Developing statistical models from low-field and high-field NMR data to discriminate *Boswellia* species

by
Trevor Fox

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR BIOCHEMISTRY 448

University of British Columbia
Okanagan, Kelowna BC

April 2024
8109 Words

Abstract

Frankincense, the sap of trees in the *Boswellia* genus, has been a valuable commodity throughout history, used in traditional medicines worldwide to treat neoplasia, fever, gastrointestinal disorders, inflammation, and wounds. This resin contains many secondary metabolites that vary depending on the species of origin. Nuclear magnetic resonance spectroscopy (NMR) can be used to obtain unique spectra from biological samples, creating a chemical "fingerprint" that encodes species-specific chemical information. Statistical modelling algorithms such as partial least squares-discriminant analysis (PLS-DA) or random forests (RF) can then be applied to these fingerprints to create predictive models. Such models determine which spectral features represent the most significant differences between samples.

This study utilized 60 and 400 MHz nuclear magnetic resonance (NMR) spectroscopy to obtain spectra from frankincense samples originating from nine different *Boswellia* species. This data was used to produce a partial least squares discrimination analysis (PLS-DA) and Random Forests (RF) models to classify withheld test data. These models were able to classify *Boswellia* species, with 400 MHz models performing with the highest accuracy. Such NMR methods are advantageous because they only require a crude sample and a robust extraction method, making them independent of chromatography and unreliable ionization. These models have various potential applications, including measuring the quality and purity of commercial frankincense products and providing insights into the complex taxonomy of the Boswellia genus based on unique secondary metabolites.

Table of Contents

1 — Literature Review 1
1.1 — *Boswellia spp.* and frankincense 1
1.1.1 — History of frankincense 1
1.1.2 — Habitat and taxonomy 2
1.1.3 — Phytochemistry of *Boswellia* spp. 3
1.1.3.1 — Terpenes and terpenoids in *Boswellia* spp. 4
1.1.3.2 — Boswellic acids 7
1.2 — Nuclear magnetic resonance spectroscopy 7
1.3 — Metabolomics 8
1.3.1 — NMR metabolomics 9
1.4 — Statistical chemometrics 11
1.4.1 — Principal component analysis 11
1.4.2 — Regression-based classification - PLS-DA 12
1.4.3 — Ensemble-based classification - Random Forests 13
1.5 — Objective 13
2 — Materials and Methods 13
2.1 — Frankincense sample material 13
2.2 — General materials 14
2.3 — Sample preparation 14
2.4 — Extraction 14
2.4.1 — Preparation of the standard extraction solvent 14
2.4.2 — Determination of extraction conditions 15
2.4.3 — Extraction of frankincense samples 15
2.5 — NMR spectroscopy 16
2.5.1 — Determination of minimum number of transients 16
2.5.2 — NMR data acquisition parameters 16
2.5.3 — Determination of instrumental detection limits 17
2.5.4 — NMR spectral processing 17
2.6 — Statistical analysis 18
2.6.1 — Data preprocessing 18
2.6.2 — Exploratory data visualization 18
2.6.3 — Preparation of models 18
3 — Results and Discussion 19
3.1 — Determination of experimental parameters 19
3.1.1 — Extraction method 20
3.2 — Statistical models 21
3.2.1 — Principal component analysis 21
3.2.2 — Partial least squares discriminant analysis 24
3.2.3 — Random Forests analysis 27
3.3 — VIP analysis 31
3.4 — Comparison between 60 and 400 MHz models 31
4 — Conclusion 32

4.1 — Future directions 32
References 34

List of Figures

Figure 1: Images of *Boswellia sacra* Roxb. ex Colebr. **A**: An incision made in the bark to reveal resin (courtesy of T. Walsh, RBG Kew); **B**: Bark peeling from the trunk (courtesy of H. Pickering, RBG Kew).....2

Figure 2: Representative structures of the boswellic acid scaffolds. **A**: ursolic acid; **B**: oleanolic acid.....7

Figure 3: Example NMR spectra taken from *B. sacra* at 60 MHz (**A**) and at 400 MHz (**B**).....19

Figure 4: Scatterplots of the intensity of separate peaks at different time intervals of extraction. Chemical shift of analyzed peaks: **A**, 5.42 ppm; **B**, 2.48 ppm, **C**, 2.02 ppm; **D**, 1.61 ppm; **E**, 1.23 ppm; **F**, 0.79 ppm.....21

Figure 5: PCA scores plots for centred 400 MHz data. Ellipses encompass the 95% confidence interval.....22

Figure 6: PCA scores plots for centered and Pareto-scaled 60 MHz data. Ellipses encompass the 95% confidence interval.....23

Figure 7: PCA scores plots for centered and Pareto-scaled 400 MHz data. Ellipses circumscribe the 95% confidence interval.....24

Figure 8: Plot of Cohen's Kappa value against number of components considered when training PLS-DA models. **A**: 60 MHz, **B**: 400 MHz.....25

Figure 9: Performance confusion matrices of PLS-DA models generated from 60 and 400 MHz spectra. **A**: 60 MHz training-set predictions; **B**: 60 MHz test-set predictions; **C**: 400 MHz training-set predictions; **D**: 400 MHz test-set predictions. Green cells represent correct predictions. 26

Figure 10: Plot of Cohen's Kappa value against number of components considered when training and cross-validating RF models. **A**: 60 MHz, **B**: 400 MHz.....28

Figure 11: Performance confusion matrices of RF models generated from 60 and 400 MHz spectra. **A**: 60 MHz training-set predictions; **B**: 60 MHz test-set predictions; **C**: 400 MHz training-set predictions; **D**: 400 MHz test-set predictions. Green cells represent correct predictions. 30

List of Tables

Table 1: List of select abundant named compounds and compound classes known to be found in frankincense extractions via various detection techniques. 6

Table 2: Summary of calculated values for LoD and LoQ.....20

Table 3: Performance metrics for test set prediction using PLS-DA models.....27

Table 4: Number of variables required per decision tree and model performance metrics of generated Random Forests models for 60 and 400 MHz spectra.....29

Table 5: Summary of VIP spectral bins and % importance (imp.) from 60 and 400 MHz RF models.....31

Abbreviations

1D	One-dimensional [spectroscopy]
2D	Two-dimensional [spectroscopy]
^{13}C	Carbon 13; spectroscopy
^{1}H	Proton; spectroscopy
B. carterii	*Boswellia carterii*
B. dalzielii	*Boswellia dalzielii*
B. elongata	*Boswellia elongata*
B. frereana	*Boswellia frereana*
B. neglecta	*Boswellia neglecta*
B. papyrifera	*Boswellia papyrifera*
B. rivae	*Boswellia rivae*
B. sacra	*Boswellia sacra*
B. serrata	*Boswellia serrata*
B. serrata	*Boswellia serrata*
$C_{\#}$	Carbon chain that is # long
$CDCl_3$	Deuterated chloroform
C	Carbon
dd	Doublet of doublets, spectroscopy
DMSO	Dimethyl sulphoxide; solvent
DMTP	Dimethyl terephthalate; chemical standard
GC-FID	Gas chromatography-flame ionization detection
HMDB	The Human Metabolome Database
hr	Hour(s)
HSQC-TOCSY	Heteronuclear single-quantum correlation and total correlation spectroscopy
J	Coupling constant; spectroscopy
LC-MS	Liquid chromatography-mass spectrometry
LoD	Limit of detection
LoQ	Limit of quantification
MEP	methylerytrhitol phosphate biosynthetic pathways
mg, g	Milligram,gram; 10^{-3}, 10^{0} grams; units of mass
MHz, Hz	Megahertz, hertz, unit of frequency; s^{-1}; 10^{6}, 10^{0}
min	Minutes
mm	Millimeter; unit of length; 10^{-3} metres
MoNA	Massbank of North America
MS	Mass spectrometry
MVA	Mevalonate biosynthetic pathway

m	Multiplet; spectroscopy
NMR	Nuclear magnetic resonance
PCA	Principal component analysis
PLS-DA	Partial least squares-discriminant analysis
ppm	Parts-per-million
p	Pentet; spectroscopy
RF	Random forests (stats) or radio-frequency (spectroscopy)
sp.	Species
SPME	Solid-phase microextraction
spp.	Species; plural
s	Singlet; spectroscopy
tq	Triplet of quartets; spectroscopy
t	Triplet; spectroscopy
°C	Degrees Celsius; temperature
δ	Chemical shift, ppm; spectroscopy
µL, mL	Microlitre, millilitre; 10^{-6}, 10^{-3} litres
µM, mM	Micromolar, millimolar; 10^{-6}, 10^{-3} molar
µs, ms, s	Microsecond, millisecond, second; 10^{-6}, 10^{-3}, 10^{0} second

Acknowledgements

I would like to express my gratitude to Dr. Paul Shipley for giving me the opportunity to conduct this research and for your incredible guidance along the way. Thank you to Dr. Susan Murch and the rest of PlantSMART for providing your support and a wonderful, welcoming lab space. I am also grateful to Emma Mitchell, Ieva Zigg, and the Department of Chemistry teaching labs for lending me time on the 60 MHz NMR.

Last but not least, thank you, reader, for taking the time to read about my work.

To my parents.

Et un nouveau soleil se lève.

1 — Literature Review

1.1 — *Boswellia spp.* and frankincense

1.1.1 — History of frankincense

Frankincense, also known as olibanum, has been a valuable commodity since the development of early civilizations.[1] This is likely due to its pleasant aroma and the simplicity of its harvest from trees of the genus *Boswellia* (family Buseraceae).[2] The resin itself is secreted from scores in the bark caused by animals, pests or humans.[3] The collected pieces are known as "tears" because of their shape. Frankincense has been used for traditional medicine, perfumes, aromatherapy, and religious ceremonies. As it was highly subject to trading, this commodity was available to civilizations worldwide, such as Northern Africa, Egypt, and parts of the Middle East.[3,4,5,6]

Most commercial frankincense originates from the species *Boswellia serrata* and *B. sacra* from which it is harvested via a wound made by a knife in the tree's bark **(Figure 1)**. The resin that seeps from the wound is then collected and hardened.[2]

The resin has historically been used in traditional medicines around the world for the treatment of neoplasia, fever, gastrointestinal disorders, inflammation, and wounds.[7,8,9] It can be assumed that this research and the promise of natural medicine contribute to the recent renewal of frankincense's popularity. This both exacerbates its overharvest and emphasizes the need to prevent it.[2] This is supported in the literature, as extracts of *Boswellia* such as the ethanolic fraction and essential oil, have been described as anti-inflammatory, antioxidative, and antineoplastic.[10,11,6]

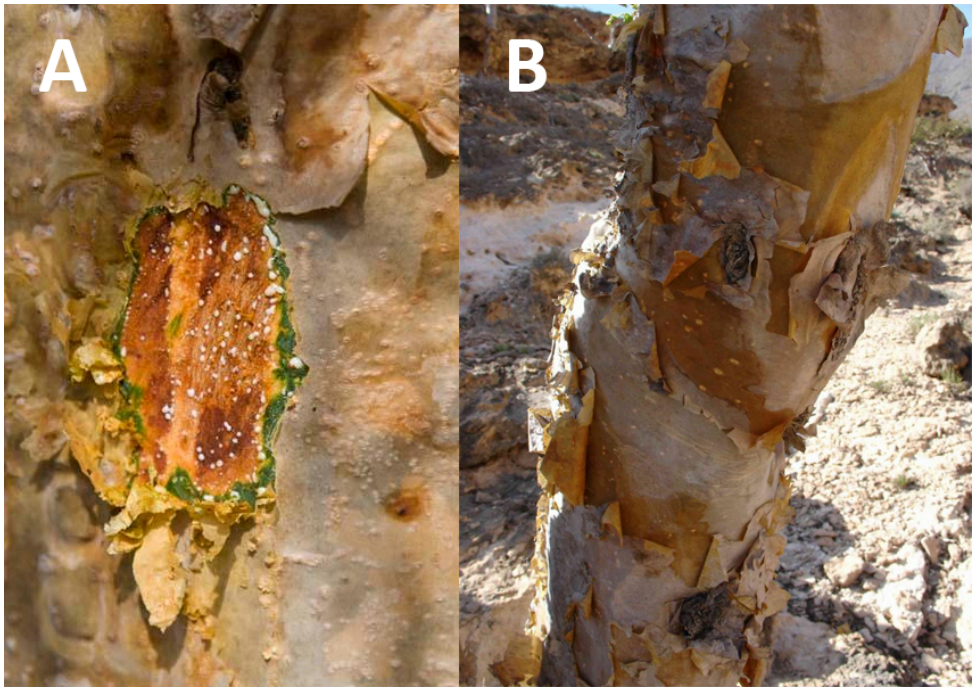

Figure 1. Images of *Boswellia sacra* Roxb. ex Colebr. **A**: An incision made in the bark to reveal resin (courtesy of T. Walsh, RBG Kew); **B**: Bark peeling from the trunk (courtesy of H. Pickering, RBG Kew).

Destructive harvesting, pests, and high demand for frankincense products are putting some species of *Boswellia* at risk of extinction.[2,12,7] As the trees are given less time to recover between harvests to maximize profit, they slowly die. Over time, as the number of trees decreases, there may no longer be enough to sustain the species. The decline in viable *Boswellia* population may encourage the adulteration of frankincense products, as has been reported in extra virgin olive oil.[13]

To prevent this from becoming prevalent, chemometrics can be used to establish statistical models to determine the purity of a frankincense sample or even the species from which it was harvested. With such safeguards, frankincense can be screened quickly using mass spectrometry or nuclear magnetic resonance spectroscopy methods.

1.1.2 — Habitat and taxonomy

The distribution of the *Boswellia* genus is wide across the globe. This distribution may have caused the misclassification of species by mere visual observation. Many different people were

discoverers of new *Boswellia* species, so it cannot be ruled out that overlap in classifications arose from separation through space and time. This separation has caused confusion in the naming of *Boswellia* spp., which has many synonyms. One constant, however, is that *Boswellia serrata* Roxb. ex Colebr. remains the type for the genus, thus sometimes referred to as simply *Boswellia*. The online plant database *World Flora Online* officially recognizes 20 species.[14] In contrast, the *Plants of the World Online* database lists 22, with outliers being *B. occulta* and *B. ruspoliana*.[15]

In the literature, there are many synonyms for *Boswellia* species. For example, *B. glabra* Roxb. is a synonym for *B. serrata* and *B. bricchettii* is synonymous to *Lanneo obovate*, which is not within the *Boswellia* genus.[6] Chemotaxonomy can be an alternative or supportive form of classifications to support the use of chemical information as a guideline for naming.

Attempts have been made to clarify the obfuscated taxonomy of the *Boswellia* genus using chemometrics.[2,16,17,18] These include using tandem gas chromatography-mass spectrometry (GC-MS) to analyze the volatile compounds and organic extracts of oils, thin-layer chromatography to assess phenolic compounds, and a novel approach using diatomite column chromatography and mass spectrometry. This information will support efforts to produce a robust method of analyzing commercial frankincense samples or other commercial products.

1.1.3 — Phytochemistry of *Boswellia* spp.

The biochemistry of frankincense has not been discussed at length in the literature. Therefore, efforts to elucidate the driving factors of chemical differences between species are essential in understanding these species. Existing research has focused on identifying which types of compounds are present in various extraction conditions.

Frankincense resin of trees within the *Boswellia* genus consists of three major fractions. The smallest of the fractions is the oil fraction, which comprises the small molecular weight lipids such as monoterpenes. The next largest fraction is the water-soluble fraction, which comprises mostly carbohydrates and gum (polysaccharide polymers).[19]

The largest fraction is the ethanolic extract, which is the most investigated of the three. This fraction contains the rest of the terpenes, terpenoids, and fatty acids.[20] Incensole is included in this fraction, which is known to be a potential anti-inflammatory, anticancer, and anti-HIV (human immunodeficiency virus) compound.[10] The ethanolic fraction potentially has clinical applications, demonstrating weak antioxidative potential and the ability to inhibit lipoxygenase, reducing inflammation.[11,21] Moreover, this fraction contains boswellic acids, a class of pentacyclic triterpenes that have garnered researchers' interest in their potential pharmacological activities (Section 1.1.3.2).[21]

1.1.3.1 — Terpenes and terpenoids in *Boswellia* spp.

Terpenes are organic compounds that are synthesized from isoprenyl 5-carbon (C_5) monomers called isopentenyl pyrophosphate and dimethylallyl pyrophosphate (IPP and DMAPP hereafter). Isoprene monomers can further be conjugated to form geranyl pyrophosphate (C_{10}), farnesyl pyrophosphate (C_{15}), geranylgeranyl pyrophosphate (C_{20}), and squalene pyrophosphate (C_{30}). These long-chain unsaturated carbon compounds are produced from the mevalonate (MVA) and methylerythritol phosphate (MEP) pathways.[22,6] The naming convention of terpenes is determined by the number of carbons in its final structure.[22]

Subsequent transformations to these long-chain carbon compounds, such as cyclization and oxidation/reduction, produce an extremely wide variety of terpene compounds. More than 35 000 terpenes and terpenoids are known, making the group one the largest of the natural product classes.[22]

A summary of compounds which are present in relatively high abundance in frankincense samples is provided below **(Table 1)**. These compounds were detected previously by analyzing the organic-soluble portion of frankincense via vapour headspace GC-MS, superheated steam extract, gas chromatography-flame ionization detection (GC-FID), and headspace solid-phase microextraction (SPME). Most of these compounds belong to the terpenes, with some presence of fatty acids.[11] It is likely that, in an organic extraction of frankincense, some high-

concentration constituents such as α-pinene, β-ocimene, and camphene are detectable by NMR.

Table 1. List of select abundant named compounds and compound classes known to be found in frankincense extractions via various detection techniques.[11,20,23]

Compound	***Class***	***Compound***	***Class***
1,8-cineole	monoterpene	borneol	monoterpene
bornyl acetate	monoterpene ester	boswellic acids, oleanane	triterpenoids
boswellic acids, ursane	triterpenoids	camphene	monoterpene
carvone	monoterpene	caryophyllene	sesquiterpene
cis-piperitol	monoterpene alcohol	d-verbenone	monoterpenoid
isopinocamphone	pinane monoterpene	limonene	monoterpene
linalool	monoterpene	m-cymene	monoterpene
myrcene	monoterpene	myrtenol	monoterpene alcohol
p-cymen-8-ol	monoterpene alcohol	p-cymene	monoterpene
sabinene	monoterpene	terpinen-4-ol	monoterpene alcohol
trans-pinocarveol	monoterpene alcohol	trans-verbenol	monoterpene alcohol
verbenene	monoterpene	α-campholenal	monoterpenoid
α-phellandrene-8-ol	monoterpene alcohol	α-pinene	monoterpene
α-thujene	monoterpene	α-thujone	monoterpene
β-cedrene	sesquiterpene	β-ocimene	monoterpene
β-pinene	monoterpene	β-thujone	monoterpene
δ-cadinene	sesquiterpene	palmitoleic acid	fatty acid
oleic acid	fatty acid	lauric acid	fatty acid
arachidonic acid	fatty acid		

1.1.3.2 — Boswellic acids

It is known that frankincense samples of all species are predominantly comprised of aliphatic terpene and terpenoid compounds such as mono-, di-, and triterpenoids **(Table 1)**. Particular attention has been paid to boswellic acids, triterpenes uniquely found in *Boswellia*. These compounds are of interest because of their clinical relevance. Boswellic acids potentially have cytotoxic, anticancer properties. A recent review revealed that several papers have been published investigating the antimicrobial, anti-inflammatory (n=132), anti-neuropathology (n=63), and anti-oxidant (n=25) effects.[24]

Boswellic acids comprise compounds with one of two carbon scaffolds. These scaffolds, ursane-type and oleanane-type, are based on the structures of ursolic and oleanolic acids.[20,25]

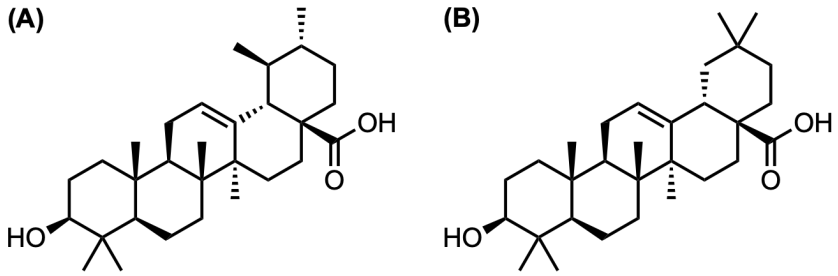

Figure 2. Representative structures of the boswellic acid scaffolds. **A**: ursolic acid; **B**: oleanolic acid.[22]

1.2 — Nuclear magnetic resonance spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is a powerful tool for various applications. The most prevalent of these applications is the structural determination of pure compounds. It operates on the premise that atoms with non-zero spins behave predictably in a strong, uniform magnetic field. Using radio-frequency pulses at or near the resonant frequency of these atoms, one can detect every chemical and magnetic environment of that particular atom. This is done by measuring the bulk precession of spins for a period after the radio-frequency pulse.[25]

The most common form of NMR spectroscopy is the one-dimensional proton (^{1}H) NMR acquisition. In this simple experiment, a radio-frequency pulse at the proton Larmor frequency generates a signal for every proton environment in an organic molecule. Many transients are collected and added together before a Fourier Transform is applied, yielding a spectrum encoding the Larmor precessional frequencies of measured protons. A chemical structure can be elucidated based on the properties that peaks exhibit in the spectrum, such as multiplicity, relative integration, and scalar coupling. Though this has been NMR's traditional purpose, an ensemble of compounds may also be measured simultaneously with no additional effort.[25,26]

NMR spectroscopy shows promise in supplementing traditional mass spectrometric metabolomics methods as a tool for untargeted chemical profiling. It can generate a "fingerprint" spectrum unique unique to a biological sample.[26] These spectra are unique because the chemical phenotype varies between biological samples, whether the species is the same or different.

1.3 — Metabolomics

> "***The general aim of metabolomics is the qualitative and quantitative analysis of all metabolites (the metabolome) present in an organism at a specific time and under specific influence***"
>
> — Yuliyana et al.[27]

Metabolomics is traditionally performed using a chromatography-based mass spectrometry method such as LC-MS.[28] This method depends on column chromatography which has drawbacks such as low throughput and column-based bias in separations. Mass spectrometry may also require derivatization to detect small compounds or compounds that are difficult to ionize reproducibly – an extra time-consuming step in methods development.[29]

The ionizability of compounds in a sample depends heavily on their functionalization, elements, and overall structure. This means that the detection of these compounds varies on a per-compound basis rather than a property of the sample or sample preparation. Possible

errors in this method include missed features via poor ionization or misclassification due to dual ionization, effectively halving the observed mass-to-charge ratio (*m/Q*). A common way around this is by derivatization, where a robust chemical reaction is applied to the sample to functionalize target compounds with a prosthetic group that is easily ionized by the instrument. While this works for compounds known not to ionize well – such as amino acids – it is not a reliable method to detect those one would not expect to see.[30,31]

Data repositories for metabolomics data contain spectra for hundreds of thousands of metabolites. Some of these databases include the Human Metabolome Database (HMDB), the MassBank of North America (MoNA), and the natural products database LOTUS.[32,33,34] A vast majority of these databases consist of mass spectrometry data. For example, only 8.87% of all spectra in HMDB are NMR spectra, and 5.83% of all compounds have an associated NMR spectrum.[34] From this, it can be seen that NMR metabolomics is a nascent field with a great need for developing new methods.

1.3.1 — NMR metabolomics

When performing NMR metabolomics, a sample extraction must be performed using an optimal solvent and method. These samples are then analyzed using the NMR instrument without processing such as derivatization. The output from a standard ^{1}H acquisition contains spectral information for all detectable organic compounds in the sample. This data can be analyzed to determine the peaks that drive statistical variance between sample treatments. However, this information is often highly obfuscated and cannot solely be used to identify compounds to which these peaks belong.

To solve this, other pulse sequences can be used post-analysis to determine which compound a peak of interest may belong to. In addition to the aforementioned ^{1}H NMR sequence, the heteronuclear single quantum correlation-total correlation spectroscopy experiment (hereafter HSQC-TOCSY) can provide this information. It encodes data that isolates total spin systems within a molecule from the other molecules in the sample. This is because

it combines the TOCSY pulse – which maps all protons in a spin system – with the HSQC, which maps all carbons to their attached protons. Thus, this two-dimensional method can be used to identify such molecules within a biological metabolomics sample.[35]

NMR methods can be quite more robust than that of mass spectrometry-based methods like liquid-chromatography mass-spectrometry (hereafter LC-MS). It can take a snapshot of the abundant compounds present in a sample with little effort and great reproducibility. It does not depend on a separation system, making detection of all types of compounds dependent only on extraction conditions such as solvent and temperature. This extraction method must take the wide variety of organic compounds that may be present. Moreover, detection of any size and class of compound is possible because NMR does not require ionization. Without the need for derivatization, samples can go straight from extraction to the instrument.

Last, NMR spectra are highly reproducible. Where mass spectrometry can see variable results from instrument to instrument, NMR data can be recreated, requiring only the same field strength of NMR instruments.[27] This requirement arises from the property of coupled peaks retaining their coupling constant despite stronger magnetic fields. For example, doublet peaks are spaced closer together in a 600 MHz spectrum than a 60 MHz spectrum.

In the literature, Kim et al. have attempted to characterize NMR-based methods for plant metabolomics.[28] Moreover, Sumner et al. have proposed a reporting standard for NMR metabolomics data.[36] The principles outlined in these previous works and others can be used as a framework to develop a pipeline for producing fit-for-purpose methods that can be adapted and validated for many other downstream uses.

A major setback of NMR spectroscopy is its sensitivity.[35] Relative to mass spectrometry, high concentrations of compounds are required to generate enough signals with the instrument. However, some measures can be taken to improve the overall signal-to-noise ratio (hereafter SNR) of low-concentration samples. For example, a 45° pulse over a 90° pulse with a higher number of transients can be used. In this case, the 45° pulse is used to reduce the

relaxation time by one half-life, allowing for more transients to be collected in the same period of time. Thus, SNR is reduced because the transients from which the baseline is averaged are more numerous. Moreover, sensitivity can be improved greatly using a supercooled cryogenic probe under a stronger magnetic field.[35] More NMR facilities contain instruments with these capabilities as time passes, making this a promising solution.[26]

1.4 — Statistical chemometrics

Chemometrics is the process of creating models from which a meaningful interpretation can be made.[37] Such models are created from data collected from various methods such as mass spectrometry, NMR spectroscopy, spectrophotometry, and more.

Statistical models are produced considering nuanced differences between biological samples to perform chemometrics. These differences are ranked based on their importance by, for example, measuring the variance that a compound contributes to the data set.[38] Compounds can be ranked based on their variance and targeted for further analysis. The statistical chemometric methods used in this study are principal component analysis, partial least squares discriminant analysis, and Random Forests.

1.4.1 — Principal component analysis

Principal component analysis (PCA) is a method that reduces a complex set of data to a collection of principal components (PCs) that describe features that give rise to the most variance. While reducing the number of dimensions to consider, the major patterns remain, allowing for simpler analysis and interpretation. The main goal of the PCA algorithm is to cluster labels together while preserving variance. This constructs a simplified description of the data set.[39,40,41]

PCA is performed by finding the components within the data set with the highest degrees of variance. The data is then recentered according to these components, and a series of matrix transformations and projections are performed.[42] The final metrics of the PCA are the

scores and loadings. Scores describe the distribution of individual points, whereas the loadings indicate the points that correlate the most with a particular principal component.

While the PCA is useful for visualizing data clustering, it cannot be used directly as a prediction model. This is because the model abstracts class away during its transformations in its effort to minimize co-variance and maximize variance.

1.4.2 — Regression-based classification - PLS-DA

Partial least squares discriminant analysis (PLS-DA) is a similar method to PCA, wherein a "supervised" model is trained from a classified two-dimensional data set. The term supervised refers to the process of providing class information to the model so that it may consider the separation of actual classes. This differs from PCA, which does not consider the real class of the data and instead focuses on the effect of the principal components. The process of training a PLS-DA model involves preprocessing data, splitting data into training and testing sets, and training with cross-validation and hyperparameter tuning to determine the number of principal components to consider. Moreover, PLS-DA aims to explain the differences between classes rather than between samples.[40]

PLS-DA is often used to create models for metabolomics. These methods analyze and compare the clustering of features in a dataset based on the variance of single features between samples. In doing so, the model maximises co-variance, such that grouped data remains highly correlated in the model. This can be the basis for deciding which compounds to analyze further, using HSQC-TOCSY to determine its structure among those present in the sample.[43]

Furthermore, when trained with a sufficient size of data, PLS-DA models can be used to predict which class an "unknown" compound belongs to by projecting it onto its latent variables, determining which class it identifies the most with.[13]

1.4.3 — Ensemble-based classification - Random Forests

PLS-DA has the tendency to overfit data. Thus, a supervised ensemble method such as Random Forests (RF) has been considered a new alternative model.[39] This is because data pruning reduces the probability of the model being dominated by the largest peaks in a spectrum. RF models are trained by preprocessing data and then constructing trees with cross-validation to determine the number of variables to include in each tree. The produced ensemble of decision trees is what is used to perform predictions.[44]

This model functions by growing a forest of many decision trees, each of which is trained on a different subset of data. These decision trees are then assessed using a test set, wherein the trees "vote" which classification is best for the incoming data. The number of votes for each class is counted, and the class with the most votes is considered the final prediction. This method handles overfitting by limiting the number of variables used for each prediction. This way, a handful of prominent variables cannot dominate the entire model.[44]

1.5 — Objective

This study aims to determine frankincense's ability to be classified from its NMR spectrum using PLS-DA and RF statistical models. Moreover, a comparison will be made regarding which model performs better and their relative tradeoffs. The results of this study will also provide valuable insights into the complex taxonomy of the Boswellia genus based on unique secondary metabolites.

2 — Materials and Methods

2.1 — Frankincense sample material

Frankincense samples from different *Boswellia* spp. were obtained from a commercial vendor (Apothecary's Garden; Hamilton, ON, Canada). Samples purportedly originated from species *B. sacra*, *B. carterii*, *B. frereana*, *B. serrata*, *B. papyrifera*, *B. neglecta*, *B. rivae*, *B. elongata*, and *B. dalzielii*.

2.2 — General materials

Both crystalline dimethyl terephthalate (DMTP) NMR reference standard and deuterated DMSO (DMSO-d_6 99.5% D) were purchased from Sigma-Aldrich (Oakville, ON, Canada). Aldrich ColorSpec NMR tubes (7 in. L x 5 mm diam., 0.38 mm wall; 400 MHz) were also purchased from Sigma-Aldrich. Twist-lock 2 mL microcentrifuge screw-top and 1.7 mL snap-top microcentrifuge tubes were purchased from VWR (Mississauga, ON).

2.3 — Sample preparation

Frankincense samples were chosen randomly, ensuring no samples had visual evidence of debris contamination (e.g. wood, dirt). Each sample, taken as an individual tear, was then fragmented into small pieces by dropping a mortar 1 inch above the table onto the sample and wrapping it in weighing paper. Fragments were then randomly selected and placed into a 2 mL screw-top microcentrifuge tube so that the total weight of the sample was approximately 0.100 g ± 0.01 g. Samples not immediately used were stored in the dark at room temperature.

2.4 — Extraction

2.4.1 — Preparation of the standard extraction solvent

The volume of DMSO-d_6 was calculated via its density, and the mass required was subsequently determined. To do so, a 50-gram bottle of DMSO-d_6 was poured into a tared 100 mL beaker. This beaker was then parafilmed to reduce exposure to water. The mass was taken, and the original bottle was dried of remaining solvent residues. Last, the solvent was decanted back into the bottle, and the mass of the residue in the beaker was taken and subtracted from the final mass. This mass was translated to volume using the recorded density of 1.190 g/cm^3 and multiplied by 5 mg to determine the mass of DMTP to dissolve. The DMTP was dissolved into the bottle and marked with its concentration.

2.4.2 — Determination of extraction conditions

Many trials were performed to determine the extraction method for this experiment. Each experimental trial was measured at 5, 20, 60, and 120 minutes (5 minutes omitted for trials 3 onward). Biological samples used for this experiment originated from tears of *B. sacra*.

Extraction trials #1 and #2 were performed under 3 mL of solvent (DMSO-d_6) in 13mm x 100mm glass culture tubes. These extractions resulted in cloudy suspensions, so a decision was made to use microcentrifuge tubes to enable centrifugation of the samples. Thus, trial #3 utilized 1.5 mL snap-lock microcentrifuge tubes, and trials #4 and #5 utilized 2 mL screw-top microcentrifuge tubes.

Each extraction was sampled at each time point by taking a 700 µL aliquot and filtering it into a clean NMR tube. The sample was left to extract until the final time when it was removed. As the 1.5 mL trial (#3) could not account for this, 500 µL aliquots were used. This, however, was not sufficient for the instrument, and this trial was not used. The same was true for the 2 mL trial (#4), as there was insufficient solvent for the final trial (600 µL). An alternative was chosen so as not to compromise the volume of the extraction solvent. This final trial (#5) used 500 µL aliquots which were diluted by DMSO-d_6 in a ratio of 1:1. This resulted in two-fold dilutions for all samples with a total volume of 1.00 mL. These samples were centrifuged, and an 800 µL aliquot of these solutions was used for NMR analysis.

Data was analyzed by plotting the integration of prominent peaks for each time-point spectrum after normalization of the DMTP peak area at 8.0678 ppm to `50 000` and processed according to section 2.5.4. These integrations were taken from peaks at δ 0.79, 1.23. 1.61, 2.02, 2.48, and 5.42 ppm and plotted.

2.4.3 — Extraction of frankincense samples

A cotton filter was prepared for each NMR tube using a 5-inch glass Pasteur pipette with a small wad of cotton packed in the taper. The small end of the pipette was then placed inside an NMR tube for later use as a funnel.

A pre-weighed frankincense sample was added to a 2 mL screw-top microcentrifuge tube. Added to this tube was 1.8 mL of DMSO-d_6. The sample was extracted for one hour at room temperature in a sonicator bath. Samples were fixed in the bath using a floating tube rack to ensure each sample was fixed at the same height. A retort stand with an attached three-pronged clamp was also used to hold the rack in the center of the bath to ensure that it did not contact the sides of the sonic bath. After extraction, each sample was centrifugated at 13,000 x g for 1 minute, gathering any remaining particulate into a pellet at the bottom of the tube. Then, 700 µL of the transparent supernatant was transferred using a micropipette into a clean 7-inch NMR tube through the cotton filter.

2.5 — NMR spectroscopy

2.5.1 — Determination of minimum number of transients

An arrayed experiment of eight consecutive acquisitions was performed using an array of 2, 4, 8, 16, 32, 64, 128, and 256 transients with all other parameters identical to those in Section 2.5.2. This data was subsequently processed as per Section 2.5.4.

2.5.2 — NMR data acquisition parameters

All 60 MHz 1-D ^{1}H NMR spectra 1-D ^{1}H NMR spectra were acquired at 25 ℃ over a spectral width of 540 Hz between 0 and 9 ppm. A relaxation delay of 1 second was used between transients. Time elapsed per transient was 5.96 seconds, yielding 2048 complex points, and 64 transients were collected from each sample.

All 400 MHz spectral data for this study was collected using a Varian MercuryPlus 400 MHz NMR instrument (Varian, Inc. Palo Alto, CA, USA), which uses a 400 MHz automated triple-broadband probe and pulse field gradient generator (1H at 400.14 MHz). Varian's proprietary desktop application VnmrJ v2.2—distributed for Red Hat Enterprise Linux v5.1—was used to control the instrument and configure experimental parameters. Spectra were acquired at 25 °C over a spectral width of 4001.6 Hz, covering a range of δ between −1.0 and 10.0 ppm with an observation pulse length of 5.90 µs (45°). Each sample was not spun, and the instrument

was maintained at 25 ℃. A relaxation delay of 1 second was used between transients. The total time elapsed per transient was 4.094 seconds which yielded 16,384 complex points. Sixty-four transients were collected from each sample.

2.5.3 — Determination of instrumental detection limits

To calculate the limit of detection and limit of quantification (hereafter LoD and LoQ), a dilution series of DMTP in DMSO-d_6 was prepared and analyzed. A 2.575 mM stock solution of DMTP was prepared as per Section 2.4.1. Aliquots of this stock solution were added to pure DMSO-d_6 and subsequently diluted in a ratio of 1:99 to produce a range of solutions from 0.002575 mM to 0.02575 mM. SNR was calculated, plotted using MNova, and correlated via concentration to peak intensity.

2.5.4 — NMR spectral processing

All NMR spectra were processed using the MNova software v15.0.0-34764 (Mestrelab Research) and its "stack" function. The processing performed consisted of per-spectrum automatic and manual phasing, automatic baseline correction via Whittaker Smoothing, and an apodization of 1 Hz.

The water peak region from 3.64 ppm to 2.93 ppm was removed from the 60 MHz spectra due to its large variance in intensity.

Spectra were referenced to the solvent peak of DMSO-D_6, and the reference peak at 8.14 ppm was used to normalize via peak-area, which was set to 50 000. MNova was then used to bin the spectra with a size of 0.01 ppm with the `sum` method.

For the integration of series experiments such as the LoD/LoQ determination (Section 2.5.3), MNova's data array and stacking features were used. This data was plotted and visualized using the R package `ggplot2` (v3.4.4).[45]

2.6 — Statistical analysis

All data analysis and processing was performed using R version `4.3.2` in R Studio version `2023.12.1+402` on MacOS (M3). All pseudorandom operations were seeded with the decimal number 14 (`set.seed(14)`).

2.6.1 — Data preprocessing

Data was centred and log-transformed using the R standard library, then Pareto scaled using the `MetabolAnalyze` (v1.3.1) R package.[46] Data sets were prepared for raw, log-transformed, and log-centered-Pareto scaled data. Data was then split into training and testing sets by randomly subtracting two samples from each species, forming the test set. The resultant data sets comprised 500 spectral bins as columns and 61 sample spectra as rows.

2.6.2 — Exploratory data visualization

Principal component analysis (PCA) scores plots were prepared using the `prcomp` function in the `stats` R package.[47] The plots were generated using `ggplot2` and `ggfortify` (v0.4.16).[48]

2.6.3 — Preparation of models

PLS-DA models were produced using the `caret` R package, which utilizes the `pls` (v2.8-3) R package.[43] Models were generated from the log-transformed data set, and a confusion matrix was produced using `ggplot2`.

RF models were also produced using the `caret` package, which uses the `randomForests` (v4.7-1.1) package for RF implementations.[49,50]

Data (N=61) was split into training (n=43) and testing sets (n=18) by drawing two spectra randomly from each species group and relegating them to the test set. For both model types, a fit tune length of 20 principal components (for PLS-DA) or number of variables (mtry, for RF) was used alongside 10-repeat, 10-fold cross-validation. Models were optimized by maximizing Cohen's Kappa metric.[51] The efficiency of the models was evaluated using confusion matrices from the test set predictions.

3 — Results and Discussion

3.1 — Determination of experimental parameters

An array experiment determined the number of transients that accomplished a good signal-to-noise while minimizing the time spent for each acquisition. In this experiment, the transients were arrayed from 2 to 256, with each step calculated as 2^n. The generated spectra were inspected visually, and it was decided that 64 transients (approximately 5-minute duration) were sufficient to see strong and distinct peaks in *Boswellia* sample spectra. To further characterize this, the LoD and LoQ metrics were calculated.

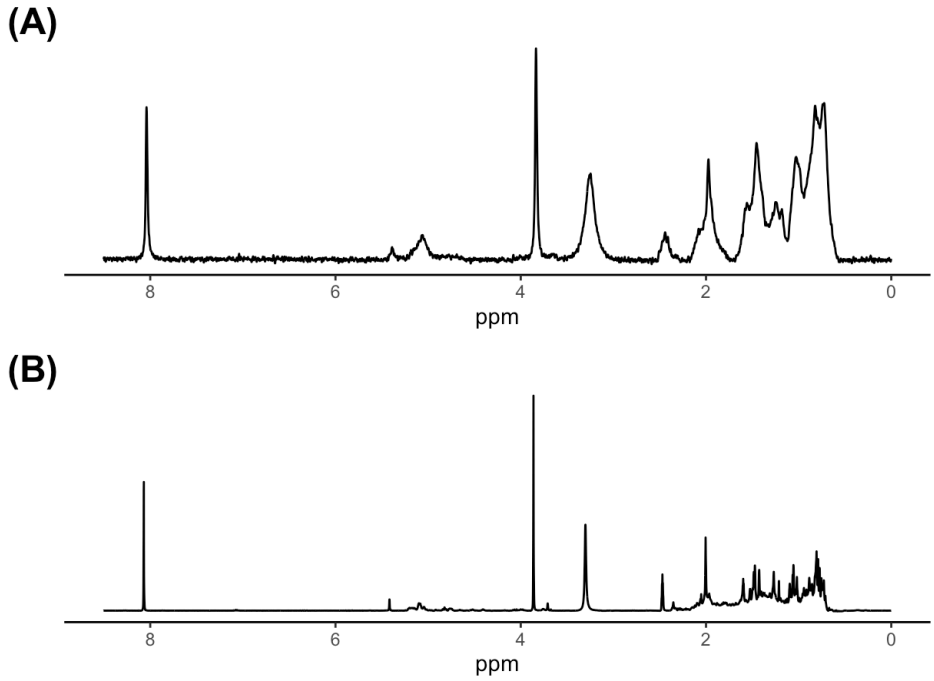

Figure 3. Example NMR spectra taken from *B. sacra* at 60 MHz (**A**) and at 400 MHz (**B**).

To calculate LoD and LoQ, another experiment wherein a series of dilutions ranging from 0.0026 mM to 0.02575 mM as per Section 2.5.3 was analyzed. The following is calculated for the 400 MHz data set. The SNR plot yielded a regression fit of $y_{\mathrm{SNR}} = 1244.5x - 0.43$. These linear regressions were then used as a system of equations to solve peak height for specific values

of SNR. According to the IUPAC guidelines for determining LoD and LoQ, the concentration required to yield an SNR of 3 and 10, respectively, was used to determine the intensity corresponding to this metric. These were then used to calculate the intensity for the same concentration. The intensity plot yielded a regression fit of $y_h = 2357.8x - 1.47$.

$$[\mathrm{DMTP}] = \frac{3 + 0.431}{1244.5} \approx 0.00276 \text{ mM}$$

$$\text{intensity} = 2357.8(0.00276) - 1.47 \approx 5.03$$

According to these calculations, all values below 1.833 should not be considered observations.

The 400 MHz LoQ was calculated similarly to yield an intensity limit of 5.547. The calculated values are summarized below **(Table 2)**.

Table 2. Summary of calculated values for LoD and LoQ.

	60 MHz		**400 MHz**	
	LoD	LoQ	LoD	LoQ
SNR Fit	$y_s = 134.6x - 6.01$		$y_s = 1244.5x - 0.4314$	
Intensity Fit	$y_i = 29.73x - 0.202$		$y_i = 2357.8x - 1.469$	
SNR	3	10	3	10
[DMTP] (mM)	0.11	0.34	0.0028	0.0084
Intensity	8.49	40.2	5.03	18.3

The limit of quantification is reported here, as it is an important metric to consider for any robust method. However, it is not taken into consideration further in this report.

3.1.1 — Extraction method

The extraction method was refined over many experimental trials, where the extraction efficacy was assessed. The first and second trials utilized 3 mL of solvent per extraction in a glass culture tube. Due to the high amount of solvent that would be required using this method, and the inaccessibility of centrifuging glass culture tubes, these trials were discarded. Thus, the next trial was conducted using 1.5 mL microcentrifuge tubes. Using 1.5 mL of solvent did not allow sufficient quantities of extract for NMR analysis (0.5 mL x 3 aliquots). Next, 2.0 mL tubes

were used with 1.8 mL of solvent, allowing for sufficient aliquots of 700 µL, however, the final aliquot was insufficient. The final protocol involved taking 350 µL of solvent and diluting it in 350 µL of standardized solvent. Since the extractions were uniformly diluted, the experimental concentration is effectively one-half of the real concentration. A plot of this data reveals that one hour of extraction is sufficient for a wide representation of compounds in the sample **(Figure 4)**.

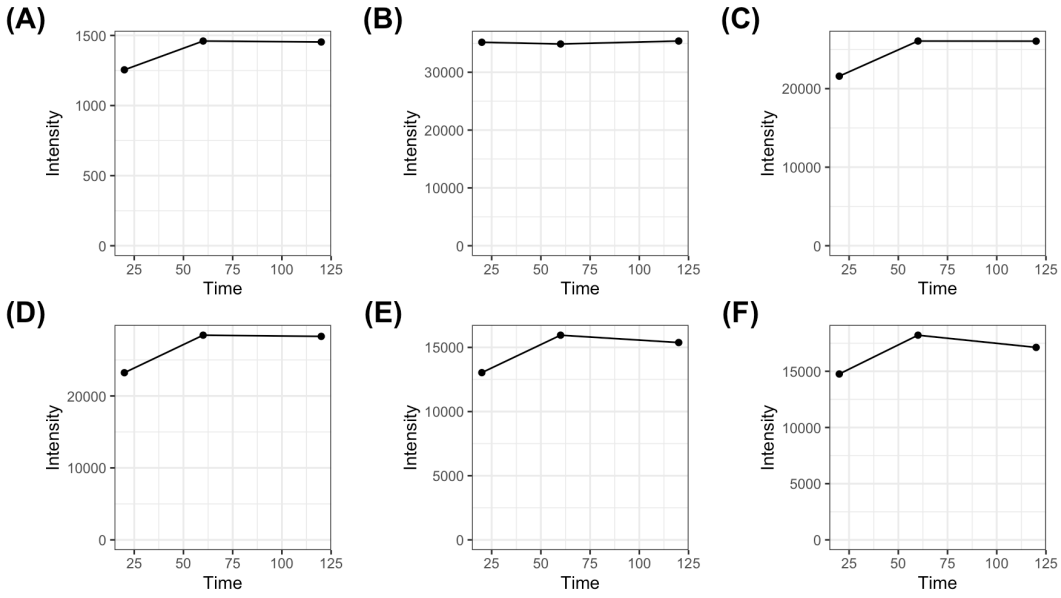

Figure 4. Scatterplots of the intensity of separate peaks at different time intervals of extraction. Chemical shift of analyzed peaks: **A**, 5.42 ppm; **B**, 2.48 ppm, **C**, 2.02 ppm; **D**, 1.61 ppm; **E**, 1.23 ppm; **F**, 0.79 ppm.

3.2 — Statistical models

3.2.1 — Principal component analysis

Principal component analysis was used for the preliminary assessment of data clustering.

Using only standard preprocessing consisting of data centring, PCA scores plots for the 400 MHz data exhibit a large amount of overlap with much of the variance explained by the first two principal components (64.779%) **(Figure 5)**. This high percentage of variance likely depends on the model, which is highly dependent on peak intensity. Without scaling, the strongest peaks are considered more often than the weakest ones.

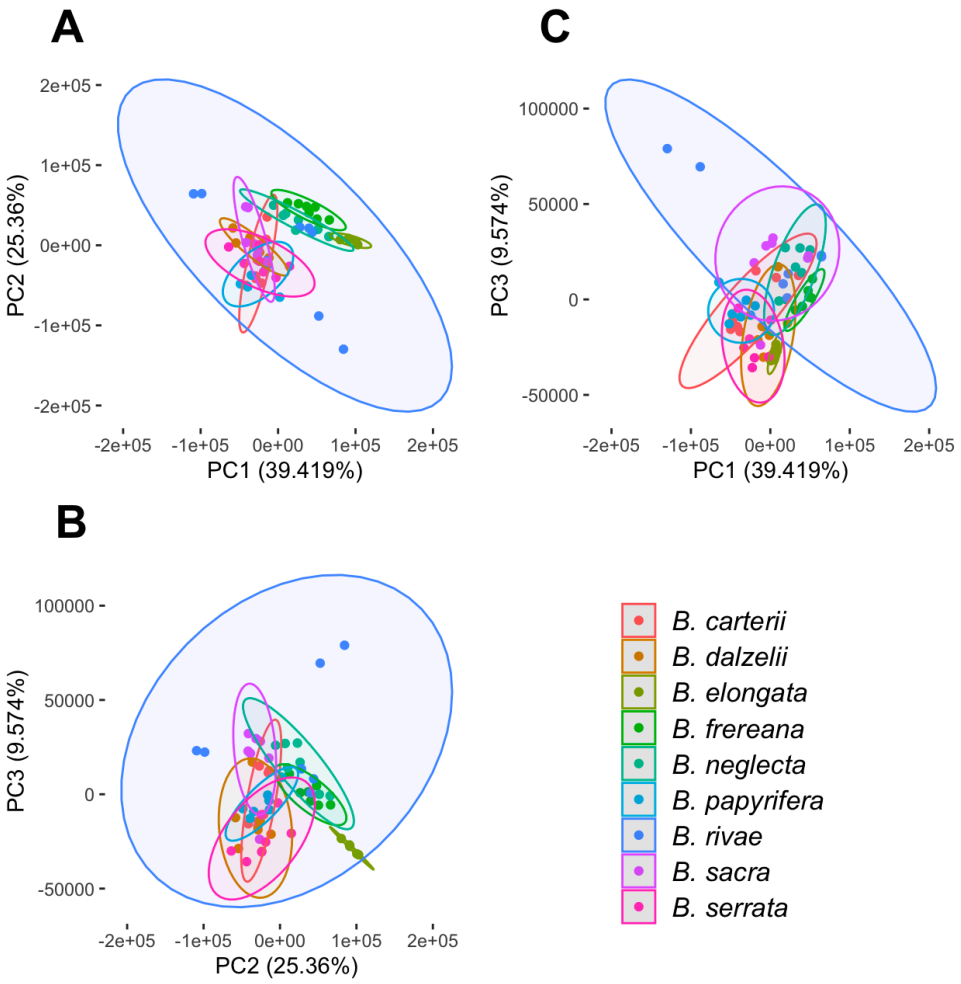

Figure 5. PCA scores plots for centred 400 MHz data. Ellipses encompass the 95% confidence interval.

Pareto scaling is a method that can be used to diminish the influence of high-intensity peaks in an NMR data set. In doing so, the statistical model becomes better at training using more subtle features than only the strongest peaks. It is performed by scaling each measurement by the square root of the standard deviation such that $I_{\text{new}} = \sqrt{\sigma} \cdot I_{\text{initial}}$ where I stands for a single value.[52]

This transformation is typically performed for NMR PLS-DA models because they are highly dependent on variations in peak intensity. The scaled data was used to generate PCA

scores plots **(Figure 6)** and **(Figure 7)**. In this model, the first two principal components did not represent most of the variance in the data, totalling 25.97%. All species exhibit sufficient clustering, with the exception of *B. rivae*, which is highly variant. Despite this, Pareto scaling and centring will likely improve the accuracy of subsequent PLS-DA models by allowing smaller but highly variant peaks to come through.[52]

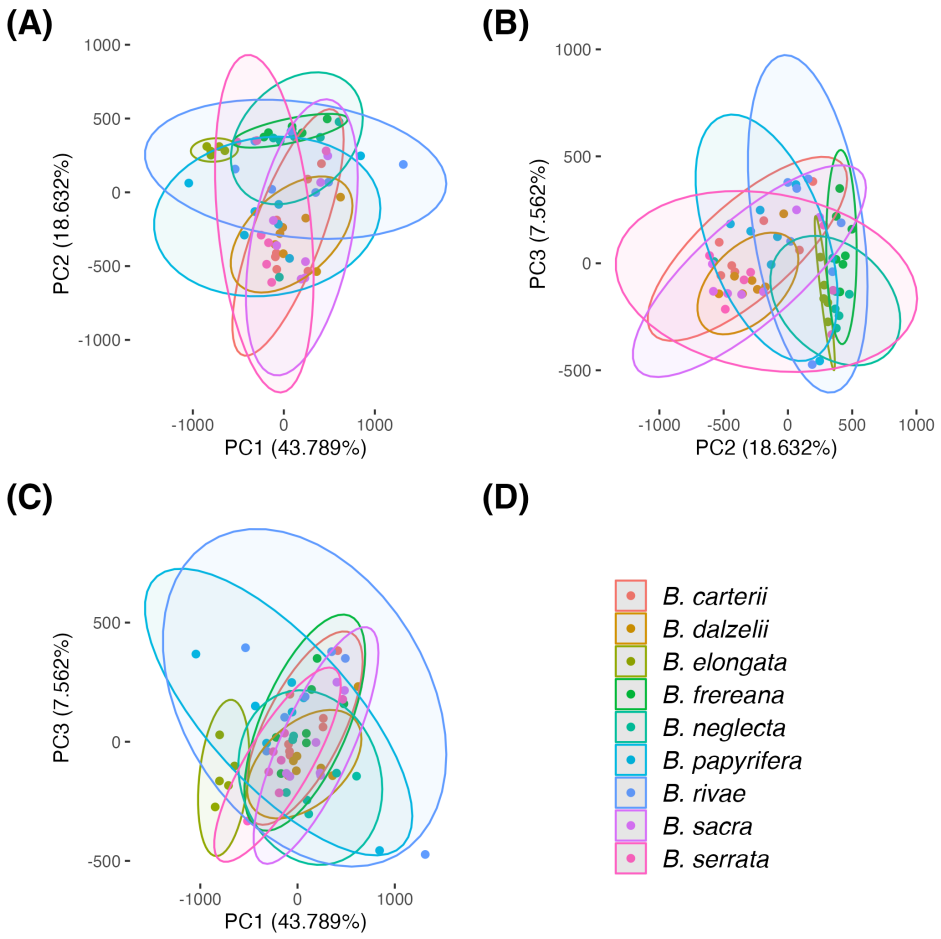

Figure 6. PCA scores plots for centered and Pareto-scaled 60 MHz data. Ellipses encompass the 95% confidence interval.

Preliminary analysis of the 60 MHz PCA plots shows that there is a high degree of overlap in the clusters. This is potentially an early sign that further classification models will have difficulty differentiating between the most important features in an unknown spectrum. This is likely due to the low resolution of the 60 MHz spectra. Since most of the spectral information is focused around the δ 0 to 3 ppm, peak overlap likely conceals smaller peaks which may vary significantly

between spectra. Possible solutions for this include sample spinning and shimming. However, the benchtop instrument used in this study did not have these capabilities.

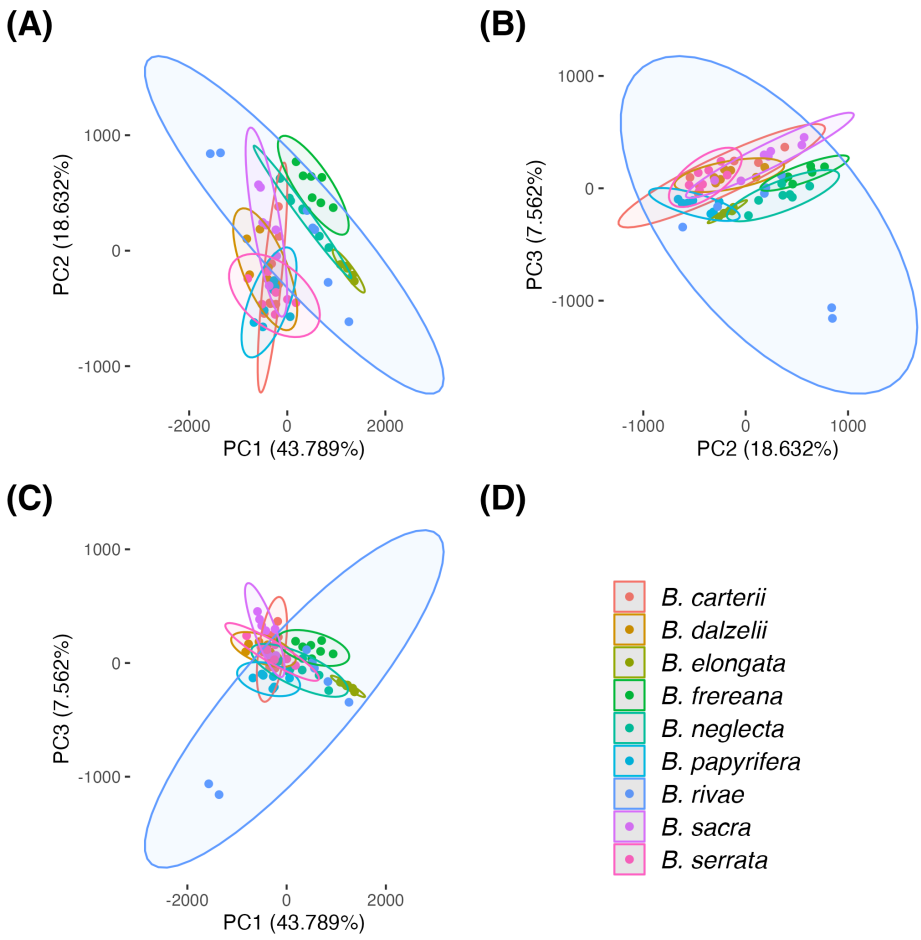

Figure 7. PCA scores plots for centered and Pareto-scaled 400 MHz data. Ellipses circumscribe the 95% confidence interval.

These 400 MHz PCA scores plots show highly discriminated clustering between species classes within the first three PCs (which represent 69.98% of the variance). This indicates a high likelihood that a subsequent PLS-DA model can classify test data correctly. This shows promise moving forward in producing prediction models.

3.2.2 — Partial least squares discriminant analysis

For this study, data with nine classes corresponding to *Boswellia* species was used to produce PLS-DA models. In turn, the model can be used to classify an "unknown" or withheld sample

as either *B. carterii*, *B. dalzelii*, *B. elongata*, *B. frereana*, *B. neglecta*, *B. papyrifera*, *B. rivae*, *B. sacra*, or *B. serrata*.

The PLS-DA model generated from 400 MHz data required consideration of 13 components to reach the optimal Cohen's Kappa value within a limit of 15 components analyzed **(Figure 8)**. The 60 MHz model required 10 components. Cohen's Kappa metric evaluates the model's accuracy while accounting for the chance that a correct classification is made randomly.[51] This metric is particularly useful for prediction models, as it considers each prediction individually.

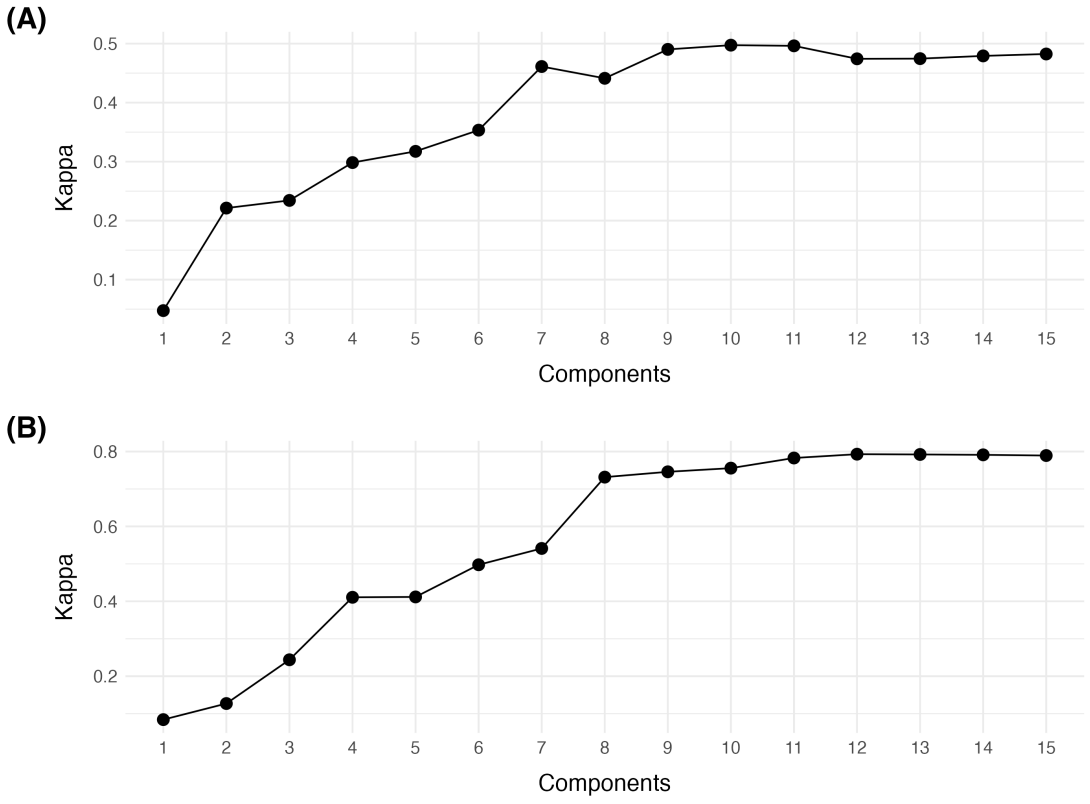

Figure 8. Plot of Cohen's Kappa value against number of components considered when training PLS-DA models. **A**: 60 MHz, **B**: 400 MHz.

As discussed previously, PLS-DA models have a high tendency to overfit data.[40] A model is likely to be overfitting when it performs well in classifying training data but fails to correctly

classify testing data with full accuracy.[13] This typically occurs for several reasons. One possible reason is the use of a high-dimensional data with many variables, such as a data set consisting of many spectral bins. This effect can be minimized by finding a binning width that does not hide potentially significant features while also reducing the dimensionality of the dataset.

Predictions using both 60 and 400 MHz data are visualized using a confusion matrix, where each cell represents the number of classifications made in that category **(Figure 9)**.

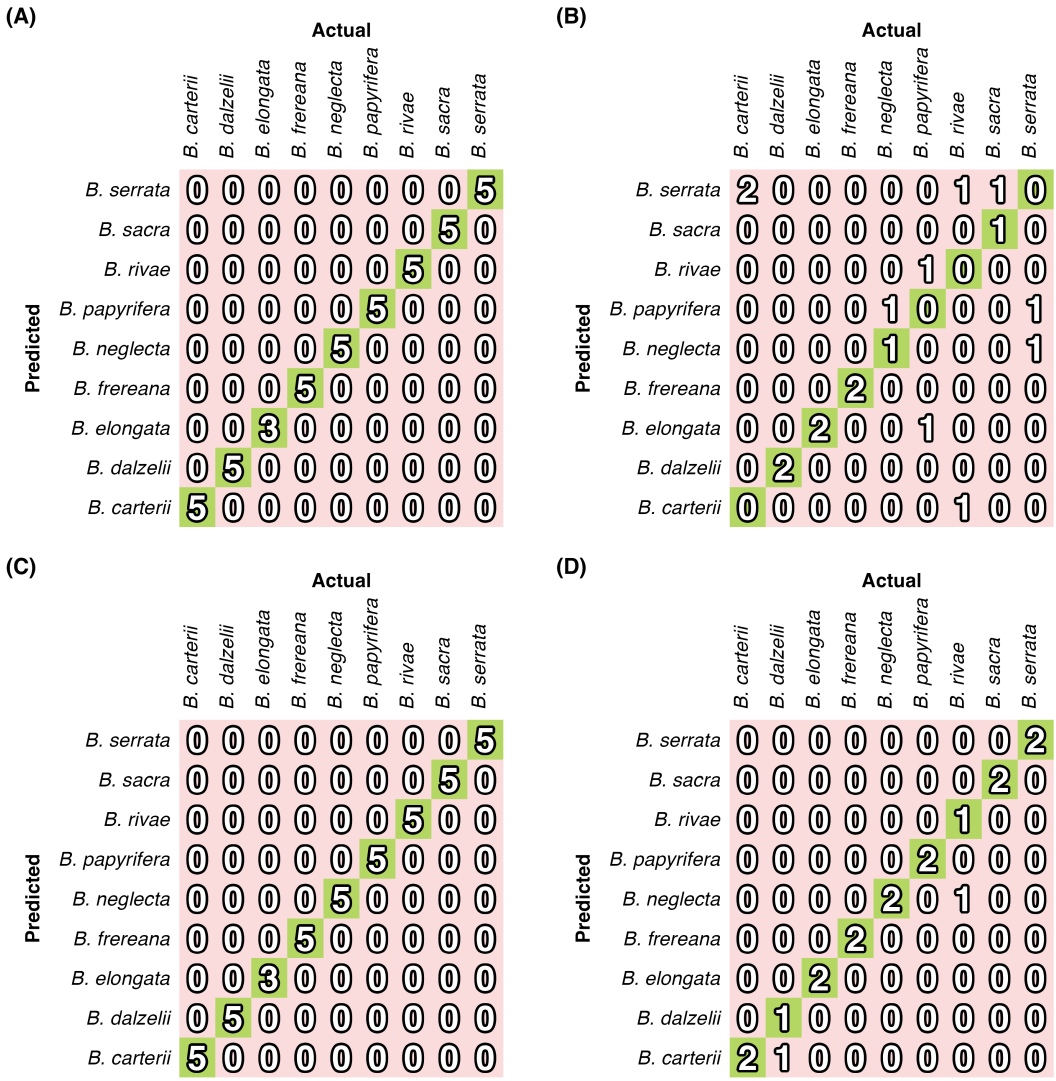

Figure 9. Performance confusion matrices of PLS-DA models generated from 60 and 400 MHz spectra. **A**: 60 MHz training-set predictions; **B**: 60 MHz test-set predictions; **C**: 400 MHz training-set predictions; **D**: 400 MHz test-set predictions. Green cells represent correct predictions.

A summary of test set prediction results is available below **(Table 3)**.

Table 3. Performance metrics for test set prediction using PLS-DA models.

	60 MHz	400 MHz
Accuracy	0.444	0.889
Cohen's Kappa	0.375	0.875
Specificity	0.976	0.980

The 60 and 400 MHz PLS-DA models showed perfect accuracy when predicting the class of training-set data as a control. However, test-set prediction results vary greatly, with particularly poor results from the 60 MHz model. The 60 MHz model had a 44.4% accuracy with 8 of 18 correct class predictions. Despite this performance, the 400 MHz model showed an accuracy of 88.9%. This shows promise for the ability of *Boswellia* species to be classified by the chemical profile of its frankincense using 400 MHz NMR data. However, overfitting was still present to a great degree in the 60 MHz models.

Profuse overfitting was avoided in the 400 by using 10-fold 10 repeats cross-validation during training, enabling the resultant model to classify over 80% correctly **(Table 3)**.

A model that reduces overfitting more strongly should be used to move forward. Thus, Random Forests models were produced from these data.

3.2.3 — Random Forests analysis

The RF training algorithm works on the premise of generating many smaller decision trees using subsets of data. Therefore, the training process aims to determine the ideal number of variables to include in each tree. The number of variables and iterations needed to select this model by optimizing Cohen's Kappa by choosing varying numbers of selected variables **(Figure 10)**.

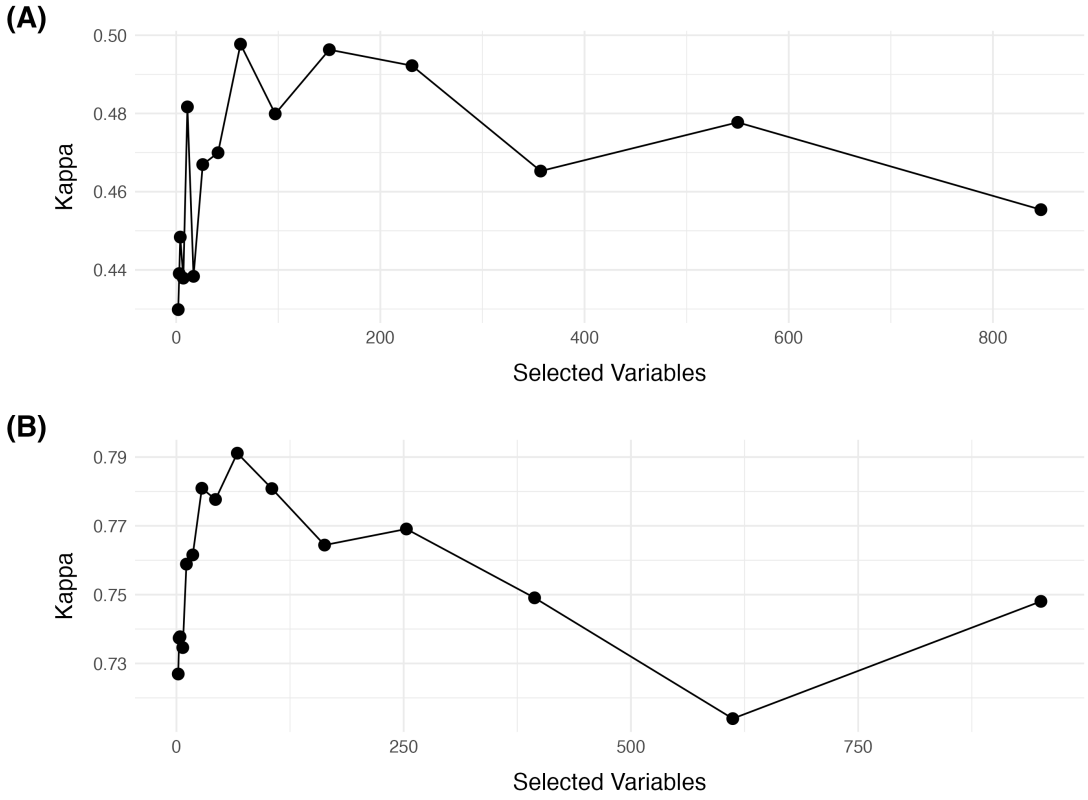

Figure 10. Plot of Cohen's Kappa value against number of components considered when training and cross-validating RF models. **A**: 60 MHz, **B**: 400 MHz.

As was done with the PLS-DA models, the ability of the RF model to predict its own training data and the testing set was assessed. The 60 MHz RF model struggled with classifying new spectra, with an accuracy of 38.9%. Just as in the PLS-DA model, this may be because important spectral features are being covered due to low spectral resolution.

RF models generated from the 400 MHz spectra demonstrated 100% accuracy when first re-predicting the species of the training set. Moreover, when applied to the test data, the same model can make accurate predictions 94.4% of the time. This indicates that there are no major signs of overfitting in this model. The metrics of the final selected model are summarized below **(Table 4)**.

Table 4. Number of variables required per decision tree and model performance metrics of generated Random Forests models for 60 and 400 MHz spectra.

	60 MHz	**400 MHz**
Variables	63	67
Iterations	9	9
Prediction Accuracy	0.389	0.944
Cohen's Kappa	0.313	0.938
Specificity	0.949	0.976

Predictions were made by applying training and test data to the RF model are also represented by a confusion matrix **(Figure 11)**.

As demonstrated before with the PLS-DA model, the 400 MHz model had significantly higher accuracy and specificity measurements. The 60 MHz model classified 7 of 18 predictions correctly, and the 400 MHz model correctly classified 17 of 18 predictions. Based on the number of correct predictions and taking into consideration the Cohen's Kappa score of 1 for the RF model, Random Forests is a promising algorithm to use for the classification of unknown *Boswellia* samples.

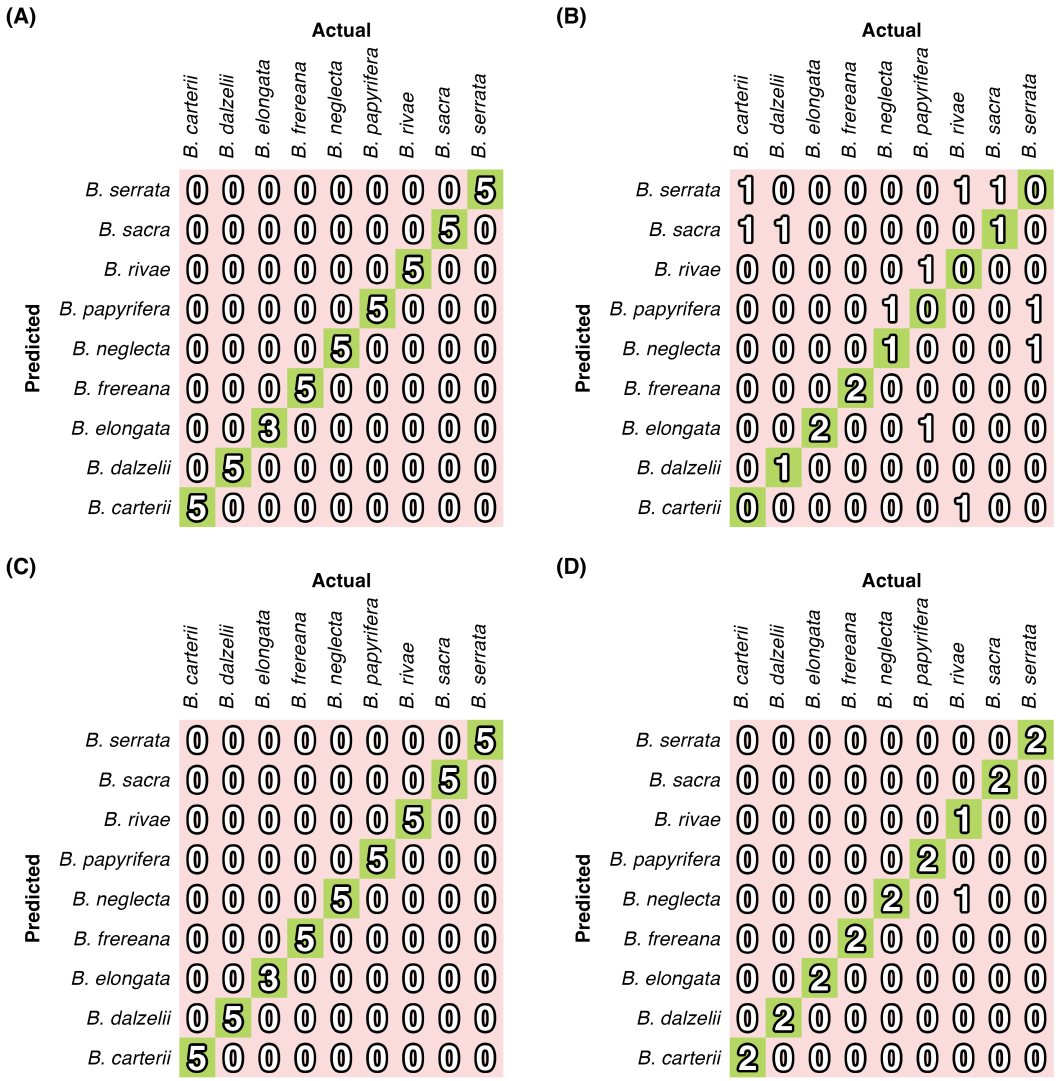

Figure 11. Performance confusion matrices of RF models generated from 60 and 400 MHz spectra. **A**: 60 MHz training-set predictions; **B**: 60 MHz test-set predictions; **C**: 400 MHz training-set predictions; **D**: 400 MHz test-set predictions. Green cells represent correct predictions.

3.3 — VIP analysis

VIP analysis can aid in selecting compounds to elucidate further and determine which compounds give rise to variance in the data set. A summary of VIPs found by the RF models is provided below. In the extension of this study, an HSQC-TOCSY NMR experiment can map cross-peaks found at the location of these VIP spectral bins to determine the structure of the compound that gives rise to the peak.

Table 5. Summary of VIP spectral bins and % importance (imp.) from 60 and 400 MHz RF models.

60 MHz		400 MHz	
ppm	**% imp.**	**ppm**	**% imp.**
5.11	100.00	7.13	100.00
5.17	97.33	7.09	87.40
5.15	95.37	3.99	77.47
2.17	94.53	6.84	76.82
5.12	94.38	3.72	76.55
5.36	91.35	4.17	74.68
1.36	87.42	3.97	71.46
5.16	84.83	3.97	70.23
5.10	77.45	4.27	67.32
5.13	76.92	6.93	65.53

3.4 — Comparison between 60 and 400 MHz models

When comparing the accuracy between 60 and 400 MHz models, it is clear that the higher-strength magnet performs better when classifying test data. The 60 MHz PLS-DA model performed with less than half the accuracy of its counterpart. Moreover, the 60 MHz RF model performed with an accuracy of just over 40% of the 400 MHz model. The gap between these two instruments is quite clear. The 400 MHz NMR is viable for creating a predictive model sufficient to extend research into this topic.

It can be argued, however, that the cost difference between the instruments justifies its potentially high throughput analysis of frankincense samples. It is possible that the inability

of the 60 MHz model to make strong predictions arises from the number of spectra used for training. With more spectra, the confidence of the model would increase. Moreover, various different types of preprocessing could be used for the spectra, in addition to or instead of Pareto scaling and centring.

4 — Conclusion

NMR spectra collected from frankincense originating from 9 *Boswellia* species were used to build chemometric models for classifying unknown spectra by predictions. Different NMR field strengths were compared, finding that 400 MHz data is more valuable for constructing these models. Despite this, it may be possible to increase the confidence of both models using a larger number of spectra in the training set.

This study demonstrated that it is possible to predict the species of origin for a *Boswellia* frankincense sample using statistical models; in particular, partial least squares discriminant analysis and Random Forests algorithms. Models produced from 400 MHz NMR performed well, with high accuracy and Cohen's Kappa values. Of the two models trained by 400 MHz data, the Random Forests model performed the best, boasting the highest accuracy in predicting species of origin for the withheld test set spectra.

4.1 — Future directions

Using the preliminary findings from this study, models trained from larger datasets should also be tested. Scaling up these models lends to a higher statistical power for that model, which would enable the end-user to be more confident in the accuracy of the predictions. Moreover, it is also important to consider the precision of these models by conducting trials in which a model is trained on a larger data set numerous times and assessing the ability of the model to predict the same result each time.

Moreover, HSQC-TOCSY NMR experiments can be used to map the chemical structure of VIP compounds to species to determine which compounds are particularly unique to certain

species. This will further help produce a metric by which the chemotaxonomy of the *Boswellia* genus can abide.

References

[1] S. J. Ivory, K. L. Cole, R. S. Anderson, A. Anderson, and J. McCorriston, “Human landscape modification and expansion of tropical woodland in southern Arabia during the mid-Holocene from rock hyrax middens,” *Journal of Biogeography*, vol. 48, no. 10, pp. 2588–2603, 2021, doi: 10.1111/jbi.14226.

[2] S. Johnson, A. Abdikadir, P. Satyal, A. Poudel, and W. N. Setzer, “Conservation Assessment and Chemistry of Boswellia ogadensis, a Critically Endangered Frankincense Tree,” *Plants*, vol. 11, no. 23, p. 3381–3382, Jan. 2022, doi: 10.3390/plants11233381.

[3] A. DeCarlo, S. Agieb, S. Johnson, P. Satyal, and W. N. Setzer, “Inter-Tree Variation in the Chemical Composition of Boswellia papyrifera Oleo-Gum-Resin,” *NATURAL PRODUCT COMMUNICATIONS*, vol. 17, no. 7, p. 1934578–1934579, Jul. 2022, doi: 10.1177/1934578X221117411.

[4] N. G. Tošić, V. D. Nikolić, V. M. Miljković, and L. B. Nikolić, “'Boswellia serrata' resin isolates: Chemical composition and pharmacological activities,” *Advanced Technologies*, vol. 11, no. 1, pp. 76–87, 2022, doi: 10.5937/savteh2201076T.

[5] M. Z. Siddiqui, “Boswellia Serrata, A Potential Antiinflammatory Agent: An Overview,” *Indian Journal of Pharmaceutical Sciences*, vol. 73, no. 3, pp. 255–261, 2011, doi: 10.4103/0250-474X.93507.

[6] A. Al-Harrasi, A. L. Khan, S. Asaf, and A. Al-Rawahi, “Taxonomy, Distribution and Ecology of Boswellia,” *Biology of Genus Boswellia*. Springer International Publishing, Cham, pp. 11–34, 2019. doi: 10.1007/978-3-030-16725-7_2.

[7] P. Sabo, A. Ouedraogo, D. S. J. C. Gbemavo, K. V. Salako, and R. G. Kakai, “Land use impacts on Boswellia dalzielii Hutch., an African frankincense tree in Burkina Faso,” *BOIS ET FORETS DES TROPIQUES*, no. 349, pp. 51–63, 2021, doi: 10.19182/bft2021.349.a31960.

[8] T. Morikawa, H. Matsuda, and M. Yoshikawa, “A Review of Anti-inflammatory Terpenoids from the Incense Gum Resins Frankincense and Myrrh,” *Journal of Oleo Science*, vol. 66, no. 8, pp. 805–814, 2017, doi: 10.5650/jos.ess16149.

[9] T. Morikawa, H. Oominami, H. Matsuda, and M. Yoshikawa, “New terpenoids, olibanumols D-G, from traditional Egyptian medicine olibanum, the gum-resin of Boswellia carterii,” *Journal of Natural Medicines*, vol. 65, no. 1, pp. 129–134, Jan. 2011, doi: 10.1007/s11418-010-0472-z.

[10] A. Al-Harrasi, R. Csuk, A. Khan, and J. Hussain, “Distribution of the anti-inflammatory and anti-depressant compounds: Incensole and incensole acetate in genus Boswellia,” *Phytochemistry*, vol. 161, pp. 28–40, May 2019, doi: 10.1016/j.phytochem.2019.01.007.

[11] K. Massei, T. Michel, G. I. Obersat, A. Al-Harrasi, and N. Baldovini, “Phytochemical study of Boswellia dalzielii oleo-gum resin and evaluation of its biological properties,”

Phytochemistry, vol. 213, p. 113751–113752, Sep. 2023, doi: 10.1016/j.phytochem.2023.113751.

[12] M. Thulin, "IUCN Red List of Threatened Species: Boswellia sacra," *IUCN Red List of Threatened Species*, Jan. 1998, doi: 10.2305/IUCN.UK.1998.RLTS.T34533A9874201.en.

[13] T. Head, R. Giebelhaus, S. L. Nam, A. P. De La Mata, J. Harynuk, and P. Shipley, "Discriminating Extra Virgin Olive Oils from Common Edible Oils: Comparable Performance of PLS-DA Models Trained on Low-Field and High-Field 1H NMR Data," Aug. 2023. doi: 10.26434/chemrxiv-2023-kt0sx.

[14] WFO, "Boswellia Roxb.." Accessed: Apr. 04, 2024. [Online]. Available: https://www.worldfloraonline.org/taxon/wfo-4000005070

[15] Plants of the World Online Kew, "POWO, ." Accessed: Apr. 03, 2024. [Online]. Available: https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:5117-1#source-KB

[16] S. Lvončík and R. Řepka, "Boswellia socotrana: One or Two Taxa?," *Novon: A Journal for Botanical Nomenclature*, vol. 28, no. 1, pp. 17–23, Feb. 2020, doi: 10.3417/2019427.

[17] M. Paul, G. Brüning, J. Bergmann, and J. Jauch, "A Thin-layer Chromatography Method for the Identification of Three Different Olibanum Resins (Boswellia serrata, Boswellia papyrifera and Boswellia carterii, respectively, Boswellia sacra)," *Phytochemical Analysis*, vol. 23, no. 2, pp. 184–189, 2012, doi: 10.1002/pca.1341.

[18] Z.-C. Chen, F.-Y. Chen, L.-L. Xu, B. Yang, and Y.-M. Luo, "Characteristic constituents with chemotaxonomic significance from the gum resin of Boswellia carterii," *Biochemical Systematics and Ecology*, vol. 104, p. 104478–104479, Oct. 2022, doi: 10.1016/j.bse.2022.104478.

[19] A. Sharma, J. Upadhyay, A. Jain, M. Kharya, A. Namdeo, and K. Mahadik, "Antioxidant activity of aqueous extract of Boswellia serrata," *J Chem Bio Phy Sci*, vol. 1, pp. 60–71, 2011.

[20] M. Miran *et al.*, "Taxonomical Investigation, Chemical Composition, Traditional Use in Medicine, and Pharmacological Activities of Boswellia sacra Flueck," *Evidence-Based Complementary and Alternative Medicine*, vol. 2022, p. e8779676, Feb. 2022, doi: 10.1155/2022/8779676.

[21] F. Iram, S. A. Khan, and A. Husain, "Phytochemistry and potential therapeutic actions of Boswellic acids: A mini-review," *Asian Pacific Journal of Tropical Biomedicine*, vol. 7, no. 6, pp. 513–523, Jun. 2017, doi: 10.1016/j.apjtb.2017.05.001.

[22] P. M. Dewick, *Medicinal natural products: a biosynthetic approach*, 3rd edition. Chichester, West Sussex, United Kingdom: Wiley, A John Wiley, Sons, Ltd., Publication, 2009.

[23] M. A. Ayub, M. A. Hanif, J. Blanchfield, M. Zubair, M. A. Abid, and M. T. Saleh, "Chemical composition and antimicrobial activity of Boswellia serrata oleo-gum-resin essential oil extracted by superheated steam," *Natural Product Research*, vol. 37, no. 14, pp. 2451–2456, Jul. 2023, doi: 10.1080/14786419.2022.2044327.

[24] K. Pilkington and G. J. Pilkington, “Boswellia: Systematically scoping the in vitro, in vivo and clinical research,” *European Journal of Integrative Medicine*, vol. 56, p. 102197–102198, Dec. 2022, doi: 10.1016/j.eujim.2022.102197.

[25] D. Marion, “An Introduction to Biological NMR Spectroscopy,” *Molecular & Cellular Proteomics : MCP*, vol. 12, no. 11, pp. 3006–3025, Nov. 2013, doi: 10.1074/mcp.O113.030239.

[26] J. L. Ward, J. M. Baker, and M. H. Beale, “Recent applications of NMR spectroscopy in plant metabolomics,” *The FEBS journal*, vol. 274, no. 5, pp. 1126–1131, 2007.

[27] N. D. Yuliana, A. Khatib, Y. H. Choi, and R. Verpoorte, “Metabolomics for bioactivity assessment of natural products,” *Phytotherapy Research*, vol. 25, no. 2, pp. 157–169, Feb. 2011, doi: 10.1002/ptr.3258.

[28] H. K. Kim, Y. H. Choi, and R. Verpoorte, “NMR-based metabolomic analysis of plants,” *Nature Protocols*, vol. 5, no. 3, pp. 536–549, Mar. 2010, doi: 10.1038/nprot.2009.237.

[29] J. L. Markley *et al.*, “The future of NMR-based metabolomics,” *Current Opinion in Biotechnology*, vol. 43, pp. 34–40, Feb. 2017, doi: 10.1016/j.copbio.2016.08.001.

[30] S. Baskal, A. Bollenbach, and D. Tsikas, “Two-Step Derivatization of Amino Acids for Stable-Isotope Dilution GC–MS Analysis: Long-Term Stability of Methyl Ester-Pentafluoropropionic Derivatives in Toluene Extracts,” *Molecules*, vol. 26, no. 6, p. 1726–1727, Mar. 2021, doi: 10.3390/molecules26061726.

[31] P. Krumpochova *et al.*, “Amino acid analysis using chromatography–mass spectrometry: An inter platform comparison study,” *Journal of Pharmaceutical and Biomedical Analysis*, vol. 114, pp. 398–407, Oct. 2015, doi: 10.1016/j.jpba.2015.06.001.

[32] A. Rutz *et al.*, “The LOTUS initiative for open knowledge management in natural products research,” *eLife*, vol. 11, p. e70780, May 2022, doi: 10.7554/eLife.70780.

[33] H. Horai *et al.*, “MassBank: a public repository for sharing mass spectral data for life sciences,” *Journal of Mass Spectrometry*, vol. 45, no. 7, pp. 703–714, 2010, doi: 10.1002/jms.1777.

[34] D. S. Wishart *et al.*, “HMDB: the Human Metabolome Database,” *Nucleic Acids Research*, vol. 35, no. suppl_1, p. D521–D526, Jan. 2007, doi: 10.1093/nar/gkl923.

[35] D. Kumar, “Nuclear Magnetic Resonance (NMR) Spectroscopy For Metabolic Profiling of Medicinal Plants and Their Products,” *Critical Reviews in Analytical Chemistry*, vol. 46, no. 5, pp. 400–412, Sep. 2016, doi: 10.1080/10408347.2015.1106932.

[36] L. W. Sumner *et al.*, “Proposed minimum reporting standards for chemical analysis,” *Metabolomics*, vol. 3, no. 3, pp. 211–221, Sep. 2007, doi: 10.1007/s11306-007-0082-2.

[37] L. Wulandari, R. Idroes, T. R. Noviandy, and G. Indrayanto, “Chapter Six - Application of chemometrics using direct spectroscopic methods as a QC tool in pharmaceutical industry and their validation,” *Profiles of Drug Substances, Excipients and Related Methodology*, vol. 47. Academic Press, pp. 327–379, Jan. 2022. doi: 10.1016/bs.podrm.2021.10.006.

[38] C. E. Turi, J. Finley, P. R. Shipley, S. J. Murch, and P. N. Brown, “Metabolomics for Phytochemical Discovery: Development of Statistical Approaches Using a Cranberry Model System,” *Journal of Natural Products*, vol. 78, no. 4, pp. 953–966, Apr. 2015, doi: 10.1021/np500667z.

[39] J. A. Lund, P. N. Brown, and P. R. Shipley, “Differentiation of Crataegus spp. guided by nuclear magnetic resonance spectrometry with chemometric analyses,” *Phytochemistry*, vol. 141, pp. 11–19, Sep. 2017, doi: 10.1016/j.phytochem.2017.05.003.

[40] D. Ruiz-Perez, H. Guan, P. Madhivanan, K. Mathee, and G. Narasimhan, “So you think you can PLS-DA?,” *BMC Bioinformatics*, vol. 21, no. 1, p. 2–3, Dec. 2020, doi: 10.1186/s12859-019-3310-7.

[41] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D’Enza, A. Markos, and E. Tuzhilina, “Principal component analysis,” *Nature Reviews Methods Primers*, vol. 2, no. 1, pp. 1–21, Dec. 2022, doi: 10.1038/s43586-022-00184-w.

[42] H. Abdi and L. J. Williams, “Principal component analysis,” *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010, doi: 10.1002/wics.101.

[43] B.-H. Mevik, R. Wehrens, and K. H. Liland, “pls: Partial least squares and principal component regression,” *R package version*, vol. 2, no. 3, 2011, [Online]. Available: https://cran.r-project.org/package=pls

[44] G. Louppe, “Understanding Random Forests: From Theory to Practice,” 2014, doi: 10.48550/ARXIV.1407.7502.

[45] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org/

[46] N. Gift, I. C. Gormley, and L. Brennan, “MetabolAnalyze: probabilistic principal components analysis for metabolomic data.” 2010.

[47] R Core Team, “R: A Language and Environment for Statistical Computing.” 2021. [Online]. Available: https://www.r-project.org/

[48] Y. Tang, M. Horikoshi, and W. Li, “ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages,” *The R Journal*, vol. 8, no. 2, pp. 474–485, 2016, doi: 10.32614/RJ-2016-060.

[49] Kuhn and Max, “Building Predictive Models in R Using the caret Package,” *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008, doi: 10.18637/jss.v028.i05.

[50] A. Liaw, M. Wiener, and others, “Classification and regression by randomForest,” *R news*, vol. 2, no. 3, pp. 18–22, 2002, [Online]. Available: https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

[51] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[52] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, “Centering, scaling, and transformations: improving the biological information

content of metabolomics data," *BMC Genomics*, vol. 7, p. 142–143, Jun. 2006, doi: 10.1186/1471-2164-7-142.