmful	11
4071777200	Tumour	Tumour	Normal	Harmful	8
4071777357	Tumour	Tumour	Normal	Harmful	5
4071777432	Normal	Tumour	Normal	Beneficial	5
4071777542	Normal	Tumour	Normal	Beneficial	10
4071777543	Normal	Tumour	Normal	Beneficial	9
4071777551	Tumour	Tumour	Normal	Harmful	8

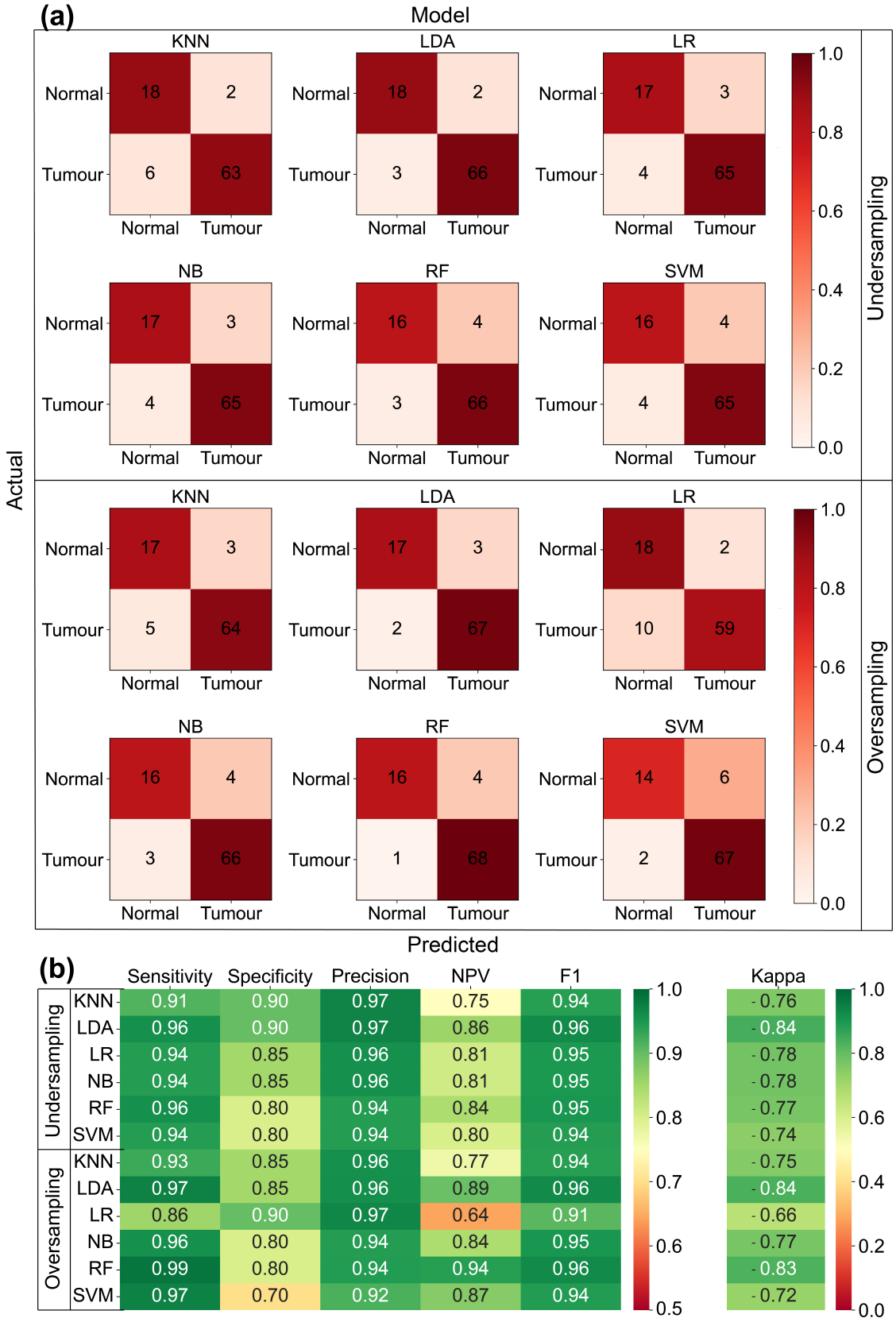

Figure D.2: Confusion matrices (a) and performance metrics (b) of classification outcomes for 12 models without batch correction