The Development of Chemical Analytical Tools for Community Drug Checking

by

Lea Gozdzialski
B.Sc., Queen's University, 2017

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Chemistry

We acknowledge and respect the lək̓ʷəŋən peoples on whose traditional territory the university stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day.

The Development of Chemical Analytical Tools for Community Drug Checking

by

Lea Gozdzialski
B.Sc., Queen's University, 2017

Supervisory committee

Dr. Dennis K. Hore, Supervisor
(Department of Chemistry)

Dr. Scott McIndoe, Academic Unit Member
(Department of Chemistry)

Dr. Margaret-Anne Storey
(Department of Computer Science)

ABSTRACT

Drugs have many uses from pleasure to pain relief, ceremony, and medicine.[1,2] Drugs also carry risks, from uncomfortable side effects, to dependence, and even death. In the case of pharmaceuticals, most people are familiar with receiving instructional notes and warnings such as "take with food" or "let your pharmacist know about other medications that might have undesirable interactions." Strict quality control means that prescribed and regulated drug mixtures are known to be as safe as possible. In the case of the illicit drug market, such assurances of quality and support are not afforded. In response, this thesis focuses on advancing the technology required for drug checking, a grassroots harm reduction initiative that aims to provide a level of quality control to the illicit drug market using various analytical approaches. Due to the unprecedented and increasing number of overdose deaths, these services have been expanding throughout North America. Drug checking empowers people who use drugs with the knowledge of what they are consuming and further provides an avenue for education and support to communities about the local drug supply. However, implementation of drug checking faces many barriers not only systemically but analytically as well, in part due to the dynamic and unpredictable drug supply and demand for simple, cost-effective, and point-of-care techniques.

This thesis explores several point-of-care analytical methods in their application to drug checking. These analytical methods include immunoassay test strips, infrared, Raman, and surface enhanced Raman spectroscopy, and gas chromatography–mass spectrometry. Notably, this research and development takes place while concurrently providing drug checking as a community harm reduction service. Most of the datasets used throughout this work are acquired at the service and reflect the local drug supply. A major focus of this research is on the detection of opioids and benzodiazepines in drug mixtures. Chemometric approaches are used to evaluate, compare, and improve the capability of multiple instruments in providing useful drug information for the local supply. This

includes classification and quantification schemes using a wide range of methods such as partial least squares regression, local outlier factor, principal component analysis, random forest classifier, least angle squares regression, correlation analysis, *k*-nearest neighbours, multivariate curve resolution, and density-based spatial clustering. Performance metrics, such as true positive rates, false positive rates, F1 scores, and receiver operating curves for qualitative detection and root mean square error and accuracy profiles for quantification, are used for evaluation. Additionally, a custom analysis platform is developed and implemented using Python and Jupyter notebooks to allow for such developments to be actualized within the service. Beyond technical evaluation, discussion of the results of this research largely considers the practical requirements of point-of-care service delivery. Within this work, technical information regarding drug checking technologies and data analysis is contextualized within harm reduction and contributes to strengthening the body of drug checking literature and resources.

Contents

List of Tables

List of Figures

List of Abbreviations and Definitions

2D-COS Two-dimensional correlation spectroscopy

2T2D two-dimensional two-trace correlation

AI artificial intelligence

ANPP 4-anilino-N-phenethylpiperidine

ATR attenuated total reflection

ATR–IR attenuated total reflection infrared

AUROC area under the receiver operating characteristic curve

CME coiled microextraction

CV cross validation

dTGS deuterated triglycine sulphate

FT Fourier transform

FTIR Fourier transform infrared

GC gas chromatography

GC–MS gas chromatography mass spectrometry

HDBSCAN hierarchical density-based spectral clustering in applications with noise

HPLC high performance liquid chromatography

IR infrared

IRE internal reflection element

KNN *k*-nearest neighbours

LARS least angle squares regression

LC–MS liquid chromatography mass spectrometry

LLOQ low limit of quantification

LOD limit of detection

MCR multivariate curve resolution

MDA 3,4-methylenedioxyamphetamine

MDMA 3,4-methylenedioxymethamphetamine

ML machine learning

MS mass spectrometry

MSM dimethylsulfone

NIR near-infrared

NIST National Institute of Standards & Technology

NMR nuclear magnetic resonance

OAT opioid agonist treatments

OPS overdose prevention site

PC principle component

PCA principle component analysis

PCC Pearson's correlation coefficient

PLS partial least squares

PLS-DA partial least square discriminant analysis

PLS-R partial least square regression

PS-MS paper spray mass spectrometry

PWUD people who use drugs

qNMR quantitative nuclear magnetic resonance

RF random forest

RIC reconstructed ion chromatogram

RMSECV root mean square error of cross validation

RMSEP root mean square error of prediction

ROC receiver operating characteristic curve

SERS surface enhanced Raman spectroscopy

SHAP shapely additive explanations

SNV standard normal variate

SORS surface offset Raman spectroscopy/spectrometer

SWGDRUG Scientific Working Group for the Analysis of Seized Drugs

TIC total ion count

XAI explainable artificial intelligence

ACKNOWLEDGEMENTS

How does one even start an acknowledgement section when no words could possibly express the depth of gratitude I have for all the people behind this work? I don't think I could have been any more lucky to end up on a project filled with the most amazing, hard-working, compassionate, supportive, and silly people. If I were to start listing names it would take up pages. You all inspire me everyday and in no way would I have ever been able to complete this without every one of you. Thank you to my family, your love has been felt from afar through it all.

Thank you to my supervisor Dr. Dennis Hore and my pseudo-supervisor (pseudovisor?) Dr. Bruce Wallace, who both started this drug checking project over 5 years ago. I am extremely grateful for your unwavering confidence and support during these years.

Mostly, thank you to all the people who have and continue to trust us to check their drugs. It is their courage, curiousity, creativity, and care above all else that fuels this movement of drug user liberation and I have learned so much from this community. Thank you to our community drug checking partners including AVI Health & Community Services and SOLID Outreach. Thank you to the National Science and Engineering Research Council of Canada for CGS-M and CGS-D scholarships, and the Department of Chemistry at the University of Victoria for a GRACE scholarship. These have allowed me to put my whole heart into this work.

In putting together this dissertation, I have also been grieving the too many lives lost along this journey—people I knew and those who I didn't get the chance to. They deserved so much more. I'd also like to specifically honour Armin Saatchi. As you will see, so much of the research I got to do would not have been possible without the on-site mass spectrometry he was an integral part of establishing. His light is missed. This is for him.

Chapter 1

Introduction

1.1 The Overdose Crisis in Canada

North America is in the midst of an overdose crisis. In Canada, over 35,000 people have died from opioid-related overdoses since 2016.[3] While this devastation has been felt across the country, the province of British Columbia (BC) has been the epicenter of the overdose emergency. In April 2016, the province of British Columbia declared illicit drug overdoses a public health emergency and that state of emergency is still in place today. From when the emergency was first declared to December 2022, over 11,000 lives have been lost to illicit drug overdoses in BC.[4] In 2023 the rate of overdose is currently at an unprecedented 46 deaths per 100,000 individuals in British Columbia.[4] Notably, the synthetic opioid fentanyl and related analogues have been increasingly involved in unregulated drug deaths, identified in over 85% of deaths in 2022 compared to 15% of deaths ten years prior in 2013.[4] In June of 2023, deaths equated to about 6 a day in the province.

Several public health services aimed at reducing drug overdoses and other associated health complications have been in place long before the current state of emergency in British Columbia. Services such as needle-exchange programs and supervised consumption sites began to emerge to support people who inject drugs and reduce the spread of blood-borne pathogens. Markedly, ongoing advocacy, activism, and education surrounding safer drug use and drug user liberation has been advanced through the initiatives and

dedication of grass-roots drug user organizations.[5,6] Peer consultation and immediate and radical responses to the needs of the community continue to address gaps in local public health efforts.[5,6] Some more recent initiatives in direct response to the opioid overdose deaths include naloxone distribution programs, expansion of opioid agonist treatment, the emergence of overdose prevention sites (OPS), and, central to this thesis, drug checking services.[7]

1.2 Harm Reduction

Services aimed to support people who use drugs have significantly increased in recent decades, not only in Canada, but around the world.[7–11] Many of the initiatives described above, including drug checking, fall under an umbrella term of "harm reduction." Since the term was first coined, the meaning of harm reduction has continually evolved. A cursory definition of harm reduction is true to the name; in general it describes initiatives, policies, and practices aimed at reducing adverse health and social consequences associated with human behaviour, which include both legal and illegal activities. Beyond the broad definition, harm reduction is a complex field that includes both practical interventions as well as more philosophical principles. Namely, in the context of drug use, five pillars of harm reduction were proposed by the Canadian Centre on Substance Abuse Working Group (1996). The fundamental principles are as follows,[12]

1. Recognize that eliminating drug use is not always attainable or desirable. Using drugs carries risks in the same way many human activities do;

2. Highlight human values of respect, worth and dignity of people, including those who use drugs;

3. Focus on reducing negative consequences, in contrast to a focus on abstinence;

4. Balance costs and benefits to individual good and societal good, and;

5. Prioritize immediate goals as identified by people who use drugs.

While these principles are the foundation of harm reduction-based service delivery, it is imperative to recognize that many harms are the product of inequities in social structures and prohibitionist policies.[12] Importantly, the motivation, work, and discussion throughout this thesis is guided by these principles of harm reduction and rooted within a social justice framework.

1.3 Drug Checking as a Harm Reduction Service

Drug checking is a community service offered under the umbrella of harm reduction services. Drug checking aims to reduce the harms associated with the unpredictable illicit drug supply by offering some degree of "quality control" in the context of drug criminalization of and stigma of individuals who use drugs. This aims to support existing harm reduction practices among people who use drugs (PWUD), such as those related to dosing drugs of unknown composition obtained from the illicit market. For instance, many individuals describe obtaining and sharing both qualitative and quantitative information through strategies such as test shots, substance descriptions, and seeking insights from those who have previously consumed substances from the same batch to enhance their safety and knowledge.[13] Free and confidential drug checking services use analytical methods to provide community members with additional qualitative and/or quantitative information about their substances, along with additional consumption information and harm reduction supports. Many early drug checking initiatives began in the festival or rave scene in response to concerns of counterfeiting, misrepresentation, or altered purity and potency of "party" drugs.[14] The current opioid overdose crisis in North America, however, has motivated scale-up and evaluation of drug checking services and technologies within community-based sites.[14] Recent research at community drug checking sites in the UK[15] and Canada[16] indicate benefits at the individual level, as services engage PWUDs and inform drug use behavior. Those studies also identified potential benefits at larger scales,

including quality control functions in the illicit market and within public health policy responses.

1.4 Drug Checking Technologies

In 2017, a global review on drug checking services surveyed the variety of service models implemented around the world.[17] The review identified many areas of differences between these services, including their modes of submission (on-site, fixed-site, postal), drug analysis methods (reagent, optical spectroscopy, mass spectrometry, or chromatography-based), wait times for results (ranging from less than 5 minutes to greater than 1 week), communication strategies with the individual seeking the testing (in-person, text, email, webpage) and engagement with the public (public webpage, reports, alerts).[17] The ideal drug analysis method is of particular debate, and heavily influences these other aspects of service. To date, many instruments have been proposed and explored within the context of drug checking[17,18] including infrared absorption spectroscopy,[19–22] Raman scattering,[23,24] surface-enhanced Raman scattering (SERS),[25–27] nuclear magnetic resonance (NMR),[28,29] ion mobility,[30] electrochemical detection methods,[31] and mass spectrometry (MS) either on its own[30] or coupled with gas chromatography (GC–MS),[23,32,33] liquid chromatography (LC–MS),[19,23,29,32] or paper spray ionization (PS-MS).[34,35] Traditional lab-based instruments, such as gas chromatography (GC) or liquid chromatography (LC) combined with mass spectrometry (MS), are highly sensitive and can be expected to report on the detailed composition of a sample. In forensic testing, these instruments are considered a gold standard; meaning they have been thoroughly tested for the application, trusted to perform well, and are often used for bench-marking and validating other methods.[18] There is an interest in portable instruments that are able to provide more immediate, point-of-care test results and that can be more easily integrated within overdose responses such as supervised consumption sites and overdose prevention services. For this reason, many drug checking projects have been exploring

instruments such as portable infrared (IR) or Raman spectrometers for on-site testing. Spectroscopy-based instruments are typically fast, reliable, and relatively straightforward to both operate and maintain, however, they often suffer from low sensitivity. To address these limitations, many point-of-care drug checking services use multiple instruments or techniques. For example, a portable instrument can be complemented with sensitive fentanyl and benzodiazepine immunoassay test strips. This multi-instrumental approach can also be supplemented with laboratory-based services such as confirmatory testing for overall monitoring of the supply. It is not yet clear which approach is best suited for responding to the opioid crisis[26,36,37] and opinions continue to shift with the evolving drug supply. The illicit opioid markets are becoming increasingly complex, challenging drug checking efforts.[38–40] In addition, the drug supply often varies between communities and therefore services might have different needs.

1.5 Gaps in Drug Checking Services and Research

Several projects have evaluated the suitability of popular point-of-care drug checking technologies, such as IR, Raman, or test strips, by comparing these methods to methods such as LC–MS,[19] GC–MS,[18,21,26] and quantitative nuclear magnetic resonance (qNMR).[41] Most of these assessments focus on IR spectroscopy and reveal that this method is suitable for detecting many compounds seen in the drug supply. Defining the limitation of a particular drug checking method has become crucial within the context of harm reduction. Recent literature has highlighted the feasibility of IR for the detection of notable substances, particularly of high-risk substances.[39,40,42,43] High-risk substances include mixtures where potent drugs are often present in concentrations below the detection limit of IR spectroscopy and/or test strips. Examples include counterfeit Xanax (alprazolam) tablets,[43] fentanyl analogues such as carfentanil,[44] and a number of adulterants found in opioids such as xylazine,[40] benzodiazepines,[42] and synthetic cannabinoids.[39] In many cases, the drug checking work to date relies on commercial software; this may be an

automated black box analysis (often in the case of handheld Raman instruments catered to law enforcement)[26] or a manual analysis by a drug checking technician.[42] Typically, when multiple techniques are used, the interpretation of test results heavily relies on a technician to consolidate information from multiple sources, assess the confidence in results based on their experience, and relay this information to the client.

Researchers are also working to optimize instrument hardware, sample preparation routines, and data analysis workflows for more accurate and sensitive drug detection. The body of related work is extensive and is commonly presented in fields such as forensics, pharmaceuticals, food science, or environmental chemistry.[45–48] Some recent efforts aimed to improve drug checking technologies include the development of novel substrates for surfaced-enhanced Raman spectroscopy (SERS) for fentanyl detection[49] and method development and validation of paper spray mass spectrometry (PS-MS) for identification and quantification of fentanyls and pharmaceuticals.[50] Novel multivariate data analysis schemes are also relevant to drug checking efforts, for example deep learning algorithms for analysis of mixtures using Raman spectroscopy,,[51] quantitative image analysis of immunoassay test strips,,[52] and data fusion techniques for improved discriminatory analysis of powders using IR and Raman.[53] Many of these extensive machine learning schemes and instrumental advances have yet to be implemented in real-time drug checking services.

In general, point-of-care drug checking technologies are expected to

1. identify both high concentration bulking agents and low concentration psychoactive components in mixtures;
2. distinguish between structurally similar drugs;
3. quantify the amount of substances in mixtures;
4. be easy to use, i.e. minimal requirement for chemistry background or extensive training;

5. offer rapid testing with a short turn-around time for results
6. be relatively affordable and accessible to enable adoption throughout communities, and
7. pair results with appropriate harm reduction messaging and resources.

1.6 Objectives and Scope

This research consists of several linked projects that aim to push the limits in one or more of these categories using a combination of Raman spectroscopy, surface enhanced Raman spectroscopy, IR absorption spectroscopy, GC–MS, immunoassay test strips and chemometric approaches. In addition, a dynamic data analysis platform was built that facilitates the examination of instrumental data. This integrates several features catered to experienced and inexperienced drug checking technicians and harm reduction workers. These tools aim to improve the accuracy and consistency of drug checking services, promote rapid implementation of research into practice, and strengthen the communication of analytical results. While some research is currently underway in comparing and developing individual instrumental techniques, minimal effort has been dedicated to automating and consolidating test results from multiple instruments and technologies for an overall interpretation. The work throughout this thesis is largely influenced by the fifth pillar of harm reduction, that is to prioritize immediate goals as identified by people who use drugs. Therefore, the nature of and the timing of each project has been influenced by drugs of concern in the local supply in Victoria, BC, the needs of diverse communities on Vancouver Island, and conversation and collaboration with the growing community of drug checkers and people engaging with drug checking.

Chapter 2 first introduces the drug checking service, the analytical instrumentation used, and the general data acquisition parameters employed. This is followed by a tutorial-style description of relevant machine learning and chemometric principles. It also

provides a step-by-step walk-through of the practical implementation of such methods using example data from the drug checking service. Details of the Python code used in these examples is exposed, and references are given to several open-source packages used throughout this thesis. This point is highlighted because initiatives such as drug checking rely on collaboration and community for growth. Open-source software reflects these same values.

The projects within this thesis are grouped into three main themes, as outlined below.

1.6.1 Investigating Instrumental Candidates for Point-of-Care Drug Checking

First, in Chapters 3–6 various portable instruments are developed in direct response to the local drug supply and community needs. During this time, drug checking with chemical instrumentation was relatively new and few evaluations existed that applied and adapted these technologies outside of a laboratory setting to the dynamic, unregulated drug market. These chapters establish the capabilities—strengths, limitations, and potential—of instrumental candidates applied in a point-of-care drug checking service.

Chapter 3 describes the development of regression models to quantify fentanyl in various drug mixtures using Raman spectroscopy. This was motivated by the growing demand for quantification information for fentanyl in the local drug supply.

Chapter 4 presents the method development and evaluation of portable GC–MS instrumentation to detect low concentration adulterants in opioid samples. This is compared to analysis of the same samples using IR spectroscopy and least angle squares regression (LARS). This project was motivated by the identification of carfentanil and etizolam through secondary testing with PS-MS, whereas the detection of these compounds were being missed at point-of-care.

Given the challenges of successfully employing portable GC–MS in real time, Chapter 5 seeks to find alternative methods to detect etizolam, a growing adulterant in the opioid supply resulting in atypical overdoses within the community. Here surface-

enhanced Raman spectroscopy (SERS) was used with a simple univariate model to detect etizolam in complex opioid mixtures. This was realized as benzodiazepine test strips, recently implemented in drug checking, were not accurately detecting etizolam. In low concentrations, identification of etizolam was also limited using other spectroscopic instrumentation such as Raman and infrared spectrometers.

Chapter 6 expands on the work done in the previous chapters using individual instruments and demonstrates a holistic approach using multiple instruments for analyzing various street samples. Here four samples which represent various degree of difficulty are evaluated, and the successes and challenges of several instruments are discussed.

1.6.2 Improving Infrared Spectral Analysis for Expanding the Reach of Drug Checking Research and Development

Second, Chapters 7–9 specifically focus on research projects aiming to develop data analysis methods applied to infrared spectroscopy. As drug checking continued to expand, infrared spectroscopy became a very common choice of instrumentation for other point-of-care services. At the same time, our project pursued ways to lower the barrier to begin drug checking in more remote communities, that may not have the resources to have a trained technician. Together, this inspired the greater focus on improving the analysis of infrared spectra, using machine learning and automation.

Chapter 7 implements a novel analysis technique, namely two-trace two-dimensional correlation analysis, with the goal of detecting low concentration substances in complex mixtures using infrared spectroscopy.

Chapter 8 demonstrates an approach for automated detection of fluorofentanyl and MDA using random forest (RF) classifiers. This work also expands from training and validation to the practical implementation of such models. This employs principles of explainable artificial intelligence (XAI).

Chapter 9 presents the most recent progress in a comprehensive machine learning pipeline for the automated analysis of infrared spectra. This work expands upon

principles discussed in Chapter 8. A multi-step classification scheme is developed where broad clusters are first established, based on natural patterns within the data. This is implemented using hierarchical density-based spectral clustering in applications with noise (HDBSCAN). Within each cluster, prediction of target compounds is achieved using a series of random forest classifiers.

1.6.3 Developing a Dynamic User Interface for Applying Research into Service

Lastly, while isolated into a single chapter in this dissertation, Chapter 10 covers the various stages of development of an interactive platform for the analysis of instrumental drug checking data. This interface was initially developed to consolidated instrumental data from multiple instruments, as reflected in early research projects emphasizing the unique information provided from multiple techniques. Later, a tailored, web-based platform was designed for infrared spectral analysis allowing for both an interactive interface for manually exploring spectral data as well as an automated analysis scheme as developed in Chapters 8 and 9. Ultimately, this platform was developed to allow for rapid translation of research into the drug checking service. Beyond our own drug checking service, this work aims to establish software catered to harm reduction-based drug checking services, that considers the nuance in spectral interpretation and harm reduction messaging, the dynamic nature of such services, and the importance of knowledge consolidation within the growing drug checking community.

Chapter 2

Background

This chapter is comprised of four main parts. First, Section 2.1 introduces a brief timeline of the community drug checking project I have been a part of and is at the core of the work in this thesis. Section 2.2 provides a background of the analytical instruments and methods used within the drug checking service and developed in various projects in this thesis. This includes scientific principles yet is structured to connect those principles to practical considerations in their application to point-of-care drug checking services. Acquisition details are also provided for each instrument. Section 2.3 introduces the field of chemometrics and related terms that are revisited throughout this work. Finally, Sections 2.4 and 2.5 include walk-through examples that apply chemometric principles (introduced in Section 2.3) to real data sets produced from drug checking instruments (introduced in Section 2.2). The examples demonstrate the functionality of open source coding packages and expose the intricacies of each step of their implementation from data set up, outlier analysis, model training and development, and performance evaluation.

2.1 Victoria, BC-Based Drug Checking Service

As a part of a harm reduction service operated as Substance Drug Checking, formerly the Vancouver Island Drug Checking Project,[37] drug samples are presented by individuals for testing at several sites that have sanctioning from public health to provide overdose prevention services including drug checking. These samples are considered unknown in

composition prior to testing with portable IR, Raman, GC–MS, and immunoassay fentanyl and benzodiazepine test strips. Later, paper spray mass spectrometry was added to the suite of instruments.

When this multi-instrument project was officially established in 2018, the drug checking service initially set up within other existing services. The drug checking instruments (and service) rotated between several sites within Victoria on a regular schedule. These sites included: (1) STS pain pharmacy, a harm reduction focused pharmacy that primarily serves unhoused populations and those struggling with substance use, (2) an overdose prevention site operated by AVI health services, (3) SOLID outreach, a drug user organization offering a cannabis substitution program amongst many other harm reduction services, and (4) Lantern services, the first official sanctioned drug checking site in Canada. When the COVID epidemic was declared in 2020, many of these essential services adapted and relocated to local encampments and temporary housing sites. The drug checking service followed, while navigating public health measures. In early 2021, a permanent drug checking storefront, named 'Substance Drug Checking', was established, again co-located with partners SOLID outreach and AVI health services. In 2022, significant efforts were made to expand the reach of drug checking to other communities on Vancouver Island. The sites established were termed 'distributed sites', where a kiosk-like software allowed for remote drug checking and interpretation from staff in Victoria, termed the central 'hub' site.[54] Mail-in drug checking was also expanded to further the accessibility of the hub site in Victoria. At the same time, a local 'envelope model' was established where envelopes were distributed to other services and sites throughout the community. Outreach workers collect drug samples and drop off the envelope in a mail slot at Substance at any time.

From an operational perspective, the workflow for drug checking has adapted to the environment and workload. In general, the service is typically provided by one or more people trained to run the analytical instruments, interpret the data, and produce the

results, and one or more harm reduction experts to facilitate communication of results and provide harm reduction advice, support, and additional resources.[37] These two roles are often referred to as 'drug checking technician' and 'harm reduction worker', respectively. However, it has become clear that the skill set for each role is not mutually exclusive. While staff typically bring an area of expertise, whether it be rooted in drug knowledge, pharmacology, analytical instrumentation, harm reduction or community support, an understanding of both the technical and harm reduction principles is critical for the success of drug checking.

This brief timeline of the practical aspects of the drug checking service provide insights into the evolving requirements of the service not only from a community perspective, but also an instrumental and technological perspective.

2.2 Instrumental Methods[1]

The following sections describe the instrumentation currently and previously used in the drug checking service and introduces their benefits, challenges, and potential as a drug checking instrument. The general acquisition parameters for each instrument is outlined as used in the service, however may have had minor changes over the years. Many of the datasets analysed in the projects that follow have been acquired during the active drug checking service provision. In some research projects sample data may be remeasured using specific methodology; this is outlined in the relevant chapters. On occasion, analytical standards are also used.

2.2.1 Immunoassay Test Strips

Test strips, also known as lateral flow immunoassays, provide a simple yes/no answer to whether a trace amount of a particular compound is present in a sample.[55,56] The test

[1] Some of the content from this section appeared in L. Gozdzialski, B. Wallace, D. Hore, "Point-of-Care Community Drug Checking Technologies: An Insider Look at the Scientific Principles and Practical Considerations". *Harm Reduct. J.*, **20**, 39 (2023).

strips commonly used for fentanyl and benzodiazepine screening in drug checking were originally designed for use with urine samples. Immunoassays offer a convenient, fast, and low-cost testing method, useful for outreach and at-home use.[55,57–59] They also remain an important tool in many drug checking services,[37,60–62] even when analytical instruments are also employed. In the case of fentanyl test strips, their on-going use is also largely attributed to their reliability and low limit of detection (about 0.150 μg/mL) compared to many analytical instruments.[26]

Basic principles. Most commercially-available lateral flow immunoassays for drug detection are based on the principles of competitive binding.[63–65] Here the absence of a test line indicates a positive result, in contrast to the sandwich assays where the appearance of a test line indicates a positive result (e.g. pregnancy tests, COVID-19 rapid tests). Immunoassays are all about the interaction between antibodies and antigens (i.e. the target molecule that binds to antibodies).[66] In this example, fentanyl is the antigen. Figure 2.1a demonstrates key components of a strip test; including the pad that contains unbound gold-labelled antibodies specific to fentanyl, a bound row of the same antibodies and a control line with bound non-specific antibody. In the scenario illustrated in Figure 2.1b., no fentanyl is present in the solution (a negative result). Upon dipping the strip in the solution the water will dissolve the antibodies from the pad. Since no fentanyl has bound, those antibodies are free to bind the test line resulting in the appearance of a red line. Since there are an excess of the antibodies flowing up with the water, they are also able to bind the control line and appear red. The presence of the control line indicates that the test strip has worked as expected; i.e. the water flowed properly and carried the antibodies. In the scenario illustrated in Figure 2.1c, fentanyl is present above a certain threshold (a positive result). Here fentanyl, the antigen, will bind the unbound gold-labelled antibodies as the water flows up the test strip. Since theses antibodies has been inhibited by fentanyl, they are not available to bind on the test line and therefore pass through. The control line, however, remains non-specific; it will bind regardless of whether fentanyl is bound to the antibodies

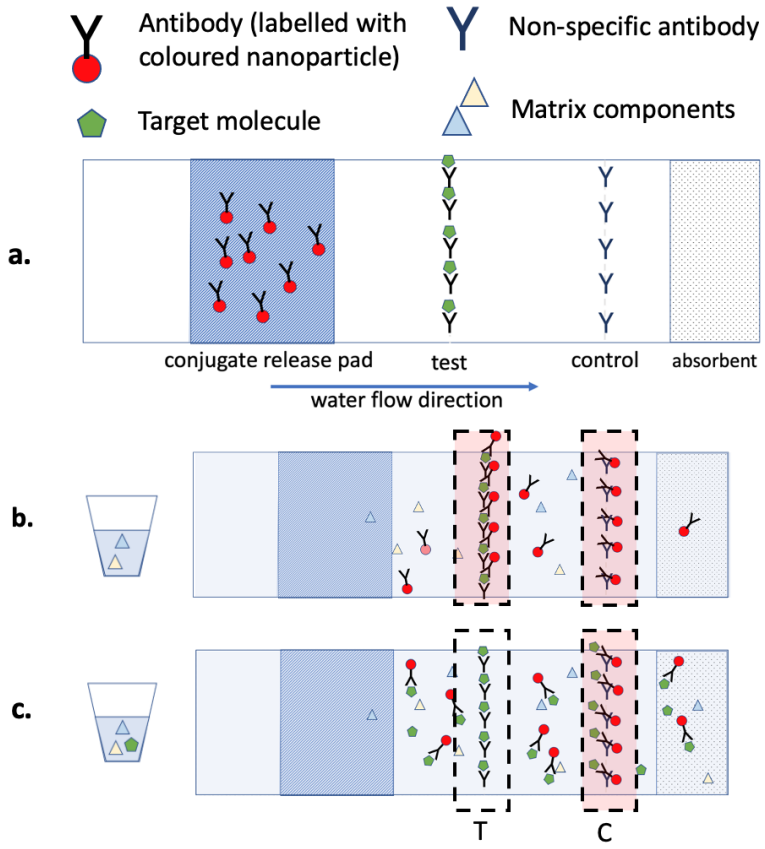

Figure 2.1: (a) Schematic of a competitive lateral flow immunnoassay dipstick, (b) Example of a fentanyl test strip where the both test and control line appears red, indicating a negative result, and (c) example of a fentanyl test strip where the control line appears red and test line is absent (inhibited), indicating a positive result due to fentanyl binding the labelled antibodies and preventing their interaction with the test line.

or not and should appear red.

Practical considerations. Immunoassay test strips have been widely adopted in many applications due to their simplicity and reliability.[67,68] However, nuances still exist in the use of test strips within the context of drug checking.[44,56,69] Some common challenges that need to be accounted for when using test strips include: (1) false positives, where out-of-class compounds with structural similarities to the target class (e.g. opioid class) can result in high frequency of false positives; (2) false negatives, where the structural diversity of drugs within the same class (e.g. benzodiazepines) make it increasingly difficult to manufacture a single assay to confidently screen for all target substances without compromising selectivity; (3) subjective results: while manufacturers recommend

to interpret the presence of a test line, regardless of its intensity, as a negative result it has been suggested that a more liberal interpretation is necessary, such as clear positive, positive, weak positive/ambiguous and negative, and; (4) communication: knowing the correct way to communicate such results may be context specific.

Many of the challenges listed above have been encountered at the drug checking service. In the case of fentanyl test strips, it is known that too much sample in solution causes false-positive results when that sample contains crystal meth or MDMA.[52] In the case of benzodiazepine test strips, etizolam, a benzodiazepine-related compound, has been shown to interact poorly,[25,70] resulting in a false-negative screen for benzodiazepines. If the concentration of the target compound approaches the limit of detection, the intensity of the test line is generally correlated with concentration[52] and therefore may appear faded. Technically, therefore test strips results can provide information on the concentration of the drug. However, That connection requires more sensitive detection of the line colour than can be gauged by eye, and a precise control of the sample mass and solution volume. Such careful preparation and digital read out is typically not the intent of test strip drug checking. Furthermore, solution pH, target drug solubility, and the presence of other drugs or cutting agents might affect binding, reaction times, and therefore the intensity of the test line.[52,71] As a result, people utilize test strips for the qualitative information they readily provide.

Data Acquisition. We use fentanyl (20 ng/mL cut-off) and benzodiazepine (300 ng/mL cut-off) test strips (BTNX, Markham, Canada). In the case of fentanyl test strips, a small amount of substance (1–2 mg) is dissolved in approximately 2 mL of water and manually agitated until dissolved. Fentanyl test strips are run on all samples presented to the drug checking service. When substances known to be MDMA or methamphetamine are presented, a larger volume of water (about 10 mL) is used due to false positives for fentanyl at concentrations greater than 1 mg/mL.[52] Benzodiazepine test strips are used for testing expected opioids, fake Xanax tablets, unknown substances, in cases when a benzodiazepine is suspected, or upon request. Here, approximately 2 mg of sample is

placed in 1–2 mL of warm water and agitated until dissolved before using the test strip according to manufacturer instructions.

2.2.2 ATR-IR Spectroscopy

Infrared (IR) absorption spectroscopy is rapidly becoming one of the most widely used instrumental methods for drug checking on account of its relatively low cost, ease of operation, speed, minimal sample preparation requirements, and the availability of libraries (open source and commercial) containing thousands of drug components including cutting agents.[18,43,72–75] IR spectroscopy can identify a wide range of compounds but it has limited sensitivity in comparison with immunoassay test strips. Recent assessments of the suitability of IR spectroscopy for community drug checking focus on the detection and quantification of fentanyl and other compounds found in the opioid drug supply.[19,20,40] People with a limited background in science have been successfully trained to operate IR spectrometers, however concerns remain around the possibility of misinterpreting data, as well as failing to recognize, and communicate, the limitations of such instruments.[19,22,76]

Basic principles. IR absorption refers to a family of techniques with the same basic operating principle. An IR light source is depicted as the lamp in Figure 2.2, and typically emits in the frequency range 500–5000 cm^{-1}, or the equivalent wavelength range of 20000–2000 nm. The ratio of the light intensity before and after sample interaction is related to how much IR light is absorbed by a drug sample. Ultimately, an IR spectrum is a plot of the degree to which the light has been absorbed (the absorbance) as a function of the IR frequency. Over the past several decades, nearly all IR absorption techniques make use of an instrument that is based on splitting and recombining the IR beam in an optical configuration known as an interferometer.[77] Subsequent Fourier transformation (FT) of the raw data provides the frequency spectrum. This is the basis for the acronym FTIR, referring to Fourier Transform IR spectroscopy. The terms IR, IR absorption, and FTIR are therefore equivalent in this context. There are several possible sampling configurations for

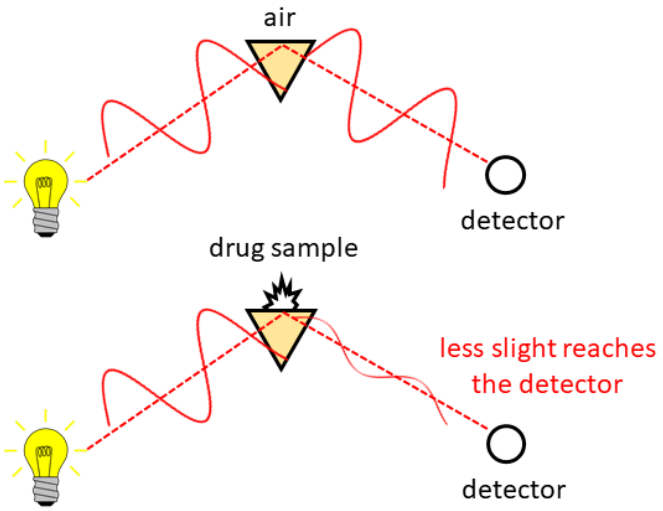

Figure 2.2: In an attenuated total internal reflection (ATR) absorption measurement, the sample is pressed against a prism, and infrared light is reflected from the surface. The ratio between measurements performed with and without the drug sample results in a spectrum that can be used to identify and quantify the components in a mixture.

an IR experiment; most spectrometers used in drug checking reflect IR light through a small IR-transparent prism against which the sample is pressed.[74,75] This method is known as ATR–IR, with the acronym describing the physical phenomena of attenuated total internal reflection (ATR).[78]

All subsequent analysis to identify the components in the drug mixture is based on characteristic molecular vibrations in this frequency spectrum, originating from the structure of chemical bonds in the constituent molecules.[79] As an example, the chemical structure of methamphetamine is shown in Figure 2.3. Other molecules with the same number of atoms, but different connectivity (for example, phentermine, Figure 2.3), will have the same number of vibrations (termed modes), but they will have different characteristic frequencies and intensities. As a result of this molecular specificity, the IR spectrum acts as a fingerprint that can be searched against a database to identify which substances are present.

Practical considerations. It is often accepted that IR absorption methods can determine most significant components (active ingredients as well as cutting agents) in

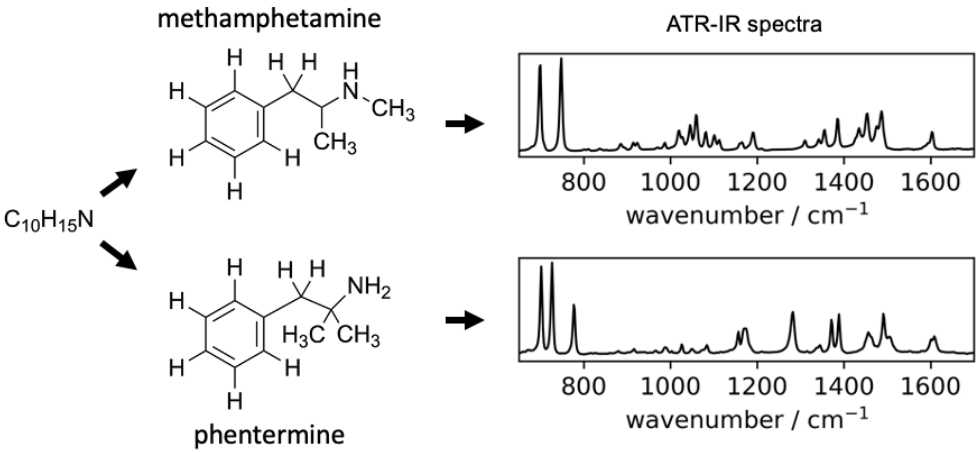

Figure 2.3: The structure of two compounds methamphetamine and phentermine, illustrating the different connectivity of the carbon framework, and location of the hydrogen and nitrogen atoms despite the same molecular formula. The similar, yet unique IR fingerprint is shown for each compound.

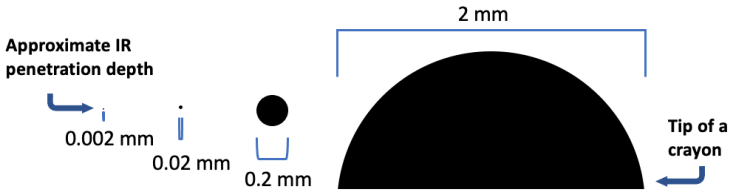

Figure 2.4: Graphical representation of the approximate sample penetration depth of 0.002 mm in ATR–IR absorption spectroscopy.

a mixture, as long as each one is at least $\approx 5\%$ of the overall mixture.[60,80,81] Although this is a reasonable estimate, in many cases such a hard-and-fast rule breaks down.

Challenges associated with detecting components of drug mixtures using IR spectroscopy include, (1) Any components present in low amounts will not absorb a significant enough fraction of the IR light to be detected. (2) Some molecules are difficult to distinguish from other species that have similar IR fingerprints, especially in low concentrations. (3) Successful identification relies on a library entry for each compound present.

Identifying compounds at concentrations near the limit of detection is subject to technician experience and confidence, ultimately resulting in subjective and variable results. The small IR beam penetration depth also poses additional sampling challenges and contributes to uncertainty for heterogeneous mixtures (i.e samples that might not be thoroughly mixed).[82] Although the sampling depth is roughly 0.002 mm, in order to ensure that the approximately 2 mm $\times$ 2 mm are of the crystal is covered, one uses approximately 1 mg of sample. However, regardless of how much sample is placed on the spectrometer, the resulting IR spectrum depends only on the very small amount of sample right at the surface of the ATR crystal. Using the numbers above provides a volume of 2 mm $\times$ 2 mm $\times$ 0.002 mm = 0.008 mm^3. Assuming a density of 1.23 g/cm^3 (based on caffeine, for example), this corresponds to a mass of approximately 10 μg. Since this probing volume is so small, obtaining quality ATR–IR spectra relies on close contact of the sample with the ATR crystal. This is why we have to apply pressure to solid samples using an anvil, but not to liquids, to collect the spectra. Figure 2.5 illustrates some scenarios at the crystal–sample interface that might affect the reproducibility of the resulting IR spectra. For example, there is a limit to how closely particles of different sizes and shapes can be packed, resulting in air gaps (Figure 2.5A).[83] The depth of the IR beam penetration into the sample also depends on the optical properties of the sample and crystal, IR wavelength (Figure 2.5B), and angle of incidence (typically fixed at 45°).[78,82] Spectral variation commonly seen might be seen

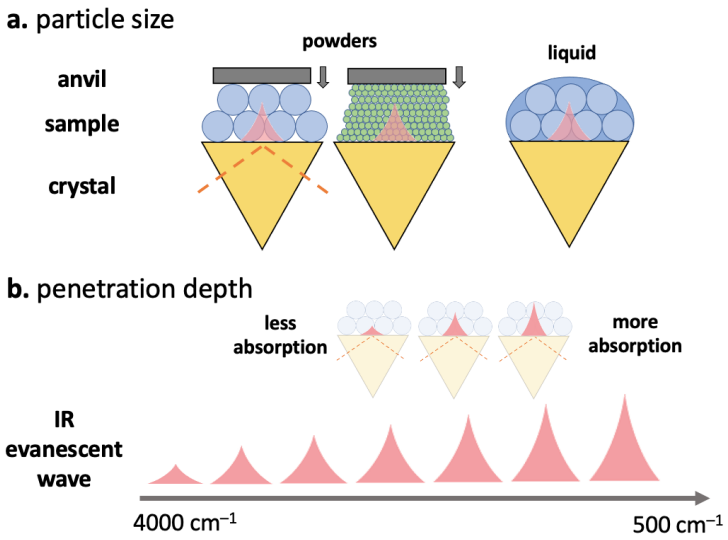

Figure 2.5: (a) A comparison of the sampling when powders and liquids are placed in contact with an ATR crystal. In the case of powered samples, there is a challenge in creating a close enough contact to ensure sufficient absorption of IR light since the particles may be loosely packed creating air gaps. Finer powders mitigate this issue as smaller particles can pack closer. (b) The wavelength dependence of the IR penetration depth further complicates the various packing scenarios, and ultimately affects the mixture analysis.

in the baseline. Although most software performs a simple correction for the wavelength-dependence of the probing IR light, the optical properties of the sample are more difficult to account for, especially with mixtures, and so the penetration depth still varies somewhat with wavelength. Furthermore, inhomogeneous powders and a distribution of particle sizes can affect the shape of the baseline, and also contribute to peak shifting and broadening.[84]

Various challenges also exist for using IR spectroscopy for determining the concentration of a particular component in a drug mixture, For starters, it may be difficult to *identify* the component as a result of any one of the reasons highlighted above. In the event of a successful library match, quantification is still challenging as spectral unmixing relies on the assumption that IR spectra of each of the pure components will be ideally superimposed to create the spectrum of the mixture. In practice, this often does not occur as a result of interactions between molecules in a mixture (often termed matrix effects), uncertainty in the optical constants of the mixture, and the various baseline artifacts described above.

These same challenges apply when discussing the purity of a given drug sample. In

many cases, only one component is identified using IR spectroscopy. However, in general the purity (e.g. how close the drug is to being 100% one ingredient) of a compound can never be quantitatively determined using techniques with limited sensitivity, such as IR spectroscopy. For example, a combination of precursors, biproducts, solvents, and other impurities may be present in low concentrations. In addition, precursors or biproducts from the drug synthesis may share many structural, and therefore spectral similarities, to the major drug product.

IR data acquisition. We employ a portable FTIR spectrometer (Agilent 4500a, Agilent Technologies, Santa Clara, California) equipped with a dTGS detector and a single-bounce diamond ATR accessory. Approximately 1–2 mg of sample is used to cover the ATR crystal. FTIR data is initially collected by co-adding 32–64 scans over the 650–4000 cm^{-1} range at a spectral resolution of 4 cm^{-1}. An open source drug library, SWGDRUG, is used in the analysis of acquired spectra.

2.2.3 Raman Spectroscopy

Raman spectroscopy is another vibrational spectroscopy method that provides molecular identity information, but based on the the way a molecule scatters light as opposed to the way a molecule absorbs light in IR spectroscopy.[85] The utility of Raman spectroscopy in drug checking is frequently compared to that of ATR–IR due to their complementary nature.[86,87] They also share many favourable characteristics: both techniques are quick, non-destructive, can be implemented with portable and robust hardware, and can identify a wide range of drugs, cutting agents and mixtures. Use of Raman spectroscopy has been reported for festival drug checking,[23,88,89] as well as in community drug checking.[17,24,90] However, similar to IR, it lacks the sensitivity required for trace detection.

Basic principles. In Raman spectroscopy, a laser that is typically in the visible region such as 488 nm (blue), 532 nm (green), or 633 nm (red) is focused onto a solid, powder, or liquid sample. Most of the light is either transmitted, absorbed, or reflected

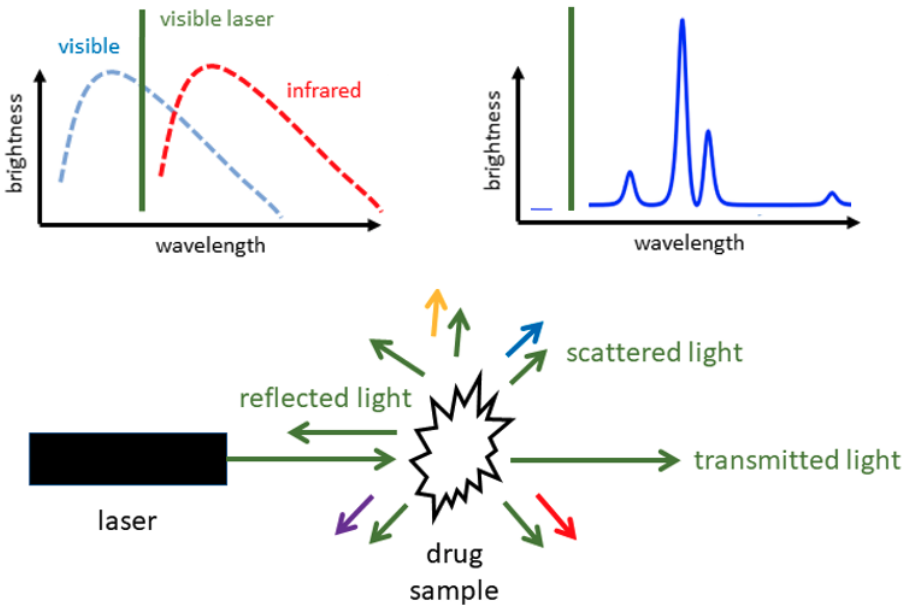

Figure 2.6: In a Raman scattering measurement a powder or liquid sample of the drug mixture is illuminated with a laser beam. Some of the reflected light appears at wavelengths (colors) different from that of the laser, and these wavelengths are characteristic of the constituent molecules in the sample.

as shown in Figure 2.6. However, a small quantity (typically less than 1%) is scattered in all directions.[85,91] The experimental challenge is to collect this scattered light and analyze its spectrum. A small portion of this scattered light will be shifted in wavelength after interacting with the molecules in the sample; this is known as Raman scattering.[92] The spectrum of the Raman scattering provides a chemical fingerprint of the molecular vibrations of the sample components in much the same way as an IR absorption spectrum and can be directly analyzed[93] and compared against a reference database for identification. The relative intensity of vibrational modes will be different than those in the IR although the frequencies are the same.[94] For this reason, one needs a database that contains dedicated Raman libraries.

Practical considerations. The appearance and quality of Raman spectra depend on the instrument configuration, such as laser power, laser wavelength, light focusing, and spot size.[91] In general, most of the main challenges listed for IR spectroscopy above also apply

(e.g. overlapping peaks, low sensitivity, requirement for libraries). Specific advantages of Raman include through-bag and through-barrier capability,[95] and minimal interference with water.[96]

Raman spectroscopy is seemingly less popular than IR spectroscopy in the drug checking community. A major shortcoming of Raman spectroscopy is that, for certain samples, the weak Raman scattered light is overwhelmed by fluorescence that is typically orders of magnitude brighter than the Raman scattering. The fluorescence of a particular molecule (such as heroin) is predicable, but significant fluorescence may also originate from unknown trace impurities. Spectral processing can sometimes successfully subtract fluorescence background, allowing accurate qualitative and quantitative detection. However, if the fluorescence is too strong, no chemical information can be obtained.

Some strategies can be employed to mitigate the effects of fluorescence. The best approach is to select a laser wavelength that does not result in significant fluorescence. For white powders, fluorescence is generally avoided by choosing a longer wavelength laser. For example, moving from 532 nm to 785 nm to 830 nm. Typically, portable Raman spectrometers are only equipped with a single-wavelength laser. One can imagine that it is near-impossible to select one wavelength that minimizes fluorescence for every single molecule when measuring such a diverse range of substances. Coloured samples (for example, opioid mixtures that have been dyed purple) are particularly challenging, as it is difficult to select a laser with a long enough excitation wavelength.[91] Near-infrared excitation at 1064 nm has been demonstrated to be a successful approach for such samples, with the compromise of lower sensitivity.[97] Other strategies to reduce fluorescence are to dilute or photo-bleach the sample.[91]

There are some instances where Raman may be advantageous over IR. In some cases, common mixtures that may be heavily overlapped in their IR spectrum might show unique and isolated peaks in the Raman spectrum, or vice versa; some of these cases will be discussed later. An obvious example is water. Water is a very weak Raman scatterer

and so drugs dissolved in solution or that have absorbed moisture can more easily be measured without significant interference. In contrast the IR spectrum suffers from strong absorption by water.[98] Since Raman spectrometers typically use a visible light source, the measurement can also be performed through transparent bags as they do not absorb visible light. Similar to water, some common and convenient materials such as polyethylene (a very common plastic) and glass are weak Raman scatterers relative to most drug molecules.[91] Spatially offsetting the collection of Raman signal has found use in measuring tablets where coatings may interfere in measuring the active ingredient,[91,99] or allow to get a more bulk sampling of drug. Many handheld Raman spectrometers benefit from a large spot size (about 2 mm in diameter) and therefore, a larger sampling area; an advantage when considering heterogeneous drug mixtures.

Raman spectroscopy can also be used for quantification, meaning the Raman signal of a substance within a mixture can be related to its concentration. The limit of detection of specific compounds depends on details of the instrument hardware, the extent to which the fluorescence can mitigated, and how much the target molecule Raman signature differs from that of other components in the mixture, both in terms of relative intensity and the characteristic frequencies of the bands. For example, studies have shown that for cocaine the limit of detection can range from 10% when cut with inositol to 40% when cut with paracetamol.[100] In another study, the LOD for heroin in a mixture was found to be as low as $\approx 5\%$.[101]

Raman data acquisition. Raman spectra were acquired using a handheld surface offset Raman spectrometer (SORS) (Resolve, Agilent Technologies, Santa Clara, California), equipped with an 830 nm laser. Scattered light with Stokes shifts in the range 200–2000 cm^{-1} is measured directly from a powder sample. About 2–3 mg of sample (more if available) is placed on a disposable aluminum tray prior to measurement. Liquid solutions may also be measured through a glass vial.

2.2.4 Surface-Enhanced Raman Spectroscopy

Surface-enhanced Raman scattering (SERS) has been developed in order to detect ultra-low concentrations of analytes.[102–105] This has enabled the detection of trace components in drug mixtures.[49,106–108] SERS may also alleviate the other primary challenge with Raman spectroscopy, fluorescence.[91] SERS has recently been explored for application in harm reduction and drug checking exclusively in the detection and quantification of trace compounds in mixtures, most commonly opioids.[25,49] Although SERS is not widely employed in drug checking, it is gaining attention on account of its potential for trace detection.

Basic principles. The general principle behind SERS is that the Raman response may be amplified by a factor ranging from ten thousand to a million when the samples are placed near metal surfaces with sharp edges.[109] A particularly simple and effective approach is to use a solution of metal nanoparticles in solution. When the nanoparticles aggregate after the addition of a suitable salt (e.g. NaCl, $MgSO_4$), the distance between neighboring particles becomes small, and the Raman signal is greatly enhanced if drug component molecules can be trapped in that region.[108,110,111] The precise mechanism by which the Raman response is amplified continues to be of research interest in the chemistry and physics communities, but appears to be sensitive to the nature of the nanoparticles, the solution conditions, and the specific target molecules.[91,109,112] For this reason, libraries prepared for standard Raman scattering are generally unsuitable for searching using SERS data, and must be tailored to the specific test conditions.

Practical considerations. SERS is poised to fill a gap in point-of-care drug checking by offering trace-level identification using relatively simple sample preparation and instrumentation. SERS, however, hass major challenges to implement as a widespread screening tool. There is a significant amount of method development and validation needed for each specific application; the behaviour of a specific analyte, or mixture of drugs is hard to predict without comprehensive testing, research and development. The optimal

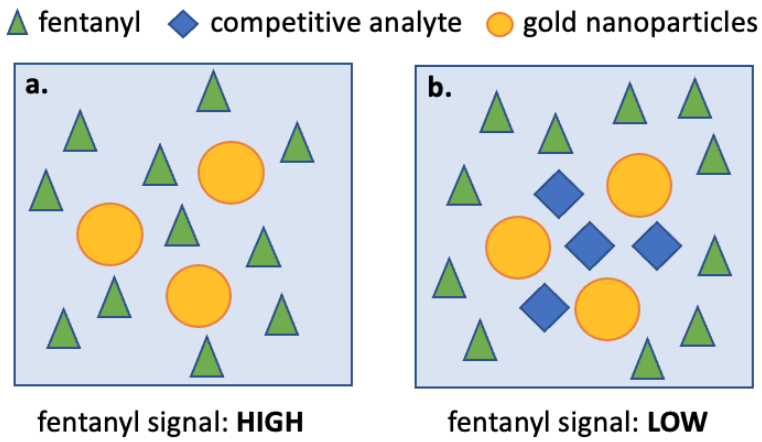

Figure 2.7: Illustration of the challenges with competitive adsorption in SERS. (a) and (b) represent two solutions with the same concentration of fentanyl, however (b) is in he presence of a competitive analyte with greater affinity to the gold nanoparticles than fentanyl. Despite having the same concentration of fentanyl in both solutions, the SERS spectrum may look significantly different. In the extreme case, fentanyl may give no signal in scenario (b).

choice of substrate, laser wavelength, aggregating agents and analyte concentration might differ for different target analytes. For example, the SERS signal for substances such as fentanyl have been shown to decrease when the concentration is too high.[108] Furthermore, as mixtures become more complicated the relationships between analyte concentration and signal intensity is less predictable. A case in point, as the relative concentration of fentanyl to etizolam increases, it becomes harder to detect etizolam.[25] Such behaviour is often attributed to competitive adsorption for the gold nanoparticles.[25,112] Figure 2.7 further illustrates this challenge, where the same concentration of fentanyl alone and in the presence of a competitive analyte (even a contaminant in trace amounts) might give dramatically different results. On the other hand, if a substance is in very low concentration relative to a bulk cutting agent, but has a significantly higher affinity for the gold nanoparticles, it could still produce a strong signal without being drowned out by the major components. There is also interest in functionalizing nanoparticles to target specific molecules to achieve selective amplification in a mixture.[113]

SERS signal from an analyte typically follows some relationship with concentration,

however it can be fairly non-linear due to the challenges described above. Some strategies have been employed for more reliable quantification with SERS, such as using internal standards or standard addition methods.[114] Again, these methods require significant method development and calibration for particular molecules of interest.

Another advantage of SERS in drug checking is that it potentially reduces the fluorescence, a major challenge in testing coloured drug samples with traditional Raman spectroscopy.[91] While SERS is not immune to fluorescence, the extremely low concentrations of analyte as well as quenching from the metal SERS substrate itself within the hot spots may result in reduced fluorescence.[91,109]

Raman spectroscopy, in combination with SERS, is uniquely positioned to provide spectral data for the identification of high concentration bulking agents, as well as actives and trace components using a single instrument.

SERS data acquisition. We use the same instrument for surface-enhanced Raman, as normal Raman, but with the sample region fitted with an accessory that enables a glass vial to be inserted and held in the focus of the laser. A small amount of substance (approximately 1 mg or less) is dissolved in 1.5 mL of a 50 nm nanoparticle gold solution (BBI solutions, Crumlin, UK) and shaken for 30 s. If the sample does not appear visibly dissolved, the vial may be briefly heated in a warm water bath. 10 μL of a concentrated (1.0 M) $MgSO_4$ solution is then added and shaken for an additional 10 s before acquisition.

2.2.5 Portable Gas Chromatography–Mass Spectrometry

A powerful separation and analysis method employs a combination of gas chromatography (GC) and mass spectrometry (MS), aptly named GC–MS.[115,116] First, the GC aims to separate the drug mixture. If successful, each component including active ingredients and cutting agents may be isolated and analyzed via MS. MS is considered to be a primary characterization technique[117,118] that can identify molecules based on their mass and fragmentation pattern. This means of analysis therefore has an advantage over the

previous spectroscopic techniques which all rely on detecting the characteristic pattern of molecular vibrations in a sea of signals associated with every molecule in a mixture. GC–MS is well known for being used as a confirmatory testing technique.[18,21,100] In many cases, these are experienced analytical laboratories and technicians working in collaboration with drug checking services to help validate point-of-care results from less sensitive techniques like IR or Raman spectroscopy. Portable GC–MS has been evaluated in forensic applications,[119,120] and in drug checking for the detection of trace substances,[121] noting the additional complexities such as maintenance and method development required for such method.

Basic principles. The general principle of gas chromatography[122] (GC) is illustrated in Figure 2.8a. The sample is first dissolved in a volatile solvent such as methanol, and then extracted into a fibre probe from either the vapor headspace above the liquid, or by immersing the probe directly into the solution.[123] The fibre is then inserted into a heated injection port, with the temperature set high enough to vaporize the components as they desorb from the fibre and enter the instrument. A carrier gas, typically helium, carries the mixture of drug molecules through a long tube (called a column) that is coiled to be more compact and fit into an oven.[124] The constituent molecules initially travel together, but increasing separation occurs as different molecules experience different degrees of attraction to the inside of the column. All molecules spend some of their time sticking to the inner walls of the column, and some of their time detached from the column. When detached, they are pushed further along the column by the helium carrier gas. Eventually, the molecules that spend the least time attached to the walls emerge at the end of the column first, while those that experience the greatest attraction come out last.[125] A plot of the output count against time (referred to as a chromatogram, Figure 2.8a) may be used to identify species based on the characteristic time at which they emerge, termed the retention time. This information is compiled using known standards, and observing their travel through the column. The temperature of the oven that surrounds the column determines

the retention time of each species. Finer control of the separation (to increase speed and enhance separation of components with otherwise similar retention times) is achieved by ramping the temperature during the separation, a technique known as temperature programming (Figure 2.8a).

In mass spectrometry a sample mixture is imparted with a charge in a process known as ionization. The molecular structure may be preserved in this process (soft ionization) or the molecule may be broken into fragments.[126] Molecular ions or fragments then enter a mass analyzer and detector that can resolve the mass-to-charge ratio characteristic of a particular molecule as shown in Figure 2.8b. In the case of fragments, the particular fragmentation pattern may be pieced together to identify the molecule from which it originated.[127] In theory, no libraries are necessary for molecular identification; the fragments may be related directly to the chemical structure. The use of databases, however, greatly accelerates this identification. MS is a trace analysis method, and so can detect low concentrations of components in a mixture. Nevertheless, analysis of a trace compound in a complex matrix using MS alone is often challenging particularly where highly sensitive detection is desired in a complex sample matrix,[122] for much the same reason as mixture analysis complicates the interpretation of IR and Raman data. As shown in Figure 2.8c, the output of the GC column may be sent directly into a mass spectrometer to achieve the combined GC–MS.

Practical considerations. Portable GC–MS is useful as a point-of-care drug checking instrument as it offers trace-level identification. However, it has a relatively high instrumental and operational complexity when compared to the other instruments discussed so far. It is not uncommon to be able to detect only one compound on IR and then see 10 or more peaks on GC–MS. This is largely related to the detection method as well as the improved resolution from having chromatographic separation. High temperatures can also result in degradation products which could explain unexpected peaks on a chromatogram—it might be unclear whether it is truly a breakdown product or from the drug mixture. The means that one can inject a pure compound into the GC–MS, and several defined peaks

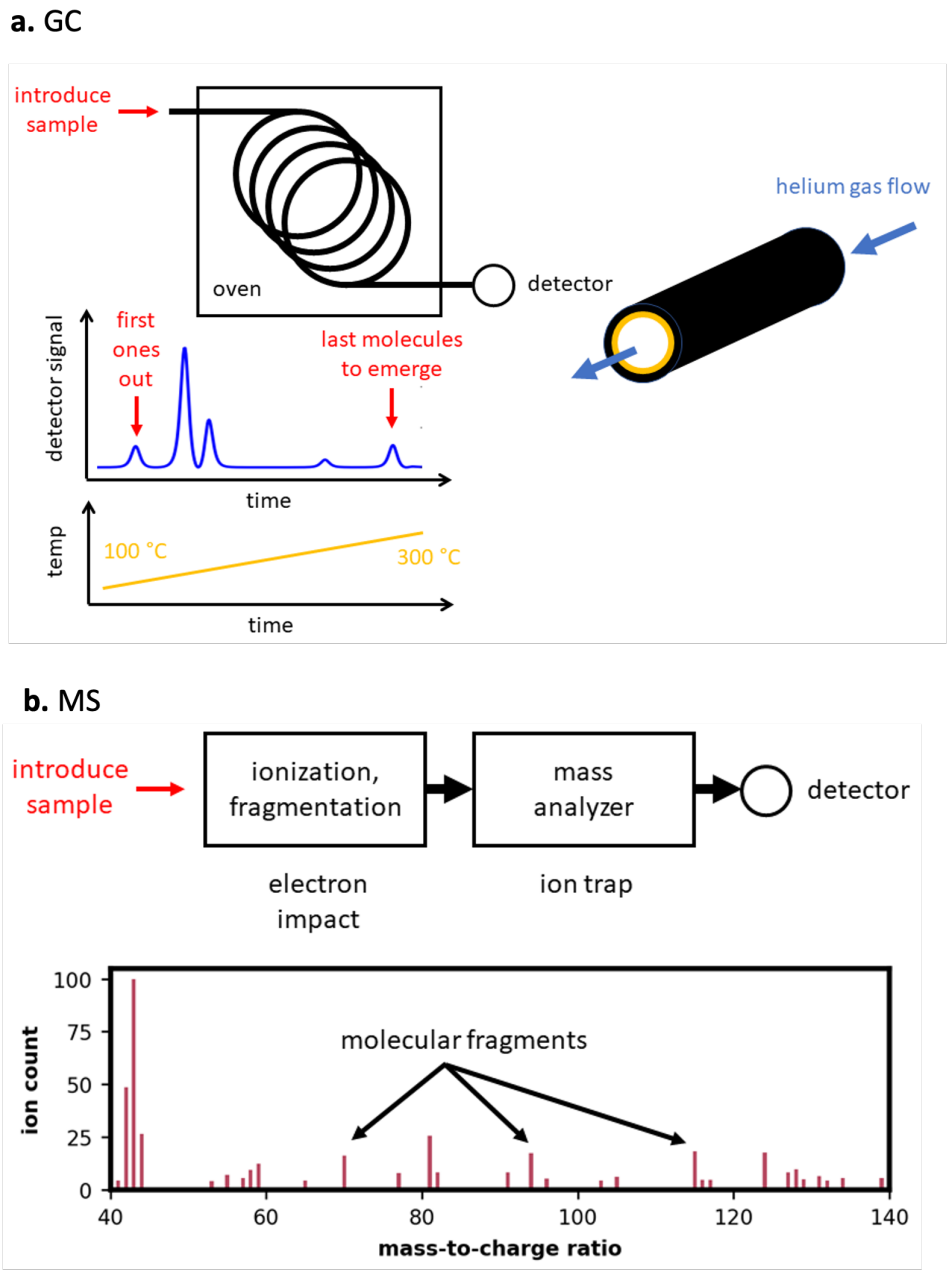

Figure 2.8: (a) In a gas chromatograph (GC), a small quantity of the sample mixture in injected into the end of a coiled tube inside an oven. Components in the mixture travel through the tube at different speeds, thereby achieving a separation of molecules for easier identification. (b) In a mass spectrometer (MS), molecules are ionized, fragmented, and then detected according to their mass and charge. A GC–MS combines gas chromatography and mass spectrometry technologies into a single instrument, with the GC output flowing directly into the MS for analysis.

are identified on the chromatogram. In addition, with such sensitive techniques, cross-contamination is a potential risk even in trace amounts.

On the other hand, occasionally some compounds do not show up on GC–MS even if they are present in high enough concentrations, For instance, some common cutting agents such as sugars have high boiling points and therefore are not volatilized to a large extent, and do not enter the column in appreciable quantities.[122] Some drug molecules are not suitable for identification with GC–MS given that they do not ionize, have limited solubility in the selected solvent, or ultimately do not have a unique retention time or mass spectrum for subsequent identification. Too high of a concentration injected into the instrument or improper temperature programming can also be detrimental resulting in significant tailing, overlapped peaks, and saturation effects in the mass analyzer and detector.[128] For drug samples, this requires a balance between preparing a high enough concentration to detect potent, trace drugs such as carfentanil, but not too high to have significant such unfavourable effects from the high amount of cutting agents such as caffeine. With careful method development such as adjusting the concentration, internal standards, gas flow rate, and temperature programming can result in reliable and sensitive drug checking.

Quantification with GC–MS depends on the relationship between the peak area of a specific analyte and the concentration. Within a certain range of concentrations this relationship is typically linear and therefore simple to model. However, parameters outlined above, such as gas flow rate, temperature programming, tailing, injection errors, changes in the mass spectrometer and column bleed will all affect this signal.[122] Common to many analytical instruments, quantification with GC–MS often uses external standards to create a calibration curve to model this relationship while accounting for some of the instrumental influences on signal through replicate measurements.[129] This model can be used to quantify future unknown samples. However, with GC–MS often the column, detector, and other instrumental factors will also change over time and using an absolute signal intensity can be misleading without re-calibration.[130] Spiking a solution with an internal standard with

a fixed and known concentration helps further account for such run-to-run variability.[130] It is important that the internal standard behaves as similar as possible to the drug molecule you are trying to quantify. Other sophisticated methods exist, such as isotope dilutions[131] and the standard addition methods (previously mentioned as methods to aid in SERS quantification).[122] In either case, accurate and precise results depend on significant method development and validation for each target molecule, as well as accurate weighing and dilution of the drug sample.

In theory, the GC and the MS components could each be employed alone as a method for drug checking. However, GC–MS offers a two-fold advantage over either method employed individually (GC or MS). First, molecules may be unambiguously identified on the basis of both their retention time and fragmentation pattern: any molecules with similar retention times will likely have different mass spectra. Second, the MS analysis is no longer of a mixture, but of a single component from that mixture, greatly simplifying the identification and quantification effort and improving its reliability.

GC–MS data acqusition. This work employs the Torion T-9 portable GC–MS (Perkin Elmer, Utah), equipped with a low thermal mass capillary gas chromatograph, in-trap electron impact ionization source, miniaturized toroidal ion trap mass analyzer, and an electron multiplier detector with a mass range of 41–500 Daltons. Approximately 1 mg of a drug sample was dissolved in 150 μL of methanol and centrifuged prior to sampling. A coiled microextraction (CME) syringe was immersed into the sample solution for approximately 10 s and allowed to dry for 1–2 min prior to injection into the GC–MS. Resulting chromatograms were analyzed by selecting or integrating the mass spectra of eluting peaks and comparing to MS libraries from SWGDRUG and NIST databases. In cases where in-house analytical references exist, targeted analysis of compounds may be achieved using a reconstructed ion chromatograph (RIC) with one or more ion fragments of interest. Any resolved peaks are then compared to the retention time of analytical standards, and their corresponding mass spectra.

2.2.6 Paper Spray Mass Spectrometry

This dissertation focuses on portable instruments being explored for community drug checking. However, confirmatory lab-based testing continues to be an integral part of an effort to provide accurate, sensitive, and quantitative results.[132] Our service utilizes paper spray mass spectrometry to fulfill this objective.[50,133] Mass spectrometry data also furthers research goals to enhance the capabilities of portable instruments as it enables us to benchmark their performance and pursue further method development.[25,121] This includes machine learning models to help automate and improve spectral analysis, target lists for GC–MS library screening, and developing strategies for communicating known limitations of individual techniques.

All PS-MS analyses were performed using a TSQ Fortis^{TM} triple quadrupole mass spectrometer and a VeriSpray^{TM} PaperSpray ion source (Thermo Fisher Scientific, San Jose CA, USA).[134] PS-MS was operated using tandem mass spectrometry and identities were further confirmed by comparison of ion ratios to certified reference standards. Details of the instrument operation, including calibration and data analysis are described in previous publications.[34,50,133,134] The limit of detection (LOD) and lower limit of quantification (LLOQ) for the analytes of interest are shown in Table 2.1. For the sample preparation, 1.3 mg of the sample (weighed with an analytical balance to enable quantitative analysis) is dissolved and vortexed in 1.3 mL methanol, from which 1 μL (volume adjusted, if necessary) is added to deuterated standards in 200 μL of methanol. From this solution, 10 μL is spotted onto the VeriSpray sample plate for MS analysis.

2.3 Chemometric Analysis

2.3.1 History and Potential of Chemometrics in Spectroscopy and Drug Detection

As introduced above, most community-based projects employing the instrumentation mostly rely on commercial software and libraries for the identification and semi-quanti-

Table 2.1: Limit of detection (LOD) and lower limit of quantification (LLOQ) for target compounds for PS-MS analysis. These values are calculated as follows: LOD = 3.3× standard deviation of the lowest calibrator, divided by the slope of the calibration curve. LLOQ = 10× standard deviation of the lowest calibrator, divided by the slope of the calibration curve.[133]

Compound	LOD (% by weight)	LLOQ (% by weight)
Caffeine	0.414	1.255
Fentanyl	0.025	0.076
Etizolam	0.039	0.118
ANPP	0.020	0.062
Carfentanil	0.018	0.056
Heroin	0.038	0.115
Cocaine	0.014	0.043

fication of drug composition, relying on experienced technicians or chemists.[42] Giskeodegard et al. argue that the human eye and brain are still unsurpassed as pattern-recognition tools and any experienced drug checker might agree that they can recognize spectral features of fentanyl better than any automated software could.[135] However, this also is a very manual and subjective process and significantly limits drug-checking in its opportunity for expansion, and reliability, consistency, and efficiency of results; all aspects which have proved vital for people who use drugs and the success of drug checking projects.[16]

Chemometrics is a general term used to describe the statistical and mathematical manipulation of chemical data.[136] Chemometrics is particularly useful for large datasets, such as those produced from drug checking, and can significantly improve the accuracy and reproducibility of spectral interpretation when compared to a manual or visual analysis. While presentation of chemometric approaches applied to drug data are prevalent in the literature, it is unclear how and if many methods make their way into real life drug checking applications. Chemometric methods are not easy to implement and have limited transferability between different instruments, which is a large barrier and undertaking for every community drug checking service. It is not surprising that small harm reduction agencies would have limited resources for hiring spectroscopists, computer scientists, or software engineers dedicated to the development and implementation of chemometric

models in a drug checking service. Therefore, this task is often left to instrument or data analysis software companies where interest in the nuances specific to drug checking needs may be limited. This following section is intended to provide a tutorial and framework to implement and familiarize the drug checking community with several chemometric methods using relevant examples. Numerous books and reviews already cover the specifics of chemometric techniques in great detail. This account aims to refer readers to those texts when necessary while directly relating the concepts well explored in other applications to common questions and challenges that arise in community drug checking. Some of the questions to ask throughout might include:

1. Does the model need to distinguish between two similar drugs? What kind of drugs are currently challenging to distinguish?

2. Are there low concentration actives in drug mixtures?

3. What methods are appropriate to predict how much of a drug there is in a mixture? Would approximate concentration be adequate?

4. Are analytical standards required to make these models? Is there spectral data with confirmatory testing that can be used as a "known" value? Is there access to a drug spectral library?

5. What uncertainties are inherent in the drug supply that need to be anticipated?

Drug detection is related to many analytical fields. Unsurprisingly, the goals of analysis in harm reduction-based drug checking overlap with that of forensic analysis or drug testing by law enforcement, and to date majority of the literature on the chemometric analysis of drugs is within that context.[118,136,137] However, many chemometric principles can also be learned from environmental,[138] clinical,[139] pharmaceutical,[140] and food analysis[141–143] and applied to drug checking. Chemometrics is broadly defined as the use of mathematical and statistical operations in the field of sciences where the output

of analytical methods are complex. In general, it uses pattern recognition methods to connect complex chemical data to valuable information to answer some sort of analytical question by extracting relevant signals and correlations within the dataset. In the case of drug checking, the analytical question is typically either, (1) "What is the substance?" and (2) "How much substance is there?". However, the question, "What is the substance?" can surprisingly be asked in a number of different ways such as, "What category does this substance fall into?" (discrimination/classification) or "Does this substance look different than what is expected?" (adulteration/counterfeiting).[136] A breakdown of various fields of chemometrics that can be applied in drug checking is illustrated in Figure 2.9.

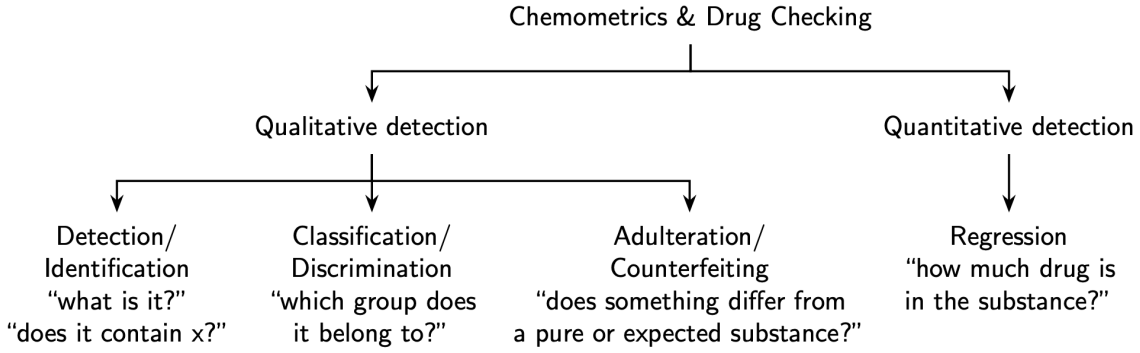

Figure 2.9: Breakdown of chemometric methods and their application to drug checking-related questions.

Prior to discussing details of particular algorithms this section will first introduce some examples in literature to inspire the scope of what can be done with spectroscopy and chemometrics. In an ideal case, the goal is to identify all the substances in a mixture (as opposed to just one particular substance of interest) such as all psychoactive components as well as relatively inert cutting agents that inform use as best as possible. However, particularly with spectroscopy-based approaches which tend to suffer from both lower sensitivity and significantly overlapping features, this is not a straightforward task. Locating approaches to solving this problem among other applications, from corn kernel discrimination[144] to distinguishing cheeses from different manufacturers with Raman,[145] can also pose a barrier.

Several works focus on automating the classification of different drug classes. Some early work uses simulated street drug mixtures of benzocaine, isoxsuprine and norephedrine with various rations of cutting agents to build a discriminatory model and proof of concept for drug detection with Raman spectroscopy.[146] The authors cite the use of drug surrogates being used due to regulatory constraints, which continues to be a major limitation for pursuing timely work within harm reduction agencies. Some research has also been done with purchased drug standards and lab-made mixtures such as for the identification fentanyl analogues in drugs using surface-enhanced Raman scattering (Figure 2.10),[49] and classification of heroin, methamphetamine, ketamine and their additives with ATR-IR (Figure 2.11).[137] Another approach to building pattern recognition models uses

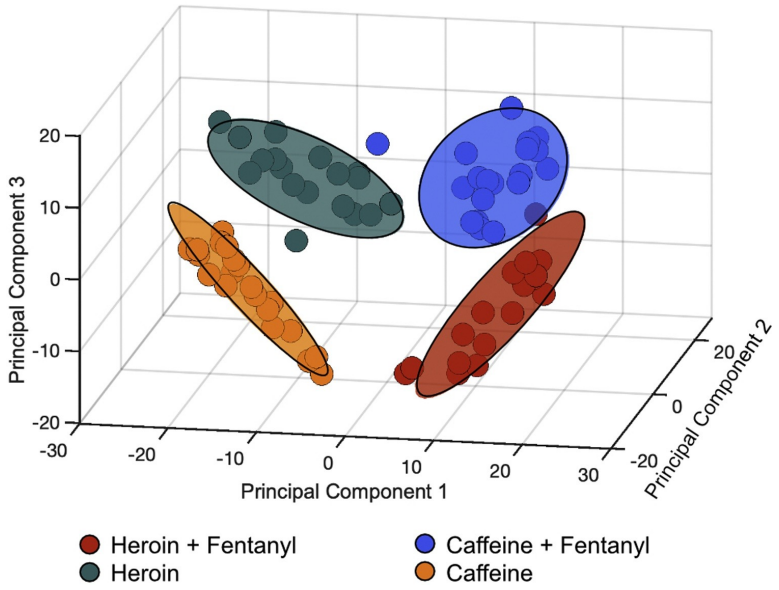

Figure 2.10: Principal component analysis on SERS data for three fentanyl analogue drug standards. Reprinted with permission from ref 49. Copyright 2021 Elsevier.

street samples with associated "confirmatory testing" performed with highly validated and sensitive techniques such as GC–MS, LC–MS, etc, as opposed to simulated standards, to both train and test the models. This approach has been adopted for the distinction between amphetamine, cocaine, ketamine in real street samples first characterized by GC–FID and GC–MS,[147] on-scene detection of cocaine in street samples using a handheld near-

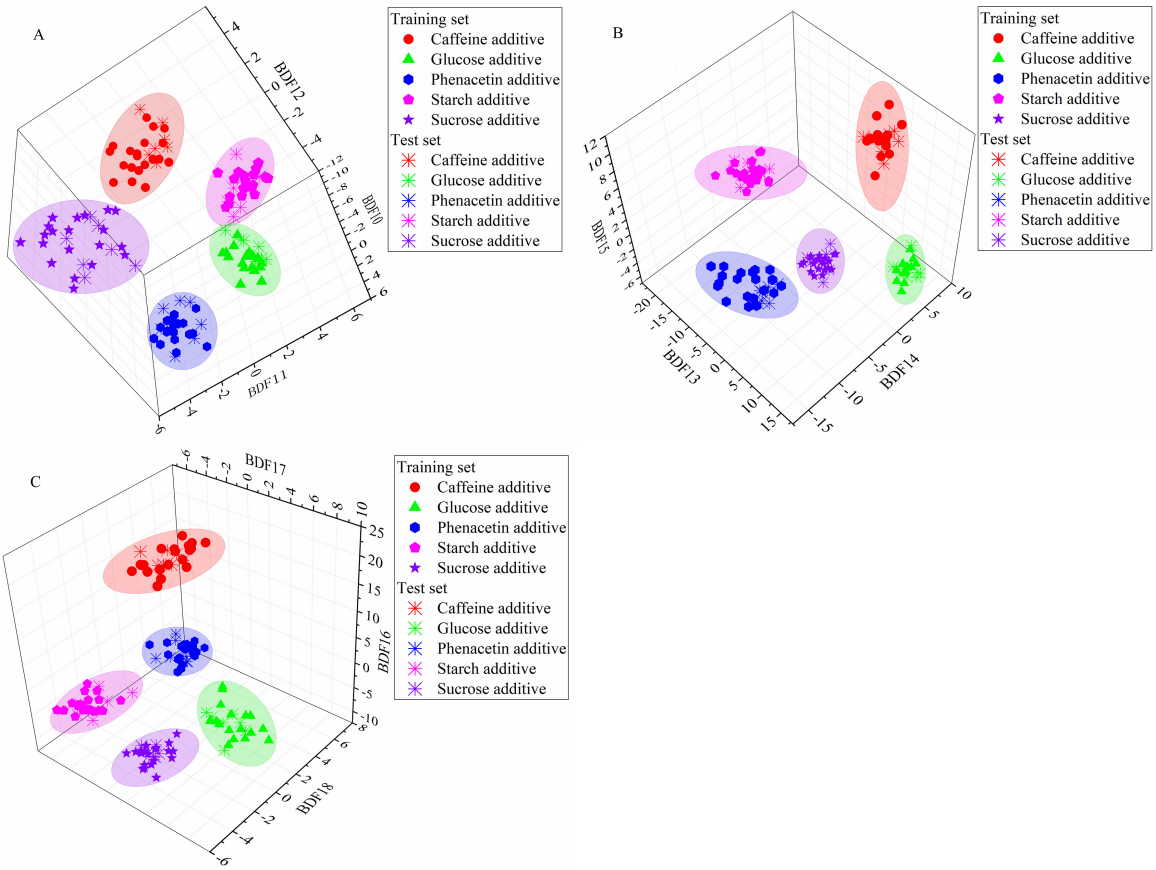

Figure 2.11: Bayes discriminant function analysis for A) heroin B) methamphetamine C) ketamine with 5 additives. Reprinted with permission from ref 137. Copyright 2020 Elsevier.

infrared spectrometer and machine learning algorithms,[148] and discrimination of synthetic cannabinoids in herbal matrices and of cathinone derivatives by Raman spectroscopy.[149] While these approaches have some uncertainty and challenges (for example, error in detecting components, inhomogeneity within samples) they also provide benefits by only including realistic compositions of samples that would truly be seen in the application. It is worth noting that in all these cases the drug subset has been created from drug seizures, and may not represent the scope of what is desired in drug checking applications. As previously mentioned, chemometrics applied in other applications such as pharmaceuticals, environmental, or food analysis shares many of the same challenges as drug checking. Many of these examples include discriminating two or more spectrally similar substances or identifying a substance or adulterant among a very complex matrices—challenges

that are shared with drug checking due to the unregulated, and complex drug supply. Most obviously, several pharmaceutical examples use chemometrics to identify counterfeit antibiotics given their ATR–IR spectrum despite the large variation that exists for excipients (filler) between manufacturers.[150] Significant work has been done in this field of detecting counterfeit pharmaceuticals with both Raman and IR spectroscopy.[151,152] Similarly, the study of environmental contaminants has shown the use of IR spectra for an automated identification and mapping of microplastics in the Mediterranean Sea.[153] The authors note that their dataset must be gradually enriched with new spectra such as PVC, PET, paints,[153] similar to how drug checking datasets and detection algorithms must adapt to what is prevalent in the drug supply.

The other major class of chemometric techniques is concerned with the quantification of substances in mixtures. Similar to the works on classification and identification, the area of applications using spectroscopy for quantitative information are broad. Several works build predictive quantitative models for drug compounds in powder mixtures using popular partial least squares regression (PLS-R), such as for cocaine using Raman spectroscopy,[154] fentanyl using IR spectroscopy,[20] and pharmaceuticals like paracetamol.[155] Others have successfully detected and quantified trace levels of fentanyl in binary mixtures with cocaine and heroin using surface-enhanced Raman spectroscopy by exploring multivariate approaches, PCA and super partial least square discriminant analysis (sPLS-DA), alongside simple univariate models.[107] In food analysis applications, Wan et al. demonstrated a comprehensive workflow including both qualitative and quantitative analysis steps.[156] Here they compared the performance of several discriminatory methods random forest, *k* nearest neighbour, linear discriminant analysis, and support vector machine, prior to using PLS-R for quantifying any combination of five artificial sweeteners.[156] The models were validated against high performance liquid chromatography (HPLC).

The following section provide a brief overview of some popular chemometric methods. There are numerous books, videos, and resources that breakdown the mathematical

principals behind these methods in great depth from experts in these fields.[157] This section focuses on a practical introduction to these methods so their application and implementation better understood in drug checking. The same methods are then applied in the analysis of spectroscopic data in Sections 2.4 and 2.5.

2.3.2 Qualitative Analysis

2.3.2.1 Unsupervised Pattern Recognition

PCA. Principal component analysis (PCA) is the most popular multivariate technique used for exploratory data analysis and is seen throughout the literature across many fields, from psychology to chemistry.[158, 159] It is primarily used for variable reduction and simplification, which is also often referred to as dimensionality reduction, and facilitates visualization of trends, detection of outliers in data, and spectral unmixing.[160] It is often used to explore relationships within a dataset since it does not require any target or known values. PCA is useful for spectroscopic drug data because spectra are multivariate; meaning that for each sample measurements are recorded at many (sometimes thousands) wavelengths. This is in contrast to a univariate data set which would simply utilize one wavelength or measurement, i.e. measurements like pH only record a single value. In this case, statistical analysis is relatively straightforward. With multivariate data, several challenges arise: (1) It is impossible to efficiently visualize and compare thousands of data points for several samples at the same time; (2) Many measurements are correlated and offer redundant information that can confuse algorithms; and (3) if each sample represents thousands of data points the dataset rapidly becomes very large, complex, and consequently, slow. This last point is often referred to as the "*curse of dimensionality*". In the work throughout this dissertation, PCA is frequently used for initial visualizations and exploratory analysis of data, as well as a dimensionality reduction/pre-processing approach when building classifiers.

The basic mathematical equation behind PCA is

$$D = SL^T \tag{2.1}$$

where D is a data matrix of dimensions $[n,m]$ where n is the number of spectra and m the number of features/wavelengths. S is a matrix of scores of $[n,f]$ where f is the number of latent factors. f must be less than the smallest dimension of D. L is a matrix of weightings of $[f,m]$, also referred to as loadings or eigenvectors.[160,161] In the simplest sense, using this equation linear algebra allows for the data matrix, D, to be decomposed into spectral quantity information (scores, S) and spectral feature information (loadings, L). The loadings can be interpreted as common building blocks for the spectral data set and are grouped by variables that covary together.[161] This transformation means that spectral variation in a multivariate data set (D) can now be described using a few principle component (PC) scores, which significantly reduces the dimensionality of the data yet minimizes information loss. Importantly, the PCs are also ordered by their contribution to data set variance in terms of magnitude and prevalence.[161] For example, most of the variance will be captured in the first few PCs and in practical applications, the hope is that this relates to an experimental/chemical phenomena of interest. Higher PCs may represent small variations such as random noise and are often ignored. Interpreting a PCA model usually involves various graphical representations. PC scores (e.g. PC score 1 vs PC score 2) for each sample are often plotted to explore groupings in the data. In addition, investigating the PC loading vectors reveals spectral relevant spectral features for each score which aids in interpreting meaning behind any observed partitioning or trends.

Cluster analysis. In contrast to PCA, where a main aim is to simplify and explore data, rather than make physical predictions, cluster analysis aims to detect similarities and group those samples based on measurements of similarity such as a correlation coefficient, Euclidean distance, or Manhattan distance.[157] In many cases, due to the high dimensionality of spectral data, it is advantageous to first perform PCA first, and then subsequently apply cluster analysis methods to the transformed data. Examples of

clustering include hierarchical and *k*-means clustering. *k*-means clustering, for example, works through several steps which aim to minimize the variances within clusters of samples. In the conventional algorithm, this occurs through three steps:

1. randomly assign a centroid for each cluster (this requires the number of clusters to be estimated)
2. assign all points to the closest centroid
3. calculate the mean of the group and assign it as the new centroid
4. repeat step 2 and 3 until it converges (the centroid does not change and the distances are therefore minimized)

2.3.2.2 Supervised Pattern Recognition

Supervised pattern recognition methods or classification methods seek to produce a mathematical model between a spectrum on a series of samples and their known groups or identity. In the context of drug checking, this translates to a question such as,"how can a Raman spectrum be related to the presence of fentanyl in a mixture?" and samples are assigned to classes, either "yes this contains fentanyl" or "no this does not contain fentanyl." When developing classification methods, there is a significant emphasis on how well or how accurately the model predicts the correct class. A popular sentiment in statistics is, "all models are wrong...but some are useful". It is nearly impossible to build a model that perfectly accommodates all of the complexities inherent to real life problems. In drug checking, complexities originate from instrumentation limitations or noise, sampling errors, and significantly, an unregulated drug supply. However, a particular degree of error might be accepted and the model still considered useful. Popular classification methods include partial least square discriminant analysis (PLS-DA), random forest (RF) classification, and support vector machines.

PLS-DA and RF classification are the main classifiers used in this work. PLS-DA is based on many of the same principles as PCA.[162] Here, dimension reduction is performed on a spectral data set, followed by predictive modelling for a categorical variable.[163] Notably, instead of the linear transformations reflecting the maximum covariance in the spectral data alone (as in PCA), PLS-DA takes into account the labels during this calculation. PLS-DA components are ranked by the maximum covariance of both the spectral (input data) and class labels (output data).[163,164] In this way, it is a "supervised" version of PCA. During training steps for PLS-DA, categorical variables are translated into numerical data, such as 0 (category 1) or 1 (category 2), and vice versa for the predicted outputs. Thresholds are established for optimally placing the numerical predictions into one of the two classes.

RF is an ensemble model, meaning that it prepares several models and consolidates their outcomes through voting schemes, for instance majority voting.[165] In RF modeling, a large number of simple decision trees are independently built using a random subset of spectral features in the data set. Many advantages of this approach are further discussed in Chapter 8.

2.3.3 Quantitative Analysis

Calibration, or quantitative analysis, is the other major class of predictive algorithms in the field of chemometrics. Calibration refers to the determination of a model between the measurement of several variables (i.e. spectra) and one or more physical parameters (i.e. concentration). While calibration can be used for a whole array of problems, in drug checking it is almost entirely used to solve problems surrounding quantification of a substance within a mixture. For example, if cocaine and phenacetin mixtures are common in the drug supply, a set of binary mixtures might be created using pure standards in the laboratory to investigate the reliability of the relationship between the IR spectra and concentration of cocaine. An even more challenging problem might be quantifying

fentanyl within complex mixtures where most of the other compounds in the matrix are vast, relatively unknown, and it might be difficult (if not impossible) to characterize them all. In this case, it is reasonable to pursue a calibration set from actual drug samples using techniques that provide independent estimates of concentration (i.e. HPLC, GC–MS).

Partial least squares regression (PLS-R) is by far one of the most implemented calibration techniques however other techniques such as random forest regression and artificial neural networks are also becoming popular. PLS-R is applied in Chapter 3 for the quantification of fentanyl using Raman spectroscopy. PLS-R again uses the same basic principles as PLS-DA (and, consequently, PCA). PLS-DA is actually an extension of PLS-R. PLS-R resolves the maximum spectral variances relevant to a continuous variable (e.g. concentration of a chemical species). In PLS-R the quantitative output itself is the target of interest.

2.3.4 Outlier Detection

One topic that has not been mentioned so far is how to handle outliers, which are samples (and their data) that differ significantly from the group or majority. Outliers are often described as, "...the bane of much experimental science"[157] and that holds true in drug checking. Outliers can occur for reasons as simple as mislabelling a sample, or containing an unexpected substance. The inclusion of outliers has consequences if they are present in training models. They also play a role in the application of models to real samples, referred to as anomaly detection. For example, if we have a model that determines whether a substance has MDMA or MDA, it would be inappropriate to apply the model to a spectrum representing cocaine. In some classification and calibration models, a sample will be forced into a class or a predicted concentration without inherent consideration if it falls in none of the categories. This would certainly give misleading results. Some algorithms are more affected by the presence of outliers than others and some algorithms are even inherently designed to flag them as such. The mention of outliers is surprisingly absent in most

articles despite it being an important consideration in models developed with the intention to be applied in real life applications, such as drug checking or quality control applications. This is likely because the implementation can be challenging and seemingly arbitrary when setting thresholds for outlier detection. The exclusion of too many samples will limit the scope of the application, however the inclusion of too many samples might contribute to increased error. This is realized in Chapter 3 where outlier detection methods are explored to make decisions to include or exclude unknown samples from prediction.

2.3.5 Pre-Processing

Another important topic that is related to chemometrics is pre-processing of data prior to input into models. In general, changes in the spectra should exclusively reflect the property of interest (e.g. concentration or class label) and the influence of instrumental noise, fluorescence background in the case of Raman, cosmic artifacts, should be minimal. Pre-processing techniques typically include a combination of background correction, filtering/smoothing, scaling/normalizing, and dimension reduction (as previously discussed with PCA).[166,167] One of three strategies are typically employed to determine which pre-processing combination is the best. This includes,

1. trial and error based on best performance according to one or more metrics;
2. visual inspection to minimize artifacts, or;
3. quality parameters, such as a measurement of "simplicity", which aim to quantify the presence of artifacts in data; based on the premise that transformations which make spectra more similar should inherently contain fewer artifacts.[167]

Most drug detection and related papers make use of the first strategy, though choice of appropriate metrics continues to be debated. In many cases, several different combinations of pre-processing give statistically insignificant differences in performance. A general rule of thumb in these cases is the simpler model, the better. Ultimately, ongoing

validation ensures the continued usefulness of the model as well as appropriateness of data pre-processing and other parameters. Optimization of spectral pre-processing plays a significant role in model development in Chapters 3 and 8.

2.4 Qualitative Pattern Recognition Example Using a Raman Spectral Dataset

Many of the concepts introduced in the previous section are revisited through several illustrated examples using spectral data acquired at the drug checking service. The workflow is implemented with open source Python code in a JupyterLab environment and is representative of the approach taken in many of the works throughout this thesis.

First, it is important to understand what kind of data is needed to answer a particular question and how to structure that data. In the case of qualitative analysis (e.g. "What class does this drug belong to?" or "What is the substance?"), Figure 2.12 reveals how the data is structured with an **X** typically representing a list or array-like object holding a list of spectral intensities for each sample, and **y** representing a list of the same length with a class label for each sample. Raman spectra of MDMA and MDA samples are

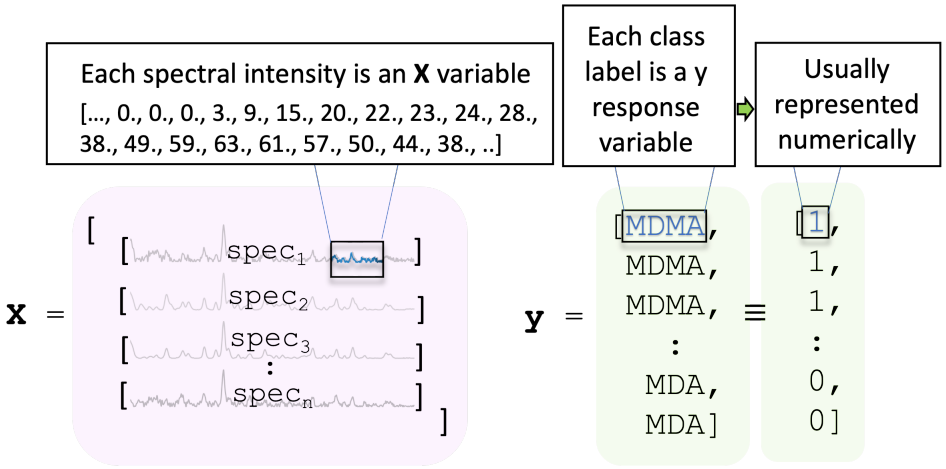

Figure 2.12: Common data structure for qualitative analysis.

used to demonstrate both exploratory and qualitative analysis in the following workflows. The first section "Data pre-processing" serves as the primary introduction to DataFrames

and plotting with `pandas` and `matplotlib` Python libraries. The following sections will primarily focus on implementation of models and workflows with `scikit-learn`.

2.4.1 Data Pre-Processing

The first step in the data analysis in all subsequent chapters is setting up and organizing the data. Within the Python language, there are multiple options for loading data from different file types (e.g. `.csv`, `.txt` files, SQL tables stored in a database, etc). `Pandas` 'DataFrames' provide an intuitive way to load, organize, filter, and manipulate tabular data. They also facilitate data transformations, troubleshooting errors in code, and rapid visualizations from histograms to line plots.

Figure 2.13 demonstrates the workflow for data set up from an initial format (here Excel) of spectral data, the transformation into a `pandas` DataFrame, and various visualizations.

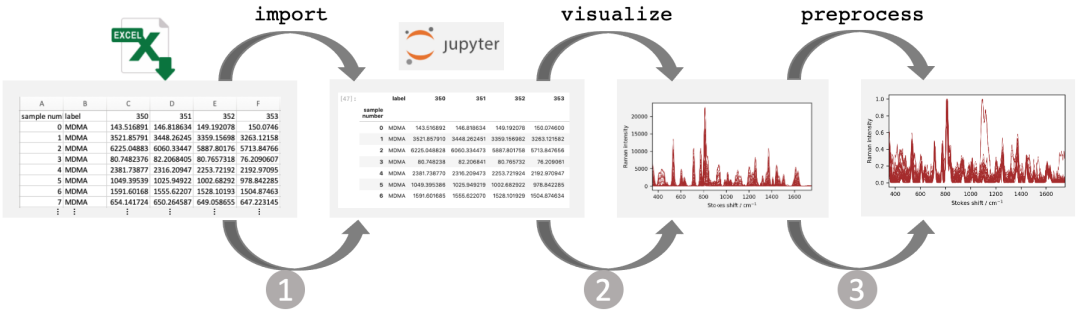

Figure 2.13: Data set up workflow including (1) importing from external file type into a DataFrame (2) visualizing with rapid plotting and (3) pre-processing the data for a different (and often better) look.

The resulting table contains information such as sample number, label, and then corresponding Raman intensity data, with each column the corresponding Stokes shift. Some manipulations of data are needed to set up the correct formatting. The code for some of these steps are shown below:

```
#step1 - import
import pandas as pd
data = pd.read_csv('classification_dataset.csv')
```

```
data['sample_number'] = data['sample_number'] + 1 # making sample
    number 1 - 50
data.set_index(['sample_number', 'label'], inplace = True) #
    setting index
data.columns = data.columns.astype(float) # column headings were
    interpretted as strings; change to float
data.columns.name = 'Stokes shift'
```

Spectral measurements are also often acquired with slightly varied parameters, such as number of data points or spectral range. These issues are illustrated in Figure 2.14. While not explicitly shown here, such differences need to be addressed prior to using the data for additional analysis. Truncation and interpolation of spectra are easily achieved through functions available in Python. When loading the data, a quick visualization ensures that

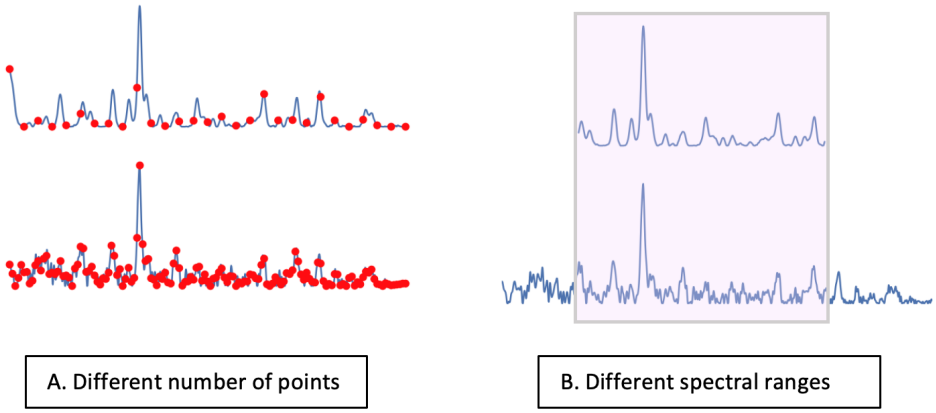

Figure 2.14: Common data problems to deal with when setting up comparable data such as (a) different number of points and (b) different spectral ranges.

(a) the dataset is correct (b) the data was imported in the expected layout and (c) a first pass at identifying data quality, outliers, and trends. `Pandas` has a built-in interface to plot data stored in a `DataFrame` with parameters such as figure size, axis labels, axis limits, linewidths (lw), and colour of the lines. The following code was used to generate the plots previously shown in Figure 2.13:

```
#step2 - visualize
ax = data.T.plot(figsize = (4.25, 2.5),
                 xlabel = 'Stokes shift / cm$^{-1}$',
                 ylabel = 'Raman intensity',
                 legend = False,
```

```
            xlim = (350,1750),
            lw = 0.5,
            c = 'brown')
```

To take the visualization a step further, importing `matplotlib` functionality with pandas allows for the creation of a figure, defining colours based on the label (MDMA or MDA) and saving the figure. In this case, the `.div()` function is used to scale all of the data by the max value of each spectrum (i.e. normalizing). Normalizing and other pre-processing is extremely important when both visualizing and analysing data.

```
#step3 - pre-process and saving
import pylab as plt

fig, axs = plt.subplots(figsize=(4.25, 2.5))
colours = ['royalblue' if i == 'MDMA' else 'tomato' for i in data.
    index.get_level_values('label')]
data.div(data.max(axis=1), axis=0).T.plot(
            xlabel = 'Stokes Shift / cm$^{-1}$',
            ylabel = 'Raman intensity',
            legend = False,
            xlim = (350,1750),
            lw = 0.5,
            color = colours,
            ax = axs)

fig.tight_layout()
fig.savefig('colour_norm.png', dpi = 200)
```

2.4.2 Exploratory Methods

Unsupervised pattern recognition is a popular approach for exploratory data analysis. In this section, exploratory analysis methods (1) PCA and (2) *k*-means clustering with the qualitative dataset are explored. These methods go beyond simple visual comparisons of plotted data. They aim to facilitate identifying trends within the spectral data (**X**) and do not take into account the class labels (**y**), when processing the data (Figure 2.15). On one hand, this is extremely useful when exploring a large amount of data without class labels. However, even in projects throughout this thesis where there is access to labelled data, unsupervised approaches are still often used as a powerful first-step because they can reveal natural trends in the data, without being biased by the suggested class labels.

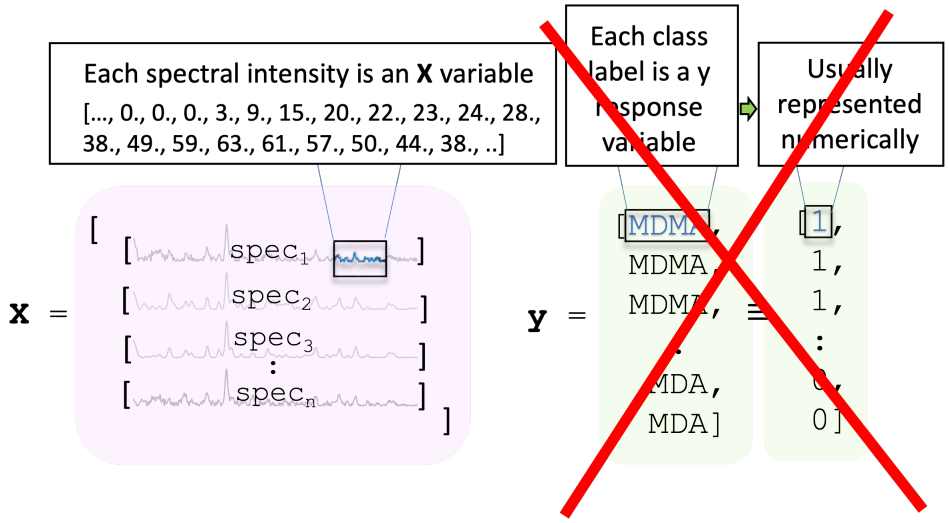

Figure 2.15: Unsupervised machine learning does not use the class labels.

Implementation of models and other transformations throughout this thesis mostly use `scikit-learn` (also abbreviated to `sklearn`), which is an open source Python library for implementation of machine learning algorithms and data processing pipelines. In some cases, custom functions are used and formatted to be compatible with the `sklearn` frameworks. `scikit-learn` functions lower the barrier-to-entry for those who are newer to chemometrics and/or coding; allowing for a practical understanding and execution of these methods before diving deep into the mathematical theory behind them. Most of the functionality offered by `sklearn` is organized with *transformer* objects and *estimator* objects. A transformer object has the primary purpose to pre-process or transform the data in some way. Transformers include methods to clean data through pre-processing such as scaling (e.g. normalizing), dimension reduction (e.g. PCA), and feature extraction. For each transformer the workflow is as follows:

1. define an instance of the algorithm
2. fit with the data
3. transform the data

Estimator objects often include some sort of transformation, however their main purpose is

to make a mathematical relationship between the data and some result (e.g. concentration, class, cluster). For each estimator the workflow is as follows:

1. define an instance of the algorithm
2. fit the algorithm with the data
3. predict new data

2.4.2.1 PCA for Visualizing Multivariate Data

As previously mentioned, PCA is used primarily as a dimension reduction technique. This allows for simpler visualization and subsequently, effective clustering. Here PCA is implemented as the first *transformer* object from `sklearn`:

```
from sklearn.decomposition import PCA

# note the three steps for a transformer object
pca = PCA(n_components=10) # 1. initialize with hyperparameters
pca.fit(data.values) # 2. fit with the data
PCS = pca.transform(data.values) # 3. transform the data
```

Figure 2.16 demonstrates this transformation from a spectral array to an array of PC component score on a dataset. These PC scores can now be used as a new representation of the spectral data. To help understand the transformation as a result of PCA, the new data (PCs) can be organized in a *pandas* DataFrame. This code is shown below:

```
#organize output into a dataframe
PCS_df = pd.DataFrame(PCS, index = pd.Index([*range(1, len(PCS)+1,
    1)], name = 'Sample number'), columns = pd.Index([*range(1,
   pca.n_components+1, 1)], name = 'PC #'))
```

Figure 2.17 then presents the display output of the first 10 rows of the Dataframe `PCS_df`. Notably, each spectrum prior to transformation had 1405 features and now majority of that information is capture within 10 features (or variables). Before visualizing the new transformed variables, it is worth introducing that pre-processing has a significant effect on the results of PCA.

The following code performs a simple normalization prior to transformation with PCA.

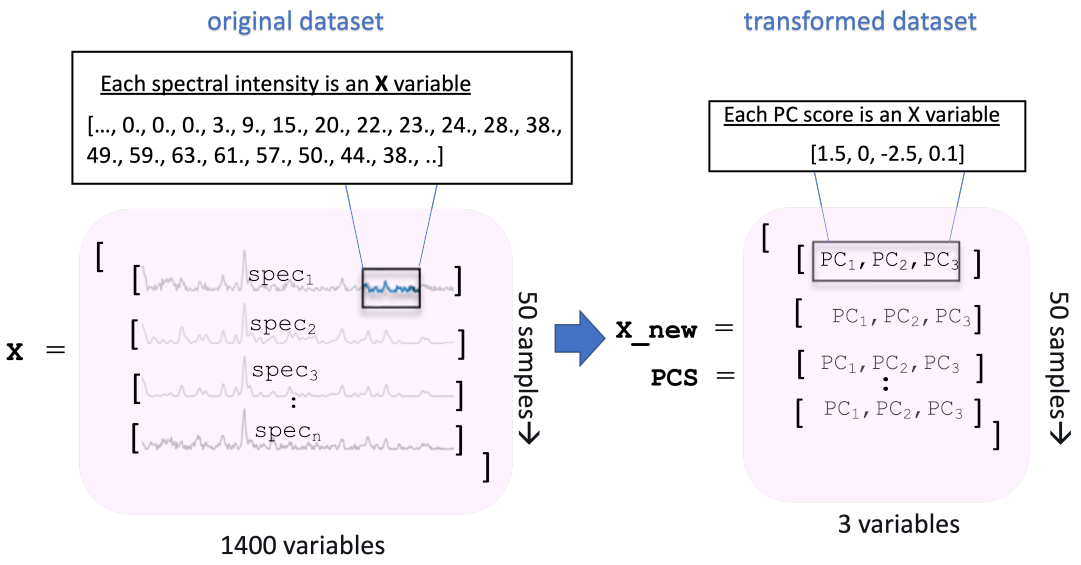

Figure 2.16: Transforming an array of spectra to an array of principal components.

PC # number	1	2	3	4	5	6	7	8	9	10
1	-36651.210073	-4945.377690	-467.231504	-406.740492	393.613415	-430.959308	235.868952	-174.078850	-54.365407	-31.474163
2	-642.230242	11681.676755	216.973412	11797.053175	-6255.440987	-1947.524338	303.922973	-58.421024	273.550958	-5.738136
3	17376.547493	27813.747430	3139.176233	-493.142375	402.462564	926.539383	781.723094	176.703383	-737.946922	-414.894278
4	-35744.646433	-4801.204466	-789.160436	-425.622461	164.156531	-301.275672	-132.731812	-118.063284	-149.417149	9.575290
5	-14943.147614	8685.536673	-1154.979704	-809.043917	-496.623639	-156.042215	-1593.536093	-1172.327486	298.283473	45.629155
6	-26203.311922	807.077492	-8.478805	-588.492751	364.221462	-85.279492	386.301940	-661.750554	321.621361	129.606488
7	-21246.183074	3882.318214	332.895847	-415.551536	701.707297	-89.075655	487.469081	-482.828399	254.707855	85.809172
8	-31560.481595	-2613.211915	295.515113	52.522333	-39.135351	-789.770901	909.945902	237.954957	-17.181581	240.366512
9	-27856.526729	539.346007	-832.522198	-430.046089	320.603651	159.675904	-911.078278	728.106703	-1121.403647	421.988301
10	-23993.573733	2528.178141	-380.831262	-825.771012	304.119926	-342.086685	-67.457544	453.387691	594.293640	350.426062

Figure 2.17: Display of dataframe of PC scores for the first ten samples.

```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer('max') # 1. initialize
X_normalized = normalizer.fit_transform(data.values) # 2. fit and
    3. transform data (note: step 2 & 3 can be combined with
    fit_transform)

pca_norm = PCA(n_components=10) # 1. initialize with
    hyperparameters
pca_norm.fit(X_normalized) # 2. fit with the data
PCS_norm = pca.transform(X_normalized) # 3. transform the data
```

A very common way to visualize the results of PCA analysis is by plotting various combinations of PC scores in a scatter plot–often shown in 2D or 3D–since it summarizes very high dimensional data (imagine trying to visualize and compare 1405 Stokes shifts

from 50 raw Raman spectra). Here only the first two PCs are plotted, since the dataset is fairly simple and expect the majority of the variance to be captured in the first few PCs. More complex problems, however, might require higher PCs to capture the features of interest. Figure 2.18 reveals PC 1 vs PC 2 scores for the data with no pre-processing and that which is performed on normalized data. The PC score plot on the raw data does show some sort of separation/trends in the data set, however it is much more scattered compared to the PC score plot with the pre-processed data. Recalling that the PC scores reflect some quantitative data (i.e. magnitude of expression of spectral features in the loadings vector), it is likely that this distribution is in part a result of fluctuations in overall spectrum intensity. Normalizing the spectra mitigates this effect. A simple normalization reveals distinct clustering and without any influence from class labels, visualization using PCA suggests that at least two or three different classes are expressed within this dataset. It is clear that pre-processing is essential for ease of interpretation of the data.

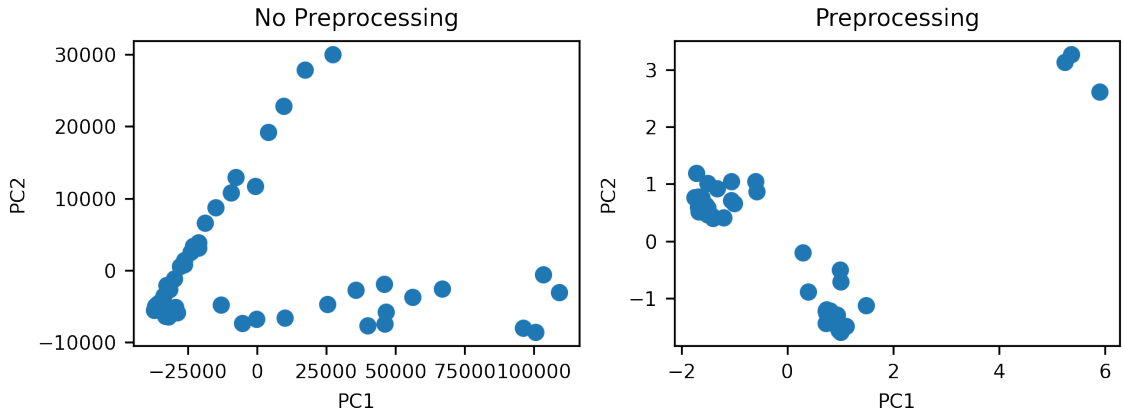

Figure 2.18: PC 1 vs PC 2 scores for the spectral dataset with no pre-processing (left) and with normalized data (right).

This here is a relatively simple problem and the success is not surprising for two unique, relatively uncut compounds (MDA and MDMA) despite their structural similarities. However, investigating spectral datasets with much more subtle differences, such as between complex multi-component opioid mixtures (see Chapter 5) reveals how PCA is a simple, yet powerful approach to investigate these types of drug checking problems.

2.4.2.2 Clustering Within the PC Space Using *k*-Means Clustering

So far in the example, the separation is visually obvious using the first two principal components. However, mathematically the cluster bounds or assignment can calculated. In clustering analysis, it is typical to use an unsupervised clustering method in combination with PCA, though they can be used alone for lower dimensional data. Again, there are a number of algorithms that use some sort of metric (distance, correlation calculation) to group samples in the PC space, as well as estimate how many significant clusters exist in the data. In the simplest sense this is like asking, "Is this new point closer to point A or point B?" In this section, the clustering of these samples mathematically is implemented using *k*-means algorithm. The code is shown below:

```
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=3) # 1. initialize
kmeans.fit(PCS_norm_df.loc[:, 1:2]) # 2. fit

# get information from the fit model
kmeans.labels_ # get the labels
kmeans.custer_centers_ # get the cluster centers
```

This code uses *attributes*, that are additional information stored within a particular model after it is fit with the data. These will differ between models, because different models calculate different information, but the attributes are easily found in the models documentation. For example, two useful attributes, `labels_` (the cluster number assigned to the training data) and the `cluster_centers_` (coordinates of the centers of the clusters) are used in the *k*-means algorithm. The cluster centers are plotted on the scatter plot of PC 1 vs PC 2 scatter plot and each sample is coloured by the label it has been assigned by the *k* means analysis (Figure 2.19). The clustering algorithm has done what was expected (from a visual perspective) for the three distinct clusters.

2.4.2.3 Pipelines: Putting Several Steps Together

Now if we have additional data and wanted to predict which group it was a part of we would simply have to treat it the same as the original data (normalize, transform in PC space, and

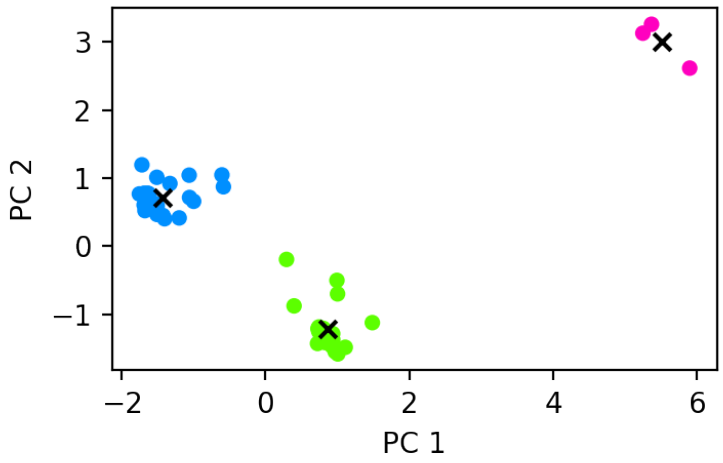

Figure 2.19: PC1 vs PC2 with clusters as determined by *k*-means algorithm. Cluster centroids are indicated by crosses.

predict with *k*-means). `Sklearn` offers the functionality to link these transformers and estimators together using a `Pipeline`.

```
from sklearn.pipeline import Pipeline

pipe = Pipeline([
    ('norm', Normalizer('max')),
    ('PCA', PCA(n_components = 2)),
    ('kmeans',  KMeans(n_clusters=3,))]) # 1. initalize all
   transformers and estimators in the pipeline

pipe.fit(data.values) # 2. fit with data
pipe.predict(unknown_spectra.values) # predict the group
   assignment of new spectra
```

The prediction results of the unknown spectra are shown in Figure 2.20. So far, the nature of each cluster has not actually been characterized beyond group 1, 2, and 3. To make this prediction meaningful, it important to now determine why certain samples are grouping together. Recall in this example here, each spectrum has some sort of label (MDA or MDMA) as previously determined by secondary techniques in combination with manual technician interpretation. However, one can imagine if this information was not available such clustering might now prompt chemists to investigate the chemical differences of each cluster to explain their separation. Figure 2.21 colours each sample according to their known label. Now it is clear that each of clusters in the example are characteristic to MDA-

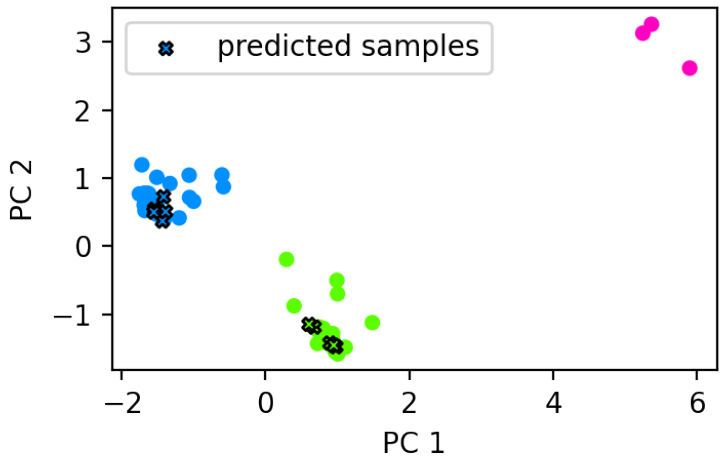

Figure 2.20: Visualizing the predicted unknown samples projected into PC space, distinctly grouping with the original two clusters. Unknown samples are represented by crosses and coloured by their predicted cluster.

labelled samples (red) or MDMA-labelled samples (blue). Notably, three of the MDA samples don't actually cluster with the others as circled in Figure 2.21. It is likely that these are outliers.

2.4.2.4 Outlier Detection Using PCA and Mahalanobis Distances

Outlier detection is introduced here as an extension of unsupervised classification methods. As mentioned previously, there are many ways to statistically detect outliers, and often there is no perfect approach. Here the Mahalanobis distance is used to measure the distance between two multivariate distributions. First, the distances for all points in the PC space are calculated by taking advantage of the functions and attributes offered by `MinCovDet` from `sklearn`.

```
from sklearn.covariance import EmpiricalCovariance, MinCovDet

robust_cov = MinCovDet(random_state = 42, assume_centered = True)
robust_cov.fit(PCS_norm_df.loc[:, 1:2])

# get the mahalanobis distances
md = robust_cov.dist_

# calculate p values
p = 1 - chi2.cdf(md, 1)
```

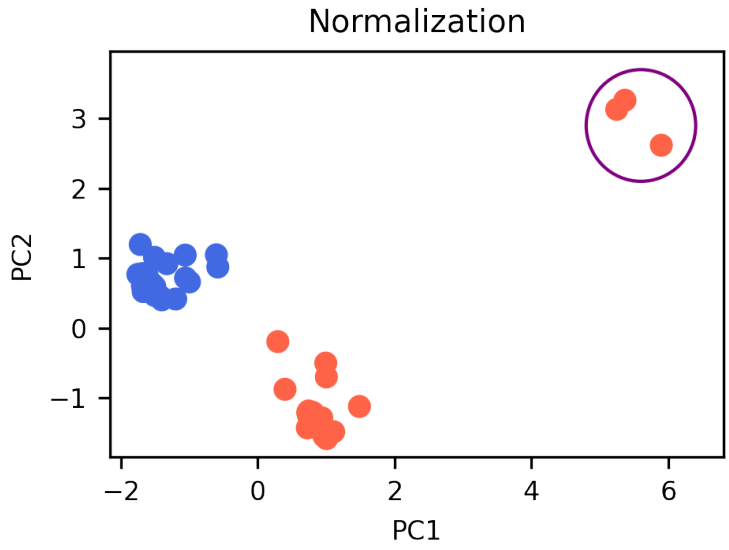

Figure 2.21: PC 1 vs PC 2 scores where blue is MDMA and red is MDA samples, which has been labelled from previous analysis. Notably three MDA (red, circled) samples do not cluster with the others, suggesting that they may be potential outliers.

Now, a threshold needs to be selected. Here a Mahalanobis distance greater than 3.29 declares a sample an outlier. This choice is based on p values at a 0.001 significance level, which is a common metric in statistics, however can be varied based on the application. By organizing the results in a dataframe those determined to be an outlier outside the threshold can be identified as shown in Figure 2.22. Adding contour lines indicating the Mahalanobis distance within the PC space helps visualize the effect different thresholds would have on the results. The resulting plot (Figure 2.23), shows the three clear samples that are determined as outliers in this example.

	PC #	1	2	mahalanobis	p	outlier?	plot colour
number	label						
34	**MDA**	5.897109	2.617210	159.807714	0.0	True	(0.6960784313725497, 0.0, 0.0, 1.0)
41	**MDA**	5.366078	3.261996	167.435213	0.0	True	(0.5, 0.0, 0.0, 1.0)
43	**MDA**	5.245849	3.130684	157.603787	0.0	True	(0.7673796791443852, 0.0, 0.0, 1.0)

Figure 2.22: Filtering the `Dataframe` to identify which samples are determined as outliers based on the calculated threshold. Three samples have a Mahalanobis distance greater than 3.29.

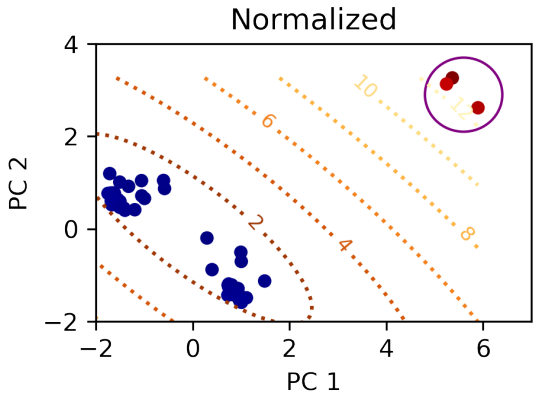

Figure 2.23: PC1 vs PC2 with contour lines indicating Mahalanobis distances.

2.4.3 Classification—Performance Evaluation of PLS-DA with Cross Validation

So far the methods used do not require any labelled data as part of their calculations. This is useful for exploratory data analysis. For example, while the initial labels (MDMA vs MDA) had suggested that only two groups would emerge, the exploratory analysis revealed a third potential group. However, without eventually having true class labels it is challenging to evaluate the performance of a model. Establishing this dataset with 'known' class values is done in a number of ways in the context of drug checking: (a) using standards made up in the laboratory (b) using samples that have some sort of established confirmatory testing such as GC–MS or (c) carefully, in some cases, using spectra that have been interpreted by an experienced person in the field. Each method has associated pros, cons, and risks. Error in the initial labelling will affect the accuracy of the model.

Supervised pattern recognition (classification) explicitly aims to optimize a mathematical relationship between spectral features ($\mathbf{X}$) and class labels ($\mathbf{y}$). There are many supervised classification algorithms it can be challenging to identify the ideal choice. It helps to understand the nature, size, and complexity of the data, and in general start with the most simple approach before considering more sophisticated and computationally expensive algorithms. In this example, the same qualitative MDMA and MDA Raman

spectral dataset, now taking advantage of the class labels, is used to implement the supervised classification method of PLS-DA. Below is a reminder of the **X** and **y** data, the three outliers as identified in the previous section have been removed:

```
X = data_rm.values # getting the spectral values from the
    dataframe with outleirs removed
y = data_rm.rename(index = {'MDMA':1, 'MDA':0}).index.
    get_level_values('label').to_numpy() # renaming the class
    labels to numerical representation 1 and 0
```

`Sklearn` has several cross validation (CV) functions that allow for preliminary evaluation of model performance. In CV the dataset is repeatedly split into a training set and tests set. The training set is used to fit the model, a test set is used to evaluate performance, and this is repeated several times with various splits of the data. Cross validation is revisited throughout this thesis to compare the influence of different combinations of pre-processing or parameter settings on model performance. `GridSearchCV` is a function that accepts a grid of parameters, performs an $n-$fold cross validation on each combination of parameters, and returns their respective scoring metrics. In the example below a `Pipeline` is again initiated, linking both `Normalizer()` transformer and `PLSDA` classifier:

```
pipe = Pipeline([
    ('norm', Normalizer()),
    ('classifier', PLSDA()),
])
```

This time, however, instead of explicitly defining certain parameters when initializing the pipeline, a dictionary with a range of parameters and their values is created (`param_plsda`). In this example, different values for the number of components in the `PLSDA()` classifier and the norm used in the `Normalizer()` transformer are explored:

```
param_plsda={'classifier__n_components': range(2, 5),
             'norm__norm': ['l1','l2','max']}
```

The grid search follows the exact same steps as used before: (1) initialize (2) fit with training data and eventually, (3) predict on unseen data. During the fit step the function automatically refits the pipeline on the parameters that achieved the best scores from the cross validation.

```
from sklearn.model_selection import GridSearchCV
# 1. initialize
gscv = GridSearchCV(
    pipe, # the pipeline as defined above
    cv=7, # number of splits for cross validation
    scoring='accuracy', # metric to calculate and rank on
    param_grid= param_plsda # parameter grid
)

gscv.fit(X, y); # 2. fit with the data
```

The complete cross validation results of the grid search can also be accessed:

```
result_df = pd.DataFrame.from_dict(gscv.cv_results_, orient='
    columns')
result_df
```

The best score from the grid search was found to be 1.0 or 100% accuracy, using parameters `n_components` is equal to 2 when computing PLS-DA and `norm` is 'l1' when computing the normalization. In this example several of the combinations give equal accuracy, therefore it is preferred to select the simplest model (`n_components`= 2). In general, cross validation is a well accepted method to get a sense of model performance and make use of smaller amounts of data. Picking the best model based on a single metric, such as accuracy, works quite well in straightforward examples as shown here. During training and cross validation, however, there is a risk of an inflated sense of performance as a result of over-fitting (i.e. the model tries to fit the data perfectly and therefore cannot generalize outside the data it has been trained on). Therefore, it is important to consider several metrics and test model performance on an external test dataset. In the following quantitative example, the discussion focuses on comprehensive optimization of pre-processing and other parameters, evaluation metrics, and testing on unseen data.

2.5 Quantitative Calibration Example with an Infrared Spectral Dataset

This section uses infrared spectra of various opioid mixtures that were made in the lab using analytical standards. This means that each sample mixture contains a known concentration of fentanyl and these labels are taken into account when training, testing and validating

the models. The goal then is to model the relationship between infrared spectra and the concentration of fentanyl in the sample, so that this model can be later used in a drug checking service to make this prediction based on an unknown sample.

2.5.1 Data Pre-Processing

In practice, most of the steps to load the quantitative dataset are the same as what was shown with the qualitative data. However, when using quantitative models the **y** response variable or "label" instead corresponds to a concentration value (Figure 2.24). In many cases, there may be more than one spectrum for each concentration (just how there was more than one spectrum from each class) which is useful for training and validating the performance of models.

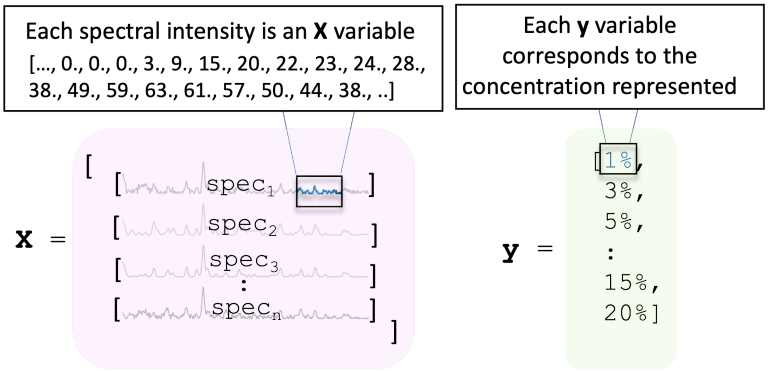

Figure 2.24: Common data structure for quantitative analysis.

The following code briefly demonstrate calling in the IR spectra for the quantitative data that was already formatted elsewhere and stored in a pickle file.

```
quant_df = pd.read_pickle('quant_df.pkl')
```

This time the Dataframe function `describe()` is used to quickly get statistics such as the average, mean, and standard deviation of each spectral variable. In Figure 2.26 the average spectrum and standard deviation (shaded in grey) of the dataset is plotted to visualize which spectral variables are changing most significantly. This is an alternate way to explore the data before further chemometrics.

```
quant_describe = quant_df.div(quant_df.max(axis = 1), axis = 0).
    describe().copy()
```

fent conc	label	percent_other	651.0	652.0	653.0	654.0	655.0
1.0	O	NaN	0.080384	0.077963	0.078082	0.078274	0.077667
		NaN	0.084552	0.082130	0.082335	0.082373	0.081255
		NaN	0.082213	0.080432	0.080783	0.080954	0.080204
		NaN	0.092118	0.089914	0.090608	0.091294	0.090782
		NaN	0.084598	0.081668	0.081552	0.081689	0.081029
...	...	...	...	...	...	...	...
20.0	DA	5.0	0.139183	0.136682	0.134893	0.134119	0.133600
		5.0	0.140089	0.138524	0.138195	0.138707	0.138621
		5.0	0.142242	0.138075	0.135109	0.133924	0.133513
		5.0	0.141857	0.137559	0.135034	0.135008	0.135821
		5.0	0.139489	0.136653	0.134571	0.133864	0.133473

325 rows × 2950 columns

Figure 2.25: Dataframe of the standard mixtures including binary mixtures of fentanyl and caffeine, and ternary mixtures with a third component of mannitol, erythritol, or heroin.

```
plt.style.use('seaborn')

fig = plt.figure()
ax = fig.add_subplot()

ax.fill_between(quant_describe.T.index.to_numpy(),
                quant_describe.T['mean']+quant_describe.T['std'],
                quant_describe.T['mean']-quant_describe.T['std'],
                color = 'grey',
                alpha = 0.5, label = 'stdev')

quant_describe.T.plot(y = 'mean',
                      ax = ax,
                      xlabel = 'wavenumber / cm$^{-1}$',
                      ylabel = 'normalized average intensity')
ax.legend()
fig.tight_layout()
# fig.savefig('quant_avg_spectra.png', dpi = 200)
```

2.5.2 Optimization of Pre-Processing and Evaluation Metrics

In this quantitative example, the workflow builds on concepts already introduced such as the utility of transformers, estimators, pipelines, and cross validation. In addition, it focus

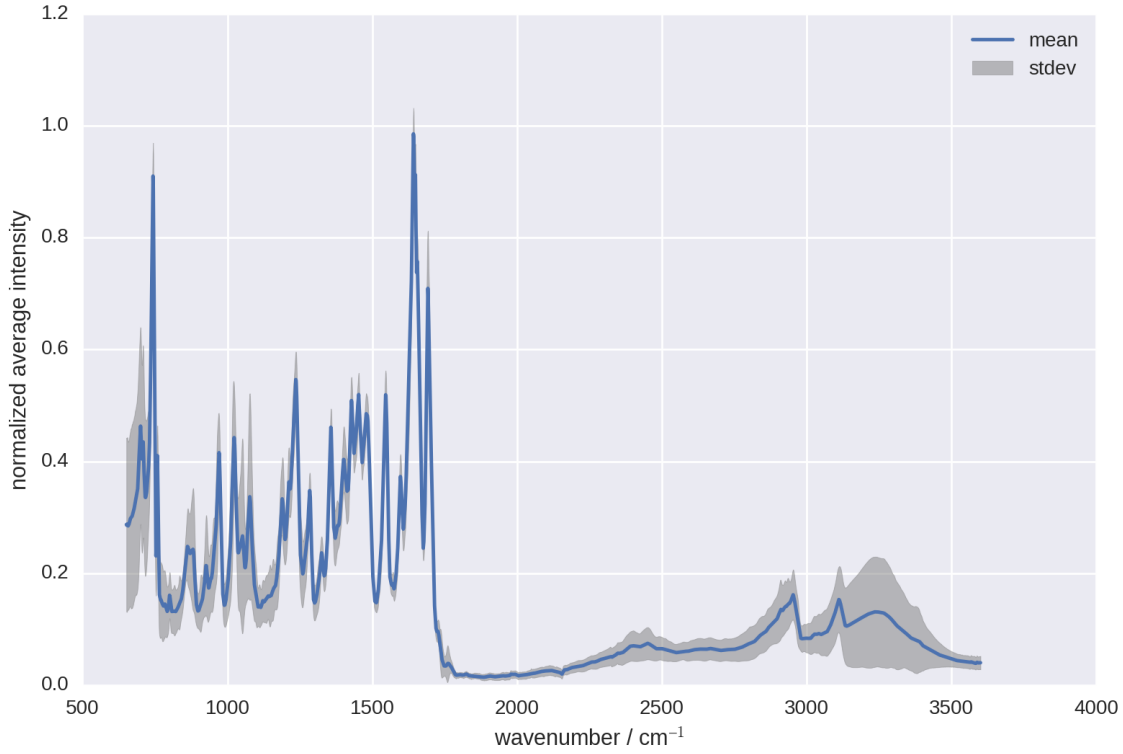

Figure 2.26: The average (blue) of all spectra in the quantitative infrared dataset for opioid mixtures with various cutting agents and concentrations of fentanyl. The standard deviation for the dataset is shaded in grey at each spectral frequency.

on exploring several combination of pre-processing and emphasize the role of validation (or test) data in model optimization. To validate the applicability of a model in predicting unknown samples it is important to create and evaluate a validation data set on trained models. In this case, the quantitative dataset has 5 replicate samples at each concentration. From each mixture, a random fraction of these samples is set aside to be the training set (also sometimes referred to as the developmental set) and another fraction for the validation set (not involved in training). A general rule of thumb is to use 2/3 of data for development and training and 1/3 for validation (also sometimes referred to as the test set).

```
X = quant_df.values.tolist()
y = quant_df.index.get_level_values('fent conc')

from sklearn.model_selection import train_test_split
# splitting the data into training and validation sets
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size
    =1/3, stratify = quant_df.index.get_level_values('label'),
```

```
    random_state=0)
```

The only pre-processing methods used so far in the data analysis pipelines are normalization and dimensionality reduction with PCA, however there are many other combinations possible. Unfortunately, there is no concrete answer as to the "perfect" pre-processing choice but there might be some hints in examples of similar applications and field or instrument specific knowledge. Determining these parameters is often done through exhaustive model tuning where models are fit and evaluated under a range of set of conditions and their performance is compared in terms of metrics such as mean error, bias, and accuracy. The following code prepares combinations of scaling options, spectral range filters, and baseline corrections for this comparison:

```
#preparing the combinations of pre-processing to explore
scaling = {('norm1', Normalizer('l1')),
           ('norm2', Normalizer('l2')),
           ('minmax', MinMaxScaler()),
           ('snv', SNVNormalizer()),
           ('stdscale', StandardScaler()),}

subset = {('range', FunctionTransformer(select_range))}

background = {('deriv1', FunctionTransformer(savgol_filter,
    validate=True, kw_args={'window_length': 5, 'polyorder': 2, '
    deriv': 1})),
              ('drpls', FunctionTransformer(baseline, validate=
    True, kw_args={'alg': 'drpls'}))}

from itertools import product

preprocessings = {(('none', None),)} | \
                 set((p,) for p in scaling) | set((p,) for p in
    background) | \
                 set(product(scaling, background)) | set(product(
    subset,scaling, background))
```

It is worth noting that the examples so far are very simple, with small datasets (50 samples) and grid space of parameters (less than 10). Additional pre-processing combinations and parameters will exponentially increase the time it takes to train and evaluate the model. In this calibration example 216 infrared spectra are in the dataset and 23 unique pre-processing combinations are explored with 10 different hyper parameters. Using a 7-fold cross validation, the regression model would have to be fit and predicted

$23 \times 10 \times 7 = 1610$ times. In this case, it might be advantageous to use `RandomizedGridCV` as a first pass approach, which offers identical functionality as `GridSearchCV` but only searches of a subset of hyper parameters to reduce fit time. If some pre-processing combinations or hyper parameters result in poor performance, subsequent testing could remove these and run a more thorough search. Several steps are put together here to demonstrate looping through pre-processing combinations, performing cross validation with `RandomizedSearchCV`, and calculating additional metrics with a validation set:

```
from sklearn.model_selection import RandomizedSearchCV
result = []
predictions = []
for i, preprocessing in enumerate(preprocessings):
    # making the pipeline with each preprocessing, ending with the
    regressor
    pipe = Pipeline(list(preprocessing) + [('plsr', PLSRegression
    ())])

    # randomized grid search to perform cross validation with
    varying number of components for PLS-R
    gscv = RandomizedSearchCV(
        pipe,
        {'plsr__n_components': range(5, 15)}, # hyper parameter
    grid
        cv=7, # 7-fold cross validation
        refit='neg_root_mean_squared_error', # refit based on
    lowest RMSE
        scoring=('neg_root_mean_squared_error', 'r2'),
        n_iter = 5 # number of times to sample hyper parameters
    ).fit(X_train, y_train) # fitting with training data

    # predict with the best model on the validation set
    pred = gscv.predict(X_val)
    predictions.append(pred)

    # append the results to be put in a dataframe
    steps = tuple(step for step, _ in preprocessing)
    result.append([
        steps,
        gscv.cv_results_['mean_test_r2'][gscv.best_index_],
        gscv.best_params_, # best parameters for each
    preprocessing
        -gscv.best_score_, # best parameters for each
    preprocessing
        r2_score(y_val, pred), # r2 on validation set
        -gscv.score(X_val, y_val), # RMSE on validation set
        np.average(abs(pred.reshape(-1) - y_val) / y_val) * 100 #
    average error calculation on validation set
        ])
```

For each pre-processing combination, the resulting model is also used to predict the previously set aside validation set. The performance when applied to the validation set is evaluated using similar metrics as in cross validation. This validation is a very important step to identify if the model performance persists for samples it has not seen before. Since multiple metrics are calculated here, ranking the results provides different answers as the "optimal" pre-processing depending on the metric used. Some common metrics include root mean square error of prediction (RMSEP),root mean square error of cross validation (RMSECV), and coefficient of determination (R^2). For simplicity, many application-focused papers tend to simply choose the model with the minimum RMSEP on the validation set. On the other hand, many statisticians argue that the optimal answer should rely on a more comprehensive set of metrics simultaneously. Ultimately, the question to ask is, "Does this model predict fentanyl with reasonable accuracy and stability to be useful in drug checking?" Ongoing validation of the model is key to answering this question, through making additional known mixtures with standards, exploring the potential for interference in accurate results, cross-lab studies, and importantly, comparing results on real samples with secondary techniques like HPLC.

Attributes of building chemometric models share many similarities with drug checking technicians; that is, more training with relevant data results in greater the confidence in performance; cross referencing of results with expert technicians and validated techniques will result in less risk; and, performance will always be limited by the data and capabilities of the instrument. Drug checking is a field where manual interpretation of spectral results, user experience, general likelihood based on supply monitoring, and harm reduction considerations all play a role in the final result. However, aspects of chemometrics are still widely untapped for harm reduction-based drug checking, and likely because its implementation is time-consuming and inaccessible to many spearheading these projects. The general workflow presented above can be used to implement hundreds of clustering, classification and calibration techniques through Python and open source packages like

`sklearn`. This background has introduced the nature of some of the numerous drug checking related questions that have been asked throughout this thesis and the wide variety of instrumental approaches and chemometric tools that are drawn from to answer such questions. With careful consideration, these automated models have the opportunity to relieve drug checking technicians of the pressures analyzing spectral signatures, memorizing trends, and presenting results with a subjective measure of certainty.

Chapter 3

Fentanyl Detection and Quantification using Portable Raman Spectroscopy in Community Drug Checking[1]

3.1 Overview

Community-based drug checking has emerged as a harm reduction practice aimed at people who use drugs. Using a portable Raman spectrometer and the statistical method of partial least squares regression, a model was developed to quantify fentanyl in both powder binary mixtures and more complex ternary mixtures. The model was then applied to samples collected over a two-year period while operating the drug checking service. As an unpredictable drug supply will always pose a risk for quantification with portable drug checking technologies, we implement check steps that guide the harm reduction decisions and conversations surrounding quantitative results.

3.2 Introduction

Among the instruments available for drug checking, spectroscopic methods are of interest due to their rapid testing, capacity for onsite use, and non-destructive nature. For these

[1]This chapter has been adapted from L. Gozdzialski, M. Ramsay, A. Larnder, B. Wallace, D. Hore. "Fentanyl Detection and Quantification using Portable Raman Spectroscopy in Community Drug Checking" *J. Raman Spectrosc.* ***52****, 1308 (2021). Acquisition of the Raman spectra was performed by LG. MR made up the standard mixtures with assistance from LG. AL provided some feedback on the discussion. Coding, writing, and data analysis was performed by LG.*

reasons, IR absorption and Raman scattering are commonly-employed techniques. Raman, in particular, is of interest due to its relative simplicity, robust hardware with no moving parts, potential for integration with other detection methods, insensitivity to water or moisture, and the opportunity for through barrier detection (e.g. vial, plastic bag, or container) using spatially-offset Raman scattering (SORS).[95] Inherent challenges are in achieving sufficient sensitivity for low concentration actives, and mitigating fluorescence. Recent interest in using the same instrumentation for surface-enhanced Raman scattering (SERS) provides the potential for ultra-trace detection which could be applied in the identification of fentanyl[108,168–170] and its analogues, such as carfentanil.[49] In the context of drug checking, developing protocols that can work directly with powdered samples and without any sample preparation is particularly useful. For this reason, direct Raman scattering remains an attractive option.

Raman spectroscopy can effectively discriminate and quantify drug molecules in a powder, crystal, tablet, or liquid form. Much of the work to date has been presented within the context of clinical,[171] forensic,[172–174] and pharmaceutical[175] testing, using a broad range of instrumentation (lab-built instruments, and commercially-available devices, including Raman microscopes). Simulated street drugs, or those seized by law enforcement, have also been tested using Raman spectroscopy.[146,148,149] Additional challenges present themselves in shifting these technologies for use in a community-based harm reduction service.[18] For example, the consequence of low-concentration adulterants are less predictable than what may be encountered in pharmaceutical quality control applications. The assessments of commercially-available Raman instruments in health-based drug checking are also limited, with some account of the limits of detection, sensitivity, and specificity of fentanyl in comparison to other popular devices.[26] The results from festival testing using Raman instrumentation have been reported and compared to gold standard methods.[23] In many such studies on-board software offers useful qualitative results. However, the implementation of portable Raman spectrometers for

further quantitative analysis has remained largely unexplored in community drug checking.

The objectives of this work are to demonstrate the ability of a portable Raman spectrometer to quantify fentanyl in relevant drug mixtures and to develop a framework for analyzing real drug checking data acquired at a point-of-care service.

3.3 Instruments and Data Acquisition

We use a portable Raman spectrometer (Agilent Resolve, Santa Clara, USA) equipped with an 830 nm laser with adjustable power over the range 100–475 mW. All spectra were acquired directly on powder samples using either a glass or aluminum sample stage. The instrument performs an internal background and cosmic ray correction, and records data for Stokes shifts of 200–2000 cm^{-1} with a resolution of approximately 12 cm^{-1}. The 350–1750 cm^{-1} region was considered in our analysis due to a lack of chemical features of interest outside this range.

Depending on several factors, such as sample size provided, Raman acquisition may be on a glass slide or aluminum substrate, or through a plastic bag. In all cases, a hand-held "point-and-shoot" mode is used for data collection. Powder drug binary mixtures were made up of fentanyl HCl (Toronto Research Chemicals, Toronto, Canada) concentrations ranging from 1–40% w/w with caffeine (Sigma Aldrich) making up the remainder of the mixture. Five spectral replicates were acquired at each concentration and were randomly split (3:2) into training and test sets. Three-component powder drug mixtures were also prepared consisting of 1, 5, 10, 15, and 20% w/w fentanyl. Here, the ternary mixture included caffeine as the second component and either mannitol (Sigma), erythritol (Sigma), or heroin base (Toronto Research Chemicals) as the third component. This third component was varied in proportion (10, 20, 30, and 40% w/w) for each fentanyl concentration to avoid any bias that may result from cross correlation. These composition choices are reflective of common cuts and mixture proportions seen in our service samples containing fentanyl. All spectra were background- and baseline-corrected using instrument firmware.

The corrected spectra were exported to an online database where they are accessible to pipelines for data processing, model development, and real-time analysis via codes written in Python. All algorithms were implemented using the Python `scikit-learn` machine learning package.[176] Quantification of the fentanyl concentration was achieved using partial least squares regression (PLS-R). The choice of spectral pre-processing and the number of latent variables used in the model are critical for optimizing performance.[166,167,177] Several combinations of pre-processing corrections were considered, including normalization (min-max, vector, standard normal variate, area), Savitsky-Golay smoothing and first and second derivatives, to correct for spectral variability and improve model performance. For a preliminary assessment of these parameters, a 10-fold cross-validation was performed.[148,178] Final choices however were guided by minimization of the root-mean-square error of prediction (RMSEP) and the principles of parsimony, where appropriate.[155,166,175,179,180] Hotelling's T^2 and Q residuals were used to explore potential calibration outliers.[180] One-class classification (OCC) algorithms, including isolation forest,[181] robust covariance[182] and local outlier factor (LOF),[183] were used for assessing the influence of anomaly detection on the distribution of prediction results of the unknown service samples. The same calibration standards used to build the PLS-R models were also used to train the respective OCC methods.

3.4 Results and Discussion

Simple binary mixtures were first considered to assess the ability of a portable Raman spectrometer to quantify fentanyl under controlled conditions with limited data variability. Figure 3.1 shows the spectral variation due to increasing fentanyl concentration through a series of the caffeine and fentanyl binary mixtures averaged at each concentration level ($n = 5$) normalized to the largest caffeine peak near 555 cm^{-1} for better visualization. The Raman spectra are dominated by caffeine features, however, evidence of the strongest fentanyl feature at 1000 cm^{-1} (inset Figure 3.1) is visible in spectra with as low as 2–

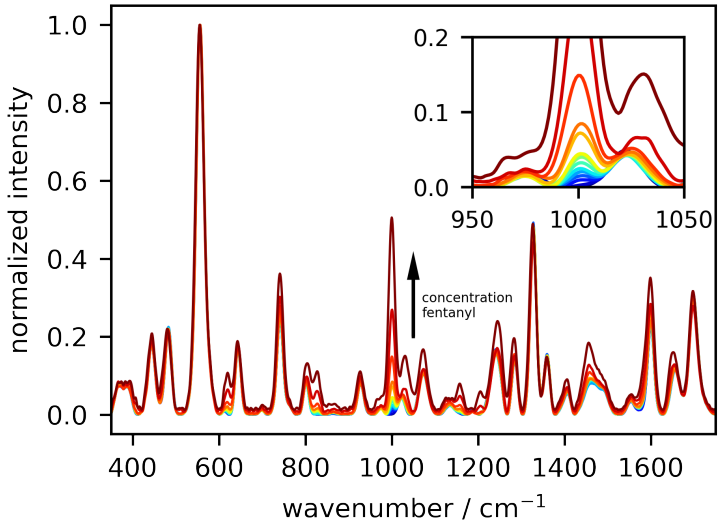

Figure 3.1: Normalized Raman scattering of binary fentanyl and caffeine mixtures ranging from 1–40% w/w fentanyl. Spectra are an average of 5 repeated measurements at each concentration.

3 %w/w fentanyl. Using PLS-R, our first model (Figure 3.2a–d) was calibrated using 3 replicates of each binary standard; we will refer to this as Model A.

We consider (1) pre-processing (2) dimensionality, i.e. number of latent variables and (3) outlier detection in optimizing model calibration. Figure 3.2a–c presents several of these steps. While not explicitly shown, the optimization procedure is an iterative process and many parameters may be considered at once. The best choice is not always obvious, however, it is common practice that decision in parameters aim to minimize the RMSEP.[166,184] After exploring several pre-processing combinations, area normalization was chosen as the pre-processing method. Four latent variables were chosen, as shown in Figure 3.2a, to achieve a global minimum of RMSEP. Using more latent variables often causes an over-fitting of the data that can be seen by the increase in prediction error when using greater than four latent variables. A plot of Hotelling's T^2 and Q residuals for the calibration samples is shown in Figure 3.2b, with 95% confidence interval thresholds indicated by the dashed lines. We investigate samples outside the lower left quadrant as potential calibration outliers. We have identified those points as originating from

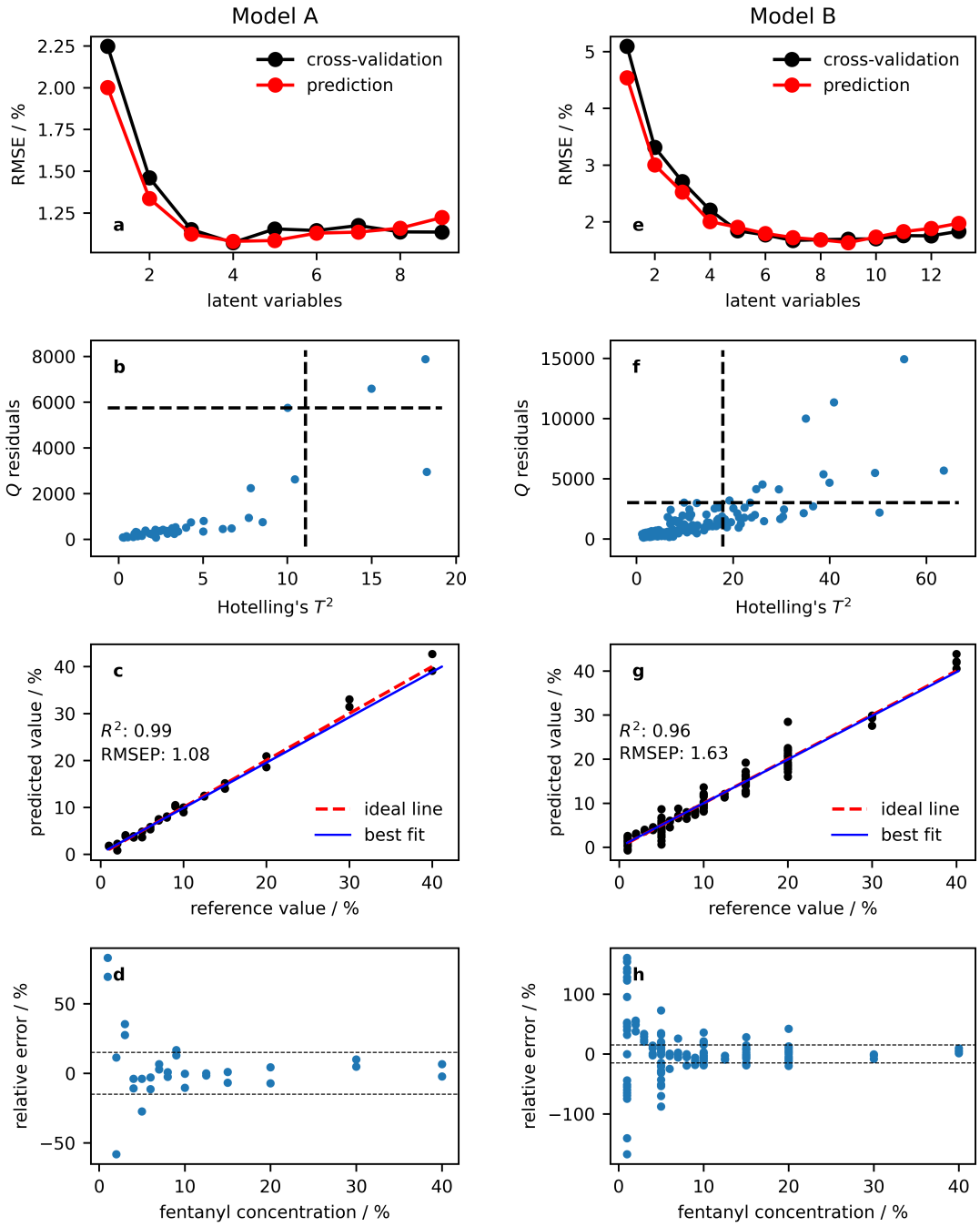

Figure 3.2: The calibration and validation steps for building Model A, trained with binary mixtures only ($n = 45$), compared to Model B, trained with additional variability based on sample composition and acquisition parameters ($n = 240$). (a,e) The root-mean-square error (RMSE) as a function of latent variables for internal validation (cross-validation, black) and external validation (prediction, red). (b,f) Hotelling's T^2 vs Q residuals. Horizontal and vertical lines mark the 95% confidence intervals. (c,f) The prediction curve for external validation test set ($n = 30$ for Model A, $n = 160$ for Model B). (d,h) Accuracy profiles for the validation sets; a $\pm 15\%$ error is shown as the horizontal dotted line for reference.

samples containing 40% w/w fentanyl; it is not uncommon for samples near the extrema in concentration to lie slightly outside this region. Excluding these high concentration samples leads to a slight improvement in RMSEP, however, their inclusion still results in a reasonable overall error and their extension of the calibration curve to higher concentrations is vital within a drug checking application. For these reasons, no samples were excluded from the calibration set. The replicates of the binary standards not included in the calibration of Model A (i.e. test/validation set) were predicted to validate the optimized PLS-R model. The resulting prediction curve is shown in Figure 3.2c. The points are well clustered along the diagonal and results in a high R^2 value of 0.99 and low RMSEP of 1.08%. The accuracy profile (Figure 3.2d) shows the relative error between the prediction test set and the known true value. While there is no set threshold for drug checking, in pharmaceutical applications there is often an allowable relative error in methods of quality control, depending on the drug. Here, a 15% threshold is shown for reference. In Model A, there is a high relative error for samples containing 5% fentanyl and below. For controlled samples in the form of analytical standards, we have shown that Model A performs well, and quantifying binary mixtures of fentanyl and caffeine is possible using a portable Raman spectrometer. As PLS models are often validated on a much smaller representation than the actual application, we stress the importance of further assessing potential sources of error. To simulate some of the variability that often occurs in community-based drug checking, an additional external validation set was acquired from the same binary powder samples. This set was acquired on different days (in attempt to control for minor fluctuations in temperature, humidity), on different instruments, with different laser power settings, and using various substrates (aluminum foil or glass slides). The aim of their inclusion is to highlight some of the challenges inherent to using portable instrumentation, especially for those with a hand-held setup. The prediction curve is shown in Figure 3.3 and the RMSEP approximately doubles from the original test set. This data emphasizes some of the risks involved in implementing a lab-based calibration curve in an application such as

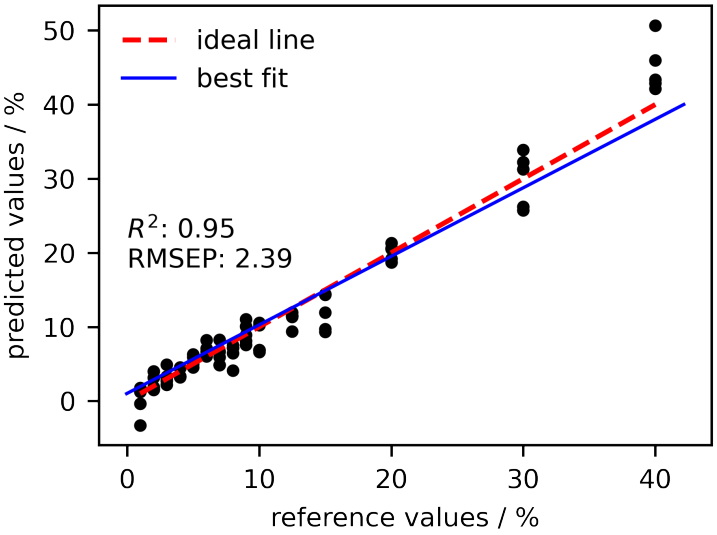

Figure 3.3: An additional external validation set was investigated to assess the risks involved with employing a perfectly lab-calibrated model in drug checking. This new validation set uses $n = 75$ samples acquired (from the same reference set) on different days, instruments, laser power settings, and sampling stages (aluminum foil or glass slide) to simulate only some of the anticipated variation that occurs during real-time drug checking. The RMSEP has approximately doubled.

drug checking where there are many circumstantial sources of variability. These figures of merit may be misleading, even prior to considering variability in composition.

We now introduce Model B (Figure 3.2e–h) that includes additional variability in the calibration set in the form of sample composition. We included ternary mixtures with an additional component of either mannitol, erythritol, or heroin. Samples were also collected with different acquisition parameters as described above. The same process for model optimization was followed as described for Model A and is shown in comparison in Figure 3.2e–h. Area normalization pre-processing and 9 latent variables are used as suggested by RMSEP values. Compared to Model A (Figure 3.2a), a higher number of latent variables are required to model the more complex dataset. Figure 3.2e shows the exploration of calibration outliers using a plot of Q residuals and Hotelling's T^2. Samples in the upper right quadrant (outside the 95% confidence intervals) are considered possible outliers. These particular samples were identified as containing high amounts of erythritol

or mannitol. This motivates future investigation into whether local models (i.e. within each class of mixtures), in combination with classification methods, may ultimately reduce error in this analysis. The exclusion of these calibration samples in this global model does not result in a reduction in RMSEP or the root-mean-square error of cross-validation (RMSECV). For this analysis we have decided to include all calibration points with this consideration for future work in mind. Compared to Model A, the RMSEP increases to 1.63% and the R^2 decreases to 0.96. The accuracy profile (Figure 3.2h) further guides comparison between the two models. Similar to Model A, Model B produces a high relative error for samples containing 5% fentanyl or less. In Model B, however, we can see a slight increase in the instability of the predicted value over the spread of the data.

A pipeline was developed for the quantification of fentanyl in opioid samples, hereafter referred to as service samples, acquired at our drug checking service. Of the 1187 service samples that had Raman spectra collected, a subset was selected for prediction using both models based on filtering using three criteria: (1) a positive fentanyl test strip result ($n = 439$), (2) a signal-to-noise above a threshold value defined by background- and baseline-corrected signals greater 500 counts (n reduced to 306), (3) outlier/anomaly detection methods (resulting in a final $n = 306$). Here, three common methods were then employed to classify service samples as suitable for prediction by the calibrated PLS-R models. Figure 3.4 plots the scores of latent variables 1 and 2 of the service samples, as transformed by the PLS model, to visualize the clustering tendencies of the service data. The anomaly scores calculated by each method are represented by the radius of the orange circle of each point in Figure 3.4. Robust covariance is the least sensitive to outliers, showing samples poorly clustered in the latent variable space to still be assigned as inliers. The resulting trends in prediction of service samples using Models A and B are shown in Figure 3.5. The number of inliers significantly increases when using Model B, which has more variability in the training data, regardless of anomaly detection method used. The distribution of predictions between the three separate anomaly-detected subsets appear very

similar but contain a few notable differences. As found with robust covariance, there are more extreme predictions. Extreme predictions are classified as greater than 40% (as we do not expect the model to perform well outside the range it has been calibrated) and less than 0% . This suggests that robust covariance may be too lenient or an inappropriate method for anomaly detection in this context.

Regardless of which model is used, the predictions over the period December 2018–September 2020 in Figure 3.5 present comparable distributions. The average fentanyl concentrations are in the range 12–17% w/w, with standard deviation of 4–12%, depending on the number of samples determined inliers. Considering that the potency of fentanyl is approximately 100 times that of morphine, this mean concentration is high.[185] As fentanyl is so potent, any variability and unpredictability in concentrations presents a challenge for people who use drugs to regulate dosing within an unregulated market. When public health, harm reduction and people involved in the drug market are able to quantify fentanyl in the unregulated supply they access new knowledge to inform responses. The potential value of our framework is to reduce the uncertainty surrounding a potent and variable drug supply as much as possible. The portability of hand-held Raman is well suited for such point-of-contact interventions to inform individual harm reduction activities such as dosing, test shots and using where naloxone is readily available. The potential of such interventions over the long term is that there will be greater common knowledge of what a typical or average concentration of fentanyl is and its impacts. As fentanyl is increasingly certain in the illicit opioid supply in North America, there is a need for interventions that can provide essential quality control measures such as the determination of potency.

A major challenge for many drug checking projects, particularly those using vibrational spectroscopy, is access to additional services or equipment for external method validation and confirmation of the identity and concentration of compounds of interest. This presents a clear challenge when trying to assess the true accuracy of these models when applying them to data acquired in a drug checking service where the composition of

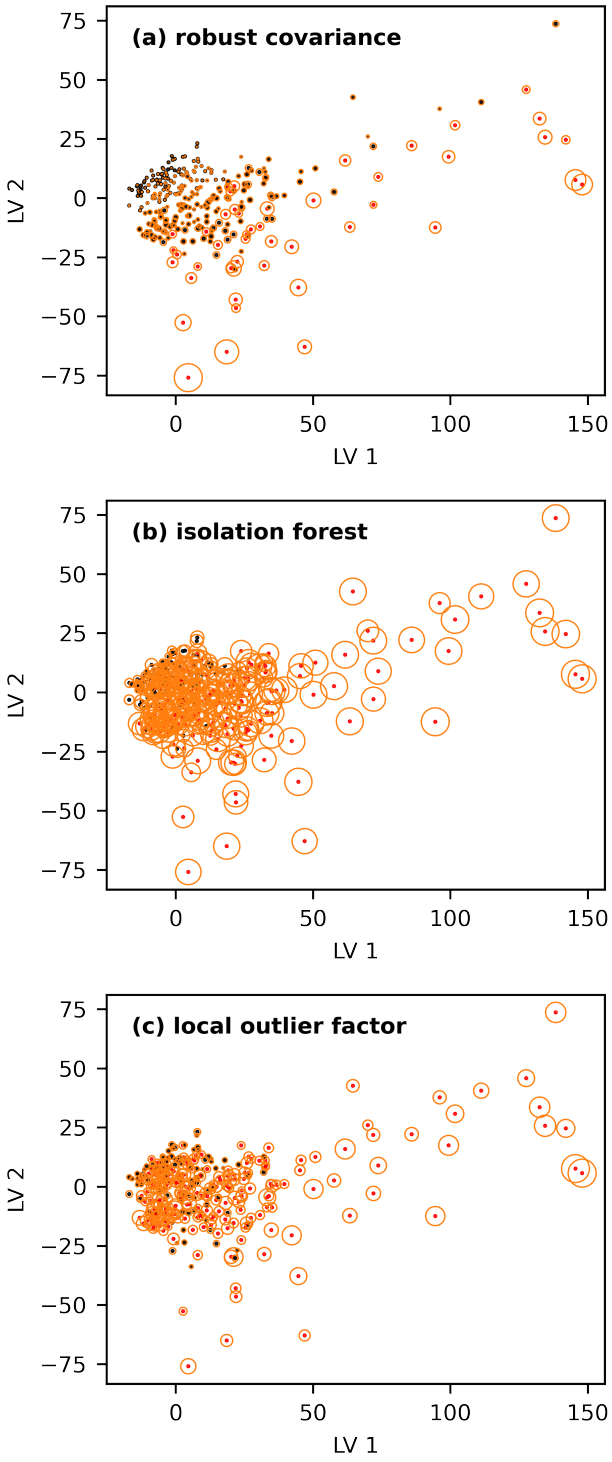

Figure 3.4: Scores of latent variables (LV) 1 and 2 of service samples ($n = 306$) as transformed by the PLS model. Samples were subjected to three outlier detection methods (a) robust covariance, (b) isolation forest and (c) local outlier factor. Outliers are shown in red and inliers in black. The anomaly scores calculated by each method are represented by the radius of the orange circle of each point.

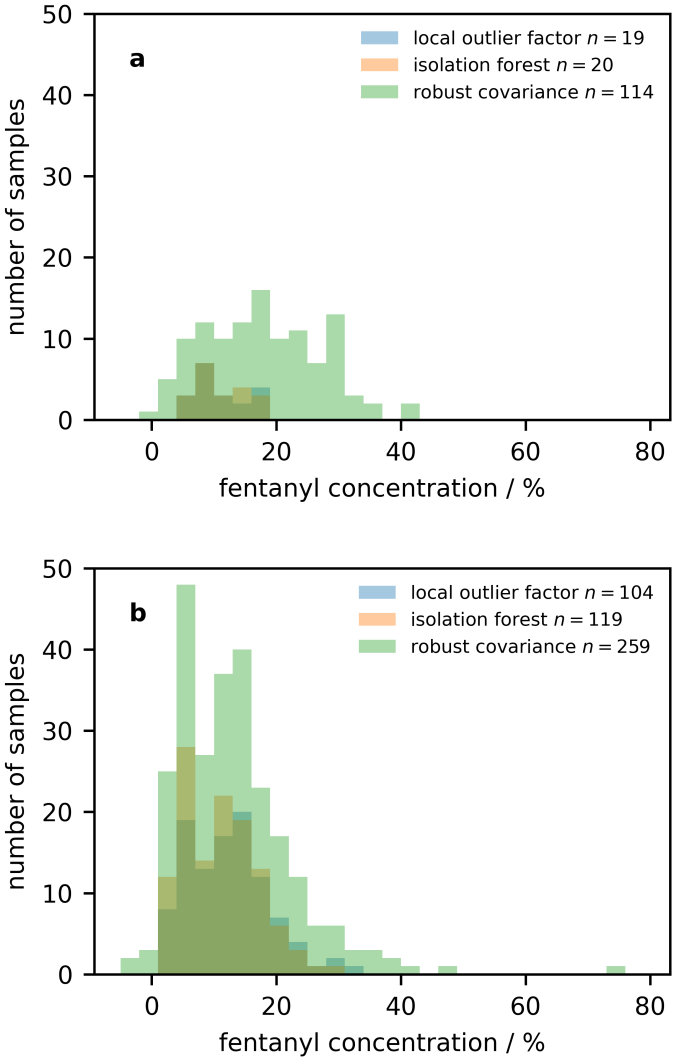

Figure 3.5: Predicted concentration of service samples deemed as inliers using (a) the optimized binary PLS-R Model A and various outlier/anomaly detection methods and (b) the optimized robust PLS-R Model B and various outlier/anomaly detection methods.

the drug mixtures is unknown. Drug checking projects exploring spectroscopic methods for quantification must prepare to adapt to the high level of variability inherent in an unregulated drug supply. The figures of merit we have calculated (RMSEP, R^2, and relative error) can provide a basis for assessing whether a portable Raman instrument is capable of quantifying mixtures using PLS-R. In the context of a drug checking service, these quantities may be used to communicate the uncertainty to service users as part of the harm reduction messaging. In particular, the accuracy profiles for known samples may be a valuable tool for communicating the anticipated degree of error depending on the predicted concentration. The use of multiple anomaly detection methods, together with visualization of the PLS-R latent variable space, can help to gauge how well-represented a particular sample is by the calibrated models. For example, if a new sample was considered an inlier by all three methods, there is increased confidence in its appropriateness for the model and its predicted value. When a sample has a calculated anomaly score near the decision boundaries, additional caution can be used in the interpretation of the results and perhaps a decision against reporting concentration. Lastly, a simple visual inspection or similarity score compared to calibration samples may aid further intuition on its reasonableness for prediction. This highlights the importance of interpretation and discussion of results when disseminating information to a service user and how it varies sample to sample.

3.5 Conclusions

A commercially available portable Raman spectrometer is capable of quantifying fentanyl in mixtures presented at a point-of-care community drug checking project. Many considerations must be addressed when using machine learning methods like PLS-R for community-based drug checking, such as modifications based on the volatile drug supply. Machine learning is an important development in drug checking for more sensitive classification and quantification models, however, it poses additional challenges when considering this variability. Here, important check steps were implemented to flag samples

as unsuitable for quantification using the models for reasons such as sample composition (e.g. additional components) or errors in acquisition (sample preparation, signal-to-noise ratio, etc). The consequences for reporting individual results or trends based on these decisions is discussed.

Chapter 4

Portable Gas Chromatography–Mass Spectrometry in Drug Checking: Detection of Carfentanil and Etizolam in Expected Opioid Samples[1]

4.1 Overview

There has been a recent increase in adulteration of opioids with low concentration actives such as fentanyl analogues and benzodiazepines. As drug checking projects using vibrational spectroscopy continue to seek confirmatory lab-based testing, the concern and reality of missing these potentially harmful substances in point-of-care testing is prevalent. A portable GC–MS was used to analyze select opioid samples acquired at a drug checking service in Victoria, Canada ($n = 59$). Certified reference standards of several fentanyl analogues and benzodiazepines were measured to guide targeted analysis of these samples. Results were compared with those obtained using a lab based paper spray mass spectrometer. Portable GC–MS was able to identify 62% of samples containing carfentanil and 36% of samples containing etizolam. In the case of etizolam, the success rate was higher for more potent samples: 78% of etizolam-containing samples were identified when

[1]This chapter has been adapted from L. Gozdzialski, J. Aasen, A. Larnder, M. Ramsay, S. Borden, A. Saatchi, C. Gill, B. Wallace, D. Hore. "Portable Gas Chromatography-Mass Spectrometry in Drug Checking: Detection of Carfentanil and Etizolam in Expected Opioid Samples" *Int. J. Drug Policy.* **97**, 103409 (2021). Acquisition of GC–MS spectra was performed by LG with assistance from JA, AL, and MR. Data analysis was done by LG. PS–MS analysis was performed by SB and AS.

the etizolam concentration was above 3% by weight. In comparison, infrared spectroscopy was able to detect etizolam in only 9% of the etizolam-containing samples, and is not sensitive enough to detect carfentanil at relevant concentrations. Portable GC–MS has potential in identifying low concentration substances in a point-of-care setting, without relying on subsequent off-site confirmatory testing.

4.2 Introduction

Laboratory-based analytical instruments are highly sensitive and can be expected to report on the full composition of a sample. However, there is an interest in portable instruments which are better suited to provide more immediate, point-of-care test results, and can be more easily integrated within supervised consumption sites and overdose prevention services. Many drug checking services seek out confirmatory laboratory testing to address any substances that may be undetectable using portable technologies such as FTIR. Mass spectrometry (MS), typically coupled with a separation technique such as gas chromatography (GC), is considered a gold standard technique for the analysis of illicit drugs.[186] This is, in part, due to its ability to confirm the presence of components based on molecular ion and/or fragment mass, without the strict requirement of library searching. Despite its excellent sensitivity and discrimination for drug identification, gas chromatography–mass spectrometry (GC–MS) is typically less considered for on-site testing due to its large size, site requirements (electricity, pump ventilation, carrier gas availability), long run-time, and dependence on significant technical knowledge and experience for operation.[18] Portable GC–MS instruments mitigate these challenges and therefore enable field use. Their development has allowed for more accurate and immediate testing in a broad range of applications such as the detection of chemical warfare agents, forensics, and environmental analysis.[120,187] Their use is also geared towards individuals who may not have a significant background in science or analytical techniques, such as emergency responders, military personnel, and law-enforcement.[120,187] Portable GC–MS

therefore holds significant promise as a portable drug checking technology. However, its effectiveness has yet to be explored within harm reduction applications.

This demonstrates the use and potential benefits of portable GC–MS in the detection of several important compounds increasingly appearing in the illicit opioid drug market linked to unprecedented rates of overdose. Opioid samples for which laboratory-based mass spectrometry identified carfentanil and/or etizolam were analyzed with portable GC–MS in order to determine whether the portable instrument was able to detect the same compounds. We also comment on the suitability of GC–MS for point-of-care drug checking with a focus on harm reduction, analogous to how FTIR is currently used.

4.3 Methods

Samples were received as a part of an ongoing drug checking project established in 2018 in Victoria, Canada.[37] For this study, a subset of samples received at the drug checking service between December 2020 and February 2021 were tested on a portable GC–MS and portable FTIR, where the presence of etizolam and/or carfentanil was identified by paper spray mass spectrometry (PS-MS).[50,133]

4.3.1 GC–MS

The Torion T9 portable GC–MS is equipped with a low thermal mass gas chromatograph and miniature toroidal ion trap mass analyzer with in-trap electron impact ionization and full scan range of 41 to 500 m/z. A solid phase microextraction (SPME) fibre-based syringe was used for internal performance validation using a manufacturer-supplied mixture[188] of standards to optimize ion target, detector voltage, and filament current. This standard mixture was sampled (sometimes multiple times) until the performance validation routine was passed. A CME fibre was used to inject drug samples into the GC–MS. Here, an aliquot of a sample collected at the point-of-care service was portioned out to approximately 1–2 mg, dissolved in 0.1 mL of methanol, and centrifuged to minimize the introduction of

particulates into the instrument. As we have not developed any protocols for quantification using GC–MS, the precise sampled mass was not critical, and we have not recorded the total mass of the drug. The CME fibre was submerged in the sample solution for approximately 10 s and left for 1–2 min to allow the solvent to evaporate from the coil. The coil was then introduced directly into the injector port on the GC–MS set to 300 °C. The column temperature was ramped from 50 °C to 290 °C at a rate of 2 °C/s. The selection of these parameters was based on a simultaneous consideration of high sensitivity and good peak resolution for a wide variety of drugs and adulterants seen in the drug supply, while maintaining a reasonably short run-time. The total run time was 300 s, and the typical time between sample injections was in the range 15–20 min, including time to clean the CME syringe. Carfentanil oxalate (0.1 mg/mL in methanol) and etizolam (1.0 mg/mL in methanol) reference standards, and an opioid analytical standard mix (containing fentanyl 10 μg/mL in methanol) were used as received (Cerilliant Corporation, Round Rock, TX) and introduced in the same manner as described above. Gas chromatograms of carfentanil and etizolam analytical standards are shown in Figure 4.1 with the mass spectrum of each compound displayed below.

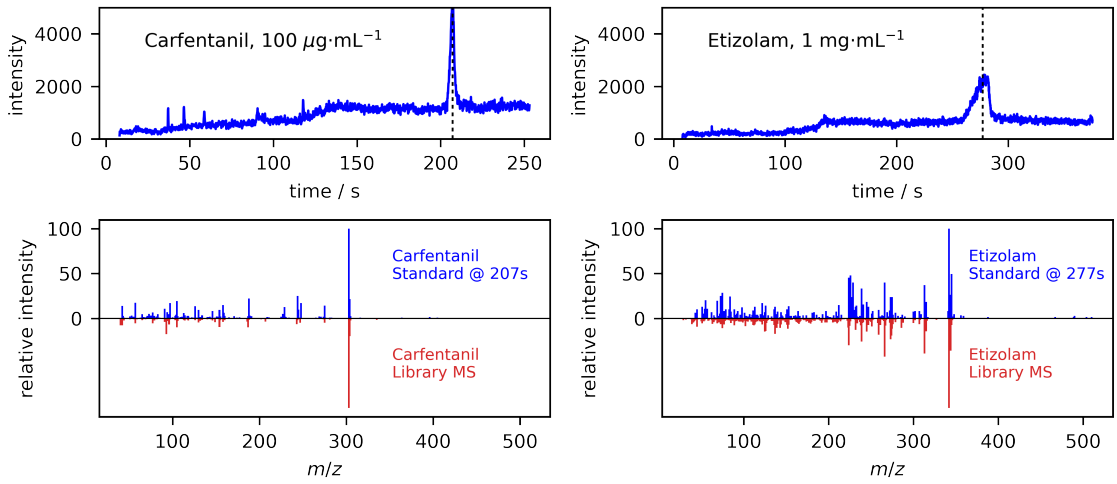

Figure 4.1: Gas chromatogram of 0.1 mg/mL carfentanil oxalate (left) and 1 mg/mL etizolam (right). The mass spectra acquired at each chromatogram peak are displayed (blue) together with the SWGDRUG MS library entries (red).

The CME fibre is cleaned using a procedure where the fibre is injected into the column at an elevated temperatures of 305 °C for 30–60 s, or longer if necessary. Following this, a system blank run is run to remove any compounds from the system that had desorbed from the fibre. The clean fibre is injected to ensure no compounds remain on the fibre. This process may be repeated until no peaks are visible in the blank run.

Heterogeneity of samples is always a concern, as the portion of the sample analyzed may not be representative of the overall composition. For the GC–MS measurements, a very small sample volume (on the order of 5 μL) is sampled, but the step of putting 1–2 mg of sample into solution homogenizes at least that portion of the sample.

4.3.2 IR Analysis

ATR-IR data was collected on an FTIR (Agilent 4500a) equipped with a single-bounce diamond internal reflection element (IRE) and a deuterated triglycine sulphate (DTGS) detector.[20] A spectral range 650–3600 cm^{-1} was recorded with a resolution of 4 cm^{-1}. 1–2 mg of sample are placed in contact with the IRE.

4.4 Compound Identification

4.4.1 Treatment of GC–MS Data

Two strategies are used for the interpretation of the resulting GC–MS data. Non-targeted searching is achieved via library matching using the SWGDRUG (Scientific Working Group for the Analysis of Seized Drugs) and 2017 NIST (National Institute of Standards & Technology) databases. In this workflow, a drug checking technician identifies peaks in the chromatogram, selects or integrates the mass spectra over a defined retention time range, and optionally subtracts a background signal. On our team, some of the technicians have backgrounds in science (chemistry, biochemistry, pharmacy) while others without prior technical education have been trained on the job. The resulting mass spectrum is then compared to database entries to retrieve a probable match based on similarity metrics

and, if possible, retention index estimations. Alternatively, targeted searching is guided by the mass spectrum, retention time and degradation products as obtained from analytical reference standards. A reconstructed ion chromatogram (RIC) is used to facilitate targeted analysis by selecting one or more mass fragments representing a particular analyte of interest. This provides a clearer identification of low concentration substances when obscured by baseline noise or other co-eluting compounds in the total-ion chromatogram (TIC). The resolved peak is then compared to the retention time of the particular analytical standard.

4.4.2 Treatment of IR Spectral Data

Automated analysis was performed using least angle squares regression (LARS)[189] and a custom IR library. The library consisted of target compounds (caffeine, fentanyl, etizolam, ANPP, heroin, and cocaine), as well as several common cutting agents, such as sugar alcohols from the SWGDRUG open-source ATR-IR library and was supplemented with in-house standards as acquired on our FTIR spectrometer. Carfentanil was excluded in this library due to the structural similarity to fentanyl (that would result in false positives) combined with the inability to detect carfentanil below 1% as a result of its comparable absorption cross section to fentanyl. False hits of target compounds were observed four times (twice with cocaine, once with fentanyl, and once with ANPP).

4.5 Results

During the study period, $n = 59$ samples that received quantitative testing using PS-MS were found to contain etizolam and/or carfentanil. In many of these samples, other notable substances were also detected. Table 4.1 presents a comparison of the substances positively identified through PS-MS analysis and those detected with portable GC–MS. In 36% of the samples, all of the substances identified by PS-MS were also detected by GC–MS analysis. If the data is reorganized according to component (Table 4.2), it reveals that 100%

of the samples containing heroin or cocaine, 95% of samples containing fentanyl, 62% of samples containing carfentanil, and 36% of samples containing etizolam were identified using portable GC–MS.

Table 4.1: Breakdown of components detected in $n = 59$ samples with portable GC–MS where target substances were identified by PS-MS. [a]For substances without an analytical standard (Cocaine, Caffeine, ANPP, Heroin), when the substance is reported as detected a top hit has been obtained in NIST database searching as described for non-targeted analysis. Any additional detected substances via portable GC–MS, such as pre-cursors, cutting agents, or breakdown products have been excluded for this comparison.

Substances detected by PS-MS	**Substances detected by portable GC–MS[a]**	**n**
ANPP, Caffeine, Cocaine, Etizolam, Fentanyl	Caffeine, Cocaine, Etizolam, Fentanyl	1
ANPP, Caffeine, Etizolam, Fentanyl, Heroin	All Detected	1
ANPP, Etizolam, Fentanyl	All Detected	1
ANPP, Caffeine, Etizolam, Fentanyl	All Detected	4
	ANPP, Caffeine, Fentanyl	6
	Caffeine, Fentanyl	2
Caffeine, Carfentanil	Caffeine	1
Caffeine, Carfentanil, Cocaine, Etizolam, Fentanyl	Caffeine, Carfentanil, Cocaine, Fentanyl	1
	Caffeine, Cocaine, Fentanyl	1
Caffeine, Cocaine, Etizolam, Fentanyl	Caffeine, Cocaine, Fentanyl	1
Caffeine, Etizolam, Fentanyl	All Detected	9
	Caffeine	2
	Caffeine, Fentanyl	14
Caffeine, Etizolam, Fentanyl, Heroin	All Detected	1
	Caffeine, Fentanyl, Heroin	4
Caffeine, Fentanyl, Carfentanil	All Detected	2
Caffeine, Carfentanil, Etizolam, Fentanyl	All Detected	3
	Caffeine	1
	Caffeine, Carfentanil, Fentanyl	2
	Caffeine, Fentanyl	2

When the samples are arranged according to etizolam concentration as determined from a PS-MS analysis (Table 4.3), this reveals that the portable GC–MS demonstrated

no evidence of etizolam within the lowest concentration bracket (0.14–0.70%). As the concentration of etizolam increases, the reliability of detecting etizolam by GC–MS also improves. However, there remains a significant unpredictability, especially within the mid concentration range (0.70–3.0%). Above 3%, the portable instrument detected 78% of the etizolam samples that were confirmed by PS-MS. Portable GC–MS detected carfentanil at much lower concentrations (0.13–0.63%). Note that 3% by weight in a typical 100 mg (one "point") opioid sample corresponds to an etizolam dose of 3 mg, approximately 3–6 times greater than a typical therapeutic dose. Again considering a 100 mg sample, the concentration range of carfentanil detected (0.13–0.63%) translates to 130–630 μg. While dosing for opioids widely varies, this dose is equivalent to roughly 1300–6300 mg of morphine. Current prescribing practices for opioid agonist treatments (OAT) with slow-release oral morphine are in the range of 235–791 mg/day.[190] For comparison, a simple IR analysis was done on the same samples and the detection of the compounds is summarized in Table 4.2. In 3% of cases, all target substances (as detected by PS-MS) were correctly identified by IR analysis. Notably, 9% of samples containing etizolam were detected with IR. No instances of carfentanil, ANPP, or heroin were detected.

Table 4.2: Summary of detection with portable GC–MS and ATR-IR of specific target compounds as identified by PS-MS for $n = 59$ samples. The GC retention time is also listed. Compounds with an asterix were not verified against an analytical standard for portable GC–MS detection.

Compound	**Detected by PS-MS (*n*)**	**Detected by portable GC–MS (*n*)**	**GC–MS retention time / s**	**Detected by ATR-IR (*n*)**
Caffeine*	58	58	115	58
Fentanyl	58	55	180	53
Etizolam	55	20	277	5
ANPP*	15	12	148	0
Carfentanil	13	8	207	0
Heroin*	6	6	170	0
Cocaine*	4	4	134	2

Table 4.3: Breakdown of concentration of etizolam (% by weight) as identified by PS-MS and detection with GC–MS.

Concentration range reported by PS-MS (% by weight)	**Etizolam detected by GC–MS?**	**n**
0.14–0.70	False	13
	True	0
0.70–1.4	False	10
	True	4
1.4–3.0	False	9
	True	5
3.0–12	False	3
	True	11

4.6 Discussion

The ability to perform on-site MS analysis enables results to be delivered directly to people who use drugs at the time of testing and detect actives at concentrations well below the limits of detection offered by other mobile technologies, such as portable FTIR. As services begin to explore quantification with spectroscopic methods as an indication of sample strength, there is the additional concern of misleading information when potent analogues such as carfentanil are going undetected. Given the preliminary data presented here, there is evidence that portable GC–MS may address many of these concerns, and further optimization and method development is encouraged.

An important consideration is the intrinsic value of the component separation that GC provides, prior to the MS analysis. Techniques such as IR and Raman spectroscopy rely on mixture analysis to deconvolute the spectral signatures of all components present in the drug sample, significantly limiting accurate discrimination of low concentration actives like fentanyl analogues. Prior separation using either GC or LC has two advantages. First,

it simplifies the subsequent MS analysis tremendously as components can be analyzed individually, increasing the accuracy of library searching routines. Second, the retention time of sample components is characteristic of the molecule, and so identification is often possible without analysis of the mass spectrum, simply based on comparison of the retention time with reference standards. Having a mass spectrometer as the detector enables further confidence in the result, as well as immediate characterization of substances with novel retention times. The inherent trade-off is speed. Samples must first be prepared for GC–MS (dissolution in methanol, occasional centrifugation, sampling with CME fibre) and then the GC run takes on the order of 5–10 min. Additional time must be allotted for cleaning the CME fibre between runs to avoid carry-over. The GC column itself is also susceptible to carry-over from components that are strongly retained. Finally, as the sample preparation typically involves placing the sample in a solvent, the detection of components is ultimately limited by their initial solubility and later volatility.

Many of the challenges described above are characteristics of GC–MS in general, and are not specifically associated with portable instruments. Operators of drug checking services need to evaluate their capacity for operating GC–MS on-site, in terms of technician training, instrument calibration, optimization, and maintenance. The results presented above demonstrate significant variability in detecting trace components using portable GC–MS. Other projects may face the same challenges as a result of: (a) the heterogeneity of the samples, (b) the lack of analytical balances at many harm reduction sites, resulting in differences in the amount of sample measured; (c) a potentially multi-component sample matrix that results in interference for certain compounds; (d) adjustments of GC–MS parameters during performance validation, and baseline fluctuations due to cleanliness of the injection port and/or CME syringe influencing the limits of detection; and (e) inconsistent sampling volume due to the capillary mechanism in the CME syringe. Many of these challenges may be addressed with further method development, including optimization of instrument parameters (temperature programming, inlet parameters, detector settings) and exploration

of techniques such as selective ion monitoring. Sample preparation steps such as the use of internal standards, solvent extractions, or derivatization of less volatile compounds could also be explored further. It is worth noting that this process of developing reliable methods relies on both access to analytical standards, as well as technicians with a comprehensive background in GC parameter optimization and validation.

4.7 Conclusions

The illicit opioid market continues to increase in complexity and unpredictability including low levels of carfentanil and etizolam. This study illustrates how a portable GC–MS performs on street opioid samples including those containing carfentanil and etizolam. It is useful in mitigating some of the challenges associated with the detection of low concentration actives within the framework of an on-site drug testing service. MS with chromatographic separation simplifies the analysis of complex mixtures and offers increased sensitivity and specificity necessary in these cases, albeit with variable results depending on the compound. Drug checking faces ongoing challenges as an overdose intervention to both provide immediate point-of-care results (typically utilizing portable instruments) while seeking to report at extremely low-levels of detection of these notable components. Given these on-going challenges, the following chapter aims to improve on the reliability and accessibility of trace detection through the use of SERS. Here, the ability to detect etizolam at point-of-care is established.

Chapter 5

Rapid and Accurate Etizolam Detection using Surface-Enhanced Raman Spectroscopy for Community Drug Checking[1]

5.1 Overview

In British Columbia, Canada, illicit opioids have been increasingly combined with etizolam, a benzodiazepine analog, that continues to challenge popular portable drug checking technologies as it is often present in low concentrations as a result of its high potency. An unknown combination of opioids and benzodiazepines may have dangerous consequences due to unpredictable dosing, increased respiratory depression, and complicated overdose response measures. Surface-enhanced Raman spectroscopy (SERS) using a portable Raman spectrometer is used to establish a univariate model for the detection of etizolam in opioid drug mixtures ($n = 100$) obtained from the Vancouver Island Drug Checking Project, where the presence of etizolam has been determined using paper-spray mass spectrometry. Benzodiazepine immunoassay test strips are also performed on all samples for comparison. SERS is shown to detect etizolam with high sensitivity

[1]This chapter has been adapted from L. Gozdzialski, A. Rowley, S. Borden, A. Saatchi, C. Gill, B. Wallace, D. Hore. "Rapid and Accurate Etizolam Detection using SERS for Community Drug Checking" *Int. J. Drug Policy*. **102**, 103611 (2022). Acquisition of SERS spectra was performed by LG and AR. Data analysis was done by LG. PS–MS analysis was done by SB and AS.

(96%) and specificity (86%). In contrast, benzodiazepine test strips demonstrate a low sensitivity (8%) for the detection of etizolam of the same samples ($n = 100$), with only small improvements when studied over a larger subset of samples ($n = 506$, sensitivity = 29%). The potential of SERS is demonstrated for trace detection of etizolam within complex sample matrices. Since SERS is one of the few portable technologies capable of trace detection, further studies on its ability for quantification and discrimination of trace adulterants in street samples is of significant interest for point-of-care applications.

5.2 Introduction

In the past five years, community drug checking efforts have been predominantly concerned with the identification and quantification of fentanyl and fentanyl analogues—the potent opioids that have virtually replaced heroin in many areas.[191] However, more recently benzodiazepines ("benzos"), and related compounds are being mixed with opioids and presenting a new, unique challenge for the effectiveness of drug checking as an overdose response.[42] Benzodiazepine and opioid co-consumption can increase the risk of drug poisoning, complicate overdose response protocols, limit access to care, and result in life-threatening withdrawal symptoms;[42] all risks which are amplified when consumption of benzodiazepines are unknown and unexpected.

Immunoassay test strips to detect fentanyl, originally intended for detecting metabolites present in urine, were implemented earlier in the overdose crisis as a drug checking method and harm reduction strategy.[59] More recently, coinciding with the increased adulteration of opioids with benzodiazepines and counterfeit Xanax, immunoassay test strips have also been used for the detection of benzodiazepines.[42,70] Despite community drug checking services shifting towards using technologies capable of more comprehensive, qualitative and quantitative information, test strips continue to be employed alongside many portable spectroscopy-based instruments in attempts to mitigate the limitations of those instruments with lower sensitivity (e.g. Fourier-transform infrared (FTIR) and Raman spectrometers).

Test strips, however, have clear limitations, such as their lack of selectivity that prevents discrimination between structurally similar compounds. At the same time, the limit of detection can vary significantly for different benzodiazepines, leading to false negatives. A notable challenge has been reported for the detection of etizolam, a benzodiazepine analog increasingly present in opioids and counterfeit Xanax tablets in BC.[42,43] Attention has recently been drawn to etizolam due to its prevalence in the supply, its complicating role in overdose responses, as well as the unreliability of detection using popular point-of-care drug checking technologies FTIR spectrometers and benzodiazepine immunoassay test strips.[42,43] In the case of FTIR, this unreliability is mostly attributed to the well-known challenge of detecting substances at concentrations less than approximately 5% within mixtures, as well as significant overlap of the infrared fingerprint of etizolam with that of common co-occurring substances and cutting agents such as fentanyl and caffeine.

Surface-enhanced Raman spectroscopy (SERS), is a method in which the analyte of interest is detected with a sensitivity that can be orders of magnitude greater than that achievable with conventional Raman on powder samples, thereby providing the potential for trace detection. This sensitivity is achieved by placing the sample in contact with metal nanoparticles or rough metal surfaces. SERS has recently been applied in the identification of fentanyl[108,168–170] and analogues such as carfentanil,[49] and the performance has been compared to that of fentanyl immunoassay test strips.[26] Previous work has also demonstrated SERS effectiveness in various applications for benzodiazepine detection and differentiation in beverages,[192] as well as identifying minor components among a high concentration of excipients such as in pressed tablets.[193] This suggests that further investigation into SERS for benzodiazepine identification in opioid samples may be worthwhile. However, many challenges with SERS are well-documented, such as the competitive absorption of substances present in a complex mixture (e.g. preferential adsorption of non-target molecules), low-affinity of some molecules to SERS substrates, poor signal reproducibility, distribution regions with intense signal enhancement (known as

"hot spots"), and variation in SERS substrates leading to a lack of standardized libraries.[194] Because of these limitations, few assessments have been done on the employment of SERS procedures on relevant real-world drug samples, despite the large body of work done on simulated lab standards.

This short report demonstrates the feasibility of SERS for the detection of etizolam in street opioid samples amongst a complex and relatively unknown sample matrix. The superior performance of SERS is demonstrated when compared to etizolam detection with benzodiazepine test strips. Presentation of these results conclude with a discussion on the potential of SERS as a rapid and robust point-of-care drug checking instrument, filling a gap in point-of-care drug checking, particularly for communities where complex opioids with trace components are prevalent.

5.3 Methods

Samples were obtained as part of a drug checking service operating in Victoria, BC since 2018. A small amount (typically less than 10 mg) of drug sample was submitted by community members and analyzed with IR spectroscopy, Raman spectroscopy, and fentanyl and benzodiazepine immunoassay test strips. Results from these technologies are delivered at point-of-care. Subsequent off-site testing was performed using PS-MS to detect target compounds with greater sensitivity.[133] Benzodiazepine test strips were performed on opioid samples ($n = 509$) received at the drug checking service from Nov 2020–Jul 2021; all samples were analyzed with PS–MS. A subset of samples that received PS-MS testing were also selected for testing using surface-enhanced Raman spectroscopy (SERS). For this study, $n = 100$ opioid samples were selected where $n = 50$ samples did not contain etizolam and $n = 50$ contained etizolam as indicated by PS-MS. All samples ($n = 100$) contained fentanyl as well as a number of other psychoactive components and cutting agents (e.g caffeine, cocaine, heroin, ANPP, sugars (Table 5.2)).

5.3.1 SERS

SERS measurements were recorded using a portable Raman spectrometer (Resolve, Agilent Technologies, Santa Clara, USA). A modified procedure of the steps illustrated in Agilent's Resolve "Trace Test" acquisition was used. Drug samples were first prepared into a 2.14 mg/mL solution in water and heated until dissolved. 70 μL of the sample solution was added to 1.42 mL of a 50 nm gold nanoparticle solution (BBI Solutions) and shaken for approximately 30 s. Following, 10 μL of a 1.0 M solution of $MgSO_4$ was added as an aggregating agent for a total volume of 1.5 mL. The solution was again shaken for approximately 10 s prior to acquisition. Spectra were recorded with Stokes shifts between 200–2000 cm^{-1} and automated baseline, fluorescence, and cosmic ray correction procedures were applied internally by instrument software. Each sample solution was measured five times and the five corrected spectra were averaged for subsequent analysis.

5.3.2 Benzodiazepine Immunoassay Test Strips

Benzodiazepine immunoassay test strips (Rapid Response, BTNX, Markham ON) with a 300 ng/mL cut-off, established using oxazepam as a calibrator, were used on all drug samples. A small amount of drug sample (1–2 mg, more if available) was placed in an Eppendorf tube with approximately 1 mL water. The tube was agitated, either manually or with a miniature vortex mixer, until dissolved (upwards of 30 s). The test strips were used and interpreted according to manufacturer instruction. A double line was recorded as a negative result and a single line indicated a positive result.

While a controlled concentration (2.14 mg/mL) of sample was used for the SERS study, the test strips were applied to a more approximate sample concentration (1–2 mg/mL). In retrospect, it would have been better to have the concentrations controlled for both tests. It was initially intended to study SERS exclusively in this application. However, later realized that, given the more widespread use of benzodiazepine test strips in drug checking services, it would be of interest to include some statistics about the success of the test strips in the

service with regards to etizolam.

5.3.3 PS-MS

For this study, all samples were sent off-site for testing using a TSQ Fortis^{TM} triple quadrupole mass spectrometer and a VeriSpray^{TM} PaperSpray ion source (Thermo Fisher Scientific, San Jose CA, USA).[134] Details of the instrument operation, including calibration and data analysis are described in previous publications[34,50,121,133,134] and introduced in Chapter 2. Briefly, 1.3 mg of the sample was dissolved and vortexed in 1.3 mL methanol, from which 1 μL (volume adjusted, if necessary) was added to isotopically labeled standards in 200 μL of methanol. The VeriSpray^{TM} sample plate was spotted with 10 μL of this spiked solution for subsequent MS analysis. The plates were air dried at ambient conditions prior to analysis, typically for 2–5 h during transport from the sample collection site to the analytical lab. For this study, this procedure was used as secondary testing to identify the presence of etizolam in opioid samples. A positive or negative result (etizolam vs no etizolam detected) was considered the "true" result for comparison with SERS and benzodiazepine test strip results.

5.4 Results

The breakdown for benzodiazepine test strip analysis on relevant opioid samples ($n = 509$) is shown in Table 5.1, where a sensitivity (true positive rate) for detection of etizolam is determined to be below 30%. The development of a simple univariate model for the detection of etizolam using SERS spectral features is outlined in Figure 5.1. The absolute intensity of a characteristic peak for etizolam at frequency 1491 cm^{-1} is used to indicate the presence ("Positive" result) or absence ("Negative" result) of etizolam given a set threshold based on peak intensity. This threshold was determined by simultaneously maximizing true positive and minimizing false negative rates. The summary of this analysis in regards to the true and predicted values, sensitivity, and specificity is shown in Table 5.1, and

highlights the results of the benzodiazepine test performed on the same subset of samples. The separation of spectral data in principal component space (Figure 5.3) suggests subtle features may be useful for more robust and comprehensive discrimination of multiple components when required for more complicated samples. While fentanyl detection with

Table 5.1: Summary of the sensitivity (true positive rate) and specificity for detection of etizolam with benzodiazepine ("benzo") test strips on opioid samples ($n = 506$) where fentanyl was present and benzo test strip tests were performed (completed Nov 2020–Jul 2021). Results are also shown for the subset tested using the method as described for SERS ($n = 100$), compared to the test strip data on those same samples. In all cases, the true label is determined using PS-MS. *Note that in two of the samples which underwent testing with SERS the benzo test strip data was absent and therefore $n = 98$ in this case.

	benzo test strips on opioid samples ($n = 509$)	**benzo test strips on subset** ($n = 98$)*	**SERS on subset** ($n = 100$)*
true positive (n)	60	4	48
true negative (n)	248	48	43
false positive (n)	29	0	7
false negative (n)	164	46	2
sensitivity (%)	26.8	8.0	96.0
specificity (%)	89.5	100.0	86.0

SERS was not a focus of this paper, notably the strongest characteristic fentanyl SERS peak (1000 cm^{-1}) was also visible in all spectra as shown in Figure 5.1a. This reveals the possibility for the simultaneous detection of fentanyl and etizolam using SERS. In these cases, it was noticed that the relative ratio of the main characteristic fentanyl peak (1000 cm^{-1}) and etizolam peak (1491 cm^{-1}) followed a general trend with their relative concentrations as shown in Figure 5.2. Notably, the two misclassified etizolam samples both had very high concentrations of fentanyl (>50%) and low concentrations of etizolam (<1%), as reported by PS-MS. The SERS method detected etizolam at concentrations between 0.17–9.63%, as determined by PS-MS. Various steps were completed to optimize the univariate model used for the detection of etizolam, shown in Figure 5.1. In addition, several supporting figures reveal the exploration of the extension of simple univariate

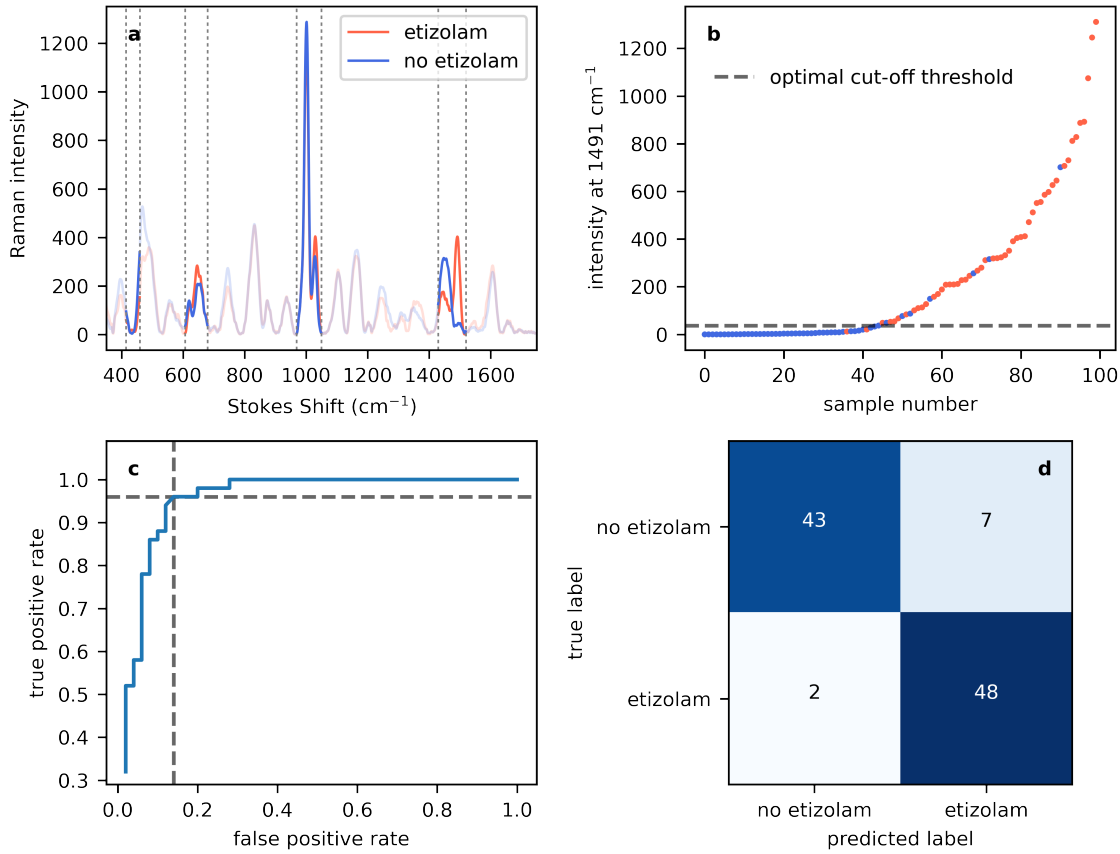

Figure 5.1: (a) Overlay of the average SERS spectrum of opioid samples containing etizolam verses samples not containing etizolam. Regions of interest are highlighted, most notably the prominent peak at 1491 cm^{-1} present in etizolam-containing samples. (b) Optimization of threshold for selection of peak height cut-off for univariate model using the intensity of the main etizolam peak at 1491 cm^{-1}. (c) Receiver operating characteristic (ROC) curve for various cut-off intensities. (d) Confusion matrix using a threshold peak height.

classification model to multivariate classification (Figure 5.3) that considers the entire SERS spectrum.

Table 5.2 breaks down the additional compounds that were identified using PS-MS for the subset of opioids samples studied with SERS ($n = 100$). In addition to the compounds listed in Table 5.2, cutting agents, mannitol ($n = 19$), erythritol ($n = 13$), and inositol ($n = 1$), were also tentatively detected using IR absorption spectroscopy in several samples. No trend was observed for interfering compounds in the SERS study to account for false positive or negative etizolam detection.

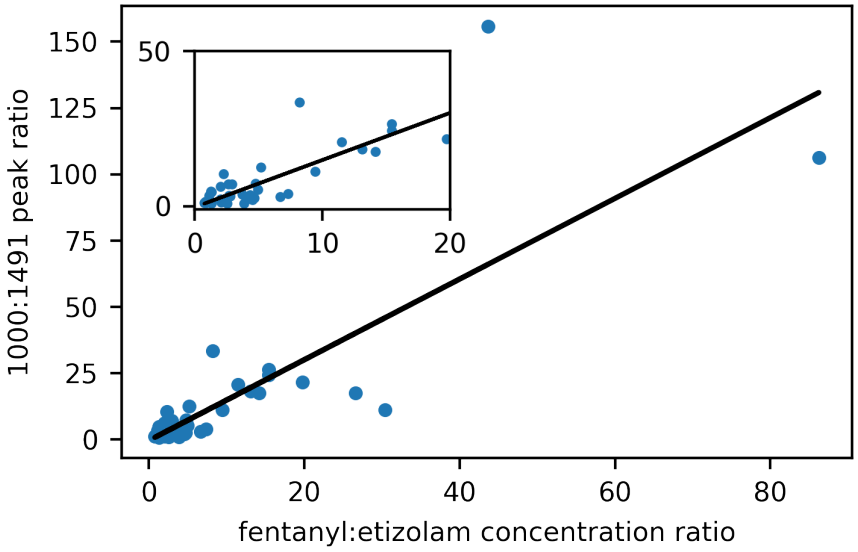

Figure 5.2: Intensity ratio of the characteristic fentanyl peak (1000 cm^{-1}) to the characteristics etizolam peaks (1491 cm^{-1}) plotted against the relative concentrations as determined by off-site PS-MS. The inset figure highlights a narrower range of the same plot where majority of the data points lie.

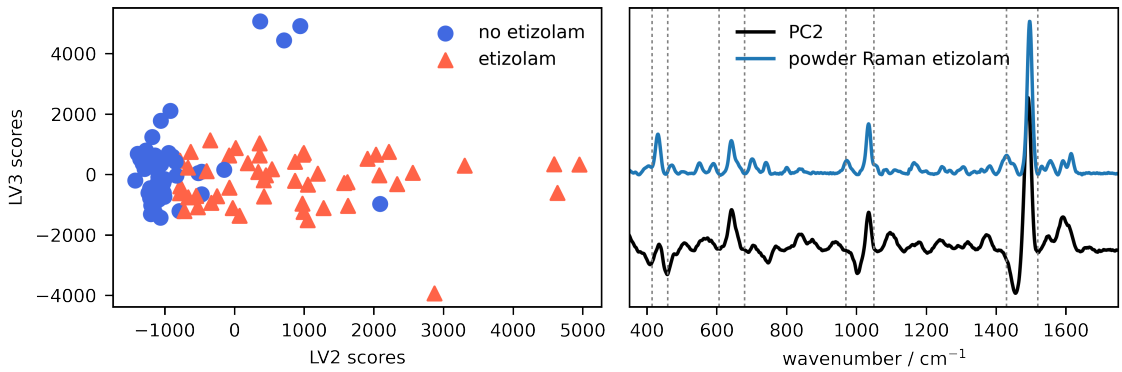

Figure 5.3: (a) Latent variable (LV) 2 and 3 plotted as calculated by PCA on the subset of opioid samples. Variation of spectra based on etizolam content appears to be represented predominantly within LV2. (b) Principal component (PC) 2 is shown in comparison to a Raman library etizolam entry. Prominent features attributed to etizolam are shown within the same regions of interest as shown in Figure 5.1a.

Table 5.2: Summary of additional compounds detected using PS-MS targeted method on the subset of samples used for the SERS study ($n = 100$).

Additional compounds identified by PS-MS ($n = 100$)
Fentanyl ($n = 100$)
Caffeine ($n = 96$)
ANPP ($n = 18$)
Carfentanil ($n = 8$)
Cocaine ($n = 7$)
Heroin ($n = 6$)
Acetyl Fentanyl ($n = 1$)

5.5 Discussion

Etizolam was successfully detected in opioid samples using SERS with high specificity (86%) and sensitivity (96%), when compared to the performance with benzodiazepine test strips. Benzodiazepine test strip implementation followed fentanyl immunoassay test strips which have largely and successfully been re-purposed and distributed for drug checking. However, drug checking projects, particularly with portable instruments like IR and Raman spectrometers, are only now beginning to recognize the limitations when it comes to trace detection of benzodiazepines in opioids using test strips. The popularity of benzodiazepine test strips in the community risks misinterpreting a negative result on the strip tests as an indication of the absence of all benzodiazepines and benzodiazepine-related substances, including etizolam. Our data supports these concerns, given the prevalence of etizolam within the drug supply. SERS presents a reliable alternative for low-concentration benzodiazepine detection and uses the same instrumentation already employed in many drug checking applications using Raman spectroscopy.

The detection of benzodiazepines is important for informing harm reduction practices. While the effect of any drug highly depends on many factors such as tolerance, age, and body weight, a 1% by weight etizolam concentration in a 100 mg (one “point”) opioid sample corresponds to 1 mg etizolam which is 1–2 times an approximate typical therapeutic

dose. However, The concomitant use of opioids and benzodiazepines is known to result in both rapid and prolonged respiratory depression, which is the primary mechanism of opioid overdose fatalities.[195] In addition, the immediate remedy for reversing opioid overdoses—naloxone—is ineffective on reversing the respiratory suppression due to benzodiazepines. Therefore, an overdose response to someone who has consumed both benzodiazepines and opioids can present as more complicated and unpredictable than cases where opioids alone are used.[195] Even in cases where overdose does not occur, it is also of concern that in absence of a safe supply people might be unaware that they are consuming benzodiazepines mixed into other drugs, including fentanyl, and it is noted that fatal seizures can result from sudden withdrawal of benzodiazepines.[195] Regardless of whether people choose to consume polysubstances, it is of upmost importance that they are aware of what is present in the drug supply to inform this use.

The observed trend of fentanyl/etizolam SERS intensity ratios with their relative concentrations will be further assessed to determine the application for quantification and reveal any matrix effects from other components in common opioid samples. Pursuit of multivariate approaches may be useful for improved specificity in detection and accuracy in quantification models. Exploration within the family of SERS techniques, such as the use of SERS chips, for targeted applications may also be considered. Further work for quantification and discrimination between other benzodiazepine-related substances using SERS will require more comprehensive studies and data processing schemes.

To date, the detection of fentanyl has been a priority for drug checking responses to the overdose crisis. However, there is an increasing need for more comprehensive results to respond to the escalating complexities in illicit opioids including the presence of benzodiazepines in expected opioids. Drug checking instrumentation and method development faces ongoing challenges with different instruments having unique benefits and limitations. Questions remain on how best to implement drug checking within public health's overdose responses and which instruments to pursue.

5.6 Conclusions

This chapter demonstrates the potential for a rapid and sensitive SERS method for use in community drug checking for the detection of etizolam within opioid mixtures, despite the complex and broad matrix found in street drug samples. A simple univariate model for a characteristic Raman peak was shown to be effective for the detection of etizolam. The detection of polysubstances in drugs, along with the capacity to report potency, is important for the safety of people who use drugs and an informed response to overdoses.

Ultimately, drug checking aims to provide a more complete characterization of a sample. The work shown so far—the quantification of fentanyl using Raman instrumentation (Chapter 3), the identification of low concentration fentanyl analogues, such as carfentanil, using portable GC–MS(Chapter 4), and the detection of trace levels of etizolam using a simple SERS protocol—suggests that multiple instruments may be able to fill in various pieces of this characterization. The following chapter presents a multi-instrumental approach to drug checking based on these results, and discusses the consequences for various types of drug samples.

Chapter 6

Characterizing Street Drug Mixtures using Multiple Technologies[1]

6.1 Overview

Drug checking is increasingly being explored outside of festivals and events to be an ongoing service within communities, frequently integrated within responses to illicit drug overdose. The choice of instrumentation is a common question and the demands on these chemical analytical instruments can be challenging as illicit substances may be more complex and include highly potent ingredients at trace levels. The answer remains nuanced as the instruments themselves are not directly comparable nor are the local demands on the service, meaning implementation factors heavily influence the assessment and effectiveness of instruments. In this perspective, we provide a technical but accessible introduction to the background of a few common drug checking methods aimed at current and potential drug checking service providers. We discuss the following tools that have been used as part of the Vancouver Island Drug Checking Project in Victoria, Canada: immunoassay test strips, attenuated total reflection IR-absorption spectroscopy, Raman spectroscopy from powder samples, surface enhanced Raman scattering in a solution of colloidal gold nanoparticles, and gas chromatography–mass spectrometry. Using four different drug mixtures received

[1]Some of the content in this chapter appeared in L. Gozdzialski, B. Wallace, D. Hore, "Point-of-Care Community Drug Checking Technologies: An Insider Look at the Scientific Principles and Practical Considerations". *Harm Reduct. J.*, **20**, 39 (2023).

and tested at the service, we illustrate the strengths, limitations, and capabilities of such instruments, and expose the scientific theory to give further insight into the answers. Each case study provides a walk-through-style analysis for a practical comparison between data from several different instruments acquired on the same sample. Ideally, a single instrument would be able to achieve all of the objectives of drug checking, however, there is no clear instrument that ticks every box; low cost, portable, rapid, easy-to-use and providing highly sensitive identification and accurate quantification. Multi-instrument approaches to drug checking may be required to effectively respond to increasingly complex and highly potent substances demanding trace level detection and the potential for quantification.

6.2 Introduction

There are significant challenges in understanding the pros, cons, and considerations for particular technologies within unique harm reduction sites, and communicating such information within a service. Further challenges arise in recognizing and predicting the ultimate potential of certain technologies given additional research and development. Recent reviews have provided a comprehensive overview of several analytical instruments and have discussed their applicability to harm reduction drug checking.[18,38,196,197] The value and, in some cases, necessity, of a multi-instrument approach to drug checking has also been recognized.[197] Yet, the right combination of instruments requires a comprehensive needs assessment that depends on many factors and stakeholders.[14,37,196,198] As drug checking expands within harm reduction interventions and as an active field of research, there is a growing need for a more in-depth explanation of the underlying technologies within this context. Furthermore, the traditional roles of operating instruments and interpreting data (i.e. technician role) and communicating results (i.e harm reduction role) are rapidly becoming blended.[76,199] Such knowledge provides the missing link between theory and practice that can best ensure service quality at drug checking sites, contribute to the body of practice-based evidence informing drug checking, and advance drug checking

research through experiential learning.[199]

This perspective is aimed at the modern drug checking service provider who has developed considerable expertise in their craft; performing measurements, interpreting data, and presenting results embedded in the principles of harm reduction.[200,201] The following work presents four case studies of different drug mixtures received and tested at the service. Here a step-by-step walk-through is provided for a practical comparison of data from several different instruments acquired on the same samples. Points highlighted throughout the examples are centred around common questions regarding the strengths, limitations, and possibilities of such instruments.

6.2.1 Methods

Sample selection. Since 2018, the Vancouver Island Drug Checking Project has operated a free and confidential service in Victoria, Canada. Prior to 2021, every sample presented was analyzed using immunoassay test strips, attenuated total reflection (ATR) IR-absorption spectroscopy, Raman spectroscopy from powder samples, surface enhanced Raman scattering (SERS) in a solution of colloidal gold nanoparticles, and GC–MS—all with portable instruments. In operating the service, we aim to analyze the data in order to report the components present in the drug mixture and, where possible, offer some information on the strength of the drugs. At the same time, we aim to evaluate these portable technologies in the context of community drug checking. We purposively selected four samples of drug mixtures that best illustrate the various degrees of challenge in the analysis with multiple instruments. The samples include a methamphetamine sample containing dimethylsulfone, a cocaine sample containing phenacetin and levamisole, and two different opioid samples. Although samples such as these four repeatably show up at the drug checking service, the selection was not based on their prevalence. Rather, the data was chosen as it most clearly demonstrates many of the principles we have introduced within the background sections and aims to connect concepts such as overlapping IR

signals, fluorescence interference in Raman spectroscopy, tailing on GC–MS, etc., to the real application of drug checking as a public health response.

6.2.2 Results

6.2.2.1 Methamphetamine & Dimethylsulfone

The first example is that of a drug mixture that contains only a single active as the major component and a smaller amount of a single cutting agent, with a negative fentanyl test strip result. Figure 6.1a shows the ATR–IR spectrum of the drug mixture (black trace) along with the library entry of methamphetamine (blue) that was identified based on its score using a correlation coefficient. Visually, one can see that the spectrum of the mixture closely resembles the spectrum of pure methamphetamine, implying that the drug is primarily methamphetamine. A weighted subtraction of the methamphetamine spectrum enables the residual to be compared against library entries again, this time producing a hit for dimethylsulfone (MSM, purple trace). Although there is little visual evidence for MSM in the spectrum of the mixture, this accounts for some of the broad features near 1300 cm^{-1} and the sharper feature near 925 cm^{-1}. A similar analysis was carried out by Raman spectroscopy from the crystalline sample. As mentioned previously, although IR and Raman spectroscopy probe the same characteristic molecular vibrations, the relative intensity of the peaks in the spectra is often different thereby providing a complementary fingerprint. For example, the sharp MSM feature with a Stokes shift just above 700 cm^{-1} (purple trace in Figure 6.1b) is more apparent in the Raman spectrum, even at the same concentration. The SERS spectrum is shown in Figure 6.1c. As a result of the variability of the SERS spectra, and their specificity for the detailed nature of the substrate, aggregating agent, and solution conditions, SERS libraries need to be tailored for the specific device and measurement protocol. However, it is common to interpret SERS spectra qualitatively in comparison with Raman library entries. One challenge associated with solution-based SERS is the large background associated with additives such as the citrate stabilizer that

accounts for many of the modes seen in Figure 6.1c. However, one can see the emergence of the 1000 cm^{-1} methamphetamine band. The GC–MS data is first presented in the form of a chromatogram using the MS total ion count (TIC) in Figure 6.1d. Several peaks are visible with retention times up to 120 s. Integrating the MS in small windows (approximately ± 2 s) around each peak enables the average MS to be displayed an analyzed in comparison to MS library entries. The MS serve as definitive compound identifications, although they are not strictly required in cases where the characteristic retention times for species are already known. The MS for the 70 s peak is shown in Figure 6.1e (blue trace, top), and has a good match to the methamphetamine library spectrum (red trace, bottom). The MS obtained by integrating the 47 s peak is shown in Figure 6.1f (purple, top), and agrees with the MSM library entry (red, bottom). Returning to the TIC chromatogram in Figure 6.1d, one notices that additional peaks are present. For example, the feature at 55 s has been tentatively identified as benzyl chloride, a suspected breakdown product of methamphetamine or a leftover precursor from the methamphetamine synthesis. As noted, molecular degradation (and the resulting mass fragments observed) due to thermal breakdown is a common occurrence with GC–MS and generally posses additional challenges in the interpretation, but may be helpful as a further indication of a component of interest. One also notices the prominent signal that ends with a sharp edge at 100 s; such tailing is commonly observed for methamphetamine however adds the challenge of potentially obscuring peaks at longer retention times.

6.2.2.2 Cocaine & Phenacetin

The second example is that of a drug mixture that contains primarily cocaine and phenacetin, again with a negative fentanyl strip test. The ATR–IR spectrum is shown in Figure 6.2a (black trace), along with the top library hit of cocaine (blue trace), and the library spectrum of phenacetin (purple) that was obtained by searching the residual following cocaine subtraction. Evidence of phenacetin in several places in the drug mixture, including the two peaks near 820 cm^{-1}, and two peaks around 1650 cm^{-1}.

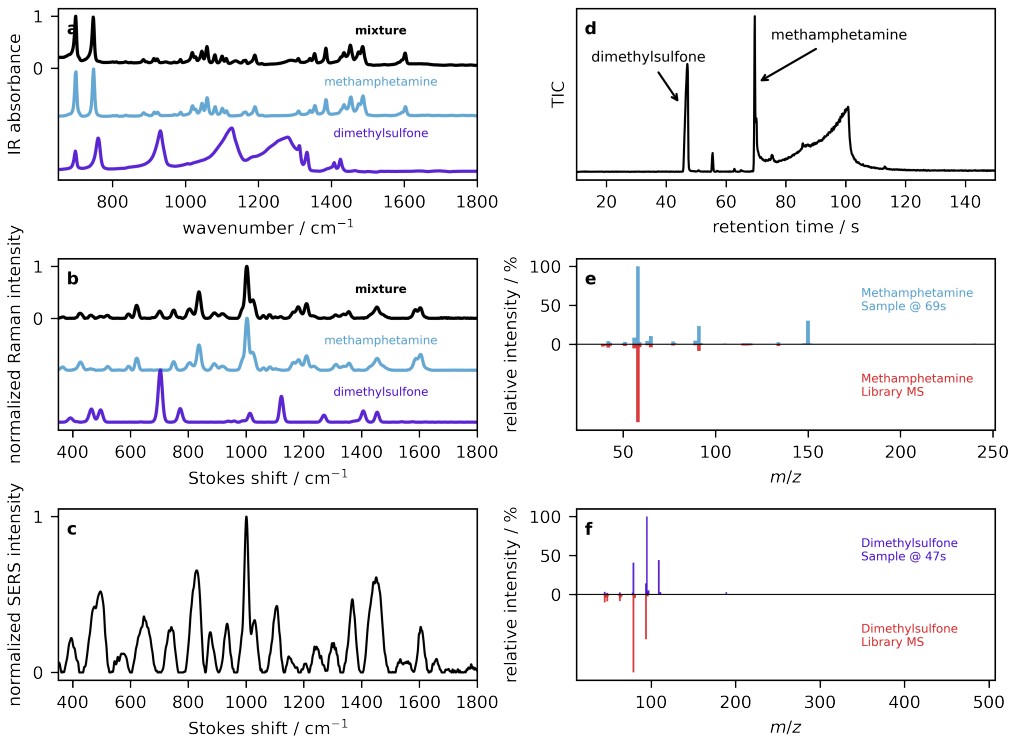

Figure 6.1: (a) ATR–IR reflection absorbance spectra of a sample containing methamphetamine (black), overlaid with library spectra of pure methamphetamine and dimethylsufone. (b) Raman spectra and the corresponding library entries. (c) SERS spectrum of the drug mixture. (d) The GC total ion count chromatogram is presented, along with mass spectra of (e) methamphetamine and (f) dimethylsulfone in the top panels, and library MS entries below in red.

Although the fingerprint region of phenacetin is more complicated than that of MSM in the previous example as a result of the larger molecular size, this also provides additional details for comparison in the library searching. In the Raman spectra (Figure 6.2b), phenacetin peaks are again apparent, this time in the sharp feature at 1340 cm^{-1}, and the broad 1620 cm^{-1} shoulder on the sharp 1600 cm^{-1} cocaine peak. The SERS spectrum shown in Figure 6.2c displays enhancements at 650 cm^{-1} and 1000 cm^{-1} attributed to cocaine.[202] The largest peaks in the TIC chromatogram (Figure 6.2d) at 133 s and 105 s have mass spectra corresponding to cocaine (Figure 6.2e) and phenacetin (Figure 6.2f). Investigation of minor peaks in the TIC chromatogram reveal that the feature eluding at 122 s has a mass spectrum that matches levamisole (Figure 6.2g), a medication used to

treat parasitic worm infections, but can have detrimental health effects resulting in serious infection with chronic use and high doses.[203] Levamisole historically has been used as a popular cutting agent in cocaine.[203] It is also noted that cocaine has many synthetic byproducts that appear in the TIC chromatogram.

6.2.2.3 Simple Opioid Mixture

The third example is an opioid sample with positive fentanyl test strip and negative benzodiazepine test strip results. The IR absorption (Figure 6.3a) and Raman (Figure 6.3b) spectra both show strong correlations with caffeine, the major component in many down samples. Both spectra have minimal visual evidence of fentanyl. However, fentanyl features (most prominently 705 cm^{-1} in the IR and 1000 cm^{-1} in the Raman) can be picked up by library searching. When those interpreting the spectral data have significant domain-knowledge, such well-known peaks are located visually to support library searching algorithms. As opioid samples are commonly presented for analysis in a community drug checking service, we have also developed quantification models based on PLS regression that can be used to estimate the fentanyl concentration.[20,24] As a safeguard, a novelty detection algorithm can be applied to ensure that the sample has sufficient similarity to the training data used in the development of the models. When our models[20] are applied to the IR spectrum (black trace in Figure 6.3a) we obtain a fentanyl concentration of 12% by weight. Applying the Raman model[24] to the data (black trace in Figure 6.3b) results in a concentration of 7%. The SERS spectrum shown in Figure 6.3c has the same enhancement of the 1000 cm^{-1} mode as seen in the previous two examples, but this time attributed to fentanyl. It is well known that many molecules produce strong Raman signals at 1000 cm^{-1} so this peak can be diagnostic even though it lacks specificity. In other words, SERS may be useful for determining low concentration actives, but one cannot distinguish fentanyl or cocaine from methamphetamine based on this peak alone. The GC–MS total ion chromatogram (Figure 6.3d) displays two prominent peaks with constituent mass spectra that are well matched to fentanyl HCl at 179 s (Figure 6.3e) and caffeine eluding at 122 s

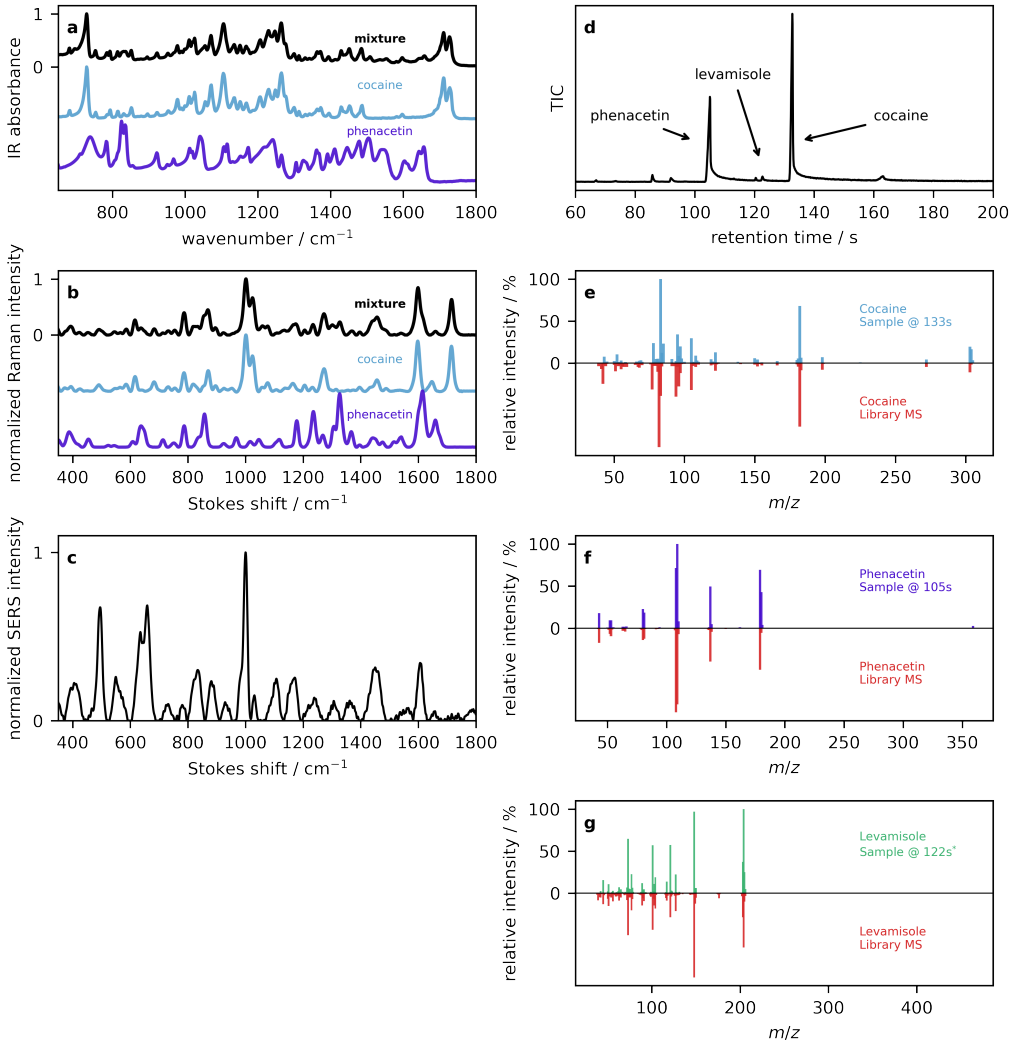

Figure 6.2: (a) ATR–IR reflection absorbance spectra of a sample containing cocaine (black), overlaid with library spectra of pure cocaine and phenacetin. (b) Powder Raman spectra and the corresponding library entries. (c) SERS spectrum of the drug mixture. (d) The GC total ion count chromatogram is presented, along with mass spectra of (e) cocaine, (f) phenacetine, and (g) levamisole in the top panels, and library MS entries below in red.

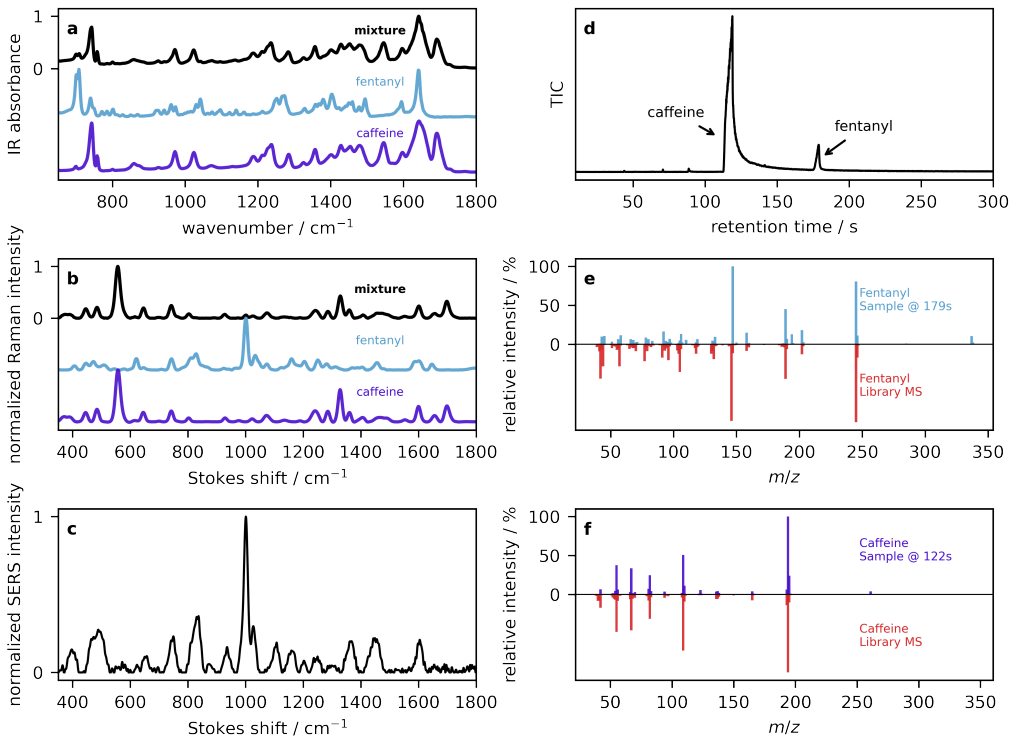

Figure 6.3: (a) ATR–IR reflection absorbance spectra of an opioid sample (black), overlaid with library spectra of pure fentanyl and caffeine. (b) Powder Raman spectra and the corresponding library entries. (c) SERS spectrum of the drug mixture. (d) The GC total ion count chromatogram is presented, along with mass spectra of (e) fentanyl and (f) caffeine in the top panels, and library MS entries below in red.

(Figure 6.3f). This helps confirm the specific fentanyl analogue that may be ambiguous based on the test strip results and IR or Raman spectra. In this example, spectroscopic instruments IR and Raman were able to rule-in the correct components (e.g fentanyl or analogue and caffeine). However, communicating such result based on those instruments alone remains complex due to the inability to rule-out certain trace compounds. While the more sensitive techniques of SERS and GC–MS do not contribute a significant amount of additional compounds to the interpretation, the ruling out of certain compounds such as carfentanil and benzodiazepines is an invaluable piece of information.

6.2.2.4 Complex Opioid Mixture

The last example is that of a challenging opioid sample with a positive fentanyl strip test, and negative benzodiazepine test. Looking at the IR (Figure 6.4a) and Raman (Figure 6.4b) spectra, the visual appearance and quantitative library match to caffeine is high in both cases, but residual searching does not produce any convincing additional results. Although IR and Raman have been shown to be capable of detecting fentanyl or other compounds at the 1–2% level, in this case, there are complicating factors that interfere with the analysis. In the IR spectra, a telltale sign of fentanyl at low concentrations (below reliable quantification) is a small peak that appears at 705 cm^{-1}. The spectra in Figure 6.4a have their caffeine peaks shifted in that region, attributed to the moisture content of the sample.[204] In the case of the Raman, this sample contained a high level of background fluorescence that degraded the signal-to-noise. However, the SERS spectrum in Figure 6.4c shows fentanyl at 1000 cm^{-1} and another peak at 1500 cm^{-1} that is characteristic of etizolam and not present in the previous example where only caffeine and fentanyl were identified.[25,192] Confirmation of low concentration actives comes from the GC–MS data. The total ion chromatogram in Figure 6.4d displays a large caffeine peak eluting at 122 s, a small but prominent peak at 181 s, and some minor baseline fluctuations that are barely noticeable and would therefore not be picked up in an untargeted analysis. Targeted analysis can be performed by looking for a specific mass fragment, and therefore species, of interest. This is called a reconstructed ion chromatogram (RIC). The inset to Figure 6.4d shows RIC traces for the 245 m/z fentanyl, 303 m/z carfentanil, and 342 m/z etizolam ions. The corresponding MS data obtained by integrating these peaks as shown in Figure 6.4e–g. In this case, less sensitive spectroscopic methods were not sufficient to identify compounds in low concentrations that are essential to people who use opioids. Some points regarding the test strips are warranted. First, the lack of specificity to a large number of fentanyl analogues means that a positive result can indicate a wide range of potency based on fentanyl analogues alone. Second, a false negative screen for benzodiazepine-related

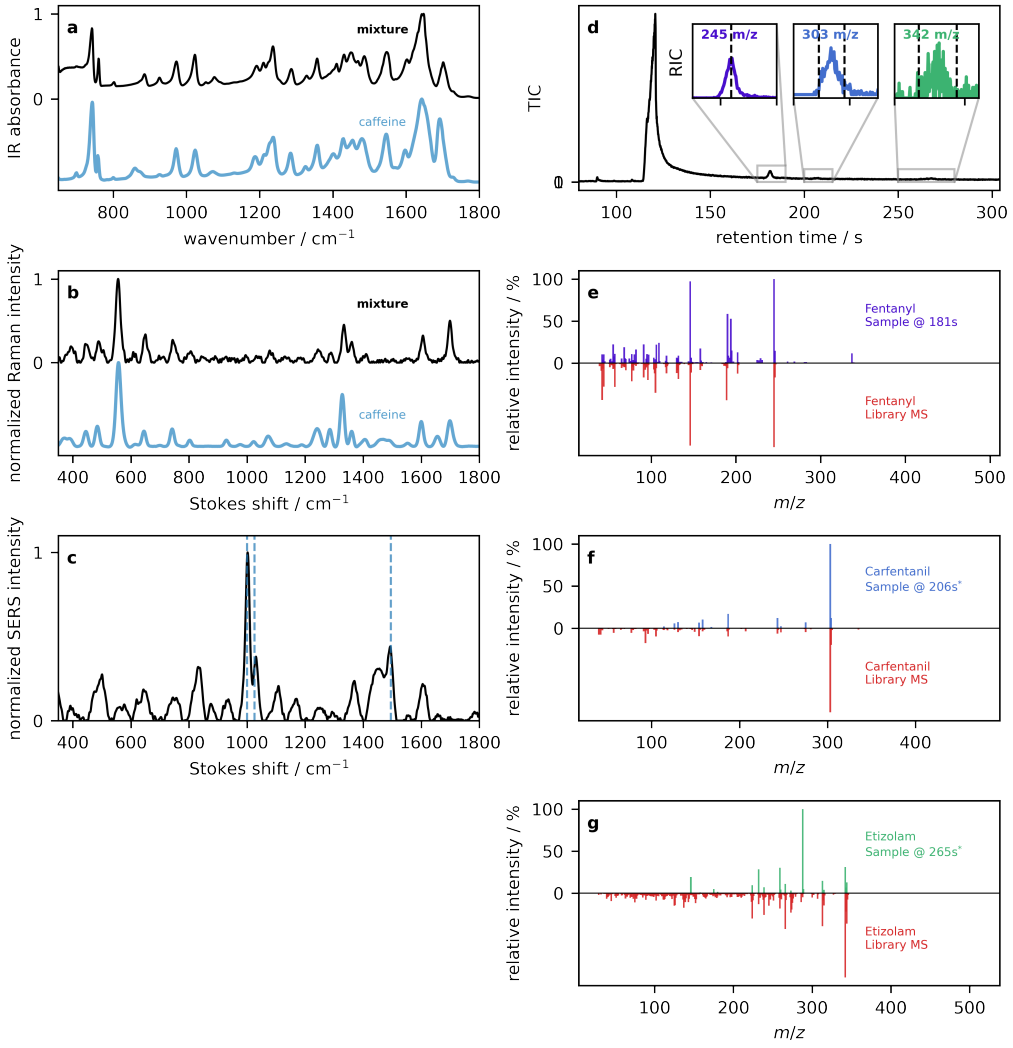

Figure 6.4: (a) ATR–IR reflection absorbance spectra of an opioid sample (black), overlaid with the library spectrum of caffeine. (b) Powder Raman spectra and the corresponding library caffeine spectrum. (c) SERS spectrum of the drug mixture. (d) The GC total ion count chromatogram is presented, along with mass spectra of (e) fentanyl, (f) carfentanil, and (g) etizolam in the top panels, and library MS entries below in red.

compounds is common when etizolam is present in a drug mixture.[25,70] As mentioned previously, this lack of positive strip test in the presence of etizolam (i.e false negative) can mostly be attributed to the structural diversity within benzodiazepines and benzodiazepine-related substances targeted with the test strips; some molecules will have poorer affinity for the immunoassay than others and therefore higher cut off limits.[70] In this case of a complex opioid mixture, IR and Raman alone are insufficient to account for all components and more sensitive technologies like SERS and GC–MS are needed for a full diagnostic.

6.3 Discussion

Prior qualitative research from the authors has confirmed that drug checking implementation will be limited unless designed to specifically address the barriers of criminalization and stigmatization.[205] The benefits of drug checking must outweigh the risks of accessing the services.[205] As a harm reduction intervention, we argue that drug checking holds potential for impact beyond an individual level, on the unregulated drug market as well as at the community and policy level.[16,206] However, public health and harm reduction organizations lack clear guidance to inform decisions on which instruments to purchase.[207] They currently do not have sufficient opportunities to assess what the operation of such instruments would actually look like in practice, with numerous potential benefits and limitations to consider. Indeed, researchers studying the implementation of drug checking in a Boston harm reduction service coined the situation to be the "Bronze Age" of drug checking.[76] We concur and contribute new discussions addressing current questions based on practice-based evidence from the Vancouver Island Drug Checking Project in Canada. Our four case studies illustrate three significant issues; the present challenges and opportunities for drug checking instrumentation, the crucial need to communicate uncertainty and limitations, and the beneficial role for confirmatory checking when utilizing more affordable portable instruments.

6.3.1 Challenges and Opportunities for Multi-Instrument Drug Checking

In the context of increasingly complex substances and unprecedented levels of illicit drug overdose, there are significant advantages to a multi-instrument approach to drug checking.[37,197,208] Using more than one method for drug identification has advantages in terms of increased confidence when a particular component is detected by more than one technique, and increased awareness when something is detected by at least but perhaps only one technique. The first three examples have illustrated scenarios where any of the methods employed are able to detect all of the major components. This is the case when the drugs are relatively uncut or the components are present in high concentrations as in the case of methamphetamine and dimethylsulfone. The cocaine and phenacetin example is similar, except that trace levamisole was not detectable by infrared or Raman spectroscopy. From this perspective, a rapid optical characterization using IR or Raman would be sufficient to confirm that the class of drug suspected is correct. This is generally true for opioid samples as well, as their major components caffeine and/or sugars are easy to detect. The challenge arises when actives are present at levels below the limit of quantification, or even below the limit of detection, of IR or Raman spectroscopy. In the most challenging situation of low-concentration actives, all portable instruments are limited in their detection. However, even in those cases, a few subtle clues from multiple sources may be enough to inform harm reduction messaging and practices. For opioid samples, such challenges are often encountered for fentanyl alone, and almost always encountered in the case of highly-potent fentanyl analogues such as carfentanil or adulterants such as benzodiazepines. An interesting circumvention of this limitation is to consider situations where certain technologies may be useful in a majority of cases. For example, opioids are rarely seen at music festivals[89] where the ideal technology is one that can provide rapid testing and results for drug classification when there is high demand. In such cases, portable IR or handheld Raman is an ideal technology, while GC–MS would be too time-consuming.

On the other hand, a community drug checking initiative typically receives a significant amount of opioid samples where the classification is often confirmed by a fentanyl strip test, and the more detailed analysis of interest to the service user is both challenging (as in Example 4) and essential due to low concentration but highly potent actives circulating in the drug supply.

6.3.2 Communicating Uncertainty and Service Limitations

Regardless of what instrument is employed, the recognition and understanding of the sources of uncertainty and limitations is at the forefront of providing a safe and reliable service model. Pragmatic knowledge translation is vital to establishing a trustworthy service.[207] For many service providers the reality is that current out-of-the-box technology is not meeting their needs,[76] particularly in the level of nuance and expertise required for the interpretation of the data. Clearly, additional research and development is needed in collaboration with existing and prospective service users and providers. Even before current drug checking technologies existed, service providers monitored the supply, witnessed the effects of the local supply, and understood and advocated for the needs of their community.[13,76,199,209,210] Integrating instrument knowledge allows the drug checking community to better consider questions such as: What is possible? What isn't possible and why? What are the fundamental challenges, and can we overcome them?

6.4 Conclusions

Ideally, a single instrument would be able to achieve all of the objectives of drug checking. Currently, there is no clear instrument that ticks every box; low cost, portable, rapid, easy-to-use and providing highly sensitive identification and accurate quantification. Multi-instrument approaches to drug checking may be required to effectively respond to increasingly complex and highly potent substances demanding trace level detection and the potential for quantification. Case studies providing practice-based evidence illustrate

limitations when seeking to directly compare and select instruments for drug checking as a harm reduction response. Comparisons of different technologies demands implementation evidence as well as consideration of the different contexts and demands. An understanding of the principles behind relevant drug checking technologies will help to link the raw data they produce and the results that are communicated to people who use drugs. We hope that the technical foundations we have provided in this perspective will enable the harm reduction community to continue to guide innovations in drug checking.

Chapter 7

Infrared Absorption Spectroscopy and Two-Trace Two-Dimensional Correlation Analysis for the Resolution of Multi-Component Drug Mixtures[1]

7.1 Overview

Community drug checking provides an essential service that responds to the unpredictable and variable supply of illicit drugs. Point of care detection of trace components using portable infrared spectrometers is a harm reduction measure to prevent overdose. This study investigates the ability of weighted subtraction and two-trace two-dimensional (2T2D) correlation analysis to reveal the presence of heroin in an opioid mixture that contains heroin and fentanyl mixed with caffeine as a cutting agent. In both methods, a spectral trace was identified that provided reasonably high correlation scores to heroin when compared to entries in drug libraries. The two-trace correlation analysis produced a higher match score, suggesting that future improvements in spectral unmixing methods may enhance the reliability of detecting trace components in drugs.

[1]The chapter has been adapted from L. Gozdzialski, B. Wallace, I. Noda, D. Hore. "Exploring the Use of IR Spectroscopy and 2T2D Correlation Analysis for the Resolution of Multi-Component Drug Mixtures" *Spectrochim. Acta, Part A.* **282**, 121684 (2022). This was a research collaboration with IN. Conceptualization and data analysis was done by LG.

7.2 Introduction

Spectroscopic instruments continue to be one of the top methods for high throughput monitoring, particularly in field applications requiring low maintenance and easy-to-use instrumentation—attributes of popular handheld and benchtop near-infrared (NIR), Raman, and IR instruments.[18,20,21] IR and Raman spectroscopy have become established as popular drug checking methods, however it is well-known that these techniques suffer from limited sensitivity, particularly in complex mixtures. Library searching, i.e. using some sort of spectral distance or other correlation metric to match against pure spectral databases, is a popular approach for untargeted identification. When two or more substances are present, however, these metrics become less reliable particularly for mixtures of components where the spectral fingerprints are heavily overlapped.[211] Sequential subtraction of library entries works with some success, especially when the constituent compounds have relatively unique IR fingerprints, are present in reasonably high concentrations, and matrix effects are minimal. However, it is interesting to consider that an experienced spectroscopist or drug checking technician might visually search for evidence of a particular compound, rather than relying on matching metrics. Another possibility is that the presence of an additional component might be recognized, yet its identity remains unclear.

Many methods have been proposed to facilitate trace detection and characterization using IR spectroscopy. Multivariate chemometric methods are increasingly popular for both drug quantification and classification using neural networks, PLS-DA, random forest, SVM, and other methods.[160,212–214] In general, these are powerful methods for identifying subtle features between samples seeking to answer questions framed as “is this A or B?” for two-class problems. In a one-class classification, this may be expressed as “is this A or not?”, however this still fails to characterize the nature of outliers or contaminants, crucial information for people who use drugs. Multivariate methods also require significant method development for each specific compound, in comparison to less sensitive library-

searching methods that require minimal or no method development to screen a wide variety of drugs.[215] Two-dimensional correlation spectroscopy (2D-COS) also has a well established history in aiding the identification of key trace features by studying a system over a known perturbation (e.g. concentration, temperature).[216] Recently a similar approach has been proposed using only two spectra, the method of two-trace two-dimensional correlation (2T2D) analysis.[217,218] Over the past three years, there have been several accounts using 2T2D techniques to aid in trace feature identification.[219–222]

Here we present some approaches to analyzing the infrared spectrum of a typical three component drug mixture composed primarily of caffeine, but with trace amounts of fentanyl and heroin. Such low concentrations of potent actives are commonly seen in the illicit drug supply.[20] We explore the potential and the challenges of identifying the trace components by weighted subtraction and 2T2D curve resolution, both paired with an IR spectral library containing entries for pure drugs and cutting agents. Progress in this area will improve harm reduction messaging efforts that are currently hindered by uncertainty in spectral analysis.[76]

7.3 Methods

7.3.1 Instruments and Data Acquisition

The sample spectrum was prepared from a ternary mixture of 90 wt/wt% caffeine (Sigma-Aldrich), 5 wt/wt% fentanyl (Toronto Research Chemicals, Toronto, Canada), and 5 wt/wt% heroin (Toronto Research Chemicals). The reference spectrum was prepared from a binary mixture of 5 wt/wt% fentanyl and 95 wt/wt% caffeine. Detailed sample preparation techniques have been described previously.[20] Infrared (IR) spectra of the drug samples were collected using a portable FTIR spectrometer (Agilent 4500a, Agilent Technologies, Santa Clara, California) equipped with a dTGS detector and a single-bounce diamond ATR accessory. The spectral range used was 650 cm–3600 cm^{-1}. All sample spectra were normalized and baseline corrected using doubly re-weighted penalized least

squares prior to analysis.[223] This method employs both first and second derivatives to address the common challenge of over- and under-subtraction when the data is noisy. All analysis and pre-processing was implemented using in-house code, utilizing the `sklearn`[176] and `pybaselines` Python packages.

7.3.2 Library Searching and Weighted Subtraction

Library searching was performed using Pearson's correlation coefficient (PCC), R, as the metric for spectral matching against a spectral database containing a combination of entries from SWGDRUG and spectra that we have measured from analytical standards. All library spectra were interpolated as needed to match sample spectra, normalized and baseline corrected prior to library matching.

Weighed subtraction was performed between the sample mixture and a pure library component using a linear least squares fit with the criterion

$$\min \sum 0.5 \cdot ||Ax - b||^2 \tag{7.1}$$

where b is the sample spectrum and A is a matrix with two rows consisting of the pure library spectrum and an offset vector (set to a constant value for each spectral coordinate). The optimization produces an output, x, for the optimal weight for spectral subtraction and constant offset term. Once the component signal is subtracted, the residual is searched again against the spectral library using PCC as the similarity metric for ranking top matches.

7.3.3 Two-Trace Two-Dimensional Correlation Analysis

In the basic 2T2D analysis,[217] a sample spectrum $s(\tilde{\nu})$ and reference spectrum $r(\tilde{\nu})$ are used to create the asynchronous Ψ and synchronous Φ correlation maps

$$\Psi(\tilde{\nu}_1, \tilde{\nu}_2) = \frac{1}{2}[r(\tilde{\nu}_1) \cdot r(\tilde{\nu}_2) - s(\tilde{\nu}_1) \cdot s(\tilde{\nu}_2)] \tag{7.2a}$$

$$\Phi(\tilde{\nu}_1, \tilde{\nu}_2) = \frac{1}{2}[r(\tilde{\nu}_1) \cdot r(\tilde{\nu}_2) + s(\tilde{\nu}_1) \cdot s(\tilde{\nu}_2)], \tag{7.2b}$$

for each pair of wavenumbers v_1 and v_2. Using the synchronous correlation map, the correlation coefficients, ρ, can then be calculated as

$$\rho(\tilde{v}_1, \tilde{v}_2) = \frac{\Phi(\tilde{v}_1, \tilde{v}_2)}{\sqrt{\Phi(\tilde{v}_1, \tilde{v}_1) \cdot \Phi(\tilde{v}_2, \tilde{v}_2)}}. \tag{7.3}$$

A recently presented procedure extends the 2T2D approach to resolve pure spectral traces within mixtures using only two spectra.[224] This starts with defining a weighting function

$$f_{jk}(\tilde{v}) = \begin{cases} \rho(\tilde{v}, \tilde{v}_k) & \text{if } j = k \\ \sqrt{1 - \rho(\tilde{v}, \tilde{v}_k)^2} & \text{otherwise} \end{cases} \tag{7.4}$$

that enables a filter to be created

$$g_j(\tilde{v}) = \prod_{k=1}^{L} f_{jk}(\tilde{v}) \tag{7.5}$$

for the set of L characteristic bands that have been identified. This quantity is then normalized to obtain the j^{th} filter function

$$h_j(\tilde{v}) = \frac{g_j(\tilde{v})}{\sum_{k=1}^{L} g_k(\tilde{v})} \tag{7.6}$$

crafted for each species by multiplying the correlation coefficient for that species by the disrelation coefficients of the other species. The component of the spectrum that displays features attributed to a specific molecule, j, can then be extracted from

$$A_j(\tilde{v}) = A(\tilde{v}) \cdot h_j(\tilde{v}). \tag{7.7}$$

where $A(\tilde{v})$ is the average spectrum. This method aims to attenuate features that are asynchronous to a selected peak while enhancing synchronous features to ultimately resolve pure spectral traces, $A_j(\tilde{v})$, contributing to the sample mixture.[224]

7.4 Results and Discussion

7.4.1 Weighted Subtraction

Since we are working with a laboratory mixture that we prepared, we know the composition of the sample in advance (90% caffeine, 5% fentanyl, and 5% heroin), but seek to determine

this result from spectral analysis. The identification of the major component caffeine is straightforward as library searching results in a 0.99 PCC (with a score of 1 representing an exact match). However, after subtracting the caffeine library spectrum using Eq. 7.1, a subsequent search fails to produce high correlation scores for fentanyl or heroin. This low correlation score is not surprising, considering their relatively low concentration. In addition, while minimizing the residual is a common approach to spectral subtraction, the path to calculating the appropriate weighting is not straightforward and typically results in over- or under-subtraction.[217,225] In the case of under subtraction, caffeine continues to dominate the residual spectral search; in the case of over-subtraction, characteristic features of minor components have been distorted resulting in spurious hits.[76] Spectral variation arising from instrument parameters also cause challenges in optimal library searching and spectral subtraction.[226] However, in our case of a hybrid library (using some common pure standard spectra acquired on our own instrument), the results still remain inconclusive using spectral subtraction methods; new strategies for library searching are needed.

Using mixtures of drug standards as library entries is intriguing as an alternative to searching for single-component matches, and can potentially reduce the propagation of error through sequential weighted subtraction. However, spanning the large and ever changing range of mixtures present in an unregulated drug supply is not practical. It is also noted that metrics such as correlation are significantly influenced by the major component (e.g. caffeine) and might be misleading when matching to libraries of drug mixtures with minor components present. For example, the sample spectrum has a correlation score > 0.99 with a standard reference mixture of fentanyl (5%) and caffeine (95%). Visually, these spectra look nearly identical as shown in Figure 7.1a. The resulting weighted subtraction between these two mixtures (purple trace in Figure 7.1b) reveals features consistent with heroin (library entry shown in the yellow trace in Figure 7.1b). However, the correlation score remains relatively low ($R = 0.67$) and amongst other substances with similar scores (Table 7.1), for example quetiapine hemifumarate (library entry shown in the

brown trace in Figure 7.1b). Based on the similarity of the two mixtures, and the uncertain results of the weighted subtraction, one might conclude that the sample is likely a two-component mixture of fentanyl and caffeine.

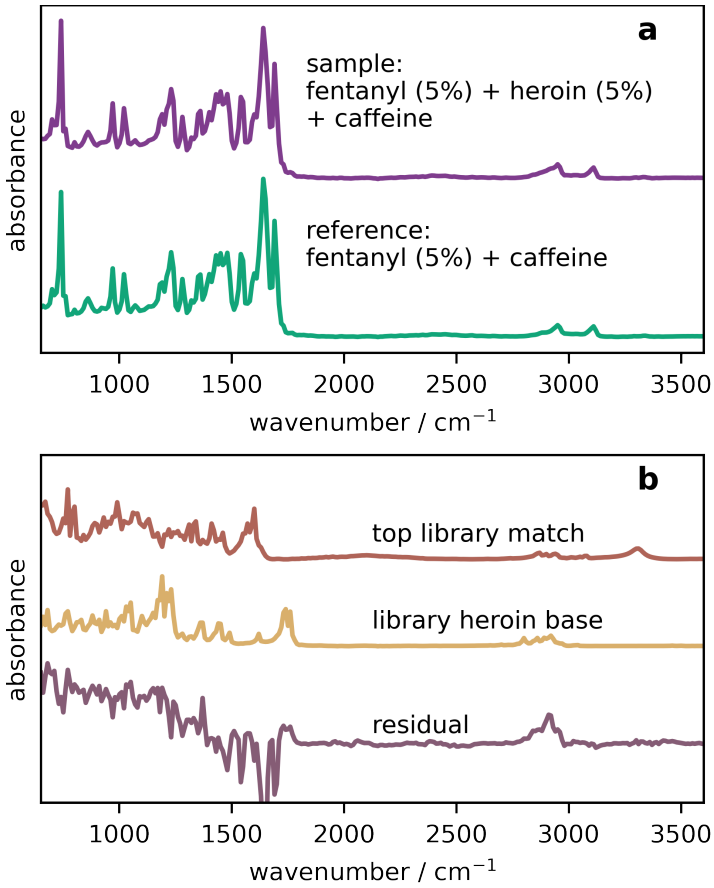

Figure 7.1: (a) Example of a drug mixture containing fentanyl (5%) and caffeine (95%) (green trace) and an "unknown" mixture containing fentanyl (5%), heroin (5%) and caffeine (90%) (purple). (b) Residual from the weighted subtraction between the reference and sample traces (purple) shown with the best-matched trace (quetiapine hemifumarate, residual $R = 0.69$, brown trace) when searching against a spectral database. The library entry of heroin (residual $R = 0.67$), the correct identity of the third component, is shown for comparison (yellow trace).

7.4.2 2T2D Analysis

2T2D is closely related to spectral weighted subtraction,[218] revealing changing spectral features upon a series of weights. It therefore presents an alternative perspective and approach to estimating the number of components in a mixture. The asynchronous 2T2D

Table 7.1: Pearson's correlation scores after the weighted subtraction of the sample mixture (5% fentanyl, 5% heroin, and 90% caffeine) and reference mixture (5% fentanyl and 95% caffeine).

Library entry	Pearson's coefficient, R
Quetiapine hemifumarate	0.69
Heroin Base	0.68
4-Hydroxy MET	0.64
Psilocin	0.64
Psilocybin	0.64

correlation map between the sample and reference spectra obtained using Eq. 7.2a is shown in Figure 7.2. The observation of strong asynchronous peaks suggests the presence of multiple components, as expected for a drug mixture. Following the established rules[217,224] for estimating the number of components within this system, we note that the most intense asynchronous cross peak is present at (1651 cm^{-1}, 741 cm^{-1}). Slice spectra at both of these indices have a mutual asynchronous peak at 2911 cm^{-1}, revealing the presence of a third component in the sample. However, the identity of the additional component remains unknown at this stage.

The curve resolution steps outlined in Eqs. 7.3–7.6 aim to extract pure component traces by attenuating spectral features out-of-sync with the identified component bands and amplifying those changing synchronously with the band of interest.[224] Previously, identifying three pure spectral bands was completed utilizing the asynchronous spectrum. Now we evaluate the correlation and disrelation coefficient for each band. For example, to resolve the pure component corresponding to the peak at 741 cm^{-1}, we first construct the filter function from Eq. 7.5

$$g(\tilde{\nu}) = \rho(\tilde{\nu}, 741) \cdot \sqrt{1 - \rho(\tilde{\nu}, 1651)^2} \cdot \sqrt{1 - \rho(\tilde{\nu}, 2911)^2} \tag{7.8}$$

that is subsequently normalized (using Eq. 7.6) to extract a pure spectral trace. In highly overlapped multi-component systems this is a challenge as pure spectral features may not entirely exist. This is revealed in Figure 7.3 where two resolved traces remain dominated

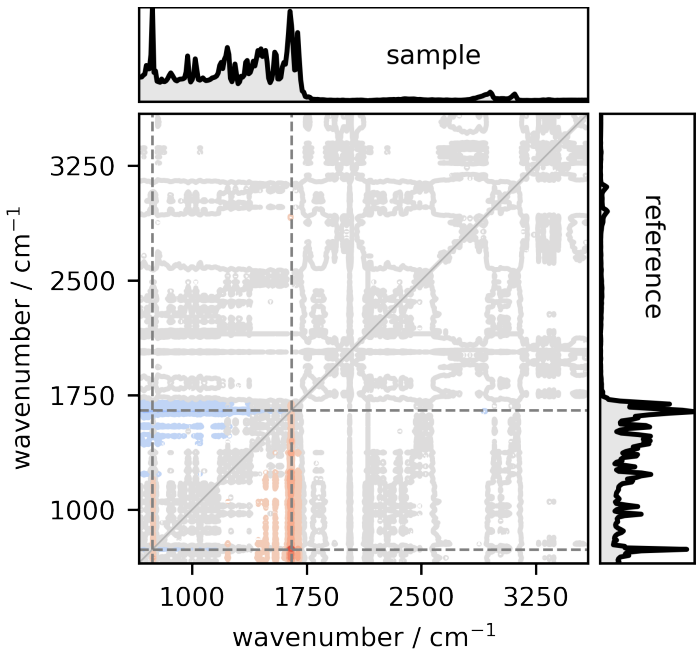

Figure 7.2: Asynchronous 2T2D spectrum calculated between the reference mixture, fentanyl (5%) and caffeine, and the sample mixture, fentanyl (5%), heroin (5%) and caffeine. Horizontal grey lines indicate locations of three mutually asynchronous peaks (741 cm^{-1}, 1651 cm^{-1}, 2911 cm^{-1}), suggesting three unique components are present.

by caffeine features. For the purpose of this example, we already know the composition of the mixture, so we expect one of the resulting pure spectral trace to match fentanyl. The identified asynchronous peaks with max intensity (741 cm^{-1}, 1651 cm^{-1}) are strong features of both fentanyl and caffeine, making their further resolution challenging. Upon initial visual inspection, we notice that the resolved trace 3 (Figure 7.3f) reveals features consistent with heroin, most noticeably in the 680 cm^{-1}, 1050 cm^{-1}, and 1720 cm^{-1} regions. Table 7.2 shows the PCCs obtained when scoring resolved trace 3 with the same library. Now heroin has dropped one spot on the list, but has a higher score. It is interesting to note that if we considered the initial weighted subtraction together with the 2T2D results, our candidate molecules would be reduced to two, including heroin.

2T2D is an intriguing approach to supplement spectral subtraction in the analysis of complex mixtures. However, in the case of significantly overlapped spectra, the application of 2T2D requires an experienced spectroscopist to guide the procedure. While it offers

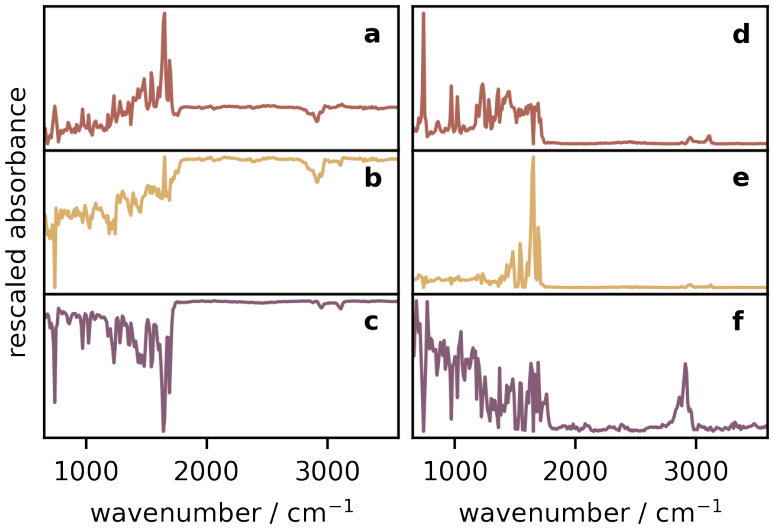

Figure 7.3: (a-c) Slices of the asynchronous 2T2D spectrum at the locations of three mutually asynchronous peaks, a) 741 cm^{-1}, b) 1651 cm^{-1}, and c) 2911 cm^{-1}. Each slice should be mostly absent of one component (e.g. fentanyl, caffeine, or heroin). (d-f) Resolved traces from 2T2D curve resolution.

Table 7.2: Top Pearson's correlation scores between the drug library and trace 3 resolved from the 2T2D analysis of the sample mixture.

Library entry	Pearson's coefficient, R
Quetiapine hemifumarate	0.78
Enalapril maleate	0.76
Heroin Base	0.75
URB-754	0.73
Yangonin	0.73

additional information when compared to weighted subtraction alone, it similarly relies on high quality reference spectra, particularly with drug mixtures that are not readily available in drug libraries. It would be interesting to further explore the use of data augmentation methods to supplement library entries of mixtures for 2T2D analysis and traditional spectral library searching.

In both the weighted subtraction and 2T2D approach, we would benefit from a tailored library that has a narrower scope. While a tailored library would not affect the correlation score, it would increase the rank of heroin among the matched library candidates. For example, the other components that outrank heroin in Tables 7.1 and 7.2 are rarely found

in opioid mixtures. In general, library searching methods need to be improved due to the similarity between drug spectral fingerprints. These need to highlight and place emphasis on peaks that are unique or mutually asynchronous; a concept central in 2T2D.

The approach taken towards resolving such mixtures is strongly influenced by the quality pure spectral libraries, access to standards of relevant drug mixtures, and resources allotted for development of machine learning models. The need for developing a highly accurate, precise, and importantly, transferable methods continues in drug checking applications. It was previously noted that mixtures of reference standards may provide only a limited representation of the broad range of drug samples encountered at a typical community drug checking service. However, their impact might be more widespread if some basic two or three component mixtures aid to facilitate detection of further minor components in mixtures.

New tools for mixture analysis are needed, as direct spectral subtraction is too reliant on the assumption of linearity in the relationship. Even when minimal interferences from matrix effects or baseline offsets occur, the route to the best spectral subtraction is often unclear. In addition to efforts in identifying the best possible resolution technique, efforts need to be dedicated to scoring algorithms that take into account high correlation between library entries themselves, making discrimination extremely challenging. It is possible that spectral subtraction or 2T2D methods have an application in fine-tuning library entries by identifying the most unique peaks for weighted correlation scoring. This fine-tuning would ensure that high correlation scores with incorrect substances are not coincidental. In community drug checking, emphasis is placed on training technicians to become more familiar with visually identifying drug patterns in IR absorption spectra instead of addressing the shortcomings of library searching approaches themselves.[76]

7.5 Conclusions

The common approach to drug mixture analysis using pure spectral drug libraries and similarity scoring based on Pearson's correlation coefficient in combination with spectral subtraction often does not lead to results that are straightforward to interpret. Using the infrared absorption spectrum of a relevant opioid drug mixture as an example, we have demonstrated the challenges associated with this approach. We have evaluated the use of two-trace two-dimensional correlation analysis as a tool for resolving multi-component mixtures. Heavily overlapped bands in the case of a mixture containing fentanyl, heroin, and caffeine affect the accuracy of both spectral subtraction and correlation-based scoring. We propose that in community drug checking applications, where the scope of drug samples received is broad and unpredictable, effort must be devoted to both tailored libraries and scoring methods that aim to reduce correlation between similar compounds. 2T2D analysis is intriguing for its ability to identify the most unique features between two given spectra that may be further pursued in developing such adjusted searching and scoring methods.

Efforts in library searching methods such as discussed in this chapter can have broad applicability. For instance, with the use of shared spectral libraries and accurate searching methods, drug checking with infrared spectroscopy can provide useful information "out-of-the-box", such that no significant data collection or method development is needed. Library searching is also important to facilitate technicians in identifying new or unseen compounds in drug mixtures. However, the growing spectral dataset of real drug mixtures provides an important opportunity for developing sensitive, tailored machine learning (ML) methods. This is now explored in the following chapter where ML models are developed for the identification of two target compounds. This work uses the infrared spectral data of drug mixtures collected at the drug checking service over the past few years. Added considerations for integrating automation in drug checking services is presented.

Chapter 8

Towards Automated IR Spectral Analysis in Community Drug Checking[1]

8.1 Overview

The body of knowledge surrounding infrared spectral analysis of drug mixtures continues to grow alongside the physical expansion of drug checking services. Technicians trained in the analysis of spectroscopic data are essential for reasons that go beyond the accuracy of the analytical results. Significant barriers faced by people who use drugs in engaging with drug checking services include the speed and accuracy of the results, and the availability and accessibility of the service. These barriers can be overcome by the automation of interpretations. A random forest model for the detection of two compounds, MDA and fluorofentanyl, was trained and optimized with drug samples acquired at a community drug checking site. This resulted in a 79% true positive and 100% true negative rate for MDA, and 61% true positive and 97% true negative rate for fluorofentanyl. The trained models were applied to selected drug samples to demonstrate a proposed workflow for interpreting and validating model predictions. The detection of MDA was demonstrated on three mixtures: (1) MDMA and MDA, (2) MDA and dimethylsulfone, and (3) fentanyl, etizolam, and benzocaine. The classification of fluorofentanyl was applied to a drug

[1]This chapter has been adapted from L. Gozdzialski, A. Hutchison, B. Wallace, C. Gill, D. Hore. "Towards Automated Infrared Spectral Analysis in Community Drug Checking" *Drug Test. Anal.* (in press). AH contributed to the ideas and discussion. Conceptualization, data analysis, and writing was done by LG.

mixture containing fentanyl, fluorofentanyl, 4-anilino-N-phenethylpiperidine, caffeine, and mannitol. Feature importance was calculated using shapely additive explanations to better explain the model predictions and *k*-nearest neighbours was used for visual comparison to labelled training data. This is a step towards building appropriate trust in computer-assisted interpretations in order to promote their use in a harm reduction context.

8.2 Introduction

Recently the term "drug checking technician" has emerged. This term typically refers to people who are trained to perform drug analysis using a range of on-site methods and instruments, such as Fourier transform infrared (FTIR) spectrometers, colorimetric reagent testing, and immunoassay and test strips, and consolidate information to make conclusions about the substance composition.[22,54] The number of drug checking technicians is rapidly multiplying as drug checking services expand both in the community, safe consumption sites, and at festivals.[54,207,209] Within communities where testing is done with FTIR, knowledge surrounding the spectral analysis of common drug mixtures within the field continues to grow alongside this expansion. Technicians learn the spectral patterns to look for when testing common drug mixtures, usually with the help of spectral libraries and software. For example, detecting a low concentration of fentanyl in a mixture of bulking agents (commonly referred to as cuts and buffs) is typically challenging for any basic library searching scheme. However, a trained drug checking technician can immediately recognize areas in the spectrum where fentanyl has minimal overlap with the cutting agents (e.g. caffeine or mannitol), manually find evidence of fentanyl in the spectrum, and consider the findings along with contextual evidence (e.g. suspected substance, appearance, anecdotal evidence if the substance has been used). At the same time, there is concern that this level of subjective interpretation in analyzing FTIR spectra may lead to misleading results.[227] Such complexity is recognized as a significant barrier to implementing drug checking services, and indirectly affects aspects such as speed, accuracy, availability,

and accessibility of drug checking.[76,81] These areas (speed, accuracy, etc.) have been recognized as significant factors affecting the willingness of people who use drugs (PWUD) to engage with drug checking.[81] It is acknowledged that current drug checking technologies, including the associated software for data analysis and interpretation, are still under development.[76]

Many of the advantages, and challenges, of drug checking with FTIR are inherent to the hardware and underlying technology (e.g. ease-of-use, non-destructive, limited capabilities for low concentration components and complex mixtures).[18,22] However, some areas of the implementation, particularly in the interpretation of the IR spectra, could possibly be improved through software. Machine learning (ML) broadly refers to a group of algorithms that reveal patterns in a set of data, connects those patterns to a meaningful result, and uses that relationship to predict future unknown data.[228] ML has been used for guiding spectroscopic interpretation for many years and the literature is rich with examples, methods, and proposed workflows.[148,228–231] For example, automation can speed up spectral analysis, alleviate the requirement of an experienced technician, and can offer a greater degree of consistency, accuracy, and precision in the reported results. However, common barriers to implementing ML include the requirement of a large quantity of high-quality data with known labels for building ML models, time and expertise required for the testing and implementation of such models, and protocols for on-going validation. Applying ML to drug checking faces additional barriers, namely obtaining exemptions for controlled substances, and coordinating off-site confirmatory testing.

Despite the many advances in ML for spectroscopic interpretation, as well as the growing body of past drug checking data (e.g. IR spectra with associated interpretations/labels), spectroscopy-based drug checking has surprisingly not been fully automated. To date most drug checking services rely on a human decision-maker (i.e. drug checking technician or harm reduction worker) for the spectral interpretation and final result,[22,60,227] even in cases where ML models exist to support that interpretation (e.g. quantification of fentanyl).

Complete automation risks the exclusion of experiential community knowledge, and the erasure of opportunities for the co-production of knowledge that combines technology and this experiential knowledge.[76,232] For instance, higher levels of satisfaction have been associated with drug checking services that operate in the contexts of clear communications and transparency.[207] Curiosity in the instruments and analytical process used in drug checking has also been noted to drive engagement of PWUD with drug checking service.[16,199] Overall, these studies motivate the consideration of contextual information and the value of curiosity when developing tools to aid in spectral interpretation.

Explainable artificial intelligence (XAI) guides explanations of predictions provided by ML models, and explores methods to present such explanations (e.g. visualizations, text, interactive tools).[232–236] Some methods include exposing feature importance, i.e. what part of the input was most influential in the prediction, and presenting "learn-by-example" cases. For example, technicians are familiar with the relationship between spectral features and the presence or absence of certain compounds, and when operating a drug checking service want to see the evidence. XAI also addresses the trust in and transparency of ML models.[235,237] In general, the pursuit of model explanations is motivated by three main purposes: model validation, model debugging, and knowledge discovery.[237,238]

This work uses a supervised machine learning algorithm trained on IR spectral data with associated PS-MS results to predict the presence of target compounds in unknown samples. XAI methods that expose the reasoning behind classification decisions are presented to (a) connect with current practices for interpreting drug checking data, (b) allow for on-going human-in-the-loop interference for improvements and quality assurance of ML models and (c) promote continuous knowledge production within drug checking services both for technicians and people engaging with the service.

8.3 Methods

8.3.1 Data Acquisition

Infrared spectra ($n = 7091$) used in this study were acquired between November 2020 and November 2022 through our service, Substance, the Vancouver Island Drug Checking Project,[239] located in Victoria, British Columbia. A portable FTIR with a 45° diamond ATR element (Agilent 4500a) was used. Spectra were acquired with 32 averages and an effective resolution of 4 cm^{-1}. Samples are received in various forms, with majority of substances tested as powders. This subset of IR spectra was chosen for building and evaluating classification models such that the same drug sample was also analyzed using PS-MS, for its ability to more unambiguously report on the presence or absence of particular trace actives.[50,133,186] Details of the PS-MS method as used for drug checking, including providing quantitative information, has been previously described in detail.[25,50,121,133] The overall composition for each sample was ultimately determined through IR analysis performed by a trained technician, using tools such as library matching, together with PS-MS in a point-of-care setting. It is noted that some of the components may be missed when they are below the LOD of IR-based methods, and not distinguishable on PS–MS.

8.3.2 Sample Selection

Two different target compounds were selected for building two classification models. The first model aimed to detect 3,4-methylenedioxyamphetamine (MDA) and the second is based on fluorofentanyl detection. These compounds were chosen based on their prevalence in and relevance to the local drug supply, and the fact that they represent simple (MDA) and more challenging (fluorofentanyl) classification problems. Furthermore, they have different potencies, with MDA usually appearing in high concentrations (or pure form) and fluorofentanyl typically being significantly cut. The MDA classification model was trained with IR absorption spectra of drug samples received at the drug checking service ($n = 4963$). Each spectrum was labeled based on whether it represented a sample

with or without MDA as 0 (not present/"negative", $n = 4804$) or 1 (present/"positive" $n = 159$), as previously determined by PS–MS, regardless of the other compounds present in the sample. For example, if a drug sample was determined to contain both 3,4-methylenedioxymethamphetamine (MDMA) and MDA through secondary testing with PS–MS, then it is labelled as "1". An external test set ($n = 2060$ without MDA and $n = 68$ with MDA) was used for validating the final, optimized classification model. Similarly, the RF model to detect fluorofentanyl was trained with a subset of IR spectra of samples determined to be within the category of opioid or "down", received at the drug checking service ($n = 2575$). Each sample was labeled based on whether it represented a sample with or without fluorofentanyl as 0 (not present/"negative", $n = 2202$) or 1 (present/"positive", $n = 373$), as previously determined by PS–MS, regardless of the other compounds present in the sample. Although MS is unable to differentiate between fluorofentanyl isomers, the IR spectral features were consistent with para-fluorofentanyl, hereafter referred to simply as fluorofentanyl. An external test set ($n = 551$ without fluorofentanyl and $n = 93$ with fluorofentanyl) was used to validate the trained model.

For the application of XAI methods, three test mixtures were chosen for prediction with the trained MDA model, representing the cases of correct positive prediction (5% MDA by weight in MDMA), incorrect negative prediction (54% MDA in dimethylsulfone), and correct negative prediction (an opioid mixture containing fentanyl, benzocaine, and etizolam). For the fluorofentanyl example, a correct positive prediction (6% fluorofentanyl in a mixture containing fentanyl HCl, 4-anilino-N-phenethylpiperidine commonly known as ANPP, caffeine, and mannitol) was chosen as an opportunity to demonstrate the ability of XAI to highlight subtle spectral features that have contributed to classification.

8.3.3 Random Forest (RF)

An RF classifier was used for binary classification. RF is an ensemble (voting) classifier that uses a series of classification trees, each built on a random subset of input features.

The RF classifiers were implemented using the scikit-learn package in Python.[176] Various combinations of preprocessing and hyperparameters were first explored using 3-fold cross validation. A random search cross validation procedure, implemented using scikit-learn, was used for the initial narrowing of the optimal preprocessing and hyperparameter space to further pursue with more comprehensive optimizations. The F1 score, a harmonic mean of the precision and recall, was used to evaluate these combinations. For the MDA model, a more comprehensive grid of hyperparameters was pursued using standard normal variate (SNV) preprocessing to maximize the F1 score. Similarly, the fluorofentanyl model was further optimized to maximize the F1 score during cross validation using min-max and second derivative spectral preprocessing. The grid of hyperparameters and preprocessing, as well as the performance metrics for both models, are shown in Tables 8.1–8.4. To simplify the model, a subset of the most important features, as calculated by the Gini index, was used to achieve similar performance with the RF. $n = 100$ and $n = 20$ features were chosen for the MDA and fluorofentanyl model, respectively, and these models were used for validation and additional analysis of selected drug mixtures.

8.3.4 *k*-Nearest Neighbours (KNN)

KNN[157] was used to find the closest matching spectra to generate examples for visualization purposes. Two models, one for positive samples and one for negative samples, were built with the respective training spectra. All spectra were pre-processed and truncated according to the optimized RF classification model. KNN was implemented using `scikit-learn` where `n_neighbors`$= 2$ and `metric='euclidean'`.[176]

8.3.5 Shapely Additive Explanations (SHAP)

The SHAP method was used to estimate the contribution of each input feature to the final prediction, therefore attempting to generate an explanation to an end-user regarding the model decision.[237,240] For the models, Kernel SHAP was implemented, a popular model-agnostic method, via the Python package `shap`.[237,240]

All pipelines for the work in this chapter were implemented using Python. Some packages of note, and the associated functions used, are

- `sklearn.pipeline.Pipeline`
- `sklearn.ensemble.RandomForestClassifier`
- `sklearn.model_selection.GridSearchCV`
- `sklearn.model_selection.RandomizedSearchCV`
- `sklearn.metrics.f1_score`
- `sklearn.decomposition.PCA`
- `sklearn.neighbors.KNeighborsClassifier`
- `shap.sample`
- `shap.KernelExplainer`

8.4 Results and Discussion

8.4.1 Model Performance and Optimization

The initial optimization of the RF model to detect MDA is illustrated in Figure 8.1a, where the colour of the grid relates to the F1 score of cross validation. The lowest F1 score of cross validation was calculated to be 0.80 with no preprocessing and the highest F1 score was calculated to be 0.86 with min-max normalization and second derivative preprocessing. Standard normal variate (SNV) pre-processing was chosen, however, as the pre-processing of choice with a similarly high F1 score of cross validation (0.85). This decision was made because spectra with less alteration are more likely to align with technicians' understanding of IR absorption and ultimately contribute to the interpretability of the model and post-hoc visualizations. To understand the model's decision-making process in a general, or

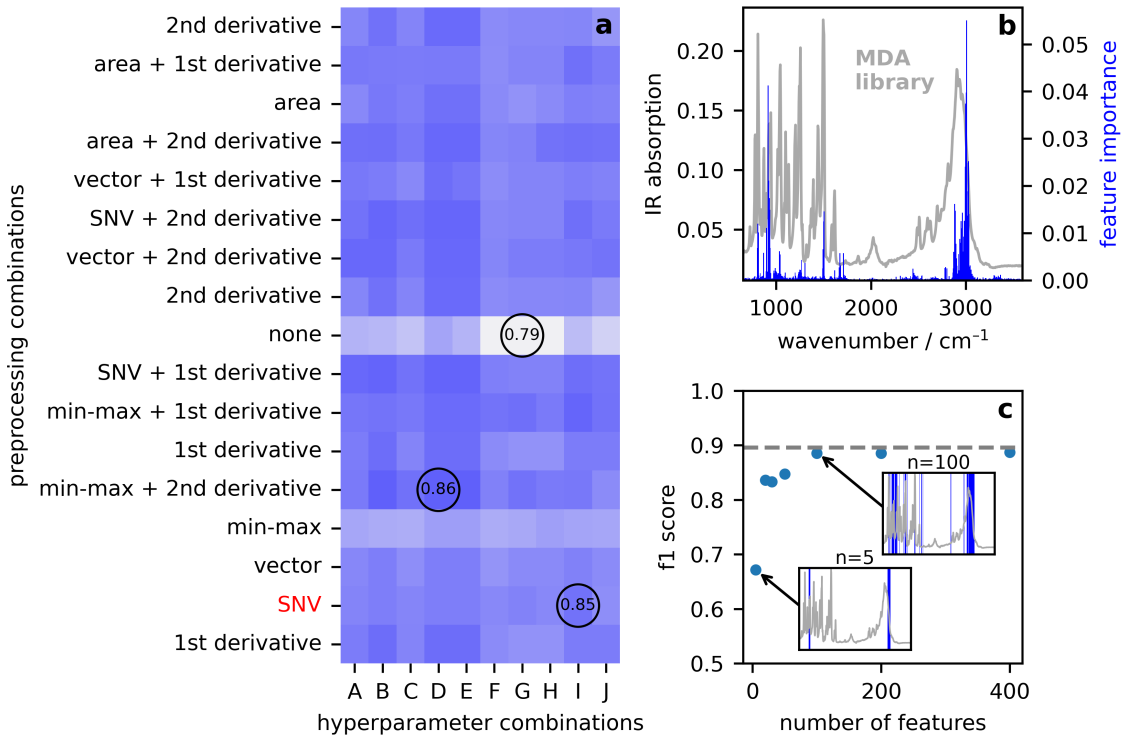

Figure 8.1: (a) Matrix representing combinations of hyperparameters and spectral preprocessing for optimizing performance of the RF model based on the F1 score. (b) Feature importance calculated within the RF model. Notably the most important features that we have identified correlate well to strong features in pure MDA. (c) F1 score on the test set confined to the n most important features as calculated from the base RF model.

"global", sense the most influential features as determined by the Gini index of the RF model are calculated and shown in Figure 8.1b. The most important features align well with strong modes shown in the library entry for MDA, which is overlaid in grey. The final classification model used in subsequent analysis uses the top 100 most important features, discarding features with minimal influence on the prediction of MDA (Figure 8.1c.). The resulting confusion matrix for the validation set is shown in Figure 8.2a. The precision was calculated on the external test set as 100% and the recall was determined to be 81%. The F1 score was calculated to be 0.9. The receiver operating characteristic (ROC) curve for the external test set is presented in Figure 8.2b. The area under the curve, which can be derived from the ROC, is a general performance measurement of how well two classes can be separated on a scale from 0–1. Here an area under the ROC (AUROC) of 0.99 was

obtained which demonstrates that there is excellent distinction between samples with MDA and samples without MDA. The complete results of the optimization of the RF model are shown in Table 8.1 and 8.2.

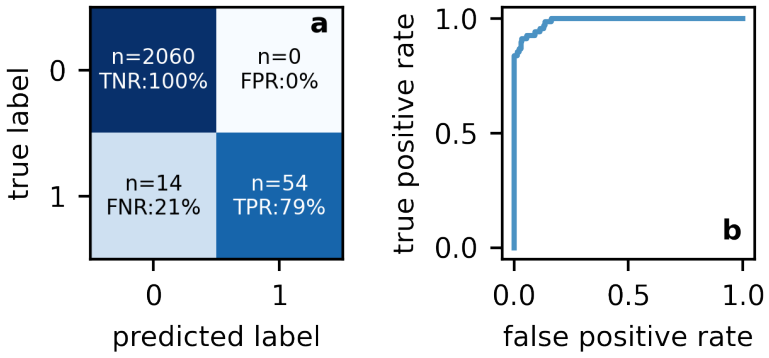

Figure 8.2: (a) Confusion matrix for external test set ($n = 2128$) using the optimized classification model from Figure 1. (b) Receiver operating characteristic curve (ROC) for the external test set, demonstrating the trade-off between true positives rates and false positive rates. Each point along the graph represents a varied decision threshold for classification.

Table 8.1: Results of initial Random Search CV with a range of hyperparameters and spectral preprocessing combinations for the MDA model. Acronyms: bs–balanced subsample, b–balanced.

	pre-processing	**boot-strap**	**class weight**	**max depth**	**min samples leaf**	**min samples split**	**n estimators**	**f1 cv**	**f1 test**
0	None + 2nd Deriv	False	bs	100	4	2	400	0.856	0.876
1	Area + 1st Deriv	False	bs	None	1	10	1000	0.853	0.885
2	Area	False	bs	100	4	2	400	0.854	0.878
3	Area + 2nd Deriv	False	bs	100	4	2	400	0.857	0.885
4	Vector + 1st Deriv	False	bs	100	4	2	400	0.855	0.885
5	SNV + 2nd Deriv	False	bs	None	4	10	200	0.858	0.885
6	Vector + 2nd Deriv	True	b	100	2	10	200	0.857	0.876
7	2nd Deriv	False	bs	100	4	2	400	0.856	0.876
8	None	False	bs	100	4	2	400	0.830	0.817
9	SNV + 1st Deriv	False	bs	None	4	10	200	0.859	0.885

Table 8.1: Cont'd

	pre-processing	boot-strap	class weight	max depth	min samples leaf	min samples split	n estimators	f1 cv	f1 test
10	Min-Max + 1st Deriv	False	bs	None	1	10	1000	0.859	0.885
11	1st Deriv	False	bs	None	4	10	200	0.855	0.885
12	Min-Max + 2nd Deriv	False	bs	None	4	10	200	0.861	0.885
13	Min-Max	False	b	60	4	10	600	0.832	0.876
14	Vector	False	bs	None	4	10	200	0.847	0.894
15	SNV	False	bs	None	1	10	1000	0.853	0.894
16	None + 1st Deriv	False	bs	None	4	10	200	0.855	0.885

Table 8.2: Grid search for optimizing hyperparameters with SNV preprocessing for MDA model. Acronyms: bs–balanced subsample, b–balanced.

	bootstrap	class weight	max depth	min samples leaf	min samples split	n estimators	mean test precision	mean test recall	mean test f1
0	True	bs	None	1	2	200	0.963	0.748	0.840
1	True	bs	None	1	2	500	0.963	0.748	0.840
2	True	bs	None	1	2	800	0.963	0.748	0.840
3	True	bs	None	1	2	1100	0.963	0.748	0.840
4	True	bs	None	1	2	1400	0.963	0.748	0.840
5	True	bs	None	1	5	200	0.963	0.748	0.840
6	True	bs	None	1	5	500	0.970	0.748	0.842
7	True	bs	None	1	5	800	0.963	0.748	0.840
8	True	bs	None	1	5	1100	0.963	0.748	0.840
9	True	bs	None	1	5	1400	0.963	0.748	0.840
10	True	bs	None	1	10	200	0.970	0.755	0.847
11	True	bs	None	1	10	500	0.963	0.748	0.840
12	True	bs	None	1	10	800	0.963	0.748	0.840
13	True	bs	None	1	10	1100	0.963	0.748	0.840
14	True	bs	None	1	10	1400	0.963	0.748	0.840
15	True	bs	None	2	2	200	0.970	0.748	0.842
16	True	bs	None	2	2	500	0.963	0.748	0.840
17	True	bs	None	2	2	800	0.963	0.748	0.840
18	True	bs	None	2	2	1100	0.963	0.748	0.840
19	True	bs	None	2	2	1400	0.963	0.748	0.840
20	True	bs	None	2	5	200	0.963	0.748	0.840
21	True	bs	None	2	5	500	0.963	0.748	0.840
22	True	bs	None	2	5	800	0.963	0.748	0.840
23	True	bs	None	2	5	1100	0.963	0.748	0.840
24	True	bs	None	2	5	1400	0.963	0.748	0.840
25	True	bs	None	2	10	200	0.963	0.748	0.840
26	True	bs	None	2	10	500	0.963	0.748	0.840
27	True	bs	None	2	10	800	0.963	0.748	0.840
28	True	bs	None	2	10	1100	0.963	0.748	0.840
29	True	bs	None	2	10	1400	0.963	0.748	0.840
30	True	bs	None	4	2	200	0.963	0.748	0.840

Table 8.2: Cont'd

	bootstrap	class weight	max depth	min samples leaf	min samples split	n estimators	mean test precision	mean test recall	mean test f1
31	True	bs	None	4	2	500	0.963	0.748	0.840
32	True	bs	None	4	2	800	0.963	0.755	0.844
33	True	bs	None	4	2	1100	0.963	0.755	0.844
34	True	bs	None	4	2	1400	0.963	0.748	0.840
35	True	bs	None	4	5	200	0.963	0.748	0.840
36	True	bs	None	4	5	500	0.963	0.748	0.840
37	True	bs	None	4	5	800	0.963	0.755	0.844
38	True	bs	None	4	5	1100	0.963	0.755	0.844
39	True	bs	None	4	5	1400	0.963	0.748	0.840
40	True	bs	None	4	10	200	0.963	0.761	0.849
41	True	bs	None	4	10	500	0.963	0.755	0.844
42	True	bs	None	4	10	800	0.963	0.755	0.844
43	True	bs	None	4	10	1100	0.963	0.761	0.849
44	True	bs	None	4	10	1400	0.963	0.755	0.844
45	False	bs	None	1	2	200	0.962	0.755	0.844
46	False	bs	None	1	2	500	0.962	0.755	0.844
47	False	bs	None	1	2	800	0.962	0.755	0.844
48	False	bs	None	1	2	1100	0.962	0.755	0.844
49	False	bs	None	1	2	1400	0.962	0.755	0.844
50	False	bs	None	1	5	200	0.962	0.755	0.844
51	False	bs	None	1	5	500	0.962	0.755	0.844
52	False	bs	None	1	5	800	0.962	0.755	0.844
53	False	bs	None	1	5	1100	0.962	0.755	0.844
54	False	bs	None	1	5	1400	0.962	0.755	0.844
55	False	bs	None	1	10	200	0.956	0.761	0.846
56	False	bs	None	1	10	500	0.956	0.767	0.850
57	False	bs	None	1	10	800	0.962	0.767	0.853
58	False	bs	None	1	10	1100	0.962	0.767	0.853
59	False	bs	None	1	10	1400	0.962	0.767	0.853
60	False	bs	None	2	2	200	0.962	0.761	0.848
61	False	bs	None	2	2	500	0.962	0.755	0.844
62	False	bs	None	2	2	800	0.962	0.761	0.848
63	False	bs	None	2	2	1100	0.962	0.755	0.844
64	False	bs	None	2	2	1400	0.962	0.755	0.844
65	False	bs	None	2	5	200	0.962	0.761	0.848
66	False	bs	None	2	5	500	0.962	0.755	0.844
67	False	bs	None	2	5	800	0.962	0.761	0.848
68	False	bs	None	2	5	1100	0.962	0.767	0.853
69	False	bs	None	2	5	1400	0.962	0.767	0.853
70	False	bs	None	2	10	200	0.949	0.767	0.847
71	False	bs	None	2	10	500	0.949	0.767	0.847
72	False	bs	None	2	10	800	0.949	0.767	0.847
73	False	bs	None	2	10	1100	0.949	0.767	0.847
74	False	bs	None	2	10	1400	0.949	0.767	0.847
75	False	bs	None	4	2	200	0.956	0.774	0.854
76	False	bs	None	4	2	500	0.949	0.774	0.851
77	False	bs	None	4	2	800	0.949	0.774	0.851
78	False	bs	None	4	2	1100	0.949	0.774	0.851
79	False	bs	None	4	2	1400	0.949	0.774	0.851
80	False	bs	None	4	5	200	0.956	0.774	0.854

Table 8.2: Cont'd

	bootstrap	class weight	max depth	min samples leaf	min samples split	n estimators	mean test precision	mean test recall	mean test f1
81	False	bs	None	4	5	500	0.949	0.774	0.851
82	False	bs	None	4	5	800	0.949	0.774	0.851
83	False	bs	None	4	5	1100	0.949	0.774	0.851
84	False	bs	None	4	5	1400	0.949	0.774	0.851
85	False	bs	None	4	10	200	0.949	0.767	0.847
86	False	bs	None	4	10	500	0.949	0.767	0.847
87	False	bs	None	4	10	800	0.949	0.767	0.847
88	False	bs	None	4	10	1100	0.949	0.774	0.851
89	False	bs	None	4	10	1400	0.949	0.767	0.847

In optimizing the second RF model for fluorofentanyl detection (Figure 8.3a), second derivative pre-processing was found to be necessary to resolve fluorofentanyl's sharp, highly overlapped and low intensity features within a crowded fingerprint region.[241] The results of the pre-proocessing and hyperparameter optimizations are shown in Table 8.3 and 8.4. The feature importance inherent to the RF model, as calculated by the Gini index, revealed that the most important features align with strong features in the library fluorofentanyl spectrum in Figure 8.3b. Again, to simplify the model, a subset of these features was used to achieve similar, and in some cases, greater performance, than using the entire spectrum. This is shown in Figure 8.3c. $n = 20$ features were chosen and that model was used for validation and additional analysis. The resulting confusion matrix is shown in Figure 8.4a. The precision and recall of the model was calculated as 80% and 61%, respectively, for the test set. The AUROC curve (Figure 8.4b) was calculated as 0.85, revealing the greater ambiguity in isolating unique spectral features between samples with and without fluorofentanyl.

Table 8.3: Results of initial Random Search CV with a range of hyperparameters and spectral preprocessing combinations for the fluorofentanyl model. Acronyms: bs–balanced subsample, b–balanced.

	pre-processing	boot-strap	class weight	max depth	min samples leaf	min samples split	n estimators	f1 cv	f1 test
0	None + 2nd Deriv	False	bs	100	4	2	400	0.615	0.652

Table 8.3: Cont'd

	pre-processing	boot-strap	class weight	max depth	min samples leaf	min samples split	n estimators	f1 cv	f1 test
1	Min-Max + 2nd Deriv	False	b	60	4	10	600	0.659	0.699
2	SNV + 2nd Deriv	False	b	60	4	10	600	0.651	0.662
3	Vector + 2nd Deriv	False	b	60	4	10	600	0.652	0.662
4	Area + 1st Deriv	False	b	60	4	10	600	0.620	0.643
5	1st Deriv	False	bs	None	4	10	200	0.601	0.602
6	SNV	False	bs	None	4	10	200	0.555	0.596
7	Vector	False	b	60	4	10	600	0.523	0.584
8	None + 1st Deriv	False	bs	None	4	10	200	0.601	0.602
9	Area	False	b	60	4	10	600	0.538	0.603
10	Min-Max + 1st Deriv	False	b	60	4	10	600	0.620	0.623
11	SNV + 1st Deriv	False	bs	100	4	2	400	0.616	0.613
12	Vector + 1st Deriv	False	b	60	4	10	600	0.612	0.613
13	2nd Deriv	False	bs	100	4	2	400	0.615	0.652
14	Area + 2nd Deriv	False	bs	None	4	10	200	0.662	0.690
15	None	False	bs	None	4	10	200	0.413	0.455
16	Min-Max	False	b	60	4	10	600	0.517	0.530

Table 8.4: Grid search for optimizing hyperparameters with Min-Max + 2nd Derivative preprocessing for fluorofentanyl model. Acronyms: bs–balanced subsample, b–balanced.

	bootstrap	class weight	max depth	min samples leaf	min samples split	n estimators	mean test precision	mean test recall	mean test f1
0	True	b	50	4	10	200	0.977	0.464	0.628
1	True	b	50	4	10	450	0.972	0.464	0.627
2	True	b	50	4	10	700	0.977	0.467	0.631
3	True	b	50	4	10	950	0.972	0.467	0.629
4	True	b	50	4	10	1200	0.972	0.467	0.629
5	True	b	80	4	10	200	0.977	0.464	0.628
6	True	b	80	4	10	450	0.972	0.464	0.627
7	True	b	80	4	10	700	0.977	0.467	0.631
8	True	b	80	4	10	950	0.972	0.467	0.629
9	True	b	80	4	10	1200	0.972	0.467	0.629
10	True	b	110	4	10	200	0.977	0.464	0.628
11	True	b	110	4	10	450	0.972	0.464	0.627
12	True	b	110	4	10	700	0.977	0.467	0.631
13	True	b	110	4	10	950	0.972	0.467	0.629
14	True	b	110	4	10	1200	0.972	0.467	0.629
15	True	b	None	4	10	200	0.977	0.464	0.628

Table 8.4: Cont'd

	bootstrap	class weight	max depth	min samples leaf	min samples split	n estimators	mean test precision	mean test recall	mean test f1
16	True	b	None	4	10	450	0.972	0.464	0.627
17	True	b	None	4	10	700	0.977	0.467	0.631
18	True	b	None	4	10	950	0.972	0.467	0.629
19	True	b	None	4	10	1200	0.972	0.467	0.629
20	True	bs	50	4	10	200	0.972	0.464	0.627
21	True	bs	50	4	10	450	0.972	0.467	0.629
22	True	bs	50	4	10	700	0.972	0.464	0.627
23	True	bs	50	4	10	950	0.972	0.467	0.629
24	True	bs	50	4	10	1200	0.972	0.469	0.632
25	True	bs	80	4	10	200	0.972	0.464	0.627
26	True	bs	80	4	10	450	0.972	0.467	0.629
27	True	bs	80	4	10	700	0.972	0.464	0.627
28	True	bs	80	4	10	950	0.972	0.467	0.629
29	True	bs	80	4	10	1200	0.972	0.469	0.632
30	True	bs	110	4	10	200	0.972	0.464	0.627
31	True	bs	110	4	10	450	0.972	0.467	0.629
32	True	bs	110	4	10	700	0.972	0.464	0.627
33	True	bs	110	4	10	950	0.972	0.467	0.629
34	True	bs	110	4	10	1200	0.972	0.469	0.632
35	True	bs	None	4	10	200	0.972	0.464	0.627
36	True	bs	None	4	10	450	0.972	0.467	0.629
37	True	bs	None	4	10	700	0.972	0.464	0.627
38	True	bs	None	4	10	950	0.972	0.467	0.629
39	True	bs	None	4	10	1200	0.972	0.469	0.632
40	False	b	50	4	10	200	0.969	0.504	0.662
41	False	b	50	4	10	450	0.974	0.501	0.661
42	False	b	50	4	10	700	0.974	0.499	0.659
43	False	b	50	4	10	950	0.974	0.499	0.659
44	False	b	50	4	10	1200	0.974	0.499	0.659
45	False	b	80	4	10	200	0.969	0.504	0.662
46	False	b	80	4	10	450	0.974	0.501	0.661
47	False	b	80	4	10	700	0.974	0.499	0.659
48	False	b	80	4	10	950	0.974	0.499	0.659
49	False	b	80	4	10	1200	0.974	0.499	0.659
50	False	b	110	4	10	200	0.969	0.504	0.662
51	False	b	110	4	10	450	0.974	0.501	0.661
52	False	b	110	4	10	700	0.974	0.499	0.659
53	False	b	110	4	10	950	0.974	0.499	0.659
54	False	b	110	4	10	1200	0.974	0.499	0.659
55	False	b	None	4	10	200	0.969	0.504	0.662
56	False	b	None	4	10	450	0.974	0.501	0.661
57	False	b	None	4	10	700	0.974	0.499	0.659
58	False	b	None	4	10	950	0.974	0.499	0.659
59	False	b	None	4	10	1200	0.974	0.499	0.659
60	False	bs	50	4	10	200	0.969	0.504	0.662
61	False	bs	50	4	10	450	0.974	0.501	0.661
62	False	bs	50	4	10	700	0.974	0.499	0.659
63	False	bs	50	4	10	950	0.974	0.499	0.659
64	False	bs	50	4	10	1200	0.974	0.499	0.659
65	False	bs	80	4	10	200	0.969	0.504	0.662

Table 8.4: Cont'd

	bootstrap	class weight	max depth	min samples leaf	min samples split	n estimators	mean test precision	mean test recall	mean test f1
66	False	bs	80	4	10	450	0.974	0.501	0.661
67	False	bs	80	4	10	700	0.974	0.499	0.659
68	False	bs	80	4	10	950	0.974	0.499	0.659
69	False	bs	80	4	10	1200	0.974	0.499	0.659
70	False	bs	110	4	10	200	0.969	0.504	0.662
71	False	bs	110	4	10	450	0.974	0.501	0.661
72	False	bs	110	4	10	700	0.974	0.499	0.659
73	False	bs	110	4	10	950	0.974	0.499	0.659
74	False	bs	110	4	10	1200	0.974	0.499	0.659
75	False	bs	None	4	10	200	0.969	0.504	0.662
76	False	bs	None	4	10	450	0.974	0.501	0.661
77	False	bs	None	4	10	700	0.974	0.499	0.659
78	False	bs	None	4	10	950	0.974	0.499	0.659
79	False	bs	None	4	10	1200	0.974	0.499	0.659

It is well known that spectroscopy-based techniques such as FTIR and Raman lack sensitivity, yet are attractive for community drug checking because they are easy-to-use, robust, portable, and lower-cost. The trade-off of this advantage is a higher limit of detection and therefore, false negatives are inevitable when low concentration components exist in the drug market. There are no formal acceptance criteria established to indicate whether a model is suitable for deployment in the context of community drug checking, and in other applications it is often stated to be "fit for purpose",[242] acknowledging associated risks of reporting on analytical results with uncertainty. In general, these classification models are expected to have high precision (confidence in positive hits if enough spectral evidence is present) but low recall (less confidence in negative hits), as seen before in drug checking applications.[25] For example, manual FTIR interpretation using spectral matching software has been shown to result in a similar outcome, where false negatives are far more prevalent than false positives due to the relatively high limit of detection of FTIR.[21,42,60] During validation, when the MDA model predicted that MDA was present in a sample, it was correct every time. However 19% were predicted as false negatives. A similar situation was found for the detection of fluorofentanyl in opioid samples. In this case, where the median concentration of fluorofentanyl in the training set was low (2.1 w/w%),

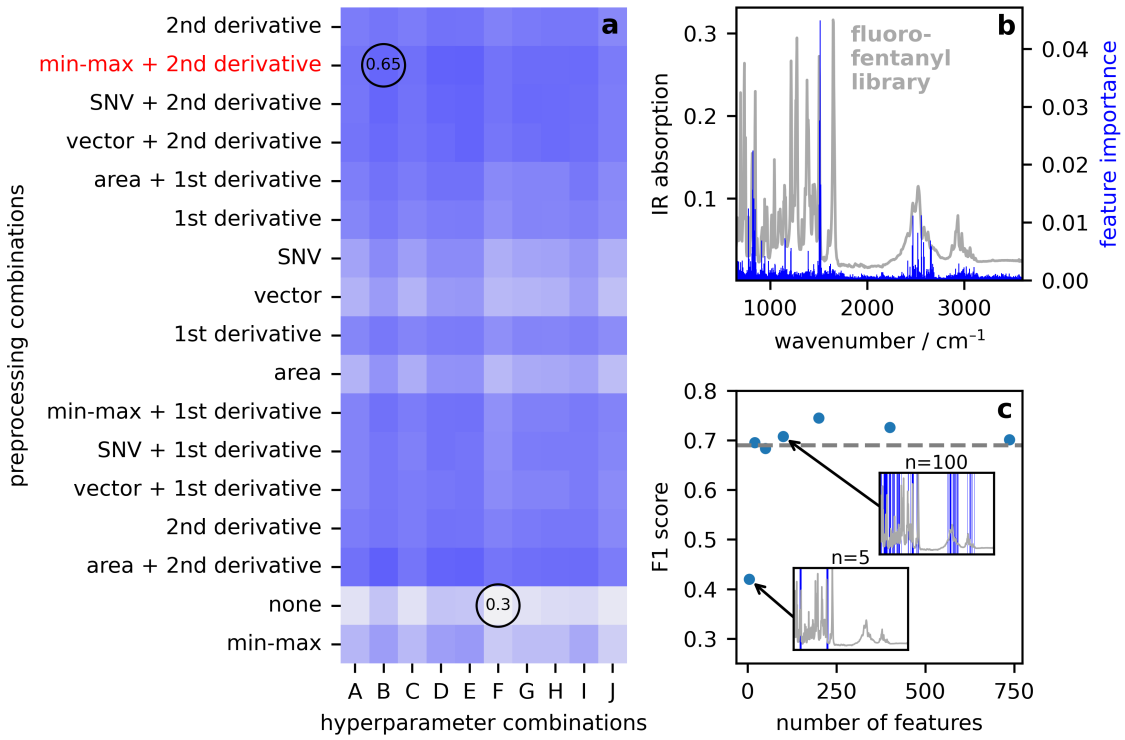

Figure 8.3: (a) Matrix representing combinations of hyperparameters and spectral preprocessing for optimizing performance of the RF model. The metric considered is the F1 score. (b) Feature importance calculated within the RF model. The few important features found correlate well to strong features in pure fluorofentanyl. Notably many strong features of fluorofentanyl have minimal importance for the prediction. (c) F1 score on the test set for n most important features as calculated from the base RF model.

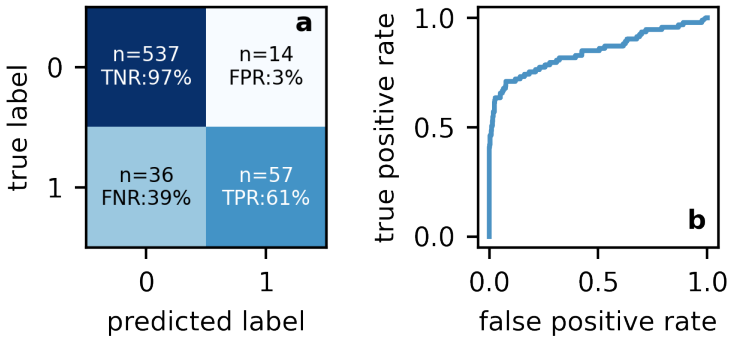

Figure 8.4: (a) Confusion matrix for external test set ($n = 644$). (b) ROC curve demonstrating the trade-off between true positives rates and false positive rates. Each point on the graph represents a varied decision threshold for classification.

IR spectroscopy approaches its limit of detection. Fourteen (3%) false positive predictions were made using the optimized model, however, fluorofentanyl was correctly detected in

only about half of the positive test cases. The ROC curves shown in Figure 8.2b and Figure 8.4b demonstrate that, if one is willing to accept an increased risk of false positives as a trade-off for improved true positive detection, such a compromise can be considered.

It is noted that in the dataset presented here, as in most drug checking datasets, the populations of the classes are typically small, significantly unbalanced, and prone to label error as the overall composition is rarely known with absolute certainty. Determining what a suitable dataset is to begin the pursuit of ML does not have a straightforward answer. This depends on the problem at hand, the quality and complexity of the data, and the algorithm of choice.[243] RF classification was explored in this application because it minimizes overfitting due to its iterative bagging and voting,[165] performs an implicit feature selection by only using features that are most influential on reducing classification error,[244] and therefore is well suited to real-world datasets with some degree of label error.[165,244] Variations of the original RF model, such as balanced RF, were also used to overcome some limitations with an extremely imbalanced training set.[165] The initial model optimization, evaluation of performance, and interrogation of the relevance of features learned by the model are important steps in determining the suitability of a drug checking dataset for ML. Metrics such as precision and recall will guide whether a particular dataset is suitable, however could possibly be further improved by using more training data, addressing label errors, data augmentation, data fusion or exploring complex neural network architectures. Eventually, for many target compounds, the error will be mostly attributed to the fact that both the training and test sets, and future samples received at the drug checking service, will contain drugs that are present below the limit of detection. Recognizing these limitations is important when considering the expansion of ML models for a range of compounds.

8.4.2 Generating Model Explanations

Three “unknown” spectra from the test set for the classification of MDA are shown in Figure 8.5a–c to demonstrate the outputs from SHAP and KNN. First, the SHAP

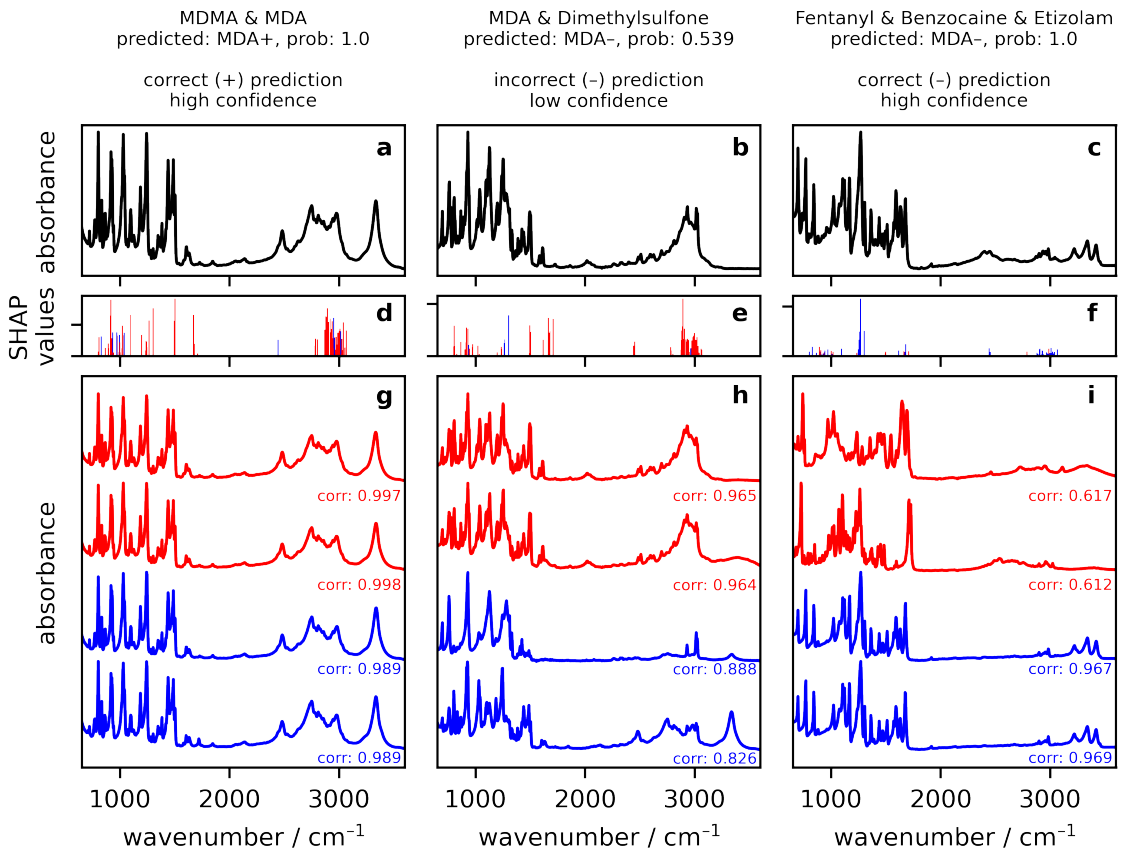

Figure 8.5: Various case examples using a combination of SHAP and KNN to aid in summarizing the RF classification results. Each column represents a different test case, with the "unknown" query spectrum shown in black (a–c). SHAP values of features that are found to contribute to a positive prediction are shown in red, and those that contribute to a negative prediction are shown in blue (d–f). Nearest neighbours traces that have MDA present are shown in red, and traces that do not have MDA present are shown in blue (g–i).

values were calculated for each test instance (Figure 8.5d–f). Second, KNN was used to retrieve factual and counterfactual explanatory cases (Figure 8.5g–i). The top four nearest neighbours from the training set, with known composition, are shown. Two nearest neighbours are from the positive class (MDA present, traces shown in red) and two nearest neighbours are from the negative class (no MDA present, trace shown in blue) with their corresponding correlation to the query spectrum. Together, these features aim to explain why a model might have predicted one class vs the other for a particular unknown sample given the IR absorption data.

The first case represents a mixture of MDMA and MDA, where MDMA is the major component, and there is minor contribution to the overall IR spectrum from MDA. The finalized model predicted the presence of MDA, with a probability of 100%. This suggests the model is highly confident in the prediction, and the SHAP values draw attention to the features in the query spectrum that most contributed to the higher confidence (red) and features that possibly did not align with the presence of MDA (blue). As a feature of XAI, this prediction is supported by the observation that intensity in these regions directly corresponds to MDA vibrational modes. The four nearest neighbours demonstrate that both samples with MDA ($+$) and without MDA ($-$) in the database have a very high correlation with the query spectrum (Pearson's correlation 0.99). With these nearest neighbours and SHAP values together, technicians may try to extract evidence that the query spectrum does in fact have features consistent with the presence of MDA.

The second case is a mixture of MDA and dimethylsulfone. The model incorrectly predicted that there is no MDA present, where the prediction probability of 55% suggests uncertainty about the presence or absence of MDA. The SHAP values reveal that there are some features that do support the presence of MDA, however it was not significant enough to result in a positive detection. The four nearest neighbours reveal that the positive nearest neighbours, both which are samples with MDA and dimethylsulfone, have much higher correlation with the query spectrum (0.96) than the nearest neighbours from the negative class (0.82–0.88). Upon further inspection, there are in fact features consistent with the presence of MDA in our query spectrum and may disagree with the model here. This brings attention to the fact that perhaps this drug combination is less frequent and was not well represented in the training.

The final case is a mixture that includes fentanyl, benzocaine and etizolam. The model predicted that there was no MDA in this sample, with a prediction probability of 100%. The SHAP values reveal that almost no features of the query spectrum are contributing to a positive prediction and highlights features that strongly suggest the absence of MDA.

The nearest neighbours from the negative class have very high correlation to the query spectrum, further instilling confidence in this prediction. The nearest neighbours from the positive class have very poor correlation scores (0.61) in contrast to the what was observed in the two previous examples (Figure 8.5g-h). Here there were no MDA positive nearest neighbours with a similar IR spectrum, further supporting that MDA in such a drug mixture was unlikely.

The same two methods for facilitating model explanations described in the previous examples were explored for fluorofentanyl classification. In this case, however, there were additional challenges because (1) the final model had poorer performance as a result of the low concentrations of fluorofentanyl and higher complexity of opioid drug mixtures, and subsequently, (2) greater spectral manipulation was required for optimal separation between the two classes (second derivative) and therefore less intuitive to use for visualizing spectral features. However, since the presence of fluorofentanyl was mostly determined from sharp features in the fingerprint region, there is a benefit from having to visualize and investigate fewer ($n = 20$) features. The implementation was demonstrated using a sample that contains fentanyl, fluorofentanyl, caffeine and mannitol (Figure 8.6). Here the fluorofentanyl model correctly predicted the presence of fluorofentanyl with a probability of 90%. Again, the positive SHAP features highlight areas that have contributed to this prediction (red), and features that, according to the trained model, countered this prediction (blue). The features that contributed to the prediction of fluorofentanyl align with characteristic modes from a pure fluorofentanyl spectrum. Figure 8.6d overlays the positive and negative nearest neighbours, both of which look very similar to the query spectrum. This implies that the features attributed to fluorofentanyl are expected to be subtle relative to features arising from other substance within the mixtures. The most influential region is highlighted on the inset, displayed both with simple min-max normalization and the second derivative pre-processing that was used in model training. This highlights that evidence of fluorofentanyl, though subtle, is present in the query

spectrum. In this case, a technician may decide to trust the model prediction based on such evidence.

8.4.3 Practical Application in Harm Reduction

In general, implementation of XAI facilitates knowledge production in a way that black-box ML models cannot.[245] It is known that curiosity about drug checking technologies, discussion around drug market trends and expectations, and integrating the personal experience of people is essential to guide drug analysis.[16,199,246] Automating FTIR-based drug checking ultimately aims to extend its reach, particularly to smaller communities and regions lacking public health and harm reduction mandates that may not have the resources to train a technician. One of the main goals of this work was to implement a framework to build adequate trust in an automated model and its predictions. The explanation produced should help present evidence and trust when in fact the prediction is correct, and hesitancy when it is incorrect. Previous studies in explainable AI have found that feature highlighting, as well as presenting explanatory factual and counterfactual cases (as we have done in the various test samples shown in Figure 8.5 and Figure 8.6), has assisted users in detecting errors while increasing their understanding of the model itself.[247] This implementation will also contribute to ongoing method re-validation, as the drug supply continues to change. Such potential was demonstrated when the MDA model incorrectly predicted that no MDA was present in an MDA–dimethylsulfone sample.

While ML offers a means to standardize the analysis of IR spectra, harm reduction messaging relies on people who can take those results, assess their validity and provide context in their interpretation. It is always important to consider how incorrect and inconsistent drug checking information might impact interpersonal relationships between the consumer, manufacturer, and distributor of illicit drugs.[16,248,249] Moreover, false positives and negatives affect trust in the service and perceptions of the utility of drug checking.[89,205,250] This could impact overall engagement with drug checking services, as

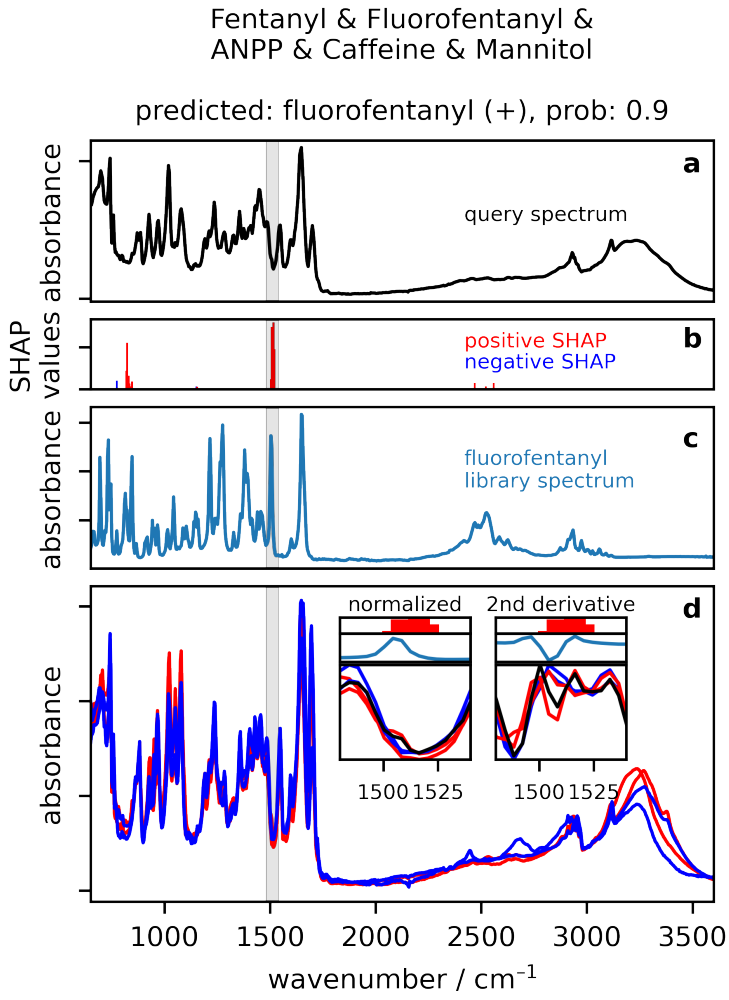

Figure 8.6: Example of a test sample spectrum (a) and explainable AI framework to build adequate trust in the model prediction. (b) SHAP values highlight the spectral regions of interest contributing to the prediction. The (c) library spectrum of fluorofentanyl and (d) the spectra of the 4 nearest neighbours are overlaid for reference. The inset in (d) combines both the k nearest neighbours from the positive class (red traces) and negative class (blue traces) and feature highlighting from SHAP in combination with the spectral library of fluorofentanyl to present visual evidence that features consistent with fluorofentanyl exist in the query spectrum (black).

PWUD often report navigating health and social support services that are stigmatizing, not culturally safe, and not relevant for their needs.[251,252] Ultimately, the optimization and evaluation of ML classification models when used in combination with XAI is poised to even better facilitate a reliable analysis and tailored discussion with service users around their drug checking results.

8.5 Conclusions

Automation has the potential to improve the speed and consistency of drug checking, offering less reliance on technician experience. This work has examined the results of classification models for MDA and fluorofentanyl. In this process, explainable AI was integrated using feature importance via SHAP values and a KNN model to retrieve related/explanatory cases from the training data. The integration of XAI methods provides a level of transparency that facilities continuous knowledge production and engagement of technicians and community members. Such methods will help to bridge the gap between the current role of a drug checking technician and the pursuit of ML methods for drug checking. The next chapter expands on the framework established here, and presents the current efforts developing a broadly applicable machine learning pipeline and automated reports.

Chapter 9

Enhanced IR Spectral Interpretation using Machine Learning Pipelines

9.1 Overview

Drug checking services using infrared spectroscopy can test a diverse range of substances. Within that list, such substance may also be in varying mixtures and relative concentrations. While there has been efforts made to apply machine learning within drug checking, it has always been limited in scope (i.e. proof-of-concept in quantifying, classifying or discriminating only a few compounds).[20,24,147,253,254] In practice, this means analysing an infrared spectrum still results in a piece-wise analysis, heavily depending on technicians ability to consolidate information. In some cases, this could even add complexity to both the training of technicians and translation of final results to a service user. There is a need to built a comprehensive framework for analysing infrared spectra acknowledging the broad categories of drugs seen in drug checking applications.

It is well-known that the performance of machine learning models is highly dependent on the training data.[255] A challenge in applications with high variability amongst all features is that simple statistical calculations, such as mean or standard deviation, do not hold much significance to the overall dataset.[256] A potential solution to this problem is to first split data into smaller and more reasonably comparable groups.[256] Clustering is an effective way to explore these natural patterns so that more nuanced features can be realized

through additional supervised or unsupervised analysis.

Expanding from the work done in Chapter 8, this chapter presents progress on the development of a comprehensive scheme using infrared data for automated drug characterization. This multi-step process includes predicting basic drug classes followed by identifying specific compounds within each class. Here, a cluster analysis method called hierarchical density-based spatial clustering of applications with noise (HDBSCAN) as an initial unsupervised method to establish major groupings of the data. A series of random forest classifiers are then developed to predict compounds frequently found within that cluster. Ten clusters were revealed, grouping by characteristics of the major compound reflected in the respective IR spectra; caffeine, water, microcrystalline cellulose, MDA, MDMA, methamphetamine, ketamine, cocaine HCl, cocaine base, and fentanyl HCl. The following steps focus on one of the found clusters where the IR spectra are characteristic to the presence of caffeine in the associated samples. Within this cluster, several compounds are targeted for random forest classification including psychoactive compounds bromazolam, carfentanil, etizolam, fentanyl, flualprazolam, flubromazepam, fluorofentanyl, heroin, and xylazine, and cuts/buffs mannitol, microcrystalline cellulose, caffeine, erythritol, and xylitol. The performance varies significantly for different compounds when applied to the test sets, with F1 scores ranging from 0.11 to 0.99. The relationship between this performance and relative concentration within the drug mixture is investigated. This exercise—evaluating model predictive performance, approximating the limit of detection, and revealing the most characteristic features—facilitates clear communication on limitations around drug checking with IR spectroscopy. This is true, and perhaps even more pertinent, in the case where the performance of a model is insufficient. As an extension to the initial cluster analysis presented here, early prospects for improved untargeted compound detection is also introduced.

9.2 Methods

9.2.1 Dataset

The IR spectral data from November 2020 to August 2023 was organized by date tested (yyyy-mm-dd) at Substance and split into a training ($n = 10447$) and test ($n = 2612$) set. In other words, IR spectra from November 2020 to April 2023 was used for training and cross validation, and the most recent data from May 2023 to August 2023 was used to test the final machine learning pipeline. This aims to simulate the realistic situation where previous drug checking data is used to inform future interpretations. This also considers the translation of such concepts to ongoing development in real-time (e.g. weekly re-validation).

9.2.2 Machine Learning Pipeline

The general proposed scheme is illustrated in Figure 9.1 and outlined as follows:

1. Unsupervised clustering to reveal natural patterns in the IR data
2. Identify the major drug class characteristic to cluster (e.g. using previous interpretations and additional library searching)
3. Within each cluster, train and optimize supervised random forest classifiers using the most frequent components as labels
4. Evaluate performance of the two-step classification on an unseen test set
5. Offer visual confirmation, probability scores, and feature highlighting as appropriate throughout the workflow

9.2.2.1 Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

HDBSCAN is a density-based clustering methods establish high-density data (points close together) as the core clusters, and treat low-density data (points sparsely distributed) as

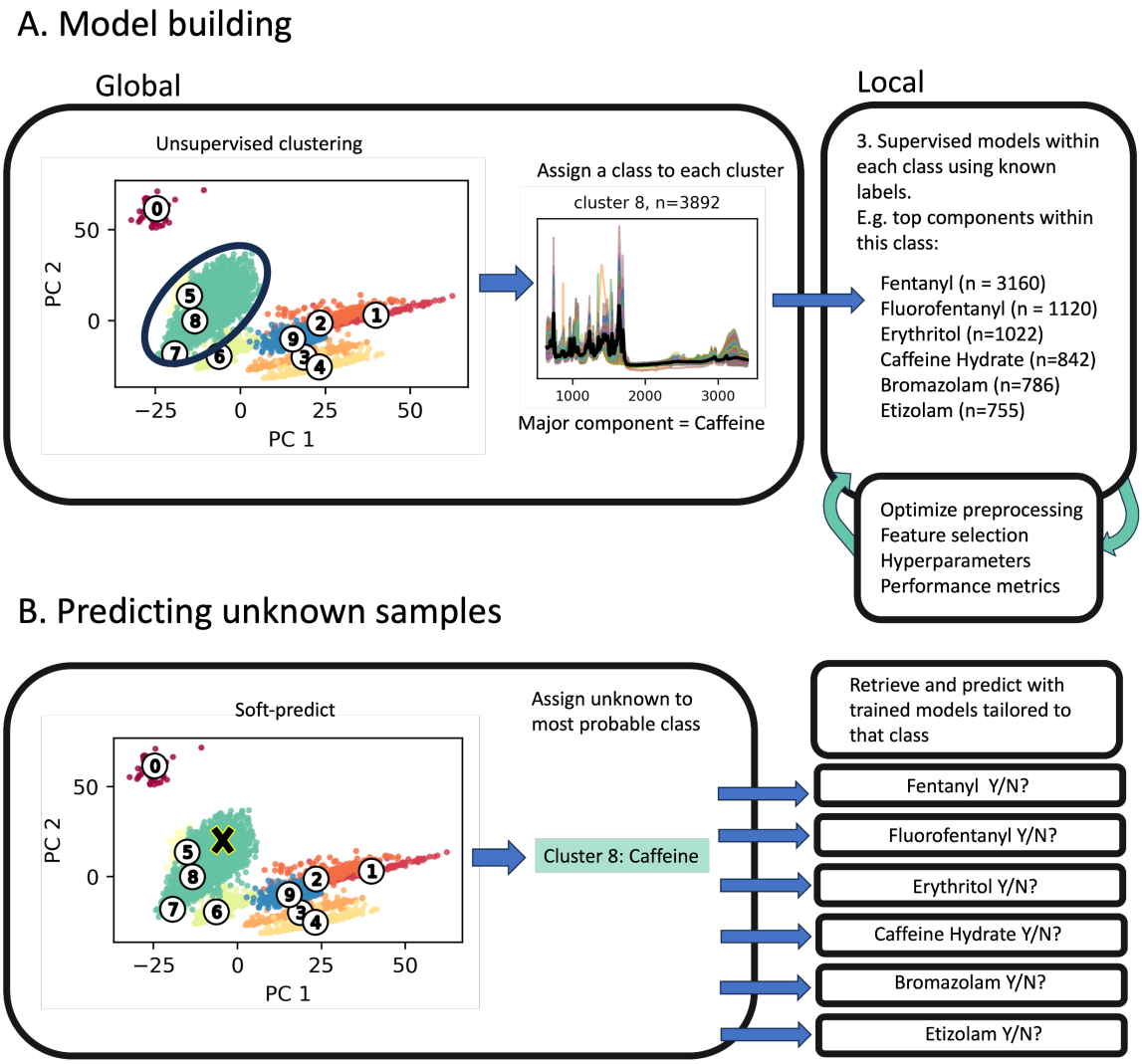

Figure 9.1: General overview of the proposed scheme for model building and automated prediction to determine the composition future unknown samples.

noise. It was implemented using `hdbscan` Python package.[257,258] There are minimal adjustable parameters within the HDBSCAN model, with the main one being minimum number of samples within a cluster (`min_cluster_size`). In general, a smaller minimum cluster size offers the opportunity for the scope of resulting clusters to be narrower and a larger minimum cluster size might provide broader groupings. For this application, aiming to use the subsets of the global clustering to train more specific supervised learning models means that the clusters need to be broad enough to have sufficient samples to train

further models, but specific enough to have certainty that they are all grouped by a similar dominant component. PCA was done before clustering for dimension reduction and ease of visualization. Clusters were also evaluated using silhouette scores. For each sample, this score is calculated using a ratio of mean distances to samples within the same cluster to the mean distances to samples outside the cluster.[259] This may be interpreted as a relative measure of clustering quality; there should be a high degree of similarity between samples assigned to the same cluster, and a high degree of dissimilarity between samples in different clusters. A silhouette score of 1.0 represents a high level of discrimination, and a silhouette score of 0 suggests overlap between clusters.

9.2.2.2 Random Forest Classifier

Random forest classifiers were implemented using `sklearn` Python package, following similar methodology as described in the previous chapter (Chapter 8). Each target compound within the clusters were used as labels for a binary classification (present or not present). This was done, as opposed to a multi-label approach, to allow for tailored pre-processing, feature selection, and to facilitate the interpretability of results.

9.3 Results and Discussion

9.3.1 Clustering

As introduced in Chapter 2, unsupervised classification algorithms do not take into account the labels associated with datasets. In this case, natural patterns are revealed irregardless of their interpretation/labels (e.g. methamphetamine vs MDMA) and instead use the features (here, infrared spectral intensities) to guide the clustering. Figures 9.2–9.3 show the raw data (not transformed into PC space) for the resulting clusters given a few varying `min_cluster_size` and `n_components` demonstrating the range of resulting cluster sizes from 6 to 19. Various approaches are worth considering in the case of exploratory analysis. However, using relative silhouette scores, and given the application of localized models,

min number of clusters was decided to be 50 with 10 PC components.

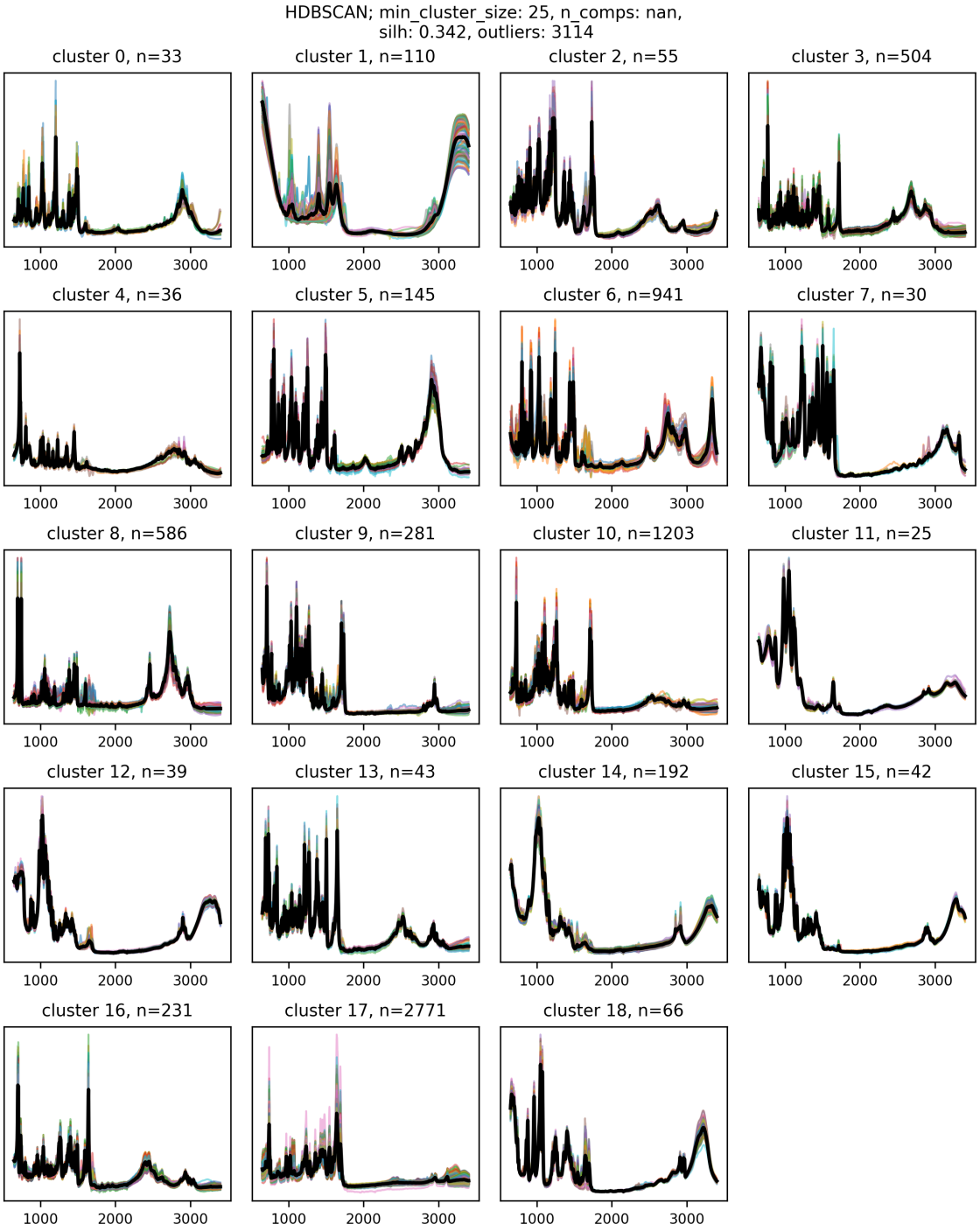

Figure 9.2: HDBSCAN clustering with minimum cluster size of 25 and high number of number of PC components, resulting in 19 clusters. The spectral traces associated with each cluster are shown, with the average of all spectra in that group overlaid in black.

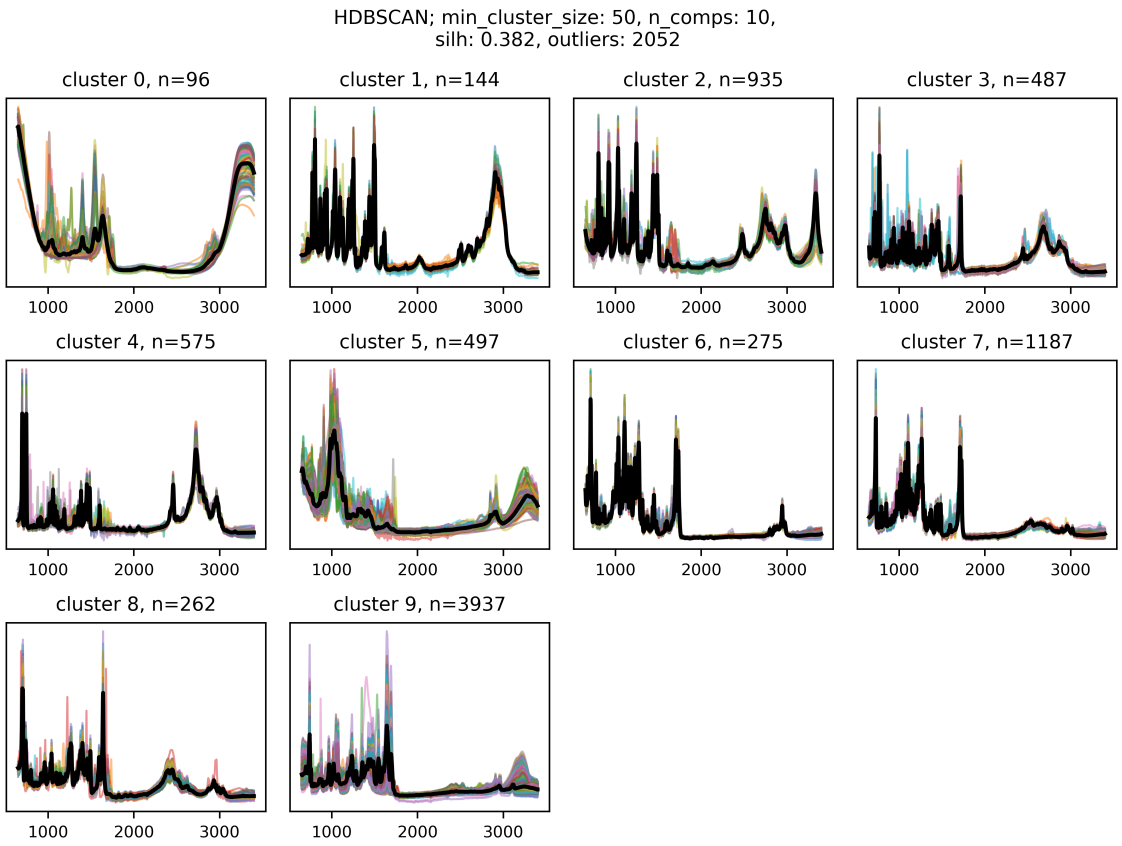

Figure 9.3: HDBSCAN clustering with minimum cluster size of 50 and low number of number of PC components, resulting in 10 clusters. These parameters were chosen for further steps. The spectral traces associated with each cluster are shown, with the average of all spectra in that group overlaid in black.

The resulting clustering is demonstrated in Figure 9.5. Using Pearson's correlation coefficient (PCC), R, searching the average spectrum within each cluster against a pure spectral library reveals the identity of the major component within each cluster. That is the clusters were determined to reflect, in order, the presence of water ($R = 0.88$), MDA ($R = 0.98$), MDMA HCl ($R = 0.99$), ketamine ($R = 0.98$), methamphetamine ($R = 0.990$), microcrystalline cellulose ($R = 0.96$), cocaine base ($R = 0.990$), cocaine HCl ($R = 0.98$), fentanyl HCl ($R = 0.97$), and caffeine ($R = 0.97$). Note that only three of the ten PCs used are demonstrated in Figure 9.5A for visualization purposes. In Figure 9.5B the silhouette coefficient values are plotted for samples within each cluster, with the average shown in the red vertical line. The silhouette coefficients can be interpreted as the degree of

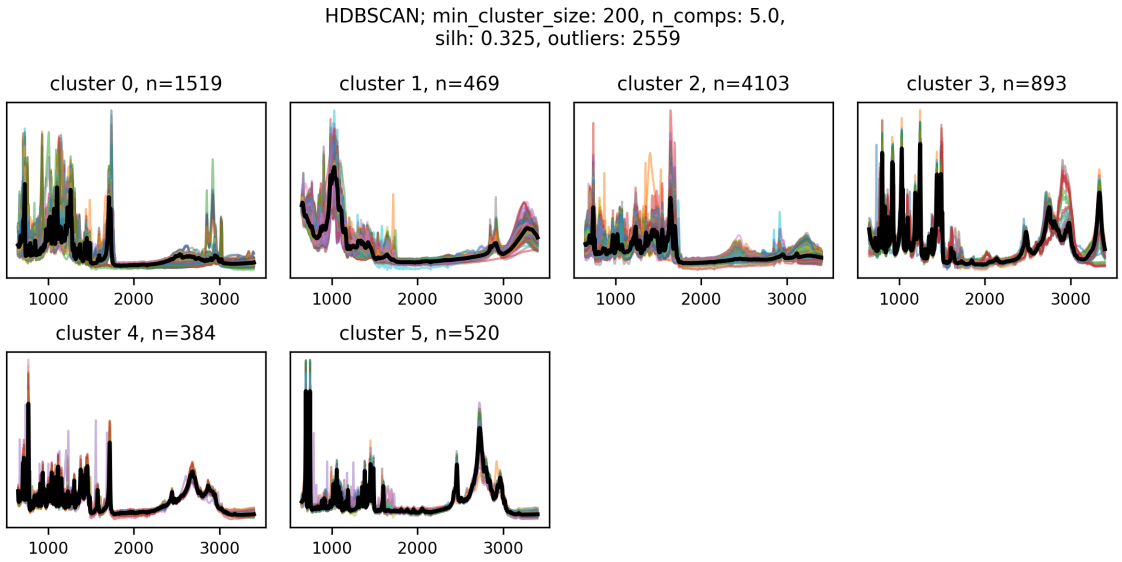

Figure 9.4: HDBSCAN clustering with minimum cluster size of 200 and low number of number of PC components, resulting in 6 broader clusters. The spectral traces associated with each cluster are shown, with the average of all spectra in that group overlaid in black.

spread within a cluster and overlap with neighbouring clusters. Cluster 9 was found to be associated with caffeine, which is most commonly present as a cutting agent in opioid or "down" samples. Notably, locally, in comparison to other drug classes, opioid samples are the most complex due to high variability in the concentration of opioids and presence of additional adulterants. In addition, caffeine may also be a cutting agent in non-opioid drugs, and therefore the presence of a wide range of drugs might result in overlap with other drug classes. In contrast, the other clusters have many samples with silhouette scores closer to 1.0 ("perfect" clustering) and visually they are much denser. In these other clusters, where the major component is typically an active and unique drug compound (e.g. MDMA, Ketamine) the spectral features are discriminatory in the absence of major cutting agents.

Notably, over 2000 sample have been classified as outliers/noise by HBDSCAN. Integrated outlier analysis in HBDSCAN is a significant advantage compared to other clustering approaches.[258] It is expected that these samples contain features characteristic to multiple clusters, are samples of infrequently seen compounds, or are of poor data quality. However for the following steps in further compound identification the outliers are forced

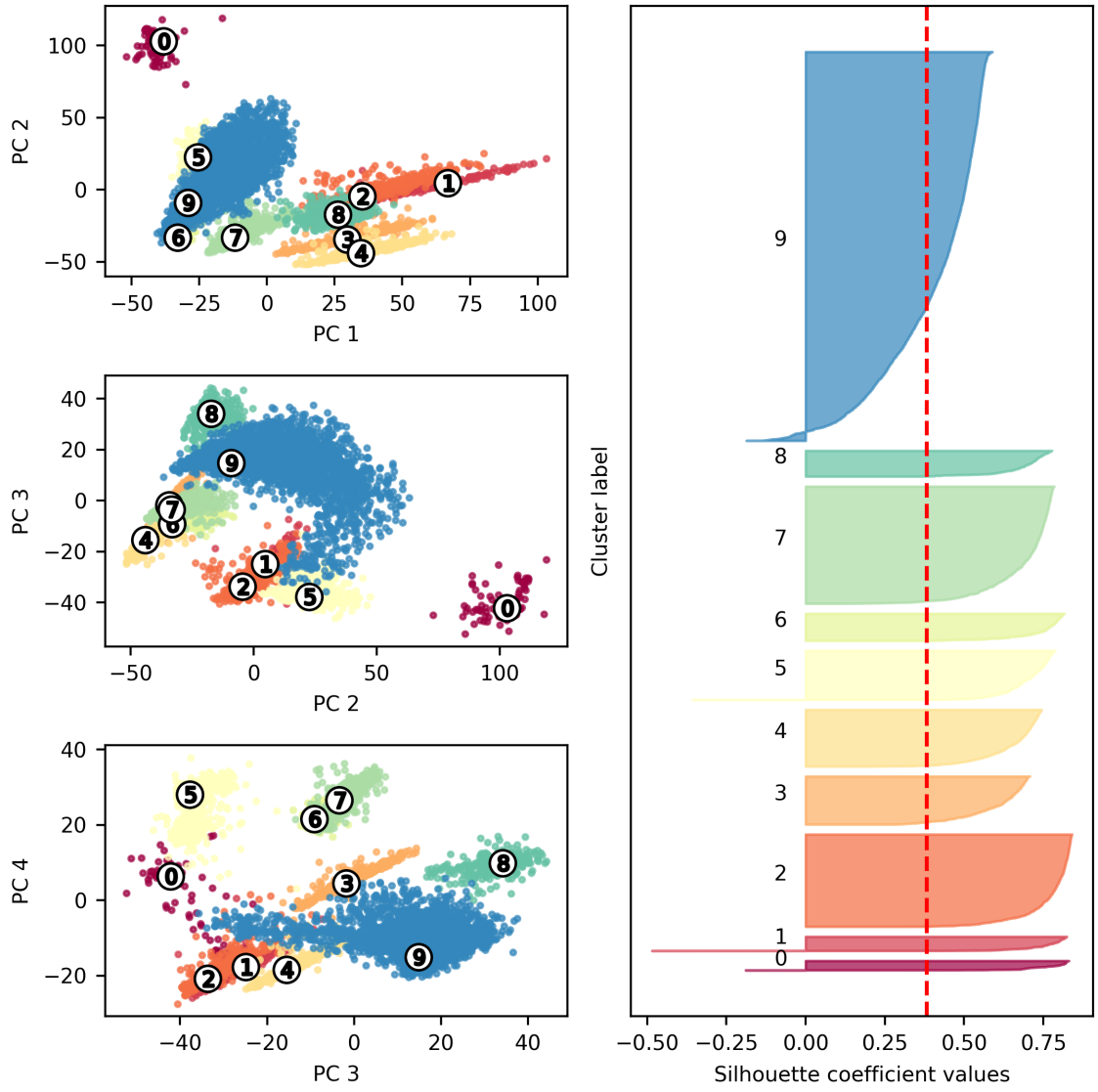

Figure 9.5: Visualization of first 4 PCs and labels from the HDBSCAN clustering. Silhouette score to evaluate the densities/degree of similarity within clusters.

(hard classification) into their most likely or closest clusters. Soft clustering (e.g. allowing samples to be assigned to multiple clusters) is also considered. This aims to have more robust subsequent classifier given some outliers (with somewhat similar features) on the cusp of clustering into a potential group. The dataset, inclusive of the outliers, is used for supervised classification steps in the following section.

9.3.2 Supervised Classification of Target Compounds

Cluster 9 (see Figure 9.3 for spectral data and Figure 9.5 for visualization of cluster in PC space) is used as an example for the following supervised classification steps. First, the labels associated with the spectra assigned to cluster 9 are investigated to get a sense of frequent components within this class. Table 9.1 shows the top components as previously determined using PS-MS and manual IR analysis. These target compounds are used as labels for training a series of binary classifiers are the results of cross validation and prediction with the a test set are how. Subsequent iterations are done with a subset of the features determined to be most influential on discriminating each particular compound. The top features are shown overlaid with a pure spectral library entry in Figure 9.6. This is presented to gain insight into whether the classifier is learning meaningful features characteristic to the target compounds.

The performance of these various models (Table 9.1) varies significantly, ranging from an F1 score of 0.99 for caffeine to 0.11 for xylazine. In most applications, 0.11 is typically not an acceptable performance for a classifier. Using flualprazolam as an example, that model resulted in an F1 score of 0.33. Recall (0.09) is the most significant challenge here compared to precision (0.89) (i.e. significant amount of false negative predictions however, few false positives). Figure 9.6 reveals the most important features weighing in on the prediction of flualprazolam. The top features align with some strong vibrational modes characteristic to flualprazolam, namely around 830 cm^{-1}. Some clarity for the resulting poor performance is realized through the data compiled in Table 9.2. Here some statistics (median, interquartile range) regarding the distribution of target compound concentrations within the dataset is reported. Compounds consistently present below 3% (see bromazolam, carfentanil, flualprazolam, and etizolam, among others) have far poorer performance. However, through this process the ability to detect such substances when present in higher concentrations can be monitored. This evaluation facilitates the development of concrete limitations that can be communicated surrounding low concentration detection

and frameworks to monitor the statistics around this.

Table 9.1: Summary of first steps of evaluating the series of random forest models for target compounds within cluster 9.

target compound	training set neg:pos	test set neg:pos	F1 test	cv mean F1	cv mean precision	cv mean recall	top feature (nm)
bromazolam	2563:626	1098:269	0.71	0.68	0.89	0.56	830.0
caffeine	3008:181	1290:77	0.99	0.99	0.99	0.99	744.0
carfentanil	3099:90	1329:38	0.23	0.24	0.88	0.14	1042.0
erythritol	2369:820	1015:352	0.90	0.91	0.90	0.92	1052.0
etizolam	2598:591	1114:253	0.56	0.61	0.83	0.49	798.0
fentanyl	2867:322	1229:138	0.97	0.97	0.96	0.98	708.0
flualprazolam	3048:141	1307:60	0.33	0.16	0.89	0.09	826.0
flubromazepam	3067:122	1315:52	0.40	0.35	0.93	0.22	816.0
fluorofentanyl	2326:863	997:370	0.77	0.76	0.89	0.66	822.0
heroin	3095:94	1327:40	0.63	0.72	1.00	0.57	1156.0
mannitol	2917:272	1251:116	0.69	0.72	0.86	0.62	1018.0
microcrystalline cellulose	3145:44	1348:19	0.56	0.65	0.85	0.54	2847.0
xylazine	3032:157	1299:68	0.11	0.08	0.55	0.05	1621.0
xylitol	3129:60	1341:26	0.63	0.80	0.75	0.87	856.0

Table 9.2: Median and interquartile range of the concentrations (w/w%) of various target compounds reflected in the training set. F1 scores and the number of features used are also shown for the final model. Note that from the database for particular compounds such as caffeine, erythritol, mannitol, xylitol and microcrystalline cellulose concentration data is unavailable as it is typical qualitatively detected using spectroscopy or above the linear range of PS–MS.

compound	median (w/w%)	IQR (w/w%-w/w%)	Not reported (n/ntotal)	F1 final	number of features
bromazolam	2.11	0.79 – 4.78	19/896	0.76	50
caffeine	50.0	40.0 – 60.0	4279/4308	0.99	200
carfentanil	0.25	0.13 – 0.54	3/128	0.44	100
erythritol	40.0	20.0 – 40.0	1137/1172	0.91	200
etizolam	2.65	0.83 – 8.6	49/844	0.63	50
fentanyl	9.74	4.55 – 15.83	87/4108	0.97	200
flualprazolam	0.48	0.21 – 1.25	18/202	0.55	20
flubromazepam	2.01	0.61 – 3.52	10/174	0.60	5
fluorofentanyl	3.49	0.64 – 9.76	21/1235	0.79	20
heroin	12.25	4.03 – 26.71	31/134	0.78	50
mannitol	20.0	19.0 – 35.0	365/388	0.70	200
microcrystalline cellulose	35.0	35.0 – 35.0	62/63	0.65	20
xylazine	0.81	0.16 – 3.16	0/225	0.22	50
xylitol	25.0	25.0 – 25.0	85/86	0.68	100

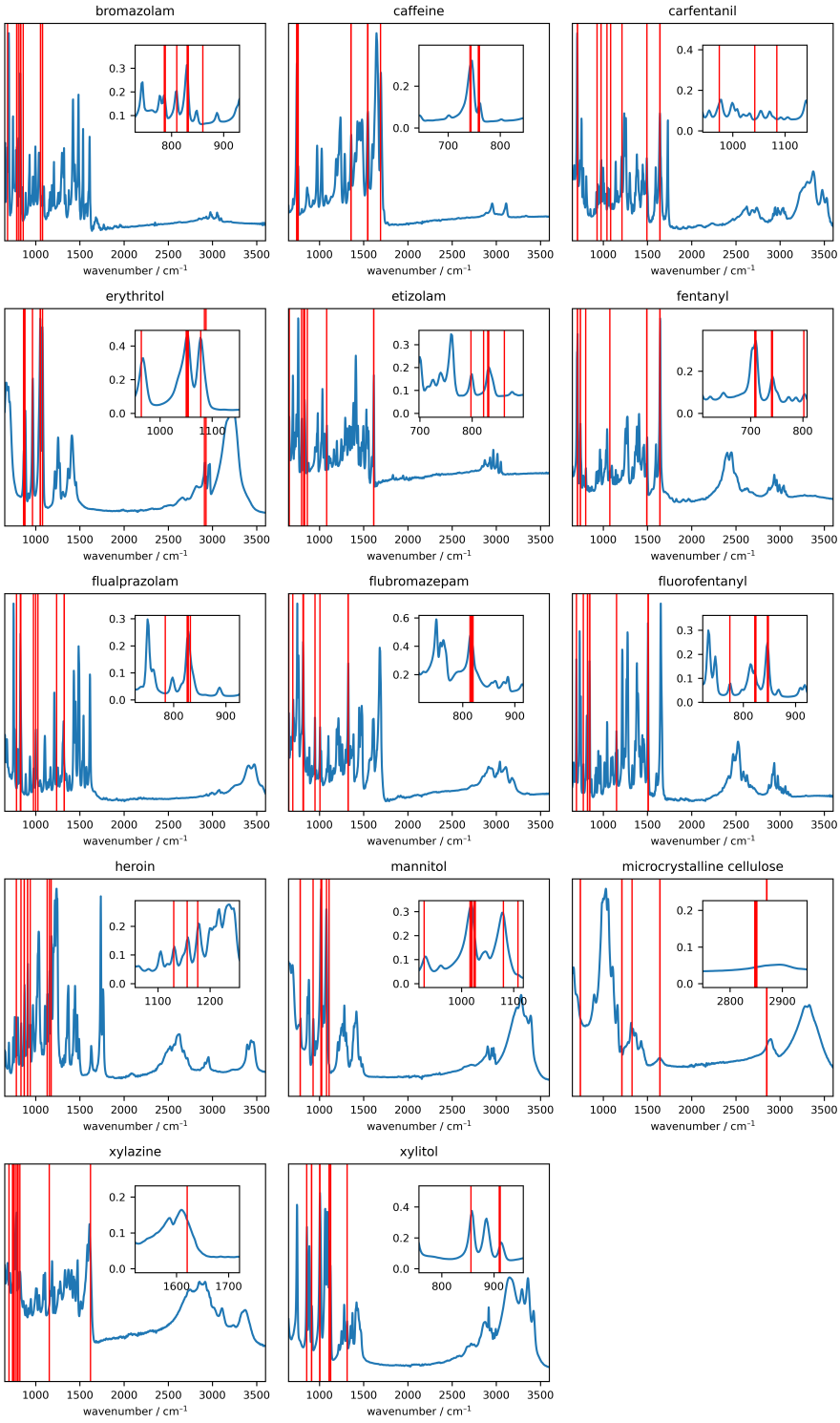

Figure 9.6: Top 10 important features as determined in training the random forest classifier for each target compound and overlaid with their library spectrum. The inset zooms into the area surrounding the most weighted feature.

9.3.3 Application to Unknown Samples

This two-step pipeline (class category followed by target compound predictions) is applied to the test set that was set aside at the beginning. This is the most recent data collected from May to August 2023.

Of the 2612 samples, 83% were predicted into a cluster, with the remaining 17% predicted as outliers. $n = 1029$ were categorized in Cluster 9, characteristic to mixtures with caffeine. Continuing with Cluster 9 as an example, the $n = 1029$ samples were then predicted with the random forest classifiers. The results are shown in Table 9.3 for each of the target compounds. In general the results on the unseen test set reasonably reflect the expected performance, based on the training steps previously shown. For example, the ability of the models to predict carfentanil with any accuracy is limited. It is interesting to consider in cases like this where very few samples exist in concentrations greater than 2 w/w%, and therefore validation of such models within an appropriate range for the technology is challenging. The true negative rate for the prediction of carfentanil is high, where only 9 out of 1022 samples had a false positive prediction. Regardless, it is reasonable to say that predicting carfentanil in the current supply using infrared spectroscopy is not possible and not recommended. This however could continue to be monitored over time to assess the ability of carfentanil prediction in higher concentrations. The concentration of the target compound within the test set plotted against their prediction of not present (0, false negative) or present (1, true positive) is shown in Figure 9.7. The purpose of this visualization is to monitor whether predictive ability trend approximately with their percent (e.g. limit of detection) and provide insight beyond the metrics in Table 9.3. For instance, 155 samples that contained bromazolam were predicted as false negatives. However, Figure 9.7 reveals that most of the false negatives are for samples with which the target compound is present in low concentrations. On the other hand, the model provides a more reliable prediction above the limit of detection. A similar trend can be seen with bromazolam, fentanyl, fluorofentanyl, and heroin. Some outliers exist, for

example for fluorofentanyl and heroin their detection in a few high concentration samples (greater than 20%) have been missed. This is surprising, and warrants further investigation and monitoring. For example, additional, or new compounds may be interfering.

Table 9.3: Results of the test samples from May–August 2023, categorized into cluster 9 and predicted with the random forest models. Metrics include F1 score, recall, precision, true negative (TN) and positive (TP), and false negative (FN) and positive (FP) for each target compound.

compound	F1	recall	precision	TN	FP	FN	TP
bromazolam	0.763	0.631	0.964	599	10.0	155.0	265.0
caffeine	0.995	0.997	0.993	1	7.0	3.0	1018.0
carfentanil	0.000	0.000	0.000	1013	9.0	7.0	0.0
erythritol	0.915	0.904	0.926	626	27.0	36.0	340.0
etizolam	0.522	0.462	0.600	1012	4.0	7.0	6.0
fentanyl	0.927	0.971	0.888	24	110.0	26.0	869.0
flualprazolam	0.400	0.333	0.500	1021	2.0	4.0	2.0
flubromazepam	0.000	0.000	0.000	1019	0.0	10.0	0.0
fluorofentanyl	0.861	0.871	0.852	353	89.0	76.0	511.0
heroin	0.600	0.529	0.692	1008	4.0	8.0	9.0
mannitol	0.787	0.729	0.854	975	6.0	13.0	35.0
microcrystalline cellulose	0.000	0.000	0.000	1029	NaN	NaN	NaN
xylazine	0.000	0.000	0.000	973	2.0	54.0	0.0
xylitol	0.959	0.921	1.000	991	0.0	3.0	35.0

9.3.4 Detection of Additional Compounds

The pipeline described so far is only suitable for detecting compounds that have been commonly seen in the local drug supply to date. By definition, this approach covers many of the samples seen on a daily basis at the drug checking service. However, investigating unaccounted for spectral signal and novel compounds is of interest. As mentioned in previous chapters, a great deal of drug checking with infrared spectroscopy takes advantage of spectral libraries and spectral matching software to identify probable components. This is often followed by attempts to remove the spectral features attributed to one compound and subsequent searching based on the remaining residual. The success of this process is subject to instrumental variation, baseline shifts, degree of overlap of the spectral features,

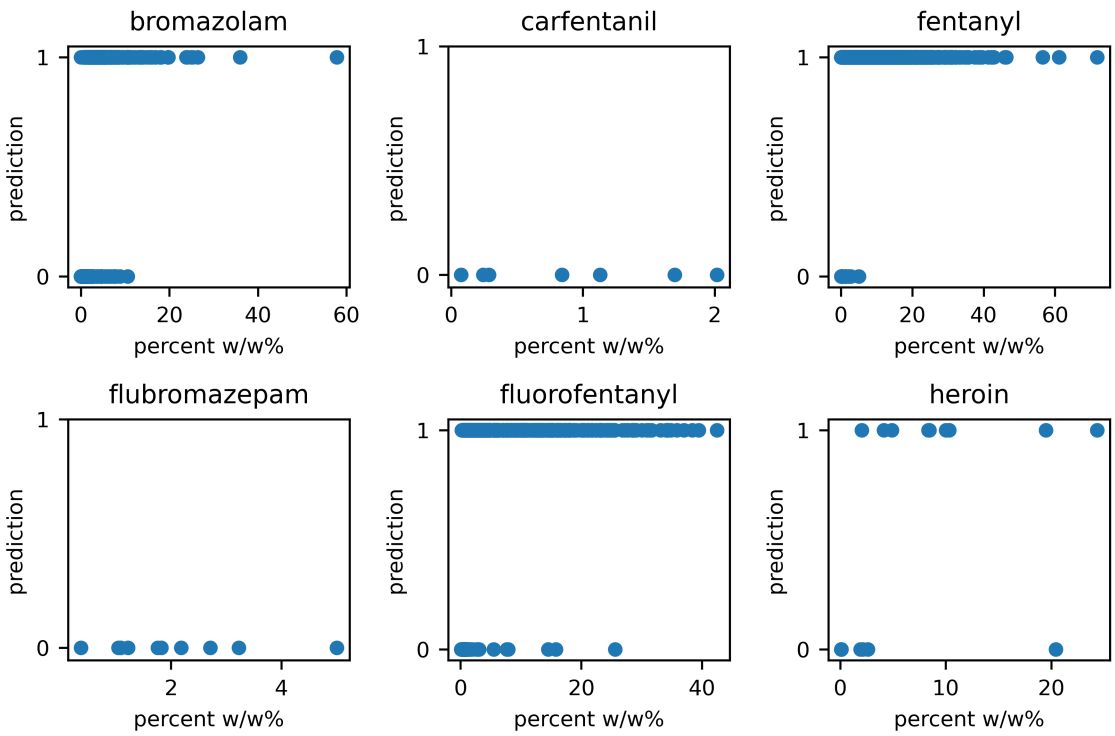

Figure 9.7: The concentration of a subset of the target compound as present in the test set plotted against the random forest model prediction (1 or 0). This reveals the approximate limit of detection for each compound.

and technician experience. Ongoing work uses the framework established so far, but focuses on the detection of additional components within the 10 major clusters. This refers to the detection of components that are not on the target list of random forest classifiers (e.g. untargeted identification). For instance, mixtures that are not common often means there is not a sufficient dataset to build supervised machine learning models. Efforts in untargeted searching also aims to identify novel or unexpected compounds in the supply a soon as possible. The current implementation to detect additional components outside of the comprehensive models aims to do this by performing a linear fit using the pure spectral traces of any identified components according to the previously presented pipeline. This was shown in Section 10.2.3.4 (Figure 10.6). The result is (a) an estimation of proportions of those substances and (b) a residual spectrum representing error between the fit and the original spectrum. The residual (remaining signal) can be used to explore library searching

for the identification of any additional compounds. While automated, this is a simple implementation which is still subject to the quality of spectral libraries and many artifacts in the residual spectrum. Now, this work extends beyond such approach to extracting pure spectral traces from our own clustered data by identifying recurring patterns within the dataset. For instance, by grouping similar spectra and applying PCA, variables that covary group into PCs and the loading plots often contain informative spectral information.[161] Such spectral information can be further refined using approaches like multivariate curve resolution (MCR) to optimize a matrix of spectral component features and vector of concentrations (or weights) fit to the spectral data matrix.[260] This concept is commonly applied to hyper-spectral image analysis, where each pixel may contain various proportions of similar compounds, and through that linear relationship the pure components and their concentrations can be estimated.[261] The interesting thing about this approach is that extracted spectral features may represent compounds or spectral variations that are not in common libraries. Some examples include minor leftover precursors, polymorphs, varying hydration states of common substances, or distortions due to matrix effects. Practically, contributions from these features can now be cleanly removed as expressed within our own dataset. The residual calculated now, is more informative, as opposed to the residual from the crude subtraction of library entries. Monitoring the mean square error of the residuals over time can also help indicate whether the remaining signal has a typical amount of noise, or indicates that an additional compound might be present in a substantial amount. The data from this approach is in the process of being explored and evaluated, with the goal to eventually be generalized for a drug checking workflow.

9.4 Conclusions

This chapter presents the most recent progress on an exhaustive machine learning pipeline for infrared spectral data. Data from November 2020 to May 2023 was used to develop a multi-step framework that was tested with the latest spectral data acquired from May 2023

to Aug 2023. During the training, the clustering method of HBDSCAN was used to initially separate the highly variable dataset into smaller, homogenous groups. The resulting ten clusters were shown to successfully group related spectra based on their dominant features. These subsets of spectra were used as a starting point to expose subtle differences related to the presence of additional compounds. Initially this was done by a series of binary random forest classifiers to target frequently seen compounds within each subset. Using the subset of samples where most the dominant features were attributed to caffeine as an example, several compounds commonly detected in opioid mixtures were targeted, including fentanyl, fluorofentanyl, heroin, and various benzodiazepines. The performance was highly varied for various compounds, and the correlation between the performance (F1 score) and concentration of target compound was revealed. This comprehensive evaluation, even in cases where the model performance is unremarkable, is important to understand and communicate the limitations of drug checking results using infrared spectroscopy and provides important quantitative metrics such as prediction probability and limit of detection to aid in this messaging. Additional untargeted analysis was proposed using multivariate curve resolution in combination with spectral library matching. A truly comprehensive analysis provides a framework for detecting frequently seen substance, as well as the ability to monitor unexpected components/error not restricted to this target list.

Chapter 10

Development of a Custom Software Platform for Point-Of-Care Data Analysis

10.1 Motivation

Throughout this work evaluating and improving various drug checking technologies, as described in the previous chapters, a main goal was to streamline implementation of research in the community drug checking service offered at the same time. Therefore, early on I proposed a tailored drug checking platform (i.e. software) to achieve such goal and be integrated within the workflow of the active service. Initially, as a multi-technology service the prototype was designed to serve as an analysis platform for several data types such as infrared, Raman, and SERS spectra. This is an alternative to the onboard (and limiting) software that came with the individual commercial instruments. To my knowledge no drug checking service has pursued developing their own analysis system. The dynamic needs of drug checking projects therefore often go unmet, or at the very least, are dependent on industry partners to facilitate changes to a software that is often catered towards a very broad range of applications (e.g. forensics, law enforcement, pharmaceuticals, environmental). This chapter describes the evolution of the drug checking platform that is now used daily in the drug checking service for analysis of spectral data by drug checking technicians and harm reduction workers.

10.2 Evolution of Features and Interface

10.2.1 Applying Research into Service

The initial implementation of the drug checking platform was very simple, and born out of need to apply the quantification models for fentanyl (e.g. Chapter 3) to spectral data acquired at the service in real-time. Technicians could select a sample, spectral data would be piped from the database into their Jupyter instance, predicted with the trained PLSR models, and the output of the model displayed.

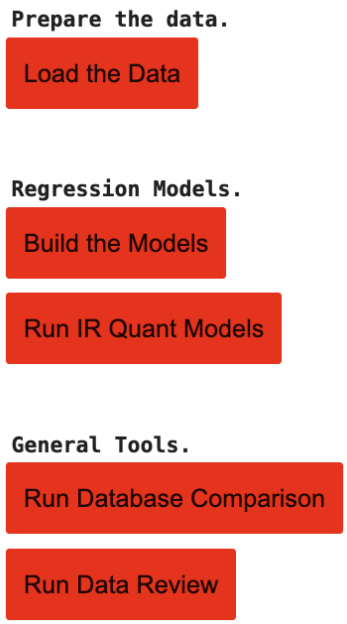

Figure 10.1: The first implementation of translating research to use in real time as a tool in the drug checking service.

10.2.2 Designing and Implementing an Enhanced Drug Checking Platform

I then implemented an interactive version, which took advantage of past spectral data to make a visual comparison. For example, most untargeted spectral analysis requires libraries of pure components, which comes with various differences such as spectral shifting, baseline shifts, etc. However, notably we were checking very similar drugs over

and over again and this data was not being used. Simply, I implemented a matching scheme that revealed the top 5 matches from real drug samples acquired at the service and their associated interpretation. Within the platform I initially considered the data from each instrument separately for data analysis. For example, an IR spectrum of a substance acquired at point-of-care is piped into a user interface prototyped in JupyterHub (using Python's `ipywidgets` functions) for data exploration. An example framework of the IR analysis is shown in Figure 10.2. This allows for (1) subtle differences to be highlighted since spectral variation from baseline and imperfect spectral library comparison was mitigated and (2) a head start on spectral interpretation. This was also expanded to several datatypes such as Raman and SERS. The interface also includes adjustable parameters, for example the selection of wavelength ranges, choice of libraries, and pre-processing schemes to influence library searching algorithms.

10.2.3 Current Implementation

10.2.3.1 Platform Architecture

The initial analysis interface was prototyped using Jupyter notebooks and deployed in a JupyterHub setting. Interactive tools were created using various Python packages such as `ipywidgets` and `plotly`. Recently, as the interface became more complex, considerations of usability, extensibility, and performance motivated the creation of an analysis platform architecture that leverage the separation of service data, spectral storage, and machine learning models. The duties of fetching data from the database, spectral storage, and model storage were distributed into separate microservices, a collection of loosely-coupled services that comprise a larger software application. By decoupling functional responsibilities, each microservice can be more efficiently maintained at an individual level, whether to modify functionality or scale up its allocated resources in real time depending on service demands. In our implementation, the available machine learning models are allocated to their own microservice, where each model is hosted at a unique

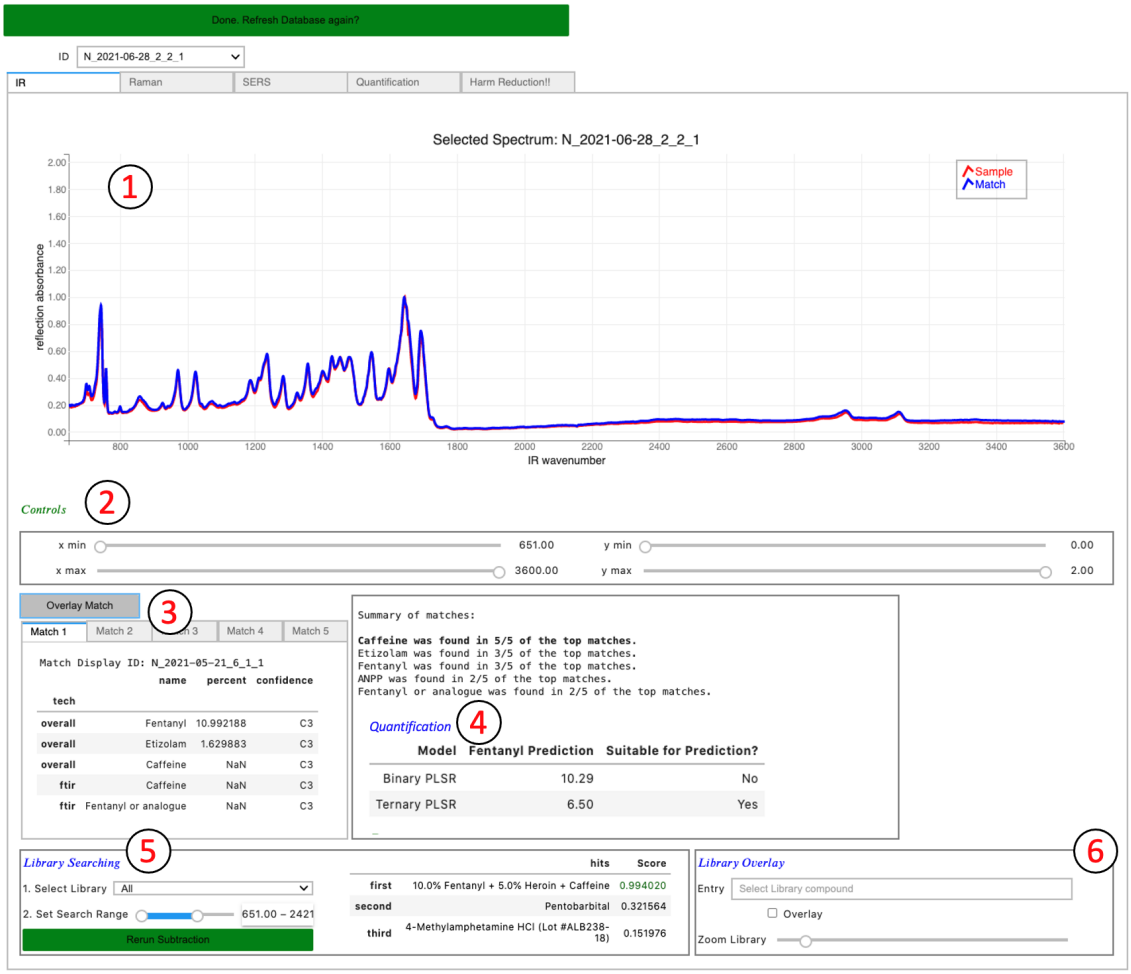

Figure 10.2: Example user interface for IR data analysis integrated into a JupyterHub tool. Features include a platform for drug checking technicians to visually explore the IR spectrum (1 & 2), a comparison to database of past street samples and their interpretations (3), quantification models for use when appropriate (4), a library searching algorithm with adjustable parameters (5), and overlaying capabilities with pure library entries, including any standard mixtures that are available (6).

endpoint on a server, and can be queried from the interface individually or as part of a cascading chain of model calls as necessary. By decoupling model storage from the rest of the application, models can be added, removed, and modified without service disruptions to the end user. This feature is essential, as the models must evolve alongside the data acquired from point-of-care sites. The interface microservice itself is a single page web application shown in Figure 10.3 that is accessible only by authorized service technicians and affiliated researchers.

10.2.3.2 Exploration of Spectral Data with Interactive Library Searching and Database Matching

Features of the interface are guided by recent and ongoing progress in quantification models, classification models, library searching schemes, and targeted analysis. The interface includes adjustable parameters to influence library searching algorithms, such as the selection of wavelength ranges and choice of libraries (Figure 10.3A).

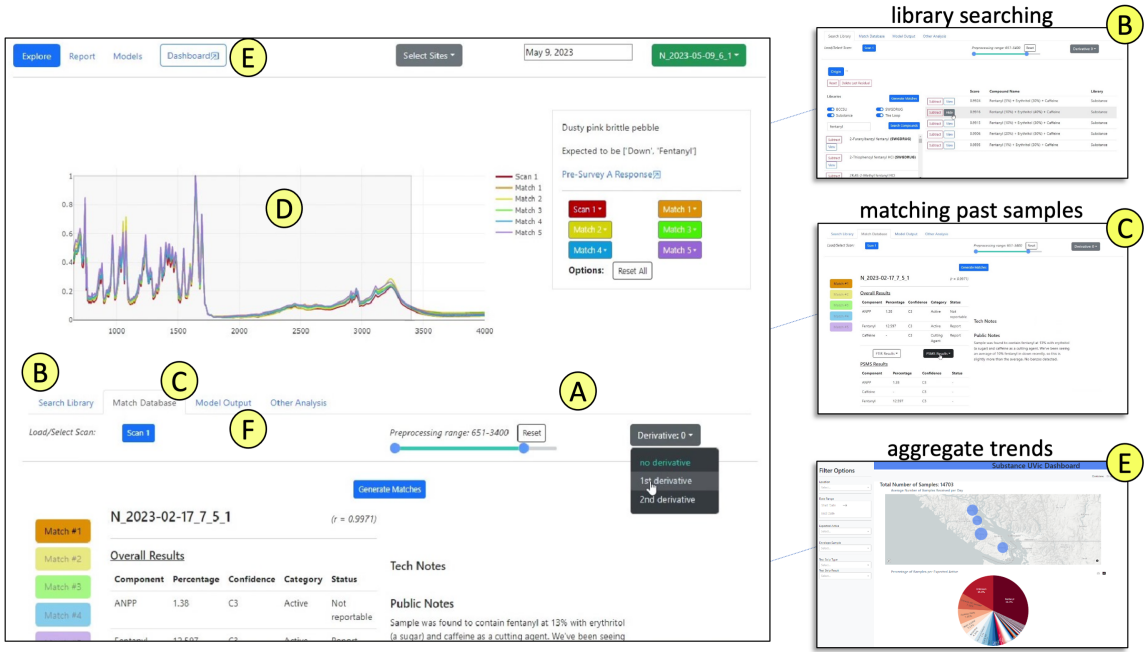

Figure 10.3: The main explore page of the spectral analysis tool. Features include pre-processing directives (A) such as the selected range and derivatives can be tweaked to produce different matches to library entries (B) or to a database of past service samples (C), and a platform for drug checking technicians to visually explore the IR spectrum (D). A separate dashboard page (E) displays aggregate drug checking statistics dynamically, according to filters and ranges set by the user. The output of quantification and classification models are also piped into the interface (F).

The spectral data of pure compounds, pulled from internal and external drug checking libraries, can both be manually searched and used as a baseline to automatically find close matches for given input samples (Figure 10.3B). Spectral data acquired from point-of-care intake is used to further enhance the spectral matching capabilities of this software. Pearson's correlation[262] is used to compare incoming samples to past samples stored in the project database, and the top matches for IR spectral data are presented to the user

(Figure 10.3C). When a close spectral match is found for a given sample, a technician may gain significant insight towards the interpretation for that sample. A simple visual comparison of the spectrum for an incoming sample to past spectra also draws attention to more subtle differences between the two mixtures (Figure 10.3D). In many cases, samples stored in the database have undergone additional investigation using GC–MS or PS-MS, and therefore offer additional information. If a sample is encountered at point-of-care for which no close spectral match exists, a technician is then able to adjust their interpretation strategies accordingly given that the specific compound or mixture in question has not been encountered before. This is important as, in many cases, portable technologies with limited sensitivity are the only instruments available at point-of-care testing. Here, the messaging of results to the service user may be extended to rely on relevant aggregate data from more sensitive confirmatory testing methods.

10.2.3.3 Application of Analysis Pipelines and Generation of Automated Reports

Given the current drug checking technologies, and training required to operate them, there is a significant challenge in addressing the increasing demand for rapid services, multiple sites and extensive hours to adequately serve the needs of diverse and dispersed populations.[205] Therefore, I proposed an additional representation that automates several decision-making steps, traditionally performed by an experienced technician.

Machine learning models developed for this context must be capable of providing accurate predictions in the face of highly variable data. Here we show two examples of analysis pipelines that currently exist in the scope of this software: a partial-least squares regression (PLS-R) model for fentanyl quantification,[20] and a random forest (RF) model used for compound classification.[253] Their training and evaluation has been previously described in detail.[20,253] The pipelines shown in Figure 10.4 are demonstrated with an IR spectrum of an expected opioid sample acquired at the point-of-care drug checking service. In the quantification pipeline (Figure 10.4A), baseline correction and area normalization were applied before prediction with the trained PLS-R model. An outlier detection

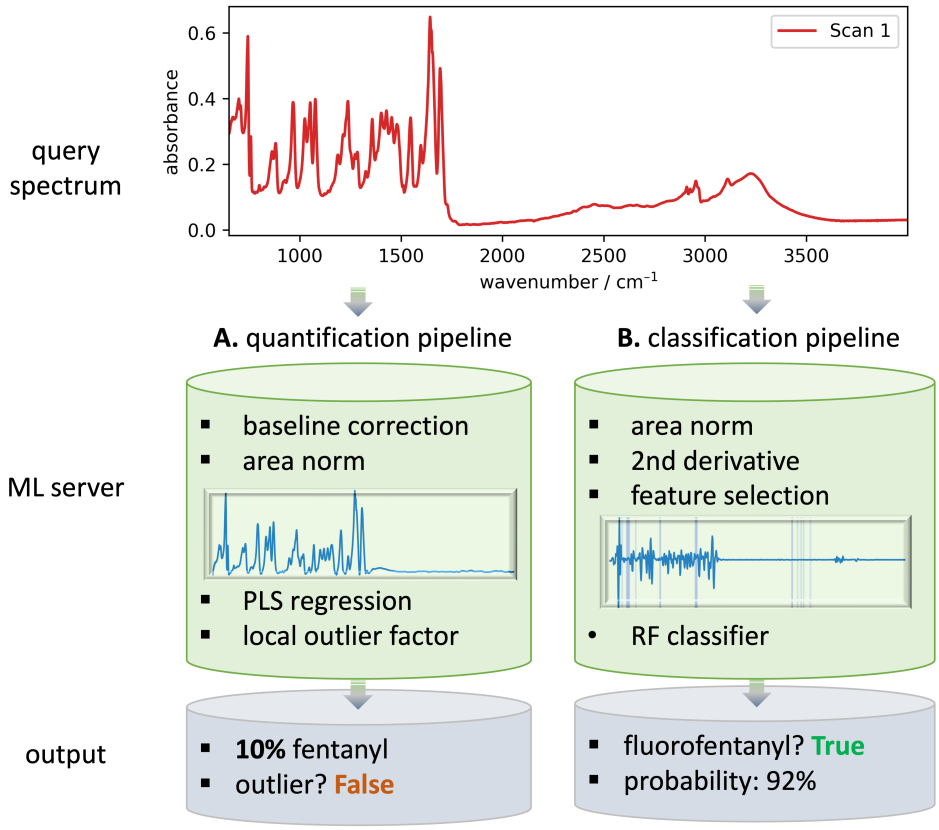

Figure 10.4: Two use-cases of the vibrational spectroscopy server, where machine learning models are stored for quantification and classification tasks. An example is shown of a spectrum representing an opioid sample, which is passed to the ML server to undergo spectral manipulation and prediction with two models: (A) quantification of fentanyl and (B) detection of fluorofentanyl (both previously described in literature in Refs. 20 and 253). The output is returned to the spectral analysis interface for use by a technician.

scheme was also used as a safeguard to ensure that only samples that fall within the scope of the model are predicted. The model predicted 10% fentanyl, and the sample was not classified as an outlier, suggesting that this is an appropriate sample for fentanyl quantification. Similarly, a classification pipeline is shown for predicting the presence of fluorofentanyl (Figure 10.4B). Here, area normalization, second derivative pre-processing, and feature selection were applied to the raw spectrum. The transformed spectra was then predicted with the RF classifier with an output of true (fluorofentanyl detected) and a prediction probability score of 92%. Outputs from both models are returned into the spectral analysis interface/toolkit (Figure 10.3E). Such models aim to improve throughput,

ensure greater consistency between technicians and sites, and reduce or eliminate repetitive tasks, enabling technicians to focus on tasks that maximize use of their skill and experience. In the current implementation, wherein a technician uses both machine learning and manual exploration, the final interpretation ultimately depends on technician interpretation and consolidation of the results from library searching, visual evidence, model outputs, and aggregate statistics.

Machine learning models, such as the two examples demonstrated here, continue to be trained, validated, and piloted prior to integrating into the spectral analysis interface at point-of-care. This flexibility and celerity in adding and improving models as drug checking evolves is essential for their ongoing relevance to analysis of the drug market. Ultimately, these machine learning-assisted analysis tools aim to address the accessibility of drug checking and be integrated in a kiosk application for a fully automated analysis. It removes the requirement of rigorous training and experience on the part of the technician and it is a step towards community drug checking by community members themselves. A current pilot implementation is show in Figure 10.5. The details of the ongoing development of the automated pipeline were presented in Chapter 9.

10.2.3.4 Integration of Explainable AI to Review Model Predictions

The concept of explainable AI (XAI) was introduced in Chapter 8. Figure 10.6 shows the implementation of such concepts within the current platform. When components are predicted, their library spectrum is automatically overlaid, along with highlighting the spectral features (vertical lines) that were important to the prediction of that compound. This is an interactive tool, where technicians can explore the spectral features and make connection between the predictions, the library spectral features, and the query spectrum. Additionally, a linear fit is performed with the library spectrum to produce a composite spectrum and residual. This is another method to demonstrate how well the query spectrum is explained by the detected compounds and make a visual comparison.

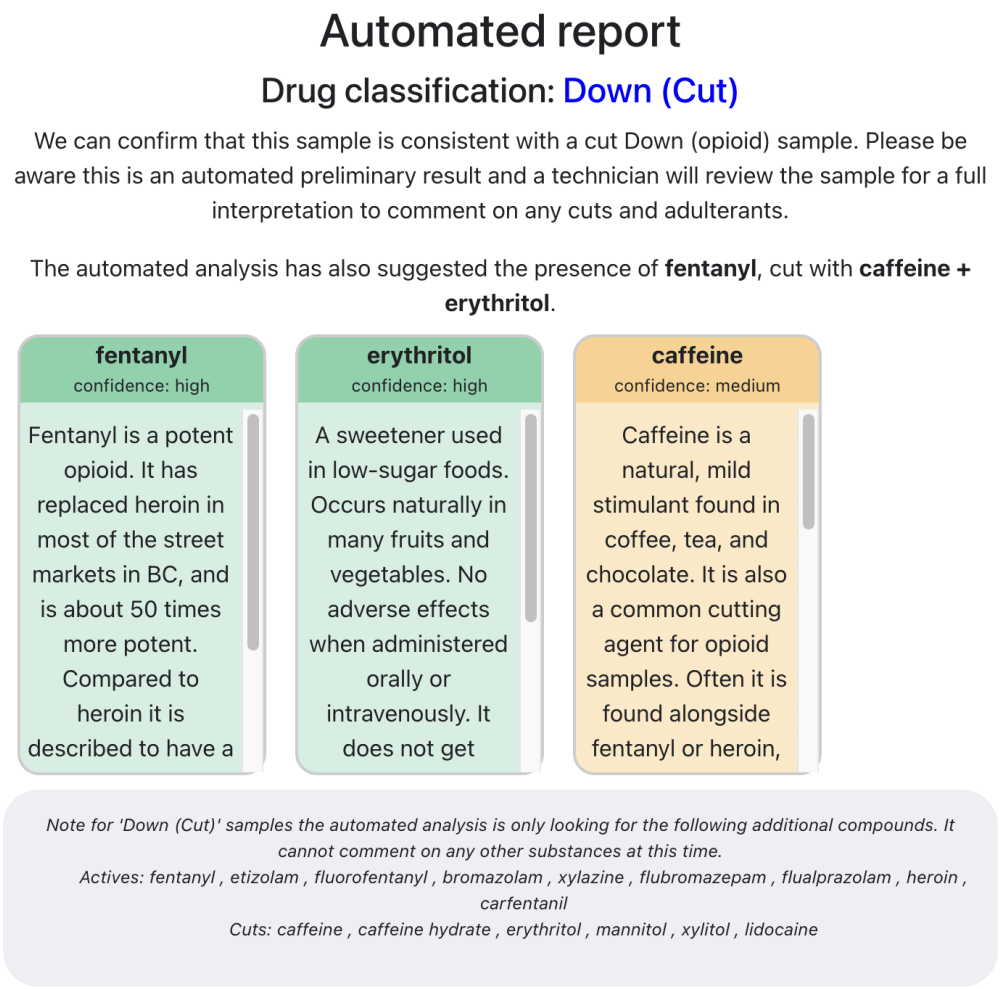

Figure 10.5: An automated report generated from the output of the classification pipeline developed for various drug classes.

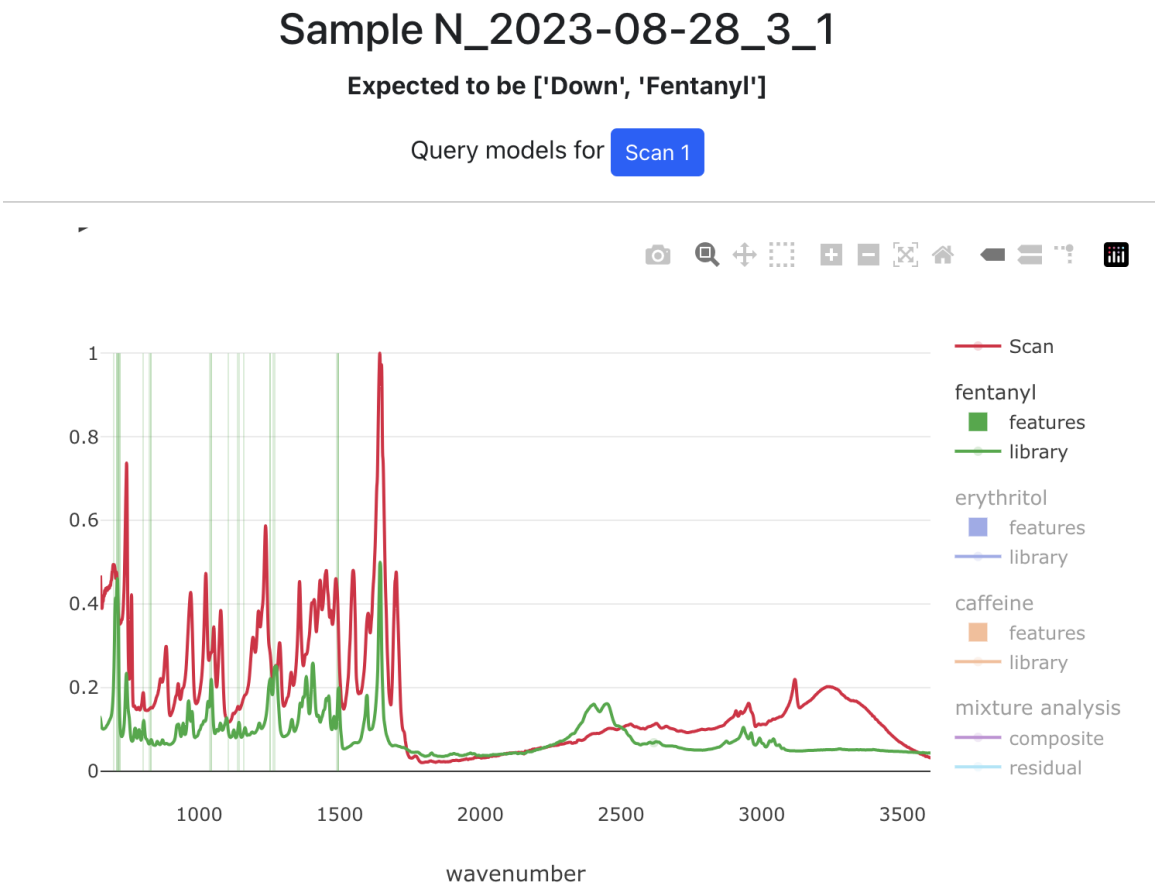

Figure 10.6: Explainable AI that expands on the automated list of components and uses the approaches as described in Chapter 8 to offer a visual aid to support and explain model predictions.

10.3 Conclusions

A flexible framework for analyzing chemical data was implemented. This platform was initially motivated by the need to streamline research, such as the developed fentanyl quantification models, into practice. Initially prototyped in Jupyter notebooks this platform has evolved throughout the course of this thesis. This began with implementing unique approaches to compare and visualize spectral data, integration of aggregate trend data, and relevant harm reduction information. To improve the usability and potential reach of this software, the features and platform architecture were redesigned as a single page web application outside of the Jupyter environment. This enabled added flexibility and efficiency in applying complex machine learning pipelines and generating an automated report based on the outputs. Through the use of tailored software and machine learning, drug checking projects can implement quick and effective harm reduction services without the same level of technical expertise required of previous drug checking software. This tool is positioned to increase the scale and reach of drug checking to remote or under-serviced communities and providers.

Chapter 11

Conclusions

11.1 Summary of Work

Chapter 1 places the objectives of this work within the principles of harm reduction and the overdose crisis. Chapter 2 provides both a technical and practical background to immunnoassay test strips, Raman, IR and SERS spectroscopy, and gas chromatography mass spectrometry. The field of chemometrics in analytical chemistry is introduced and important concepts in model building for classification, data exploration, and quantification are explained with example code included.

In Chapter 3, a regression models was trained and optimized for the quantification of fentanyl HCl in lab-made mixtures of various cutting agents. A pipeline was then implemented for the application of the model to real drug samples from the service, including fine-tuning outlier detection protocols. The prediction pipeline was retroactively applied to samples collected between December 2018–September 2020. This project was in response to demand from service-users for more accurate fentanyl quantification shortly after the service began operating.

Beginning in late 2020, a subset of samples were tested with a lab-based mass spectrometer that revealed the presence of highly potent substances, particularly fentanyl analogues and benzodiazepines (and related substances) in the opioid supply. This was a similar experience throughout the province, where services using infrared spectroscopy

alone were unable to reliably detect these compounds.[42,43] In Chapter 4, using a combination of drug standards and a subset of opioid samples collected at service, the capability of portable GC–MS to detect etizolam and carfentanil at point-of-care was investigated and compared to analysis with infrared spectroscopy.

The growing presence of low-concentration benzodiazepines in the opioid supply continued to challenge drug checking instruments. Namely, benzodiazepine test strips were beginning to be trialed as a binary yes/no screening for the presence of benzodiazepines. It was soon realized that etizolam, a benzodiazepine-related drug, was unreliably detected with these immunoassay strips. In Chapter 5, a protocol was presented for surface enhanced Raman spectroscopy and a simple univariate model was built for the reliable detection of etizolam in drug mixtures. Further potential for multivariate techniques like principal component analysis in this application were demonstrated.

As drug checking continued to expand across North America, many questions remained unanswered on the "ideal" point-of-care technology. Despite the need for this knowledge, the most recent review of potential drug checking technologies done was in 2018.[18] In Chapter 6 a holistic view was presented of the current capabilities of several instruments when used together: immunoassay test strips, Raman, IR, SERS, and GC–MS. Here four drug samples were studied in-depth and a walk-through-style analysis was provided for a practical comparison between data from several different instruments acquired on the same sample. This multi-technique analysis applied many of the models and methods presented in the previous chapters, which in contrast focused on capabilities of individual instruments alone.

The following chapters shift to a strong focus on infrared spectroscopy and applying machine learning and other novel data analytic to improve performance and consistency of results, reduce barriers for drug checking, and expand the reach of this work to contribute to the many services that currently implement infrared spectroscopy for drug checking. Chapter 7 implements a novel approach called two-trace two-dimensional

correlation spectroscopy in resolving pure components in complex spectral signals. In Chapter 8 thousands of spectra collected over the past several years were used to train, optimize, and validate classifiers for the detection of fluorofentanyl and MDA. Here attention is drawn to the importance of transparency and knowledge-building in drug checking services/community and in response implemented interactive examples and feature highlighting to "explain" the model predictions. In Chapter 9 current work in building a comprehensive prediction pipeline is presented. This combines unsupervised clustering, outlier detection, supervised classification modeling, and library searching. Finally, Chapter 10 describes the evolution of a tailored analysis platform that has been a part of every step of this work. This was built to allow for the translation of research and development into direct service and change dynamically with the needs of drug checkers and services.

11.2 Recommendations for Future Work

While this thesis uses many analytical instruments, most drug checking services currently only use infrared spectroscopy. Despite the widespread use of IR, there is still much room for improvement in the analysis. Future work primarily focuses on improving the current framework for automated IR analysis initiated in this dissertation. Some practical steps can be taken to enhance the models that have been established. This includes, for example, addressing label errors within the dataset and working around small data. Some methods exist to mathematically reveal samples that are likely to be labelled incorrectly, given a large enough dataset.[263] On one hand these instances could be excluded, such as in outlier analysis, however some more sophisticated methods exist to re-assign labels.[263] The improvement in labels accuracy in the training set could ultimately result in an improvement in model performance. In addition, the infrared dataset used in this work consists of around 15000 spectra. In a machine learning context, this is still a small dataset. Previous work in this field suggests that data augmentation can be extremely

useful to improve the scope and accuracy of such machine learning models.[51,264] For instance, creating new "mixtures" from weighted sums of existing samples could allow for (1) the anticipation of mixtures never before received at the service (2) improved learning of relevant spectral features and (3) more comprehensive limit of detection studies. For example, in Chapter 9 it was mentioned that it is challenging to assess machine learning models in cases where training data reflects the target compound within a very narrow range of concentrations. An example was discussed for flualprazolam; it is unknown whether the model would be able to accurately predict cases where the target compound is present at an unusually high concentration since minimal real data of such nature exists. From a harm reduction lens, the detection of such a highly potent substance is critical. Data augmentation (e.g. creating artificial mixtures of varying concentrations of flualprazolam) may give better insight into these limit of detection considerations and should be investigated. Improvements can be made on the front end of automated IR analysis (e.g. technician platform) as well. Additional metrics, such as measurements of certainty, need to be integrated into the final report to aid in communicating the confidence in simple and complex classification tasks. Given the prevalence of IR in drug checking services and a global need for improved methods, steps are being taken to share developed software outside of this drug checking project. As this moves forward, it is crucial to ensure that the tools and models developed for IR spectral analysis remain valid given variation between different instruments and minor differences in acquisition parameters and various manufacturers.

In Chapter 9, preliminary work was presented on the combination of unsupervised curve resolution techniques and library searching methods. Additional evaluation is needed to assess whether this approach can be useful for a range of drug classes, and whether this technique can be applied in an automated way or if it is better suited for exploratory/novel compound analysis. Similarly, work on various methods for improved library searching, such as barcoding and moving-window correlation, has yet to be thoroughly evaluated or

implemented. It is further acknowledged that classification has been a major focus of this thesis as opposed to quantification efforts. Chapter 10 revealed that a linear fit of pure compound spectra (library spectra) of the detected compounds provide a rough estimate of their respective proportions. These concentration estimates need to be formally validated against mass spectroscopy data before quantitative models can be implemented into the drug checking platform.

Finally, ATR-IR is not the ideal drug checking technology for complex mixtures. If a goal of drug checking is to provide a complete characterization of a substance, including the detection and quantification of potent, low concentration drugs, development of other instrumentation is needed. One proposal to improve the limit of detection of FTIR is through the use of a conventional diffuse reflectance sampling technique. This sampling configuration (as opposed to ATR) provides an opportunity to amplify the overall signal due to a much larger sampling volume. Consider a mixture of carfentanil, present in trace amounts ($<1\%$), cut with caffeine. The anticipated outcome of a diffuse reflectance FTIR measurement is an over-saturation of the spectrum (e.g. no light reaches the detector) in spectral regions characteristic to the dominant compound in the mixture (i.e. caffeine). However, because of the amplified signal, characteristic modes of carfentanil might now be identifiable in regions where the spectra of the two components do not overlap. NIR spectroscopy is also a possible non-destructive method to explore in drug checking, as it is increasingly popular in process analysis technology (PAT) applications in pharmacy and food quality monitoring.[265] Borrowing principles from PAT, there is an opportunity to set up real-time monitoring systems to see how data changes over time to identify unusual samples and more quickly detect emerging patterns.[265] As another application of PAT, efforts could continue to engage with dealers and drug manufacturers who may want to assess the quality of compound mixing and dosing. PAT could provide real-time feedback on mixing techniques prior to pursuing quantitative characterization with mass spectrometry so that the analytical results is more representative of an entire sample or

dose.

The current model of drug checking was initially established to test a small quantity of a substance, typically at the consumer-level. This system is insufficient for reliable quality assurance and supporting the autonomy of PWUD because of the significant uncertainties in drug heterogeneity and dosing. There are limited practical interventions or alternative options available for PWUD to minimize these uncertainties. Engaging with a larger supply provides an opportunity for the reach of drug checking to be realized throughout the broader community.

References

[1] Boyd, S. C. *Busted : an illustrated history of drug prohibition in Canada;* Fernwood Publishing: Winnipeg, 2017.

[2] Malleck, D. *When good drugs go bad : opium, medicine, and the origins of Canada's drug laws;* UBC Press: Vancouver, 2015.

[3] Federal, provincial, and territorial Special Advisory Committee on the Epidemic of Opioid Overdoses, "Opioid- and Stimulant-related Harms in Canada", Technical Report, Public Health Agency of Canada, Ottawa, ON, 2023.

[4] BC Coroners Service, "Illicit Drug Toxicity Deaths in BC", Technical Report, Ministry of Public Safety & Soliciter General, Victoria, BC, 2023.

[5] Kerr, T.; Small, W.; Peeace, W.; Douglas, D.; Pierre, A.; Wood, E. *Int. J. Drug Policy* **2006,** *17*, 61–69.

[6] Klein, A. *Health Care Anal.* **2020,** *28*, 404–414.

[7] Strike, C.; Watson, T. M. *Int. J. Drug Policy* **2019,** *71*, 178–182.

[8] Fairbairn, N.; Coffin, P. O.; Walley, A. Y. *Int. J. Drug Pol.* **2017,** *46*, 172–179.

[9] Kerr, T. *J. Epidemiol. Community Health* **2019,** *73*, 377–378.

[10] Kerr, T.; Mitra, S.; Kennedy, M. C.; McNeil, R. *Harm Reduct. J* **2017,** *14*, 28.

[11] Wallace, B.; Pagan, F.; Pauly, B. *Inter. J. Drug Policy* **2019,** *66*, 64–72.

[12] Pauly, B. *Int. J. Drug Policy* **2008,** *19,* 4–10.

[13] Mars, S. G.; Ondocsin, J.; Ciccarone, D. *Harm Reduct. J.* **2018,** *15,* 26.

[14] Laing, M. K.; Tupper, K. W.; Fairbairn, N. *Int. J. Drug Policy* **2018,** *62,* 59–66.

[15] Measham, F. *Br. J. Clin. Pharmacol.* **2020,** *86,* 420–428.

[16] Wallace, B.; van Roode, T.; Pagan, F.; Hore, D.; Pauly, B. *BMC Public Health* **2021,** *21,* 1156.

[17] Barratt, M. J.; Kowalski, M.; Maier, L. J.; Ritter, A. "Global Review of Drug Checking Services Operating in 2017", Technical Report, National Drug and Alcohol Research Centre, Syndey, 2018.

[18] Harper, L.; Powell, J.; Pijl, E. M. *Harm Reduct. J* **2017,** *14,* 52.

[19] Gonzales, R.; Titier, K.; Latour, V.; Peyre, A.; Castaing, N.; Daveluy, A.; Molimard, M. *Int. J. Drug Policy* **2021,** *88,* 103037.

[20] Ramsay, M.; Gozdzialski, L.; Larnder, A.; Wallace, B.; Hore, D. K. *Vib. Spectrosc.* **2021,** *114,* 103243.

[21] Ti, L.; Tobias, S.; Lysyshyn, M.; Laing, R.; Nosova, E.; Choi, J.; Arredondo, J.; McCrae, K.; Tupper, K.; Wood, E. *Drug Alcohol Depend.* **2020,** *212,* 108006.

[22] Tupper, K. W.; McCrae, K.; Garber, I.; Lysyshyn, M.; Wood, E. *Drug Alcohol Depend.* **2018,** *190,* 242–245.

[23] Gerace, E.; Seganti, F.; Luciano, C.; Lombardo, T.; Di Corcia, D.; Teifel, H.; Vincenti, M.; Salomone, A. *Drug Alcohol Rev.* **2019,** *38,* 50–56.

[24] Gozdzialski, L.; Ramsay, M.; Larnder, A.; Wallace, B.; Hore, D. K. *J. Raman Spectrosc.* **2021,** *52,* 1308–1316.

[25] Gozdzialski, L.; Rowley, A.; Borden, S.; Saatchi, A.; Gill, C. G.; Wallace, B.; Hore, D. K. *Int. J. Drug Policy* **2022,** *102,* 103611.

[26] Green, T. C.; Park, J. N.; Gilbert, M.; McKenzie, M.; Struth, E.; Lucas, R.; Clarke, W.; Sherman, S. G. *Int. J. Drug Policy* **2020,** *77,* 102661.

[27] Dies, H.; Raveendran, J.; Escobedo, C.; Docoslis, A. *Sens. Actuators, B* **2018,** *257,* 382–388.

[28] Mehr, S. H. M.; Tang, A. W.; Laing, R. R. *Magn. Reson. Chem.* **2022,** *1,* 1–11.

[29] Magnolini, R.; Schneider, M.; Schori, D.; Trachsel, D.; Bruggmann, P. *Int. J. Drug Policy* **2023,** *114,* 103972.

[30] Gwak, S.; Almirall, J. R. *Drug Test. Anal.* **2015,** *7,* 884–893.

[31] De Rycke, E.; Stove, C.; Dubruel, P.; De Saeger, S.; Beloglazova, N. *Biosens. Bioelectron.* **2020,** *169,* 112579.

[32] Scarfone, K. M.; Maghousi, N.; McDonald, K.; Stefan, C.; Beriault, D. R.; Wong, E.; Evert, M.; Hopkins, S.; Leslie, P.; Watson, T. M.; and, D. W. *Harm Reduct. J* **2022,** *19,* 3.

[33] Valdez, C. A. *Crit. Rev. Anal. Chem.* **2021,** 1–31.

[34] Vandergrift, G. W.; Gill, C. G. *J. Mass Spectrom.* **2019,** *54,* 729–737.

[35] Kennedy, J. H.; Palaty, J.; Gill, C. G.; Wiseman, J. M. *Rapid Commun. Mass Spectrom.* **2018,** *32,* 1280–1286.

[36] Kerr, T.; Tupper, K. "Drug Checking as a Harm Reduction Intervention", Technical Report, British Columbia Centre on Substance Use, Vancouver, 2017.

[37] Wallace, B. *et al. Drug Test. Anal.* **2021,** *13,* 734–746.

[38] Palamar, J. J.; Salomone, A.; Barratt, M. J. *Curr. Opin. Psychiatry* **2020,** *33,* 301–305.

[39] Ti, L.; Tobias, S.; Maghsoudi, N.; Milloy, M.-J.; McDonald, K.; Shapiro, A.; Beriault, D.; Stefan, C.; Lysyshyn, M.; Werb, D. *Drug Alcohol Rev.* **2020,** *40,* 580–585.

[40] Tobias, S.; Shapiro, A. M.; Wu, H.; Ti, L. *Can. J. Addiction* **2020,** *11,* 28–32.

[41] Holzgrabe, U.; Deubner, R.; Schollmayer, C.; Waibel, B. *J. Pharm. Biomed. Anal.* **2005,** *38,* 806–812.

[42] Laing, M. K.; Ti, L.; Marmel, A.; Tobias, S.; Shapiro, A. M.; Laing, R.; Lysyshyn, M.; Socias, M. E. *Int. J. Drug Policy* **2021,** *1,* 103169.

[43] Tobias, S.; Shapiro, A. M.; Grant, C. J.; Patel, P.; Lysychyn, M.; Ti, L. *Drug Alcohol Depend.* **2021,** *218,* 108300.

[44] Bergh, M. S.-S.; Å. M. L. Øiestad,; Baumann, M. H.; Bogen, I. L. *Int. J. Drug Pol.* **2021,** *90,* 103065.

[45] Workman Jr., J.; Koch, M.; Lavine, B.; Chrisman, R. *Anal. Chem.* **2009,** *81,* 4623–4643.

[46] Ferreira, C.; Paulino, C.; Quintas, A. *Chem. Res. Toxicol.* **2019,** *32,* 2367–2381.

[47] McGorrin, R. J. *J. Agric. Food Chem.* **2009,** *57,* 8076–8088.

[48] Cai, W.; Huang, H.; Li, Z.; Li, X.; Fan, J.; Zhang, S.; Feng, G.; Chen, J. *Anal. Chem.* **2023,** *995,* 14228–14234.

[49] Wilson, N. G.; Raveendran, J.; Docoslis, A. *Sens. Actuators, B* **2021,** *330,* 129303.

[50] Borden, S. A.; Saatchi, A.; Krogh, E. T.; Gill, C. G. *Anal. Sci. Adv.* **2020,** *1,* 97–108.

[51] Fan, X.; Ming, W.; Zeng, H.; Zhang, Z.; Lu, H. *Analyst* **2019,** *144,* 1789–1798.

[52] Lockwood, T.-L. E.; Vervoordt, A.; Lieberman, M. *Harm Reduct. J.* **2021,** *18,* 30–30.

[53] Bueno, J.; Lednev, I. K. *Anal. Methods* **2013,** *5,* 6292–6296.

[54] Wallace, B.; Gozdzialski, L.; Qbaich, A.; Azam, S.; Burek, P.; Hutchison, A.; Teal, T.; Louw, R.; Kielty, C.; Robinson, D.; Moa, B.; Storey, M.-A.; Gill, C.; Hore, D. *Drugs Habits and Social Policy* **2022,** *23,* 220–231.

[55] Peiper, N. C.; Clarke, S. D.; Vincent, L. B.; Ciccarone, D.; Kral, A. H.; Zibbell, J. E. *Int. J. Drug Policy* **2019,** *63,* 122–128.

[56] Goldman, J. E.; Waye, K. M.; Periera, K. A.; Krieger, M. S.; Yedinak, J. L.; Marshall, B. D. L. *Harm Reduct. J.* **2019,** *16,* 3.

[57] Weicker, N. P.; Owczarzak, J.; Urquhart, G.; Park, J. N.; Rouhani, S.; Ling, R.; Morris, M.; Sherman, S. G. *Int. J. Drug Policy* **2020,** *84,* 102900.

[58] Krieger, M. S.; Goedel, W. C.; Buxton, J. A.; Lysyshyn, M.; Bernstein, E.; Sherman, S. G.; Rich, J. D.; Hadland, S. E.; Green, T. C.; Marshall, B. D. L. *Int. J. Drug Policy* **2018,** *61,* 52–58.

[59] Park, J. N.; Frankel, S.; Morris, M.; Dieni, O.; Fahey-Morrison, L.; Lutta, M.; Hunt, D.; Long, J.; Sherman, S. G. *Int. J. Drug Pol.* **2021,** *94,* 103196.

[60] McCrae, K.; Tobias, S.; Grant, C.; Lysyshyn, M.; Laing, R.; Wood, E.; Ti, L. *Drug Alcohol Rev.* **2020,** *39,* 98–102.

[61] McCrae, K.; Tobias, S.; Tupper, K.; anad B. Henry, J. A.; Mema, S.; Wood, E.; Ti, L. *Drug Alcohol Depend.* **2019,** *205,* 107589.

[62] Karamouzian, M.; Dohoo, C.; Forsting, S.; McNeil, R.; Kerr, T. *Harm Reduct. J* **2018,** *15,* 46.

[63] Koczula, K. M.; Gallotta, A. *Essays in Biochem.* **2016,** *60,* 111–120.

[64] Angelini, D. J.; Biggs, T. D.; Prugh, A. M.; Smith, J. A.; Hanburger, J. A.; Llano, B.; Avelar, R.; Ellis, A.; Lusk, B.; Naanaa, A.; Feasel, M. G.; Sekowski, J. W. *Forensic Chem.* **2021,** *23,* 100309.

[65] Angelinia, D. J.; Biggs, T. D.; Maughan, M. N.; Feasel, M. G.; Sisco, E.; Sekowski, J. W. *Forensic Sci. Int.* **2019,** *300,* 75–81.

[66] Qian, S.; Bau, H. H. *Anal. Biochem.* **2004,** *326,* 211–224.

[67] Li, F.; You, M.; Li, S.; Hu, J.; Liu, C.; Gong, Y.; gYang, H.; Xu, F. *Biotechnol. Adv.* **2020,** *39,* 107442.

[68] Xiao, X.; Hu, X.; Lai, X.; Peng, J.; Lai, W. *Trends Food Sci. Technol.* **2021,** *111,* 68–88.

[69] Krasowski, M. D.; Pizon, A. F.; Siam, M. G.; Giannoutsos, S.; Iyer, M.; Ekins, S. *BMC Emergency Med.* **2009,** *9,* 5.

[70] Shapiro, S. A.; Sim, D.; Wu, H.; Mogg, M.; Tobias, S.; Patel, P.; Ti, L. "Detection of etizolam, flualprazolam, and flubromazolam by benzodiazepine-specific lateral flow immunoassay test strips", Technical Report, BC Centre on Substance Use, 2020.

[71] Mikkelsen, S.; Ash, K. *Clin. Chem.* **1988,** *34,* 2333–2336.

[72] Bunaciu, A. A.; Aboul-Enein, H. Y.; Fleschin, S. *Appl. Spectrosc. Rev.* **2010,** *45,* 206–219.

[73] Bunaciu, A. A.; Aboul-Enein, H. Y.; Fleschin, S. *Appl. Spectrosc. Rev.* **2011,** *46,* 251–260.

[74] Blum, M.-M.; John, H. *Drug Test. Analysis* **2012,** *4,* 298–302.

[75] Hans, K. M.-C.; Müller, S.; Sigrist, M. W. *Drug Test. Analysis* **2012,** *4,* 420–429.

[76] Carroll, J. J.; Mackin, S.; Schmidt, C.; McKenzie, M.; Green, T. C. *Harm Reduct. J* **2022,** *19,* 9.

[77] Griffiths, P. R.; de Haseth, J. D. *Fourier Transform Infrared Spectroscopy;* John Wiley & Sons, Inc.: Toronto, 2007.

[78] Harrick, N. J. *Internal Reflection Spectroscopy;* Interscience Publisher, John Wiley & Sons.: New York, NY, 1967.

[79] Wilson, E. Bright, J.; Decius, J. C.; Cross, P. C. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra;* Dover Publications: New York, 1980.

[80] Karch, L.; Tobias, S.; Schmidt, C.; Doe-Simkins, M.; Carter, N.; Salisbury-Afshar, E.; Carlberg-Racich, S. *Drug Alcohol Depend.* **2021,** *228,* 108976.

[81] Bardwell, G.; Boyd, J.; Tupper, K. W.; Kerr, T. *Int. J. Drug Policy* **2019,** *71,* 125–132.

[82] Milosevic, M. *Internal Reflection and ATR Spectroscopy;* John Wiley & Sons, Inc.: Hoboken, NJ, 2012.

[83] Udvardi, B.; Kovacs, I. J.; Fancsik, T.; Konya, P.; Batori, M.; Stercel, F.; Falus, G.; Szalai, Z. *Appl. Spectrosc.* **2017,** *71,* 1157–1168.

[84] Planinsek, O.; Planinsek, D.; Zega, A.; Breznik, M.; Srcic, S. *Int. J. Pharm.* **2006,** *319,* 13–19.

[85] Long, D. A. *The Raman Effect: A Unified Treatment of The Theory of Raman Scattering by Molecules;* John Wiley & Sons: New Jersey, USA, 2002.

[86] Vankeirsbilck, T.; Vercauteren, A.; Baeyens, W.; der Weken, G. V.; Verpoort, F.; Vergote, G.; Remon, J. P. *Trends Anal. Chem.* **2002,** *21,* 869–877.

[87] Paudela, A.; Raijadab, D.; Rantanen, J. *Adv. Drug Deliv. Rev.* **2015,** *89,* 3–20.

[88] Johnson, C. S.; Stansfield, C. R.; Hassan, V. R. *Forensic Sci. Int.* **2020,** *313,* 110367.

[89] Mema, S. C.; Sage, C.; Xu, Y.; Tupper, K.; Ziemianowicz, D.; McCrae, K.; M. Leigh, M.; Munn, B.; Taylor, D.; Corneil, T. *Can. J. Public Health* **2018,** *109,* 740–744.

[90] Guirguis, A.; Gittins, R.; Schifano, F. *Behav. Sci.* **2020,** *10,* 121–141.

[91] Smith, E.; Dent, G. *Modern Raman Spectroscopy – A Practical Approach;* John Wiley & Sons, Ltd.: West Sussex, 2005.

[92] Raman, C. V. *Ind. J. Phys.* **1927,** *2,* 387–398.

[93] Harris, D. C.; Bertolucci, M. D. *Symmetry and Spectroscopy: An Introduction to Vibrational and Electronic Spectroscopy;* Dover Publications: New York, 1989.

[94] Ferraro, J. R.; Nakamoto, Z.; Brown, C. W. *Introductory Raman Spectroscopy;* Academic Press: San Diego, 2003.

[95] Matousek, P.; Clark, I. P.; Draper, E. R.; Morris, M. D.; Goodship, A. E.; Everall, N.; Towrie, M.; Finney, W. F.; Parker, A. W. *Appl. Spectrosc.* **2005,** *59,* 393–400.

[96] Eberhardt, K.; Stiebing, C.; Matthäus, C.; Schmitt, M.; Popp, J. *Expert Rev. Mol. Diagn.* **2015,** *15,* 773–787.

[97] Guirguis, A.; Girotto, S.; Berti, B.; Stair, J. L. *Forensic Sci. Int.* **2017,** *273,* 113–123.

[98] Hale, G. M.; Querry, M. R. *Appl. Opt.* **1973,** *12,* 555–563.

[99] Eliasson, C.; Matousek, P. *Anal. Chem.* **2007,** *79,* 1696–1701.

[100] Kranenburg, R. F.; Verdiun, J.; de Ridder, R.; Weesepoel, Y.; Alewijn, M.; Heerschop, M.; Keizers, P. H.; van Esch, A.; van Asten, A. C. *Drug Test. Anal.* **2021,** *13,* 1054–1067.

[101] Liu, C.-M.; He, H.-Y.; Xu, L.; Hua, Z.-D. *Drug Test. Anal.* **2020,** *1,* 1–9.

[102] McNay, G.; Eustace, D.; Smith, W. E. *Appl. Spectrosc.* **2011,** *65,* 825–837.

[103] Fan, M.; Andrade, G. F. S.; Brolo, A. G. *Anal. Chim. Acta* **2020,** *1097,* 1–29.

[104] Lim, C.; Hong, J.; Chung, B. G.; deMello, A. J.; Choo, J. *Analyst* **2010,** *135,* 837–844.

[105] Pilot, R.; Signorini, R.; Durante, C.; Orian, L.; Bhamidipati, M.; Fabris, L. *Biosensors* **2019,** *9,* 57.

[106] Cailletauda, J.; De Bleyea, C.; Dumonta, E.; Sacréa, P.-Y.; Netchacovitcha, L.; Gutb, Y.; Boiretc, M.; Ginotb, Y.-M.; Huberta, P.; Ziemonsa, E. *J. Pharm. Biomed. Anal.* **2018,** *147,* 458–472.

[107] Wang, L.; Vendrell-Dones, M. O.; Deriu, C.; Dogruuer, S.; Harrington, P. B.; McCord, B. *Appl. Spectrosc.* **2021,** *75,* 1225–1236.

[108] Smith, M.; Logan, M.; Bazley, M.; Blanchfield, J.; Stokes, R.; Blanco, A.; McGee, R. *J. Forensic Sci.* **2020,** *66,* 505–519.

[109] Pérez-Jiménez, A. I.; Lyu, D.; Lu, Z.; Liu, G.; Ren, B. *Chem. Sci.* **2020,** *11,* 4563–4577.

[110] Farquharson, S.; Brouillette, C.; Smith, W.; Shende, C. *Front. Chem.* **2019,** *7,* 706.

[111] Leonard, J.; Haddad, A.; Green, O.; Birke, R. L.; fand A. Kocka, T. K.; Lombardi, J. R. *J. Raman Spectrosc.* **2017,** *48,* 1323–1329.

[112] Xie, L.; Lu, J.; Liu, T.; Chen, G.; Liu, G.; Ren, B.; Tian, Z. *J. Phys. Chem. Lett.* **2020,** *112,* 1022–1029.

[113] Tay, L.-L.; Poirier, S.; Ghaemi, A.; Hulse, J.; Wang, S. *Frontiers Chem.* **2021,** *9,* 680556.

[114] Goodacre, R.; Graham, D.; Faulds, K. *Trends Anal. Chem.* **2018,** *102,* 359–368.

[115] Brettell, T. A.; Lum, B. J. Analysis of Drugs of Abuse by Gas Chromatography–Mass Spectrometry (GC-MS). In *Analysis of Drugs of Abuse: Methods in Molecular Biology*, Vol. 1810; Humana Press: New York, NY, 2018.

[116] D'Atri, V.; Fekete, S.; Clarke, A.; Veuthey, J.-L.; Guillarme, D. *Anal. Chem.* **2019,** *91,* 210–239.

[117] Garg, U., Ed.; *Clinical Applications of Mass Spectrometry in Drug Analysis;* Humana Press: New York, 2016.

[118] Ojanperä, I.; Kolmonen, M.; Pelander, A. *Anal. Bioanal. Chem.* **2012,** *403,* 1203–1220.

[119] Eckenrode, B. A. *J. Am. Soc. Mass Spectrom.* **2001,** *12,* 683–693.

[120] Leary, P. E.; Kammrath, B. W.; Lattman, K. J.; Beals, G. L. *Appl. Spectrosc.* **2019,** *73,* 841–858.

[121] Gozdzialski, L.; Aasen, J.; Larnder, A.; Ramsay, M.; Borden, S. A.; Saatchi, A.; Gill, C. G.; Wallace, B.; Hore, D. K. *Int. J. Drug Policy* **2021,** *97,* 103409.

[122] Hübschmann, H.-J. Quantitation. In *Handbook of GC/MS: Fundamentals and Applications*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 2008.

[123] Schomburg, G.; anadd R. Dielmann, H. B.; Weeke, F.; Husmann, H. *J. Chromatogr.* **1977,** *142,* 87–102.

[124] Regmi, B. P.; Agah, M. *Anal. Chem.* **2018,** *90,* 13133–13150.

[125] Eiceman, G. A.; Gardea-Torresdey, J.; Overton, E.; Carney, K.; Dorman, K. *Anal. Chem.* **2004,** *76,* 3387–3394.

[126] Awad, H.; Khamis, M. M.; El-Aneed, A. *Appl. Spectrosc. Rev.* **2015,** *50,* 158–175.

[127] Willard, H. H.; Merritt, Lynne L., J.; Dean, J. A.; Settle, Frank A., J. *Instrumental Methods of Analysis;* Wadsworth Publishing Company: Belmont, 7th ed.; 1989.

[128] Wei, A. A. J.; Joshi, A.; Chen, Y.; McIndoe, J. S. *Int. J. Mass Spectrom.* **2020,** *450,* 116306.

[129] lavagnini, I.; Magno, F. *Mass Spectrom. Rev.* **2007,** *26,* 1–18.

[130] HimaBindu, M. R.; Parameswari, A.; Gopinath, C. *Int. J. Pharm. Qual.* **2013,** *4,* 42–51.

[131] Whiting, T. C.; Liu, R. H.; Change, W.-T.; Bodapati, M. R. *J. Anal. Toxicol.* **2001,** *25,* 179–189.

[132] Crepeault, H.; Socias, M. E.; Tobias, S.; Lysyshyn, M.; Custance, A.; Shapiro, A.; Ti, L. *Drug Alcohol Rev.* **2022,** *0,* 13580.

[133] Borden, S. A.; Saatchi, A.; Vandergrift, G. W.; Palaty, J.; Lyshyshyn, M.; Gill, C. G. *Drug Alcohol Rev.* **2022,** *41,* 410–418.

[134] Borden, S. A.; Saatchi, A.; Gill, C. G.; Wijeratne, N. R. "Quantitation of Drugs of Abuse and Their Metabolites in Urine Using PaperSpray Tandem Mass

Spectrometry for Clinical Reserach and Forensic Toxicology. Technical Note 73467", Technical Report, ThermoFisher Scientific, 2020.

[135] Giskeodegard, G.; Bloemberg, T. G.; Postma, G.; Sitter, B.; Tessem, M.-B.; Gribbestad, I. S.; Bathen, T. F.; Buydens, L. M. C. *Anal. Chim. Acta.* **2010,** *683,* 1–11.

[136] Kumar, R.; Sharma, V. *Trends Anal. Chem.* **2018,** *105,* 191–201.

[137] He, X.; Wang, J.; You, X.; Niu, F.; Fan, L.; Lv, Y. *Spectrochim. Acta, Part A* **2020,** *241,* 118665.

[138] Dupont, M. F.; Elbourne, A.; Cozzolino, D.; Chapman, J.; Truong, V. K.; Crawford, R. J.; Latham, K. *Anal. Methods* **2020,** *12,* 4597–4620.

[139] Tortorella, S.; Cinti, S. *Anal. Chem.* **2021,** *93,* 2713–2722.

[140] Jent, Y. R. . P. C. . L. M. . C. L.-M. . A. E. . N. *J. Pharm. Biomed. Anal.* **2007,** *44,* 683–700.

[141] Zhang, S.; Tan, Z.; Liu, J.; Xu, Z.; Du, Z. *Spectrochim. Acta Part A* **2020,** *227,* 117551.

[142] Karunathilaka, S. R.; Mossoba, M. M.; Chung, J. K.; Haile, E. A.; Srigley, C. T. *J. Agric. Food Chem.* **2017,** *65,* 224–233.

[143] Granato, D.; Putnik, P.; Kovacevic, D. B.; Santos, J. S.; Calado, V.; Rocha, R. S.; Da Cruz, A. G.; Jarviz, B.; Ye Rodionova, O.; Pomerrantsev, A. *Compr. Rev. Food Sci. Food Saf.* **2018,** *17,* 663–677.

[144] Farber, C.; Kurouski, D. *Anal. Chem.* **2018,** *90,* 3009–3012.

[145] Ostovar pour, S.; Afshari, R.; Landry, J.; Pillidge, C.; Gill, H.; Blanch, E. *J. Raman Spectrosc.* **2021,** *52,* 1705–1711.

[146] Noonan, K. Y.; Tonge, L. A.; Fenton, O. S.; Damiano, D. B.; Frederick, K. A. *Appl. Spectrosc.* **2009,** *63,* 742–747.

[147] Deconinck, E.; Aït-Kaci, C.; Raes, A.; Canfyn, M.; Bothy, J.-L.; Duchateau, C.; Mees, C.; De Baekeleer, K.; Gremaux, L.; Blanckaert, P. *Drug Test. Anal.* **2021,** *13,* 679–693.

[148] Kranenburg, R. F.; Verduin, J.; Weesepoel, Y.; Alewijn, M.; Heerschop, M.; Koomen, G.; Keizers, P.; Bakker, F.; Wallace, F.; v. Esch, A.; Hulsbergen, A.; v. Asten, A. C. *Drug Test. Anal.* **2020,** *12,* 1404–1418.

[149] Metternich, S.; Fischmann, S.; Münster-Müller, S.; Pütz, M.; Westphal, F.; Schönberger, T.; Lyczkowski, M.; Zörntlein, S.; Huhn, C. *Forensic Chem.* **2020,** *19,* 100241.

[150] Mittal, M.; Sharma, K.; Rathore, A. *Spectrochim. Acta, Part A* **2021,** *255,* 119710.

[151] D'egardin, K.; Roggo, Y.; Been, F.; Margot, P. *Anal. Chim. Acta.* **2011,** *705,* 334–341.

[152] Custers, D.; Cauwenbergh, T.; Bothy, J. L.; Courselle, P.; De Beer, J.; Apers, S.; Deconinck, E. *J. Pharm. Biomed. Anal.* **2015,** *112,* 181–189.

[153] Kedzierski, M.; Falcou-Pr"efol, M.; Kerros, M. E.; Henry, M.; Pedrotti, M. L.; Bruzaud, S. *Chemosphere* **2019,** *234,* 242–251.

[154] Ryder, A. G.; O'Connor, G. M.; Glynn, T. J. *J. Raman. Spectrosc.* **2000,** *31,* 221–227.

[155] Kachrimanis, K.; Braun, D. E.; Griesser, U. J. *J. Pharm. Biomed. Anal.* **2007,** *43,* 407–412.

[156] Wang, Y.-T.; Li, B.; Xu, X.-J.; Ren, H.-B.; Yin, J.-Y.; Zhu, H.; Zhang, Y.-H. *Food Chem.* **2020,** *303,* 125404.

[157] Brereton, R. G. *Applied Chemometrics for Scientists;* John Wiley & Sons: Chichester, 2007.

[158] Salih Hasan, B. M.; Abdulazeez, A. M. *JSCDM* **2021,** *2,* 20–30.

[159] Gewers, F. L.; Ferreira, G. R.; Arruda, H. F. D.; Silva, F. N.; Comin, C. H.; Amancio, D. R.; Costa, L. D. F. *ACM Comput. Surv.* **2021,** *54,* 70.

[160] Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987,** *2,* 37–52.

[161] Beattie, J. R.; Esmonde-White, F. W. L. *Appl. Spectroscp.* **2021,** *75,* 361–375.

[162] Ruiz-Perez, D.; Guan, H.; Madhivanan, P.; Mathee, K.; Narasimhan, G. *BMC Bioinf.* **2020,** *21,* 2.

[163] Lee, L. C.; Liong, C.-Y.; Jemain, A. A. *Analyst* **2018,** *143,* 3526–3539.

[164] Ståhle, L.; Wold, S. *J. Chemom.* **1987,** *1,* 185–196.

[165] Breiman, L. *Mach. Learn.* **2001,** *45,* 5–32.

[166] Bocklitz, T.; Walter, A.; Hartmann, K.; Rösch, P.; Popp, J. *Anal. Chim. Acta* **2011,** *704,* 47–56.

[167] Engel, J.; Gerretzen, J.; Szymanska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. C. *Trends Anal. Chem.* **2013,** *50,* 96–106.

[168] Mirsafavi, R.; Moskovits, M.; Meinhart, C. *Analyst* **2020,** *145,* 3440–3446.

[169] Kimani, M. M.; Lanzarotta, A.; Batson, J. S. *J. Forensic Sci.* **2020,** *66,* 491–504.

[170] Haddad, A.; Comanescu, M. A.; Green, O.; Kubic, T. A.; Lombardi, J. R. *Anal. Chem.* **2018,** *90,* 12678–12685.

[171] Stone, N.; Kendall, C.; Smith, J.; Crow, P.; Barr, H. *Faraday Discuss.* **2004,** *126,* 141–157.

[172] Virkler, K.; Lednev, I. K. *Forensic Sci. Int.* **2008,** *181,* e1–e5.

[173] Virkler, K.; Lednev, I. K. *Anal. Bioanal. Chem.* **2010,** *396,* 525–534.

[174] Penido, C.; Pacheco, M.; Lednev, I.; Silveira, L. *J. Raman Spectrosc.* **2016,** *47,* 28–38.

[175] Mansouri, M. A.; Sacré, P.-Y.; Coïc, L.; De Bleye, C.; Dumont, E.; Bouklouze, A.; Hubert, P.; Marini, R. D.; Ziemons, E. *Talanta* **2020,** *207,* 120306.

[176] Pedregosa, F. *et al. J. Mach. Learn. Res.* **2011,** *12,* 2825–2830.

[177] Gerretzen, J.; Szymanska, E.; Jansen, J. J.; Bart, J.; van Manen, H.-J.; van den Heuvel, E. R.; Buydens, L. M. C. *Anal. Chem.* **2015,** *87,* 12096–12103.

[178] Olds, W. J.; Sundarajoo, S.; Selby, M.; Cletus, B.; Fredericks, P. M.; Izake, E. L. *Appl. Spectrosc.* **2012,** *66,* 530–537.

[179] Jesus, J. I. S. d. S. d.; Löbenberg, R.; Bou-Chacra, N. A. *J. Pharm. Pharm. Sci.* **2020,** *23,* 24–46.

[180] Andrade, J. M.; Carballo-Paradelo, S.; Teran-Baamonde, J.; Carlosena, A.; Soto-Ferreiro, R. M.; Prada-Rodriguez, D. *J. Anal. At. Spectrom.* **2013,** *28,* 1911–1918.

[181] Liu, F. T.; Ting, K. M.; Zhou, Z.-H. Isolation Forest. In *Eight IEEE International Conference on Data Mining*; IEEE: Pisa, Italy, 2008.

[182] Hubert, M.; Debruyne, M.; Rousseeuw, P. J. *WIREs Comput. Stat.* **2018,** *10,* e1421.

[183] Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. *SIGMOD Rec.* **2000,** *29,* 93–104.

[184] Zhao, N.; Wu, X.-s.; Zhang, Q.; Shi, X.-y.; Ma, Q.; Yan-jiang, Q. *Sci. Reports* **2015,** *5,* 11647.

[185] Vardanyan, R. S.; Hruby, V. J. *Future Med. Chem.* **2014,** *6,* 385–412.

[186] Borden, S. A.; Palaty, J.; Termopoli, V.; Famiglini, G.; Cappiello, A.; Gill, C. G.; Palma, P. *Mass Spectrom. Rev.* **2020,** *39,* 703–744.

[187] Leary, P. E.; Dobson, G. S.; Reffner, J. A. *Appl. Spectrosc.* **2016,** *70,* 888–896.

[188] https://www.perkinelmer.com/product/calion-pv-mix-std-mininert-19ga-pkg-3-ntssmix031019.

[189] Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. *Ann. Stat.* **2004,** *32,* 407–499.

[190] British Columbia Centre on Substance Use,; B.C. Ministry of Health, "A Guideline for the Clinical Management of Opioid Use Disorder", Technical Report, British Columbia Centre on Substance Use, Vancouver, BC, 2017.

[191] Misailidi, N.; Papoutsis, I.; Nikalaou, P.; Dona, A.; Spiliopoulou, C.; Athanaselis, S. *Forensic Toxicol.* **2018,** *36,* 12–32.

[192] Segawa, H.; Fukuoka, T.; Itoh, T.; Imai, Y.; Iwata, Y. T.; Yamamuro, T.; Kuwayama, K.; Tsijikawa, K.; Kanamori, T.; Inoue, H. *Analyst* **2019,** *144,* 2158–2165.

[193] Bell, S. E. J.; Fido, L. A.; Sirimuthu, N. M. S.; Speers, S. J.; Peters, K. L.; Cosbey, S. H. *J. Forensic Sci* **2007,** *52,* 1063–1067.

[194] Aoki, P. H. B.; Furini, L. N.; Alessio, P.; Aliaga, A. E.; Constantino, C. J. L. *Rev. Anal. Chem* **2013,** *32,* 55–76.

[195] Jones, J. D.; Mogali, S.; Comer, S. D. *Drug Alcohol Depend.* **2012,** *125,* 8–18.

[196] Alonzo, M.; Alder, R.; Clancy, L.; Fu, S. *WIREs Forensic Sci.* **2022,** *4,* e1461.

[197] Giulini, F.; Keenan, E.; Killeen, N.; Ivers, J.-H. *J. Psychoact. Drugs* **2022,** *55,* 85–93.

[198] Trans European Drug Information TEDI, "TEDI Guidelines: Drug Checking Methodology", https://www.tedinetwork.org/wp-content/uploads/2022/03/TEDI_Guidelines_final.pdf, 2022.

[199] Betsos, A.; Valleriani, J.; Boyd, J.; McNeil, R. *Social Science & Medicine* **2022,** *314,* 115229.

[200] Rhodes, T. *Int. J. Drug Pol.* **2002,** *13,* 85–94.

[201] Rhodes, T. *Int. J. Drug Policy* **2009,** *20,* 193–201.

[202] Dana, K.; Shende, C.; Huang, H.; Farquharson, S. *J. Anal. Bioanal. Tech.* **2015,** *6,* 1000289.

[203] Brunt, T.; van den Berg, J.; Pennings, E.; Venhuis, B. *Arch. Toxicol.* **2017,** *91,* 2303–2313.

[204] Suzuki, E.; Shirotani, K.; Tsuda, Y.; Sekiguchi, K. *Chem. Pharm. Bull.* **1985,** *33,* 5028–5035.

[205] Wallace, B.; van Roode, T.; Pagan, F.; Phillips, P.; Wagner, H.; Calder, S.; Aasen, J.; Pauly, B.; Hore, D. *Harm. Reduct. J* **2020,** *17,* 29.

[206] Wallace, B.; van Roode, T.; Burek, P.; Pauly, B.; Hore, D. *Drugs Education Prevention Policy* **2023,** *30,* 443–452.

[207] Masterton, W.; Falzon, D.; Burton, G.; Carver, H.; Wallace, B.; Aston, E. V.; Sumnall, H.; Measham, F.; Gittins, R.; Craik, V.; Schofield, J.; Little, S.; Parkes, T. *Int. J. Environ. Res. Public Health* **2022,** *19,* 11960.

[208] Whitehead, H. D.; Hayes, K. L.; Swartz, J. A.; Prete, E.; Robison-Taylor, L.; Mackesy-Amiti, M. E.; Jimenez, A. D.; Lieberman, M. *Forensic Chem.* **2023,** *33,* 100475.

[209] Maghsoudi, N.; Tanguay, J.; Scarfone, K.; Rammohan, I.; Ziegler, C.; Werb, D.; Scheim, A. I. *Addiction* **2022,** *117,* 532–544.

[210] Clarke, S. E. D.; Kral, A. H.; Zibbell, J. E. *Int. J. Drug Policy* **2022,** *99,* 103467.

[211] Rodriguez, J. D.; Westenberger, B. J.; Buhse, L. F.; Kauffman, J. F. *Anal. Chem.* **2011,** *83,* 4061–4067.

[212] Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986,** *185,* 1–17.

[213] Abdi, H.; Williams, L. J. *WIREs Comput. Stat.* **2010,** *2,* 433–459.

[214] Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 1947–1958.

[215] Lawson, L. S.; Rodriguez, J. D. *Anal. Chem.* **2016,** *88,* 4706–4713.

[216] Noda, I.; Ozaki, Y. *Two-Dimensional Correlation Spectroscopy: Applications in Vibrational and Optical Spectroscopy;* John Wiley & Sons, Ltd: San Francisco, 2004.

[217] Noda, I. *J. Mol. Struct.* **2018,** *1160,* 471–478.

[218] Noda, I. *J. Mol. Struct.* **2020,** *1213,* 128194.

[219] Yang, R.-J.; Liu, C.-Y.; Yang, Y.-R.; Wu, H.-Y.; Jin, H.; Shan, H.-Y. *J. Mol. Struct.* **2020,** *1214,* 128219.

[220] Sohng, W.; Eum, C.; Chung, H. *Anal. Chim. Acta* **2021,** *1152,* 338255.

[221] Kavitha, E.; Stephen, L. D.; Brishti, F. H.; Karthikeyan, S. *J. Mol. Struct.* **2021,** *1244,* 130964.

[222] Walkowiak, A.; Wnuk, K.; Cyrankiewicz, M.; Kupcewicz, B. *Molecules* **2022,** *27,* 433.

[223] Xu, D.; Liu, S.; Cai, Y.; Yang, C. *Appl. Opt.* **2019,** *58,* 3913–3920.

[224] Noda, I. *Spectrochim. Acta A* **2022,** *276,* 121221.

[225] Lanzarotta, A.; Lakes, K.; Marcott, C. A.; Witkowski, M. R.; Sommer, A. J. *Anal. Chem.* **2011,** *83,* 5972–5978.

[226] Zhang, X.; He, A.; Guo, R.; Zhao, Y.; Yang, L.; Morita, S.; Xu, Y.; Noda, I.; Ozaki, Y. *Spectrochim. Acta, Part A* **2022,** *265,* 120373.

[227] Dasgupta, N.; Figgatt, M. C. *Am. J. Epidemiol.* **2022,** *191,* 248–252.

[228] Meza Ramirez, C. A.; Greenop, M.; Ashton, L.; u. Rehman, I. *Appl. Spectrosc. Rev.* **2021,** *56,* 733–763.

[229] Angulo, A.; Yang, L.; Aydil, E. S.; Modestino, M. A. *Digital Discovery* **2022,** *1,* 35–44.

[230] Ren, H.; Li, H.; Zhang, Q.; Liang, L.; Guo, W.; Huang, F.; Luo, Y.; Jiang, J. *Fundam. Res.* **2021,** *1,* 488–494.

[231] Guo, S.; Popp, J.; Bocklitz, T. *Nat. Protoc.* **2021,** *16,* 5426–5459.

[232] Kaluarachchi, T.; annd S. Nanayakkara, A. R. *Sensors* **2021,** *21,* 2514.

[233] Miller, T. *Artif. Intell.* **2019,** *267,* 1–38.

[234] Ribeiro, M. T.; Singh, S.; Guestrin, C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016.

[235] Yang, F.; Huang, Z.; Scholtz, J.; Arendt, D. L. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In *Proceedings of the 25th*

International Conference on Intelligent User Interfaces; IUI '20 Association for Computing Machinery: New York, NY, USA, 2020.

[236] Cai, C. J.; Reif, E.; Hegde, N.; Hipp, J.; Kim, B.; Smilkov, D.; Wattenberg, M.; Viegas, F.; Corrado, G. S.; Stumpe, M. C.; Terry, M. "Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making", 2019.

[237] Gianfagna, L. *Explainable AI with Python;* Springer: Cham, Switzerland, 2021.

[238] Du, M.; Liu, N.; Hu, X. *Commun. ACM* **2019,** *63*, 68–77.

[239] "Substance: Vancouver Island Drug Checking Project", `https://substance.uvic.ca`,.

[240] Lundberg, S. M.; Lee, S. *CoRR* **2017,** *abs/1705.07874*,.

[241] Butler, H. J.; Smith, B. R.; Fritzsch, R.; Radhakrishnan, P.; Palmer, D. S.; Baker, M. J. *Analyst* **2018,** *143*, 6121–6134.

[242] on Drugs, U. O.; Crime, "Guidance for the validation of analytical methodology and calibration of equipment used for testing of illicit drugs in seized materials and biological specimens", 2009.

[243] Ayres, L. B.; Gomez, F. J. V.; Linton, J. R.; Silva, M. F.; Garciia, C. D. *Anal. Chim. Acta.* **2021,** *1161*, 338403.

[244] Menze, B.; Kelm, B.; Masuch, R.; Himmelreich, U.; Bachert, P.; Wolfgang, P.; Hamprecht, F. *BMC Bioinf.* **2009,** *10*, 213.

[245] Yang, G.; Ye, Q.; Xia, J. *Information Fusion* **2022,** *77*, 29–52.

[246] Barratt, M. J.; Measham, F. *Drugs, Habits and Social Policy* **2022,** *23*, 176–187.

[247] Kenny, E. M.; Ford, C.; Quinn, M.; Keane, M. T. *Artif. Intell.* **2021,** *294*, 103459.

[248] Bardwell, G.; Boyd, J.; Arredondo, J.; McNeil, R.; Kerr, T. *Drug Alcohol Depend.* **2019,** *198,* 1–6.

[249] Carroll, J. J. *Contemp. Drug Probl.* **2021,** *48,* 327–345.

[250] Betzler, F.; Helbig, J.; Viohl, L.; Ernst, F.; Roediger, L.; Gutwinski, S.; Ströhle, A.; Köhler, S. *Eur. Addict. Res.* **2021,** *27,* 25–32.

[251] Murney, M. A.; Sapag, J. C.; Bobbili, S. J.; Khenti, A. *Int. J. Qual. Stud. Health Well-being* **2020,** *15,* 1744926.

[252] Neufeld, S. D.; Chapman, J.; Crier, N.; Marsh, S.; McLeod, J.; Deane, L. A. *Harm Reduct. J* **2019,** *16,* 41.

[253] Gozdzialski, L.; Hutchison, A.; Wallace, B.; Gill, C.; Hore, D. *Drug Test. Anal.* **2023,** *online ahead of print,* DOI 10.1002/dta.3520.

[254] Ti, L.; Grant, C. J.; Tobias, S.; Hore, D. K.; Laing, R.; Marshall, B. D. L. *PLoS ONE* **2023,** *18,* 0288656.

[255] Anderson, R. B.; III, J. F. B.; Wiens, R. C.; v. Morris, R.; Clegg, S. M. *Spectrochim. Acta, Part B* **2012,** *70,* 24–32.

[256] Backhaus, K.; Erichson, B.; Gensler, S.; Weiber, R.; Weiber, T. Cluster Analysis. In *Multivariate Analysis: An Application-Oriented Introduction*; Springer Fachmedien Wiesbaden: Wiesbaden, 2023.

[257] McInnes, L.; Healy, J.; Astels, S. *J. Open Source Software* **2017,** *2,* 205.

[258] McInnes, L.; Healy, J. Accelerated Hierarchical Density Based Clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*; IEEE: New Orleans, Lousiana, 2017.

[259] Rousseeuw, P. J. *J. Comput. Appl. Math* **19987,** *20,* 53–65.

[260] de Juan, A.; Tauler, R. *Anal. Chim. Acta.* **2021,** *1145,* 59–78.

[261] Rebiere, H.; Martin, M.; Ghyselinck, C.; Bonnet, P.-A. *J. Pharm. Biomed. Anal.* **2018,** *148,* 316–323.

[262] Ng, P. H. R.; Walker, S.; Tahtouh, M.; Reedy, B. *Anal. Bioanal. Chem.* **2009,** *394,* 2039–2048.

[263] Frénay, B.; Kabán, A. A Comprehensive Introduction to Label Noise. In *Twenty-Second European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*; ESANN: Bruges, Belgium, 2014.

[264] Blazhko, U.; Shapaval, V.; Kovalev, V.; Kohler, A. *Chemom. Intell. Lab. Syst.* **2021,** *215,* 104367.

[265] Mazivila, S. J.; Olivieri, A. C. *Trends Anal. Chem.* **2018,** *108,* 74–87.