USING MID-INFRARED SPECTROSCOPIC FINGERPRINTING, MULTIVARIATE ANALYSIS AND MACHINE LEARNING TO DIFFERENTIATE TRADITIONAL HERBAL MEDICINE

YEAP ZHAO QIN

UNIVERSITI SAINS MALAYSIA

2022

USING MID-INFRARED SPECTROSCOPIC FINGERPRINTING, MULTIVARIATE ANALYSIS AND MACHINE LEARNING TO DIFFERENTIATE TRADITIONAL HERBAL MEDICINE

by

YEAP ZHAO QIN

Thesis submitted in fulfilment of the requirement
for the degree of
Master of Science

November 2022

ACKNOWLEDGEMENT

Thank you to my supervisor, Associate Professor Dr. Yam Mun Fei, my lab associates, Dr. Tan Chu Shan, Dr. Loh Yean Chun, Ch'ng Yung Sing, Dr. Ng Chiew Hoong and Chen Ying. Without their support, this study would not have been possible.

Heartfelt gratitude to co-supervisor Associate Professor Dr Yoon Tiem Leong for his immense assistance in the completion of this study.

Secondly, thanks to School of Pharmaceutical Sciences, USM for this once in a lifetime opportunity and also to the USM Graduate Assistance Scheme for the tremendous financial assistance provided.

I am also grateful to family and friends that have been essential through the course of this run.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

%T	percentage of transmittance
°C	degree Celsius
^{1}H-	proton
2D	two dimensional
A	absorbance unit
ATR-	attenuated total reflection
BernoulliNB	Bernoulli-Naïve Bayes
BP	British Pharmacopoeia
CA/MCA	correspondence analysis/multiple correspondence analysis
CAGR	compound annual growth rate
CCA	canonical correlation analysis
ChP	Chinese Pharmacopoeia
cm^{-1}	reciprocal centimeter
DTSG	dielectric thermal smart glass
EU	European Union
EUP	European Pharmacopoeia
FTIR	Fourier transform infrared radiation
GaussianNB	Gaussian-Naïve Bayes
GC x GC - TOFMS	comprehensive two-dimensional gas chromatography - time-of-flight mass spectrometry
GC-MS	gas chromatography - mass spectrometry
GPA	generalized procrustes analysis
h	hour
HPLC	high performance liquid chromatography
HPTLC	high performance thin layer chromatography
ICA	independent component analysis
IR	infrared radiation
KBr	potassium bromide
kPa	kilopascal
LC	liquid chromatography

MANOVA	multivariate analysis of variance
MDS	multidimensional scaling
mg	milligram
min	minute
mm	millimeter
MRA	multiple regression analysis
MVA	multivariate analysis
NCD	non-communicable disease
NMR	nuclear magnetic resonance
PC	principal component
PCA	principal component analysis
PLS	partial least square regression
RDA	redundancy analysis
S/N	signal-to-noise
SARS	severe acute respiratory syndrome
SIMCA	soft independent modeling of class analogy
SMOTE	synthetic minority oversampling technique
SVC	support vector classification
T%	total transmittance percentage
THM	traditional herbal medicine
USP	United States Pharmacopoeia
UV-vis	ultraviolet-visible
VOC	volatile organic compound
WHO	World Health Organization

PENGGUNAAN CAP JARI SPEKTROSKOPI INFRAMERAH ANALISA MULTIVARIAT DAN PEMBELAJARAN MESIN UNTUK MEMBEZAKAN UBATAN HERBA TRADITIONAL

ABSTRAK

Perubatan herba tradisional adalah satu bahagian penting dalam sistem kesihatan global dengan isu berterusan seperti pengodiaan dan pengenalan salah. Walaupun piawaian global semasa menggunakan pengenalan kromatografi untuk memerangi isu ini, terdapat beberapa kelemahan dengan kaedah sedemikian. Kajian ini melihat lima ubat herba tradisional, iaitu *Anoectochilus roxburghii*, *Aristolochia manshuriensis*, *Dioscorea hamiltonii*, *Gelsemium elegans* dan *Alisma orientalis*, serta cara mengklasifikasikannya. Spektrum inframerah herba dikumpulkan daripada sejumlah 200 sampel dari 20 °C hingga 120 °C dengan selang 10 °C. Isyarat inframerah kumpulan berfungsi juzuk kimia utama setiap herba hadir dalam spektrum inframerah yang dikumpul. Analisis komponen utama spektrum ini mendapati batasan di mana kejayaan analisis mungkin memerlukan lebih kurang jenis herba disertakan. Gabungan pengiraan spektrum gangguan haba telah dilakukan untuk mendapatkan cap jari kimia dua-dimensi bagi setiap sampel. Pengelas pembelajaran mesin telah dilatih untuk menjana model. Model ini mengklasifikasikan herba dengan ketepatan 87.9% apabila semua herba disertakan dalam latihan. Apabila *Alisma orientalis* dikecualikan daripada latihan untuk mengukur kekukuhan model, 91.3 % daripada sampel telah dikelaskan dengan betul.

USING MID-INFRARED SPECTROSCOPIC FINGERPRINTING, MULTIVARIATE ANALYSIS AND MACHINE LEARNING TO DIFFERENTIATE TRADITIONAL HERBAL MEDICINE

ABSTRACT

Traditional herbal medicine is an important part of the global health system with persistent issues like adulteration and misidentification. While current global standards employ chromatographic identification to combat this issue, there are some disadvantages with such methods. This study looked at five traditional herbal medicine, namely *Anoectochilus roxburghii, Aristolochia manshuriensis, Dioscorea hamiltonii, Gelsemium elegans* and *Alisma orientalis,* and ways to classify them. Infrared spectra of the herbs were collected from a total of 200 samples from 20 °C to 120 °C at 10 °C intervals. Infrared signals of functional groups on the main chemical constituents for every herb were present in the infrared spectra collected. Principal component analysis of these spectra found a limitation where the success of the analysis might require less types of herbs included. Computational combination of thermally perturbed spectra was performed to obtain two-dimensional chemical fingerprints for every sample. Machine Learning Classifier were trained to generate models. The model classified the herbs with an accuracy of 87.9 % when all the herbs were included in training. When *Alisma orientalis* was excluded from training to measure the robustness of the model, 91.3 % of the samples were classified correctly.

CHAPTER 1

INTRODUCTION

1.1 General Introduction

With the affordability of healthcare becoming more and more challenging, traditional herbal medicine (THM) has carried the role of supplementing and in rare cases, replacing modern medicine for a growing proportion of the global population. (Hartman et al., 2018; World Health Organization, 2013) At the same time, monotherapy is observing a decline in their effectiveness to control non-communicable diseases such as hypertension and diabetes while THM and their extracts are reported to be multi-pathway elicitors while generally being perceived as having less harmful side effects. (Guerrero-Garcia & Rubio-Guerra, 2018; He et al, 2011; Intergovernmental Bioethics Committee, 2015; Jeon et al., 2018; Khan et al., 2011; Materson, 1995; Shah & Gilani, 2010; Tan et al, 2018) Not to mention the growing distrusts in modern medicine due to misdiagnosis, over-treatment and over-prescription that has led to the opioid crisis in the United States had caused more patients to look into THM as an alternative or complementary solution to their medical problems. (Armstrong, 2006) In short, THM industry is expected to continue growing due to its affordability, the weaknesses in monotherapy, modern medicine and its practices.

The THM industry is plagued with the problem of adulteration as reported by independent research across the world. In Iran, a taxonomical evaluation found adulterants in majority of the THM products that were purchased from local markets and the author urged governing bodies to take action on the matter. (Joharchi & Amiri,

2012) A French university in 2010 detected sibutramine, sometimes together with phenolphthalein, in more than half of the body-slimming supplements tested there. (Vaysse, 2010) Sibutramine was a widely distributed appetite suppressant that has been discontinued in various countries since 2010 due to the causal link found between its consumption with increased cardiovascular events and strokes, while phenolphthalein is a laxative that has been removed from over-the-counter purchase due to concerns for its carcinogenicity. (Citronberg et al., 2018; Von Haehling et al, 2007) The presence of adulterant in THM is a public health concern as demonstrated by Vanherweghem in 1992 where 9 female patients under the age of 50 with rapidly progressive interstitial renal fibrosis reportedly followed a traditional slimming regiment where *Stephania tetrandra*, has been adulterated with *Aristolochia fangji*. (Cosyns, 2013: Vanhaelen et al, 1994; Vanherweghem et al., 1993) The aristolochic acid from *Aristolochia fangji* is a known nephrotoxic and a listed suspected carcinogen by the World Health Organization International Agency for Research on Cancer. (International Agency for Research on Cancer, 2021)

European Pharmacopoeia has since published guideline to conduct adulterant detection for THM as a quality control using techniques such as high performance liquid chromatography (HPLC) and high performance thin layer chromatography (HPTLC) which have been adopted worldwide. However, the Committee on Herbal Medicinal Products of the European Medicines Agency states that marker compounds selected for any product are preferably but not necessarily the one that elicits the claimed therapeutic effects due to the chemical complexity of extracts. In other words, chromatography techniques do not give information on the efficacy of the herbal product. Besides that, chromatography techniques require precise control of

experimental conditions, intermediate level technicians to operate and the use of mobile phase that could potentially harm the environmental.

Spectroscopy techniques capture the electromagnetic interactions between molecules and light, and therefore can remain consistent at higher variation in experimental conditions. The consistency of the spectra is a key factor in the process of chemical fingerprinting where the general pattern of the spectra is considered as a unique identifier to a specific THM. Fourier transform infrared radiation (FTIR) spectroscopy in particular also does not need as much sample preparation for analysis and require relatively minimal training to execute. FTIR also serves as a holistic method that can potentially give efficacy information on a THM since the presence or absence of a functional group can be observed directly from the spectra. In addition, device and equipment for FTIR are much more economical than chromatography methods and nuclear magnetic resonance (NMR) spectroscopy. In summary, FTIR is a chemical fingerprinting method that require little sample preparation and training while producing consistent results that can be correlated with efficacy of THM while being more economical.

Multivariate analysis is a method that is suitable for post-analysis due to the complexity of signals that can be obtained. Principal component analysis (PCA) is a type of unsupervised multivariate analysis where the dimensionality of the spectra could be reduced. The projection highlights the variations among the data to show grouping trends where spectra that are similar cluster together while spectra of different THM or adulterant tend to scatter.

Due to the presence of hundreds or thousands of molecules present in THM, signals can easily overlap in an FTIR spectrum. This can cause larger signals to overlap with crucial smaller signals, causing information lost. However, taking FTIR spectra

of the same sample at a consistent temperature interval and then combining the spectra could probe into the molecular response to heat, creating a two dimentional (2D) FTIR chemical fingerprint. This chemical fingerprint lets us extract another dimension of information from the same sample which can further distinguish one sample from another due to the unique response of different compounds that may or may not present in the sample.

While statistical models such as PCA has conventionally been sufficient for smaller studies, having a more robust classifying algorithm for datasets that are ever growingly complex such as a 2D-FTIR spectra is inevitable. The application of Machine Learning Classifier is the ideal choice here because recognizing patterns and differentiating feature is one of the tasks that machine learning excels at. Previous studies have used machine learning to recognize THM through taxonomical data or in conjunction with some form of chemical data such as visible/near-infrared. (Begue et al., 2017; Xue et al., 2019) Both prior studies have the same goal with this study in that machine learning can be employed to differentiate, discriminate, identify or classify THM. However, the information that are being fed into their respective algorithms are vastly different than this study, not to mention the machine learning algorithm employed are slightly dissimilar in their own rights. Therefore, there is novelty in the approach of this study to combine 2D-FTIR chemical fingerprint with Machine Learning Classifiers.

The application of Machine Learning Classifier here is slightly modified, instead of using just one classifier, multiple classifiers are to first “learn” from randomly selected 2D-FTIR chemical fingerprint. The classifiers will then form independent models that can be used to decide if a chemical fingerprint that the classifier has never seen before belongs to one or none of the classes of THM. These

decisions are then put through a collective voting scheme that helps us minimize misidentification that might be present in any one of the classifiers.

1.2 Problem Statement

In brief, THM industry will continually grow as issues emerge in modern medicine while its affordability keeps decreasing. Research has shown adulteration is a major problem in THM that needs to be tackled with global effort and consensus. Although major pharmacopoeias have adopted chromatography techniques for adulterant detection, there are innate problems with the techniques particularly with the dependency on marker compounds. FTIR offers a holistic overview of the chemical properties for fingerprinting, as well as being consistent, require little sample preparation, and relatively little training while being economical. The combination of FTIR fingerprint with PCA and Machine Learning Classifiers could potentially be the solution to these issues.

1.3 Objectives

Therefore, the aims of this study are:

a) To collect Conventional FTIR spectra of five THMs, namely *Anoectochilus roxburghii* (38 samples), *Aristolochia manshuriensis* (54 samples), *Dioscorea hamiltonii* (25 samples), *Gelsemium elegans* (50 samples), and *Alisma orientalis* (33 samples).

b) To differentiate *Anoectochilus roxburghii*, *Aristolochia manshuriensis*, *Dioscorea hamiltonii*, *Gelsemium elegans*, and *Alisma orientalis* through PCA analysis on their Conventional FTIR spectra.

c) To generate 2D-FTIR spectra from the Conventional FTIR spectra collected for *Anoectochilus roxburghii*, *Aristolochia manshuriensis*, *Dioscorea hamiltonii*, *Gelsemium elegans*, and *Alisma orientalis*.

d) To classify 2D-FTIR spectra of *Anoectochilus roxburghii*, *Aristolochia manshuriensis*, *Dioscorea hamiltonii*, *Gelsemium elegans*, and *Alisma orientalis* by utilizing Machine Learning Classifiers and collective voting scheme.

CHAPTER 2

LITERATURE REVIEW

2.1 Traditional Herbal Medicine

2.1.1 Non-communicable Diseases

Diseases that cannot be transmitted directly from one person to another are called non-communicable diseases (NCD) and they account for over 71 % of global deaths or 41 million people each year. The most common NCD deaths are from cardiovascular diseases, cancers, respiratory diseases and diabetes, making up for over 80 % of all NCD deaths. (World Health Organization, 2008) Prevalence of NCDs are commonly linked to poor lifestyle choices such as a lack of exercise, unhealthy diet and harmful use of alcohol or tobacco. (Ministry of Health, 2016) The treatments for NCDs vary due to the broad type of diseases but the saying "prevention is better than cure" holds true here. In 2018, World Health Organization (WHO) estimates that every one dollar invested in implementing recommended interventions will yield seven dollars return by 2030 or equivalent of 350 billion dollars in economic growth. (World Health Organization, 2018) Therefore, it is apparent that the strategy to combat NCDs lies more in the prevention and intervention rather than curing them. Although most intervention focuses on lifestyle and habitual intervention, pharmaceutical intervention is essential in cases where other forms of interventions might not be ideal due to reasons such as pre-existing health conditions and the socio-economic status and determination of a patient. It is not surprising that pharmaceutical intervention is especially popular in demand these days and the most common form of pharmaceutical intervention comes from monotherapy.

2.1.2 Modern Medicine and Traditional Herbal Medicine

Currently, monotherapies for NCDs are seeing a steady decrease in their effectiveness especially for hypertension and diabetes. (Guerrero-Garcia & Rubio-Guerra, 2018; Jeon et al., 2018; Materson et al., 1995) This combined with rising cases of over-prescription, or over-treatment and misdiagnosis has contributed to the growing distrusts for modern medicine. (Armstrong et al., 2006) With the skyrocketing costs of healthcare in virtually every country, there are substantial demand and potential for THM to step up and fill in the gaps of modern medicine. In 2011, the United Nations General Assembly recognizes the potential of THM in the prevention of NCDs. (United Nations, 2011) According to some estimates, THM industry is expecting a steady 5.88 % compound annual growth rate (CAGR) between 2018 and 2023, reaching more than USD 129 billion. (Market Research Future, 2019) In Malaysia alone, the Ministry of Entrepreneur Development said that they are looking at RM 32 billion in 2020 for the market value of THM industry. (Naharul, 2019) A survey study documented THM use in Mauritius for the treatment and prevention of NCDs and concluded that the dependency of the local population on THM is significant, citing weakened confidence in modern medicine. (Chintamunnee & Mahomoodally, 2012) THM are not just frequently used for NCDs but other ailments and even during epidemics as well. A clinical trial by the WHO demonstrated the ability of THM in tandem with modern medicine preventing infection during the severe acute respiratory syndrome (SARS) outbreak. (World Health Organization, 2004) The continued success of THM industry come from the decline in monotherapy efficacy, general public's distrust in modern medicine and the financial burden associated with modern medicine.

2.1.3 *Anoectochilus roxburghii, Aristolochia manshuriensis, Dioscorea hamiltonii, Gelsemium elegans* and *Alisma orientalis*

2.1.3(a) *Anoectochilus roxburghii*

Anoectochilus roxburghii can be found in Eastern India, Nepal, Sri Lanka, southeastern mainland China, Taiwan and Japan. Traditionally, *Anoectochilus roxburghii* is a highly prized plant used for a variety of ailments. It was listed as a protected species in "Convention on International Trade in Endangered Species of Wild Fauna and Flora". Efforts to artificially cultivate and seedling breeding of this plant have been conducted. Figure 2.1 show the plant of *Anoectochilus roxburghii*.

Figure 2.1 Photo of *Anoectochilus roxburghii.* (Li, 2021)

The plants are usually found in warm and humid forest floors around decaying organic matter such as dead tree trunks or leaf litter. They also prefer to grow in shaded area without too much sunlight.

Various studies have been conducted to investigate its potential in treating diabetes. The plant is also studied for its role in preventing and treating hepatitis, hyperliposis and some tumors. There are various products of *Anoectochilus roxburghii* marketed towards patients with hyperuricemia, diabetes, hepatitis, *Helicobacter pylori* infection and asthma. (Ye et al., 2017)

2.1.3(b) *Aristolochia manshuriensis*

Aristolochia manshuriensis is a type of deciduous and climbing shrub that can be found natively in the northeastern region of China formerly known as Manchuria, some forests in Korea and eastern Siberia. Moist and sunny or partially shaded environment is best for its growth in the wild. ("Aristolochia manshuriensis", n.d.) Figure 2.2 shows the flower of *Aristolochia manshuriensis* hanging off its shrub branches.

Figure 2.2 Photo of *Aristolochia manshuriensis*. (Qian, 2022)

Aristolochia manshuriensis (Mandarin name: Guan Mu Tong) is often confused and wrongly used in place of *Akebia Quinata* (Mandarin name: Mu Tong) as a traditional analgesic, antiphlogistic and diuretic agent. (Wu et al., 2016)

Aristolochia manchuriensis contains aristolochic acid which is a known nephrotoxic which has caused well documented progressive interstitial renal fibrosis in young women. (Vanherweghem et al., 1993) The alcoholic extract of *Aristolochia manshuriensis* was found to be lethal to mice. (Zhu, 2002)

2.1.3(c) *Dioscorea hamiltonii*

Dioscorea hamiltonii can be found in the higher altitudes (100 m to 2,000 m) of slightly warm to hot climate with humid conditions. These regions include the Western Ghats in India, the mountains of the Himalayas bordering north-eastern India, Nepal, Bhutan and the Bengals, the mountains from southern China, Thailand, Vietnam and Laos. (Lim, 2016) Figure 2.3 below shows *Dioscorea hamiltonii* growing at a hill.

Figure 2.3 Photo of *Dioscorea hamiltonii*. (Itariajin, 2011)

The rhizomes of the plant are consumed as food and traditionally for medicinal purposes. There are studies that show *Dioscorea hamiltonii* ameliorating spleen asthenia in mice, exhibiting antimicrobial activities against bacterias and yeasts and some anabolic and gonadotropic activities. (Kaladhar et al., 2010; National Institute of Materia Medica, 1999; Qin et al., 2003) *Dioscorea hamiltonii* were also found to have antioxidant, anti-inflammatory and immune regulation effect. (Zhao et al., 2019)

2.1.3(d) *Gelsemium elegans*

Gelsemium elegans can be found bordering East India and southeastern China, including Laos, Vietnam, Myanmar, Thailand, and Malaysia. They can also be found in the island nation of Indonesia. The shrubs grow in the 200 m to 2,000 m elevated forests in mountains and can grow up to 12 m in height as a twining vine. (Zhou et al., 2017) Flower of *Gelsemium elegans* is shown in Figure 2.4.

Figure 2.4 Photo of *Gelsemium elegans.* (Qiao, 2004)

Despite known as a toxic plant, the leaves have been traditionally used in the treatment of eczema, bruises, rheumatoid arthritis and skin ulcers. (Wang et al., 2019) The main active constituents and toxic constituents are under the alkaloid family. Studies have shown that compounds extracted from the plant have some ethnopharmacological value as anti-tumor, anti-inflammatory and analgesic agent, exhibiting immunomodulatory and anti-anxiety effect, and playing some roles in cardiovascular repair and hematopoietic protection. (Lin et al., 2021)

2.1.3(e) *Alisma orientalis*

Alisma orientalis is native to China, Korea and Japan. They are commonly found around the edges or slightly submerged in water feature such as lakes and rivers under 800 m. Climate conditions that are preferable for *Alisma orientalis* includes humid, shaded and warm. Figure 2.5 show *Alisma orientalis* plant in its natural habitat.

Figure 2.5 Photo of *Alisma orientalis.* (Wu, 2018)

The roots of the plant are traditionally used for conditions such as oliguria, edema, gonorrhea, leukorrhea, diarrhea and dizziness. *Alisma orientalis* has been studied as diuretic, anti-urolithiatic, antinephritic, antiatherosclerotic, immunomodulatory and hepatoprotective agent. (Shu et al., 2016)

2.1.4 Challenges of Traditional Herbal Medicine Industry

However, it is important to recognize that the THM industry is certainly not without flaws. First, the THM industry was built on non-scientific data driven claims, anecdotal experiences and sometimes unverifiable myths. (Xu & Xia, 2019) This has left the industry with a reputation of being inconsistent in their therapeutic efficacy. However, various committees and organizations from local and international governments as well as key industry players have pushed for more scientific approach

for THM in quantifying their contents, investigating mechanism of actions, measuring effectiveness and verifying the safety of THM products for decades. The efforts have helped improve the overall outlook and perception of the general public towards THM.

2.1.4(a) Adulterants and Misidentifications

With more and more investment in legitimizing THM, various products have been flooding the market. The main problem that comes with such a vast market potential is adulterants and misidentified products for various reasons. Adulterants are defined precisely as any material introduced when making a product that are not being disclosed as part of the formulation or incidental introduction of a material into the product in the manufacturing process. (National Cancer Institute, n.d.) The institution goes on to add that adulterants potentially render a product to be hazardous, less expensive to produce or not deliver its intended purposes. It is not hard to see that misidentifying THM may also cause similar problems with consuming adulterated THM. The significance of adulterants in THM can be gauged from various publications such as in 2012, researchers in Iran performed taxonomical evaluations on the THM products acquired from their local markets and found adulterants in most of the products while the rest are misidentified. (Joharchi & Amiri, 2012) In 2010, more than half of the body-slimming supplements sampled by a French university contained traces of sibutramine or a combination of sibutramine and phenolphthalein. (Vaysse et al., 2010) Sibutramine and phenolphthalein are both drugs that has been removed due to health concern. The former was marketed as appetite suppressant and its consumption was found to cause increase cardiovascular events and stroke while the latter was marketed as a laxative and removed for its carcinogenicity. (Citronberg et al., 2018; von Haehling et al., 2007) Perhaps the most famous case of adulterants or

misidentification of THM causing a public health crisis happened almost three decades ago documented in a publication on Lancet in 1993. (Vanherweghem et al, 1993) The publication reported rapidly progressive interstitial renal fibrosis in 9 young female patients. The publication found that all 9 patients have been following a traditional slimming regiment which contains *Stephania tetrandra* which was thought to be the cause for the kidney failure. However, follow-up investigations identified that *Stephania tetrandra* was not the cause and that they were adulterated or substituted with *Aristolochia fangji* which contains aristolochic acid, a known nephrotoxic as well as a suspected carcinogen listed by the World Health Organization International Agency for Research on Cancer. (Cogliano et al., 2011; Cosyns, 2003; Vanhaelen et al., 1994) It is apparent that the issue of adulterant or misidentification in THM can easily create dire consequences for the public health and needs oversight.

2.1.5 Current Global Standard for Traditional Herbal Medicine Quality Assessment

The first step in establishing a system to oversee the THM industry is to understand that most countries and economic regions employ vastly different classifications, systems and standards. (Heinrich, 2015) The increasingly globalized market will need one or at most, just a few standards to ensure the quality and safety of THM product as well as to remove the international trade barrier. (Ekor, 2014) Multiple pharmacopoeias such as the European Pharmacopoeia (EUP), the Chinese Pharmacopoeia (ChP), the United States Pharmacopoeia (USP) and the British Pharmacopoeia (BP) have since introduced regulatory guidelines and monographs. EUP for example recommends chromatography techniques with marker compounds to identify THM and to detect adulterants in THM products using a reference

monograph for member states in the European Union (EU). (Bouin & Wierer, 2014; European Medicines Agency, 2006a; European Medicines Agency, 2006b) However, chromatography techniques are unable to provide efficacy representation from the markers for THM products as they are often times poorly studied and the component that elicits claimed therapeutic effects are often not known. Furthermore, the complexity of the chemistry in a THM product in forms such as raw powder, crude extracts and sometimes minimally processed parts of a plant such as dried leaves/flowers makes it difficult to identify the ideal marker compound. (European Medicines Agency, 2008)

2.2 Chemical Fingerprinting

In recent years, there has been a shift from using marker compounds for samples with complex chemical signals in chromatography analysis and instead, assume the chromatogram as the chemical fingerprint unique to these samples. Since any sample, specifically complex mixtures such as THM have unique chemical profiles, these collected chemical profiles are assumed to holistically represent the samples. Therefore, further interpretation of peaks or components is not necessarily required. Chemical fingerprinting is often used together with statistical and analytical models to compare the fingerprint of a sample to a reference fingerprint for identification or to simply distinguish between multiple classifications of samples. Chemical fingerprinting using gas chromatography - mass spectrometry (GC-MS) is widely used in the oil and gas industry to evaluate the distribution profiles of oil spills in the environment to identify their source. (Stout & Wang, 2016; Yang et al., 2017) Comprehensive two-dimensional gas chromatography - time-of-flight mass spectrometry (GC x GC-TOFMS) was used by Ueland et al. (2016) to chemically

fingerprint volatile organic compound (VOC) from rhinoceros' horns and successfully distinguished between two species of rhinoceros. High performance liquid chromatography with fluorescence detection, HPLC, liquid chromatography (LC) with fluorescence and LC has respectively been used as the chemical fingerprinting tool for distinguishing nut and detecting adulterant in almond-based product, assess food quality and evaluate food authenticity, distinguishing normal and flavanols-rich cranberry-based pharmaceuticals, and for differentiating avocados of different geographical origin and botanical variety. (Bakhytkyzya et al., 2018; Campmajóa et al., 2020; Esteki et al., 2019; Martín-Torres et al, 2020)

However, chemical fingerprinting is not limited to only chromatography techniques. Spectroscopy techniques are also explored extensively for this purpose. ultraviolet-visible (UV-vis) spectroscopy was used to fingerprint and discriminate *Coffea arabica* beans under conditions with restricted water and raised carbon dioxide levels. (Marcheafave et al., 2019) Crawford et al. (2020) compared untargeted proton-nuclear magnetic resonance (^{1}H-NMR) spectroscopy as a chemical fingerprinting tool with phenolic profiling, fatty acid profiling and microsatellite analysis for discriminating variety of processed olives. ^{1}H-NMR was also chosen as chemical fingerprinting tool by Windarsih et al. (2019) and Sun et al. (2018) for identifying *Curcuma longa* that are adulterated with *Curcuma heyneana* and differentiation of herbal medicine respectively. Attenuated total reflection (ATR-) FTIR spectroscopy techniques have also been adopted to fingerprint asphalt samples. (Ren et al., 2019)

2.2.1 Chromatographic and Spectroscopic Chemical Fingerprinting

Chromatography relies upon the interactions between constituents with the stationary phase and mobile phase. The interactions can be influenced by solvent choice (polarity/number of solvent/isocratic or gradient flow/boiling point/pH), flow rate (pump strength/column packing material), diameter and length of column, temperature, injection method, volume of injection, humidity, and the concentration and type of sample. Therefore, chromatography requires careful settings of protocol, sometimes expensive parts and consumables, controlled instrumentation environment and technical training for operators which translates to time and financial challenges. Spectroscopy illustrates the interaction between atoms with the electromagnetic spectrum which are affected by the type of wavelengths used for analysis. Any wavelength chosen presents unique challenges sometimes not experienced by another choice of wavelengths. For example, concentration of sample (Beer-Lambert Law) and excitation wavelengths are considerations for UV-vis spectrometry, magnetic field strength, temperature and choice of isotope are important for NMR spectroscopy and sample ionization method is crucial to mass spectroscopy. However, all techniques regardless need extensive sample preparation when compared to infrared radiation (IR) spectroscopy.

2.3 Infrared Radiation Spectroscopy

Solid samples are common with any IR spectroscopy and Conventional IR require only grounded dry powder mixed with a holding material, usually an inorganic salt to form a pellet for analysis while ATR-FTIR does not even require a proper pellet for analysis. IR spectroscopy can also be used with liquid and gas samples as well but

requires more advanced setup and instrumentation. The minimal sample preparation required for IR spectroscopy when compared to any other spectroscopy and chromatography methods is an advantage that can save time, money and reduce harmful chemical use, since most THM can be found already in dry powder form or easily made into dried powder without much further processing. Furthermore, significant amount of THM products are crude extracts that can be readily concentrated down into dry powder form ready for IR spectroscopy analysis.

2.3.1 Near-, Mid- and Far-Infrared Radiation

Although not his intention, Sir William Herschel – a British astronomer - is often credited for the discovery of infrared. (White, 2012) The infrared is a part of the electromagnetic spectrum. There are three infrared regions named in relation to their position to the visible lights on the electromagnetic spectrum with loosely defined boundaries using wavenumbers or wavelengths. Near-IR region contains higher energy and picks up right after the red color that our human eyes can see, around wavenumber 14,000 cm^{-1} to 4,000 cm^{-1} or wavelength 0.8 μm to 2.5 μm. The mid-IR region continues at wavenumber 4,000 cm^{-1} to 400 cm^{-1} or wavelength 2.5 μm to 25 μm while the far-IR region covers the rest of the region up to the microwave region from wavenumber 400 cm^{-1} to 10 cm^{-1} or wavelength 25 μm to 1,000 μm. The photon energy equation below demonstrates that IR contains energy that is proportionate to its wavenumber.

$$E=hf \quad (1)$$

Where E is the energy of the radiation

h is the Planck's constant = 6.626 x 10^{-34} Js

f is the frequency of radiation

$$c = f\lambda \quad (2)$$

$$f = \frac{c}{\lambda} \quad (3)$$

c is the speed of light in vacuum = 2.998 x10^{8} ms^{-1}

λ is the wavelength of the radiation

$$v = \frac{1}{\lambda} \quad (4)$$

v is the wavenumber of radiation

Substituting (4) into (3):

$$f = cv \quad (5)$$

Substituting (5) into (1):

$$E = hcv \quad (6)$$

$$\therefore E \propto v \quad (7)$$

For a diatomic bond system, the system can exhibit motions such as oscillation, vibration and rotation when given a certain amount of energy in its limited degree of freedom while more complex bond systems with higher degrees of freedoms can exhibit more modes of movements. Since IR carries different amounts of energy at different wavenumbers, some of that energy can be absorbed by the bonds in a sample. This principle is used in IR spectroscopy by passing a range of IR through a sample and detecting the loss of intensity at specific wavenumbers to generate a spectrum or

multiple spectra. For near-IR region, the spectra describe the overtone and harmonic vibrations, generally being featureless and clean while mid-IR spectra give information on the vibrations and other modes of movements of a structure. Spectra from the far-IR region are significantly lower in energy and correspond to the more intricate structural information of a sample. Although the information from near-IR and far-IR are useful in their own rights, for the purpose of chemical fingerprinting of THM which contains mixtures of complex and large organic molecules, mid-IR is especially great at providing just enough details as a fingerprint region for organic molecules in this type of analysis.

2.3.2 Fourier Transform Infrared Radiation Spectroscopy

Traditional IR were dispersive measurements where a monochromator grating or a prism was used to select the wavelengths of the beam and the scanning times were particularly slow due to the large range of wavelengths that needed to be covered. Nowadays, they are replaced by FTIR spectroscopy which uses interferograms that can measure all IR wavelengths simultaneously, significantly improving scanning times. The source of infrared light is first passed through a beam splitter which splits the light into two rays. The two rays then travel along their paths with distances of very small differences and then are reflected back along the path by a mirror at the end of each path. The two rays that are now slightly out of phase due to the differences in their path lengths then recombine and pass through the sample to the detector. This signal is called the interferogram and a Fourier transformation function is then applied to generate the spectra. The process is demonstrated in Figure 2.6.

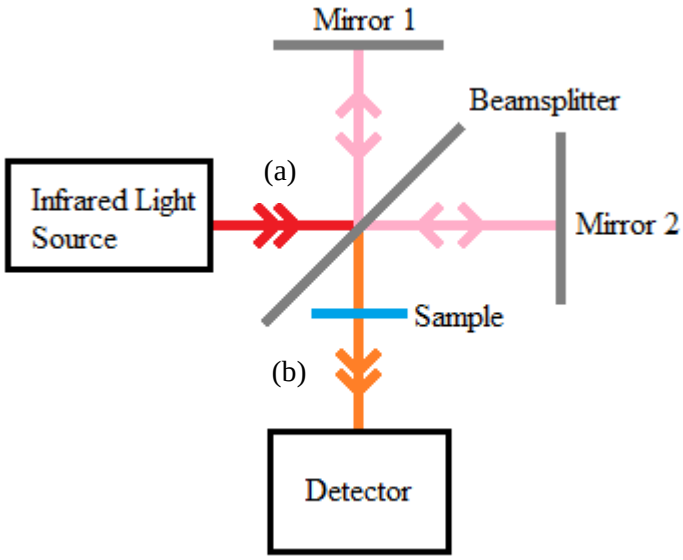

Figure 2.6 (a) Infrared light source split by a beam splitter towards two mirrors with slight differences in path lengths. (b) Reflected split beams recombine with slight phase shift from differences in path lengths to generate interferograms detected by the detector.

2.3.3 Signal-to-noise Ratio

Although FTIR in the mid-IR region addresses a lot of the problems faced by other analytical methods and has its own inherent advantages as a chemical fingerprinting tool, there are important parameters in the scans and spectral processing that needs to be considered carefully since differences in these parameters can render collected spectra unusable. The goal of collecting chemical fingerprint data is to maximize the signal-to-noise (S/N) ratio since the datasets are usually vastly similar with just minute differences. Enhancing the S/N ratio can greatly accentuate the differences for discrimination between samples. This can be achieved through increasing the scan number and resolution of the scan. Signals are consistent while noises are random occurrences during scanning. Therefore, increasing the scan number records several interferograms and averages the signals collected, eliminating the noise from the spectrum. However, the S/N ratio improves at the rate of the square root of

the scan number, giving diminishing return at high scan number but significantly longer time investment. The scan number used for this research lands on 16 as the middle ground between S/N ratio and scan time.

2.3.4 Resolution

The resolution of the spectrum essentially translates to the amount of data points present on the spectrum. A typical mid-IR spectrum with resolution of 1.0 cm^{-1} has 3601 data points while resolution of 4.0 cm^{-1} will yield 901 data points. Low amount of data points has smaller file sizes and is easier to handle but will generally present an over-smooth spectrum where important features are lost. This research used resolution of 1.0 cm^{-1} to have at least 1 data point for each wavenumber with a total data point of 3601 per spectrum.

2.3.5 Total Transmittance Percentage

The total transmittance percentage (T%) of a spectrum before spectral processing is the difference between the maximum and minimum transmittance. T% of over 60 % is the range where a spectrum is considered acceptable for analytical purposes. A lower than 60 T% indicates a sample that is too concentrated and small peaks can be easily obscured by larger or broader peaks.

2.3.6 Spectral Processing

After a spectrum has passed the T% criteria, the spectrum has to go through a few processing to be meaningful. A baseline correction is performed on the spectrum

to accommodate for instrument noise and light scattering. The spectrum will also require Savitzky-Golay smoothing to further enhance the S/N ratio. Spectral subtraction for spectrum shifted upwards and spectral addition for spectrum shifted downwards are performed when necessary to obtain a 0.00 A minimum point. Following that, the minor inconsistences in the thickness of sample pellets are corrected by using a simple min-max normalisation from the range of 0.00 A to 1.50 A to enable comparison between spectra. Going through all these crucial steps will ensure the spectra collected are feasible for the next steps in the analysis.

2.4 Fingerprint Analysis

2.4.1 Multivariate Analysis

Multivariate analysis (MVA) is a statistical method where multiple correlated or uncorrelated dependent variables are collected and the dependencies of the collected variables are analyzed. (Long, 2013) MVA is suitable for analyzing data when there is a complex number of variables that may or may not correlate with one another to produce a responding variable. MVA has become a common tool in the field of chemometrics to analyze chemical or spectral data that are multivariate in nature. Besides that, the robustness of the statistical model generated in MVA are crucial to find underlying information or subgroups that are usually obscured such as those that are in spectra. There are however, many types of MVA such as additive tree, canonical correlation analysis (CCA), cluster analysis, correspondence analysis/multiple correspondence analysis (CA/MCA), factor analysis, generalized procrustes analysis (GPA), independent component analysis (ICA), multivariate analysis of variance (MANOVA), multidimensional scaling (MDS), multiple regression analysis (MRA),