Untargeted Flavoromics to Identify Flavor Active Compounds

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Ian Guiles Ronningen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Devin G. Peterson

March 2016

Acknowledgements

I would like to acknowledge the members of the flavor research and education center which have long provided funding for the development of this work. Working in this this academic-industry partnership was incredibly valuable. The feedback, reality checks, conversations and brainstorming was wildly valuable to my development as a scientist and as a person. And to all those who helped directly and indirectly, your time and thoughts were very valuable.

Dedication

To those who spent their valuable time, dedication and energy towards my education, encouragement and inspiration, I thank you.

Summary

Flavor is a multi-model sensation consisting of a complex set of chemical stimuli that are perceived as a mixture, and has largely been characterized using targeted techniques. While targeted techniques have historically provided knowledge, it is not without limitation. Increasing the number of analytes allows for more comprehensive understanding of food systems which can lead to new discoveries. The goal of this dissertation is to develop untargeted analytical methods to identify flavor active materials in citrus extracts that relate to aging. Increased characterization of flavor systems provides a foundation to better understand complex chemistry and can lead to new insight. This dissertation illustrates how an untargeted workflow was developed to identify age related compounds, and establish their sensory significance through recombination modeling and structural elucidation.

In this dissertation preprocessing methods were optimized using an uneven multi-level two factor design of experiment and further untargeted modeling. Signal to noise ratio and thresholding were varied during pre-processing which produced data sets varying in size from >50,000 features to 500. The produced data sets were further investigated by unsupervised multivariate approaches, looking at model fit and data utilization. Part per billion differentiation was achieved using both unsupervised (principle component analysis) and supervised (projection to latent structures) multivariate modeling, indicating a high quality analytical and statistical framework.

To understand how a food system ages ethanol extracts from different citrus varietals (Navel, Mineola, and Valencia) were aged and chemically profiled using Ultra

Performance Liquid Chromatography Mass Spectrometry. Machine learning (Random forest) and multivariate (projection to latent structures) models were implemented to model the age of the extracts. Varietal difference was leveraged and models were adapted to understand the age of the samples, rather than model the varietal differences. Statistically important compounds were isolated using food grade mass spectrometry directed fractionation. These isolated compounds underwent sensory evaluation using descriptive analysis in both a Solvent Assisted Flavor Evaporation (SAFE) extract of orange juice and a volatile orange flavor (VOF) model beverage. The isolated compounds showed significant impact in both tasting mediums. All compounds identified increased with sample age, and when evaluated in the SAFE extract showed significant decrease in orange character. Compound 413 showed a significant increase in cooked character and green bean character and a suppression of floral character. While compound 383 showed a significant increase in green bean character. In the volatile orange flavor (VOF) compound 413E2 and compound 457 showed an increase in sweetness over the control, which was the only noted change to the 'taste' attributes noted among the compounds isolated. Compound 383 showed a decrease in cooked character over the control. Compound 661 showed suppression of floral aroma over the sample blank. Compounds that positively correlated with age were reported to increase the cooked and green bean character and suppression of floral and orange character notes, which indicate degradation in flavor quality, or a deviation from fresh character. Structural elucidation using Nuclear Magnetic Resonance (^{1}H, HMBC, TOCSY, HSQC) and accurate mass (TOF) revealed compound 693 was Nomilin 17-O-beta-D-

glucopyranoside, while compounds 383 and 661 were shown to be novel compounds. The systematic name of compound 661 was (5-(((2R,3S,4R,5R)-4,5-dihydroxy-3-((3-hydroxy-3-methyl-5-oxohexanoyl)oxy)-6-(4-(3-hydroxypropyl)-2,6-dimethoxyphenoxy)tetrahydro-2H-pyran-2-yl)methoxy)-3,3-dimethyl-5-oxopentanoic acid. There are two 3-Hydroxy-3-methylglutaric acid units and a dihydrosinapyl alcohol moiety bonded to a sugar backbone. Compound 383 was identified as (3,5,5-trimethyl-4-((E)-3-oxo-4-(((2S,3S,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)oxy)but-1-en-1-yl)cyclohex-2-en-1-one) and suggested reaction product of a sugar and terpene moiety.

The final piece of this research was to characterize aging in lemon extracts and identify contextual variable interactions associated statistically significant compounds. This chapter (chapter 5) works to generate more value from untargeted analysis, which are historically outcome sparse. In order to identify contextual interactions a machine learning workflow was optimized to model age. Six types of lemon extracts were aged and profiled using Ultra Performance Liquid Chromatography Mass Spectrometry. After preprocessing and data filtering the data was modeled using 9 different machine learning approaches. Random Forest produced the highest quality initial model, which went through further development and tuning. Within the random forest model number of trees, features tried at each node, max number of terminal samples were all optimized, as was the final model using both gini and entropy for decision criteria. The final model had a training fit of 0.951 and test score of 0.928(+/- 0.0049). Statistically important variables from this analysis were investigated for contextual data interactions that would

illuminate data trends for further study. This approach aims to help provide additional value from untargeted flavoromics and better understand contextual interactions in data sets.

In summary this dissertation illustrated the value of applying data driven methods to understand flavor and further validation of identified compounds through recombination sensory panel testing. As there is a wealth of information mined from chemical data sets, the ability to demonstrate how to filter flavor active information provided a proof of concept for the utilization of untargeted methods for flavor discovery.

Table of Contents

List of Tables

List of Figures

Chapter 1: Research Introduction

The first chapter is written to provide an overview of previous research utilized in this dissertation, including: flavor chemistry, citrus flavor, analytical chromatography, multivariate statistics, data science and machine learning. Following an introduction is the research hypothesis, an experimental outline, discussion on the novelty of research, and areas of limitations and areas for expansion of this research. Additional chapters expand on specific literature surrounding this dissertation, optimization of data extraction, identification of age related flavor compounds, and machine learning to identify data interactions. This thesis aims to understand how all of these fields work together to identify novel flavor information.

The world of flavor research continuously works to identify chemical drivers for flavor perception with an emphasis on aroma, taste, and trigeminal compounds. Identifying which chemical species directly contribute to flavor perception provides food manufacturers a better understanding on how to produce more palatable food through encouraging desirable flavor attributes and reducing undesirable attributes. As more information is learned about the drivers of food flavor, the more dynamic this research becomes. The historic emphasis on aroma and the basic senses has now evolved into areas understanding mouth feel, modulation, and chemesthetic senses to name a few. In order to manage the complexity of food, reductionist approaches are frequently employed. To reduce the complexity of food systems it is very common for food processors to identify a few “quality” related compounds. These compounds ideally help

food processors infer whether a food product is acceptable or not. These compounds are not explicitly flavor active as they may act as a proxy for sensory outcomes (e.g. Appearance). When product or process variation takes place, like reformulation or deviation in processing, understanding how these "quality" compounds change help food processors predict product performance in the market place. Knowledge around chemical drivers of flavor and how they relate to product change or processing lead to improved consumer experience. By only selecting a few compounds to understand food chemistry a large amount of valuable information is overlooked. Identifying new chemical species that impact flavor helps food processors produce food with improved sensory attributes. A more complete picture on how variables in food lead to changed flavor perception is important, but has historically been a challenge due to the complexity of food.

To more effectively monitor the diverse chemical classes that make up a foods flavor sensitive and holistic methods are desirable. Since flavor is not perceived as single compounds individually out of context, methods that are more comprehensive better express food chemistry. The depth of small molecules in food can range from hundreds to thousands and so limiting analytical focus can exclude useful information. Traditionally methods of flavor analysis reduce the complexity of a food to better enable researchers to establish causative relationships. These reductionist methods help to manage the complexity, but ignore a massive amount of chemical information.

Model systems are a common reductionist method to understand how chemical constituents impact food flavor. These model systems often involve flavor reconstitution but ignore other parts of food flavor such as non-volatiles. Model systems are highly

utilized to understand specific roles of reactants on character in flavor research, but lack translation as models limit complexity. The chemical complexities of recombination models rarely reach the complexity of food systems and often use only a small number of flavor compounds (Schieberle, 2001). For reaction models that aim to understand how flavor precursors react to form flavor added chemical complexity leads to potential side reactions. Simple models may be used to avoid side reactions but can lack translation to real food systems. Additionally, analytical flavor profiling and quantitation have been utilized to reconstitute the volatile flavor fingerprints and taste profiles in an attempt to recreate the most impactful constituents but often do not recreate the targeted character. The challenge is how to reconstruct a flavor profile comprised of tens to hundreds of chemical species in a medium that will reproduce the same sensory response. These approaches both try to reduce the chemical complexity of food; model systems try to focus on identifying which precursors are drive flavor formation, and aroma and taste reconstitution try to recreate analytical observations by removing compounds that researchers do not consider to be flavor active. Each of these research methodologies uses a reductionist approach to simplify the chemical complexity allowing researchers to establish more causative relationships. While allowing for progress to be made the limited scope ignores significant amounts of information.

The limitations of analytical methods were recognized and new approaches to relate analytical signal with sensory results using statistics was present in the literature even in the 1970s (Persson, 1973). These simplistic statistical models hoped to address a small amount ($n<20$) of volatile compounds and then fit the analytical data to a specific sensory

profile. These approaches were limited by the number of variables evaluated, painstaking time needed to model data, and modeling approaches available. A historic challenge with multivariate methods is that without commercially available packages and powerful computing stations the calculations were very laborious, leaving the scope of investigation dramatically smaller than what is now possible. Over the years there have been numerous applications of statistical modeling in flavor chemistry. It was not until the implementation of higher power computation and the advent of chemometrics that the complexity of data began to reach levels that could investigate hundreds to thousands of compounds and adequately address the complexity of food. Further attempts to fit analytical data to sensory data led to the application of chemometrics research in the 1990s, where more powerful computers and modeling approaches (principle component regression, multi linear regression and projection to latent structures regression) could better handle complex data (77 variables) (Togari, 1995). These studies attempted to correlate sensory data with analytical data to better identify drivers of flavor. Relating instrumental to sensory results is highly desirable, as instrumentation has increased throughput compared to sensory panels at significantly reduced cost. At the same time instruments produce data with less variation than human beings. Proponents of these approaches tout how models can replace human evaluation, and can lead to the prediction of consumer liking and acceptability by instrumental evaluation. However these approaches have largely been unsuccessful, likely due to the multi-modal nature of flavor, the chemical complexity and lack of methods to adequately characterize flavor stimuli and discern differences in sensory attributes (e.g. numerous compounds within a

food could be “fruity” or “bitter”). These methods also suffer from not maintaining context of the food during analysis and eliminating the impact of flavor delivery.

A fundamental limitation of targeted methods (e.g. GC-Olfactometry, LC-Taste, Sensomics) is the evaluation of flavor compounds singularly and not as a complex mixture as they are perceived by people. Sections or individual compounds are evaluated while ignoring the rest of the food components, limiting observations and conclusions. Targeted methods emphasize direct flavor activity and ignore the potential for flavor modulation or other complex perception relationships. Therefore, targeted methods lack the ability to identify flavor materials that are not directly flavor active, which are of great interest to industry. While targeted methods have long made advancements to flavor research, there is a wealth of perceptual mechanisms that get ignored. Targeted methods are poorly suited to investigate these flavor aspects, while untargeted methods have the ability to investigate a more comprehensive chemical dataset and capture the chemistry behind these complex flavor attributes. The development of untargeted methods therefore provides the opportunity to identify novel flavor materials by investigating chemical species in food that are historically overlooked. Discovery of novel compounds provides new information about flavor profiles and leads to better understanding of food chemistry.

Untargeted methods are effective in identifying chemical changes in complex systems and are well suited for understanding how food ages (Madrera, 2003). Studying flavor chemistry as food ages holds challenges since having enough sampling points to get good reaction data is convoluted without extensive pre-experimentation. These

studies often model a small number of drivers and rely heavily on sensory studies (Hough, 2003). One of the foundations of untargeted methods aims to remove biases associated with assumptions made during sampling and experimentation (Kell, 2004). This coupled with the reality of financial and time cost associated with more intensive sampling plans can make early work challenging in the untargeted space. However, being able to understand food aging instrumentally would largely increase the ability of food processors to understand how their food platform changes. Better understanding process variances and aging helps maintain flavor quality longer. Instrumental analysis is more rapid than sensory panels at profiling food, while less variable. Additionally, analytical methods are able to see minute chemistry that sensory panels would not. Combining powerful analytical techniques with comprehensive methodology, new flavor attributes are illuminated which may normally be overlooked in traditional targeted work. Combination of traditional targeted methods and untargeted comprehensive methods would allow for a more insightful and translatable understanding of food systems. By coupling known flavor attributes with data science approaches can lead to an understanding of how flavor compounds are modulated or impacted by other chemical species. The hypothesis at hand is: *Can multivariate methodology and data science identify flavor active chemical changes as food ages?* The specific area of interest is the freshness of the food system, as food ages there is degradation in flavor quality (or a food reaches an optimal after some time and moves away from that flavor optimal). The bulk of research in this field focuses on the aroma of a food system, often ignoring the non-volatile aspects of flavor chemistry. As such, this dissertation will specifically emphasize

the non-volatile aspect of aging.

This thesis introduces a number of novel aspects to what has previously been called Flavoromics, a type of chemometrics that emphasizes flavor active compounds. Flavoromics was coined by Reineccius (2008) at the 235th ACS meeting, discussed by Vos et al (2008) at the 12th Weurman symposium, and expanded on by Charve et al (2011). Flavoromics was initially purposed for insightful correlation of instrumental response with sensory response, but is generally referenced as the linking of instrumental insight with sensory results in an untargeted, mostly by comprehensively chemical fingerprinting methods. This approach uses the entire chemical makeup of a food system (as much as analyzed) to investigate flavor drivers. Flavoromics uses approaches primarily stemming from the field of metabolomics, but also applies new areas of data science and analytics to create a more insightful statistical framework. In this dissertation, flavoromics is applied to understand the aging of citrus extracts with an emphasis on flavor discovery as opposed to a correlation between sensory and instrumental data. A descriptive sensory panel was further utilized to evaluate isolated compounds. To the best of the authors knowledge this is the first time untargeted methods were utilized to identify flavor active materials that were validated by sensory recombination analysis.

1.1 Research Hypothesis and Objective

From the outset the research strategy of this thesis was to use a comprehensive (as much as possible), untargeted statistical approach to investigate the non-volatile changes in model orange systems. An important distinction between traditional research and "data-driven" research is the hypothesis is generated after the data is collected. The

hypothesis commonly becomes that the data identified by the model is relevant to the research subject, the study then shifts to understand the role of the data and how the identified variables fit into the research field. How this takes form is that the statistically identified compounds are hypothesized to be important to the researcher's goal, and then the relationship is established. One can think of this as identifying novel data structures and creating a hypothesis around the identified data, this is called inductive research or commonly data driven (Kell, 2004.). In essence our operating hypothesis for the study is that flavoromics will allow us novel insight into how flavor attributes change as a model food system ages, a more specific hypothesis is that the identified variables are important to food aging and contribute sensory active outcomes to the food system. Once data is collected, analyzed, and screened a further "data-driven" hypothesis is selected, in essence ignoring biases provided by a researcher adding in their own assumptions. This hypothesis is generated on the identified variables of importance and their relation to the research question at hand. Although, this discussion is a bit simplistic as the field of data analytics relies heavily on the use of assumptions for data sets and populations. Comprehensive approaches to characterized flavor aim to identify chemical information that would otherwise be overlooked by the traditional targeted methods, to identify compounds that relate to the sensory changes experienced as food systems age. Once chemical trends are modeled further experimentation (sensory validation) to evaluate identified compounds will establish the relationship of the identified compounds to the flavor profile. This approach differs from other applications of flavoromics in that the sensory piece acts as validation rather than drawing corollaries between sensory

responses and analytical trends.

The aim of this research is to develop multivariate statistics, data handling approaches, and modeling to focus chemometric modeling on flavor active compounds associated with aging of citrus model systems (orange and lemon). Analytically the study uses an untargeted chemical fingerprinting approach and data modeling will use a combination of supervised and unsupervised methods.

1.1.1 Hypotheses

1) Targeted methods do not provide complete insight into the complex chemistry as food ages and a more comprehensive untargeted analytical approach will provide novel flavor information. Critical to identifying flavor active compounds by untargeted methods are approaches to variable reduction and selection, as modeling approaches do not know if variables are flavor active so methods that better focus flavor activity are needed.
2) Singular modeling methods for untargeted approaches may be poorly suited to aging trends so multiple modeling approaches with orthogonal criteria for modeling can help investigate collected data and identify new variables of interest that would otherwise be ignored by a single modeling approach. Shortcomings of traditional projection based multivariate methods are offset by new machine learning methods.
3) Isolation of compounds identified by untargeted methods with subsequent purification will lead to flavor activity (directly or indirectly). Screening and validation can help identify the context of the identified feature to the original

food system.

4) Developing samples that allow for identification of unique concentration dependent data relationships can lead to new methods to understand and investigate flavor interactions. This will provide insight into how compounds may be coupled together in reaction mechanisms or may interact from a sensory perspective.

1.1.2 Objectives

1) Development of highly sensitive sample preparation, cleanup and Liquid Chromatography-Mass Spectrometry chemical fingerprinting methods to define the non-volatile chemical composition of citrus model systems.

2) Implementation of sensitive and selective raw-data preprocessing, data handling and data modeling framework to model chemical composition differences as citrus model systems age. Further develop methods for data reduction and data selection criteria to emphasize flavor active chemical differences.

3) Isolation of identified chemical features from food systems with further sample purification, sensory validation through screening experiments, large scale sensory difference testing and chemical characterization (Tandem Mass Spectrometry and Nuclear Magnetic Resonance) of identified and sensory active compounds.

4) Develop an approach to screen and understand concentration specific data interactions for further investigation via sensory panel to understand how data based interactions translate to actual tasting scenarios.

1.2 Experimental Approach

1.2.1. Experimental Approach: Overview

A schematic overview of the strategy of this project is present in Figure 1.1. To summarize, after chemical finger printing and analytical signal preprocessing an un-targeted chemometric approach is applied to understand the aging of citrus extracts through multivariate statistical investigation and machine learning approaches.

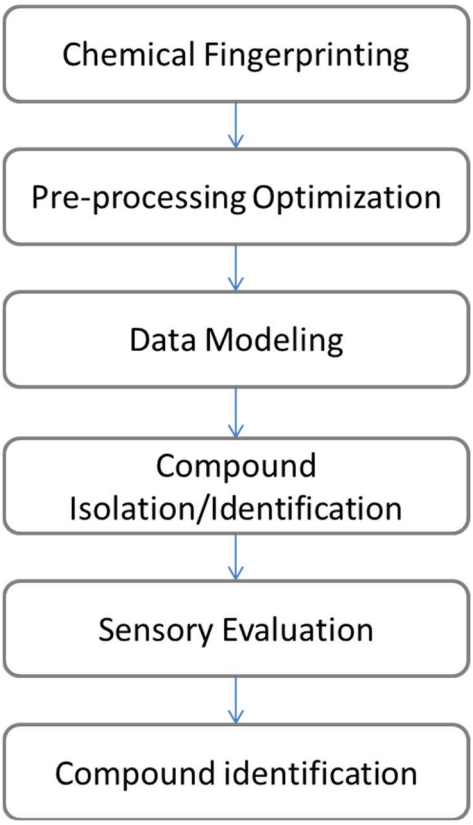


Figure 1: The general workflow for this dissertation, from chemical fingerprinting to validation using sensory panels and structural elucidation

Within these models, approaches are taken to provide an emphasis on chemical species that relate with age to identify flavor active materials. This is in contrast to a targeted approach, which would normally pre-select a very small number of compounds for investigation, only tracking a small number of species relative to the overall chemical complexity of food. The proposed investigation, using more comprehensive methods, utilizes the entire chemical fingerprint.

The benefit of an untargeted approach over traditional targeted approaches leads to better addressing chemical and data complexity and provides additional benefits from a data context perspective, which will be discussed in detail later. By including more chemical species there is greater potential for novel flavor compound discovery. By investigating more compounds there is more potential for discovery while having more data has statistical implications. Food industry often tries to understand how flavor information can be leveraged to provide improved experiences to consumers and increase the desirability of their products. By better understanding what contributes to the flavor profile of citrus products (focus of the current thesis) a variety of value streams may be impacted. Knowing what citrus non-volatiles lead to altered sensory perception can lead to understanding how to develop improved citrus flavors, citrus beverages and citrus fruit. In this work citrus extracts of different varietals were aged over various times, with each sampling point showing a change in flavor character as established by descriptive panel. UPLC-MS was utilized in order to chemically characterize the food systems and numerous data handling and processing techniques were used to identify what compounds best associated with

age. This work approaches flavoromics from a different approach that other bodies of research. Novel modeling approaches (e.g. Machine Learning) were applied to produce a flavoromic investigation of citrus extract aging. Sensory evaluation will take place after statically important compounds are isolated and purified, and act as a validation of the compounds flavor activity. These purified compounds are then used for compound elucidation through Nuclear Magnetic Resonance. This approach stresses recombination as an evaluation of sensory contribution rather than correlation as the end goal. Taking this approach establishes a causative relationship between the isolated compounds and the food system rather than the prediction of future flavor quality.

The bulk of flavor research using untargeted methods has worked to correlate sensory data with analytical data, a challenging prospect even with targeted methods. This work aims to isolate compounds that are relevant to food aging and understand their sensory relevance. Food aging is often a complex set of varied chemical reactions that can be utilized to understanding how chemistry changes relate to flavor changes (e.g. loss of freshness). Through using multivariate and machine learning, tools that are well adapted to complex data, and data handling methods that orient modeling towards flavor relevant information, flavor chemistry associated with product aging can be understood. To accomplish this goal multivariate methods are expanded upon through including second order analysis for variable selection. The bulk of untargeted research uses Partial Least Squares (PLS) and Principle Component Analysis (PCA) to model chemical complexity. While they are useful

tools, recent advances in modeling approaches lead to an extremely robust set of algorithms for analysis including machine learning and ensemble techniques. Ensemble techniques, like Random Forest analysis, hold powerful classification and regression abilities and are especially suited for the use of prediction. Through combining new modeling techniques (Random Forest) with more classical projection approaches (PLS, PCA) data utility is predicted to increase and lead to more robust characterization of food chemistry. The result of this is novel information discovery and understanding how these materials contribute to the quality attributes of a food platform.

1.2.2 Research Plan

To achieve the goals outlined above the following research plan was developed and executed:

1) Develop sensitive reverse phase separation techniques to gain sensitive and robust chemical insight into food systems. Methods emphasize analytical robustness, reproducibility, sensitivity and speed, while minimizing the time spent for sample preparation.
    a. Methods emphasize speed, throughput while maintaining analytical sensitivity, reproducibility and robustness. The goal is to gain chemical insight into chemical composition with one method, since additional methods require significant investment of time to gain additional insight (instrumental and computation). Using multiple phases and ionization methods lead to significant data redundancy,

which is a large detriment to some machine learning approaches, increasing computational resources and often increasing complexity of model interpretation.

b. Experiments are run to ensure that small chemical differences were reliably seen, and methods could reproducibly achieve this type of close to baseline detection and reliability.

2) Develop data handling methods that extract as much usable data from chromatographic separations as possible while reducing the amount of included noise.

a. Create data preprocessing methods that were able to robustly extract true chemical information from chromatographic data while limiting the presence of noise or other extraneous data.

b. Develop further data cleaning methods to better focus modeling approaches on understanding chemistry associated with aging. Identify appropriate checks that allows for construction of robust data structures that effectively capture chemical species relating to aging of food systems.

3) Develop multivariate infrastructure for effectively modeling compounds relating to aging of food systems.

a. Generate multivariate models that provide insight into quality of data structure and can be used to eliminate any compounds that may over leverage data modeling.

b. Model collected data and produce high quality models with appropriate and systematic tuning of relevant parameters.

c. Identify most appropriate set of validation criteria for model based on data structure, previously tuned modeling parameters and goal of modeling approach.

d. Develop variable selection approaches to narrow the emphasis of the compounds to be isolated for sensory evaluation. Ensuring that approaches are meant to emphasize the flavor activity as much as possible, in this case aging of a food platform not what makes different varieties of citrus taste differently.

4) Isolate identified chemical species using food grade mass spectrometry directed fractionation at a high level of purity and conduct further sensory characterization and validation of identified compounds.

a. Develop chromatographic methodology that allows for isolation of identified chemical species from a food system in a directed and systematic way.

i. Reinjection of isolated compounds onto original UPLC analytical column for validation that isolated compound is original analytical signal, validating that the isolated compounds were the original analytic trace.

b. Conduct sensory characterization and validation of isolated compounds to identify taste activity and contribution to the flavor profile with

aging.

c. Conduct structural chemical characterization (Tandem mass spectrometry and nuclear magnetic resonance) of sensory active compounds.

5) Screen untargeted machine learning approaches to understand which approach best suits the data. Further tune the initial model to optimize training and test model fit.

a. Develop bivariate scatterplot matrices to understand which of the top statistical features have contextual data relationships.

1.3 Novelty And Outcomes of Presented Thesis

The following aspects of the research contribute individually and as a whole to novel areas of research in food and flavor research:

- The comprehensive approach in the current thesis is a data driven, untargeted method to identify flavor active compounds associated with changes in flavor quality during aging of food products. Prior untargeted research on flavor and food focused on identifying sensory correlated compounds by comparing foods. The food systems varied flavor character and an analytical model was developed to fit a block of instrumental data to a block of sensory data. This thesis identifies compounds that chemically define food age and further isolates these compounds for flavor activity. This work also progresses flavoromics away from fitting models of analytical data to sensory data, illustrating additional value streams for untargeted analytics for flavor compound discovery.

- Non-traditional data management and modeling approaches advance the knowledge base of multivariate analysis in food and flavor and incorporate newer methods of modeling data. By expanding modeling and data filtering beyond traditional methods, such as principle component analysis and projection to latent structures, new information and interpretations of data is possible. Development of filtering approaches and variable selection criteria provides additional opportunity to create a more pointed and selective approach for investigating statistically important compounds. The departure beyond principle component analysis and projection to latent structures incorporates a number of newer and complex algorithms. These new approaches include ensemble tree techniques, data quality based variable reduction and filter approaches (second order analysis), and complex combinatorial approaches to understand flavor interactions. Through proof of concept these methods will show value and be easier to implement for researchers working in untargeted analysis and wish to gain more value from their existing workflows.
- An untargeted approach, as outlined in the current thesis, provides a basis for systematic flavor compound discovery. Through using approaches that use sensory validation on the back end, rather than solely for screening for compounds, confounding sensory phenomena are better addressed. The presented methodology is a more implementable and industry applicable approach that emphasizes chemically important and sensory important compounds, as opposed to identifying correlated compounds. Through flavor recombination studies an

understanding of how food aging impacts endogenous flavor chemistry is gained. A chemical understanding of flavor drivers can help food industry better preserve flavor quality, which helps add commercial value. At the same time if flavor modulators are the target for discovery this work illustrates how data science approaches can lead to new sensory combinations that show interaction or modulation.

- The project focuses on non-volatile analysis, an analytical platform that has historically not been used to understand fresh flavor. This work can act as a proving ground for non-volatile liquid chromatographic methods and their use in understanding the flavor contribution of non-volatiles. Expanding on how individual compounds can modulate complex food systems can help push our understanding of complex flavor chemistry forward.

1.3.1 Outcomes

As flavoromics develops there is great opportunity for customization of outcomes depending on the desired information and available resources and computation infrastructure. Generally untargeted methods are identified for either prediction or discovery, and implementation is either classification or regression.

1.3.1.1 Discovery

In this thesis models are generated to identify statistically relevant information that would typically be obscured or ignored by targeted flavor discovery models. Models build association levels for features, presence/ absence and intensity, and differences between sample groups are generated. The statistical reasons for sample differentiation

in this thesis relate back to food age, and flavor activity. This new understanding can help understand how chemical make-up can lead to an altered flavor profile. For example, will a change in a newly optimized manufacturing process lead to a different flavor profile, and will that change be important to the consumer? Will sourcing ingredients from a new supplier lead to cost savings and change in consumer acceptability? How is the flavor of roasted coffee different with a more rapid temperature ramp? Does the higher throughput in manufacturing offset the change in consumer liking? Applying comprehensive methods allows for more thorough and exhaustive investigation of chemical species in food, which provides an orthogonal approach to existing methods for flavor discovery. Identifying natural products that deliver desirable flavor attributes or help modulate existing off flavors deliver new methods of changing and protecting endogenous flavor while delivering on consumer preference for natural labeling.

Are there flavor active aspects of one food that are highly desirable that can be enhanced in another, or can be utilized in a flavor material? Through understanding how two products vary can lead to value added products. If one could understand why one source of stevia was sweeter and less bitter than another extract, there would be large commercial implications. At the same time understanding how minute variation in formulation or manufacturing can lead to drastic changes in liking can aid in optimization of consumer liking and minimization of cost.

1.3.2.2 Prediction

The outcome of prediction is to be able to forecast a food quality attribute given a

certain set of data. In an ideal application this would only use analytical systems and sensory as a step for validation. Multivariate statistics lends itself to prediction as it is able to handle a large number of inputs, thereby being able to handle the chemical makeup of food (composition, manufacturing parameters, etc.). For flavor, the goal would be to predict flavor and sensory qualities through chemical analysis, reducing the need for sensory to be part of an already established product line. For example, comparing a produced food to a gold standard or evaluating whether certain changes of a food will lead to a deviation in product quality. Being able to predict changes on a large manufacturing scale, either in ingredient variation or process variation, can help food processors better accommodate changes and keep consistency. Maintaining quality and consistency are key to meeting customer expectations and help ensure repeat purchases. Prediction can also help food processors understand what quality metrics are critical for their foods, and how changes in ingredients (source, environmental, storage, etc.) may lead to changes in consumer preference and market performance.

1.4 Research Challenges

This research is not without its challenges, although multivariate methods have shown extensive success in other fields, flavor adds a level of complexity that is not present in other applications. Developing rigorous data driven hypothesis which are devoid of bias are challenging and not all academics look favorably on this research approach. Outside of the difficulty to develop an inductive hypothesis that does not just validate itself the following are challenging and interwoven aspects of this research piece:

- In order for an untargeted approach to model an attribute of a system, variation

around that attribute must be present, and there must be adequate chemical diversity to differentiate that attribute from the rest of the changing chemistry.

- Even though the methodology strives to analyze as much chemical diversity as possible, there is no singular method that effectively analyzes every compound and is able to report the concentration in a manner that is representative in a food system. Increasing the number of methods run on samples leads to increased time of analysis, data structure size, and data redundancy. Data redundancy is where one chemical feature is reported multiple times leading to over reporting along with other issues. Additionally, increasing methods of analysis leads to larger data size which can overwhelm computing infrastructure and computer limitations, especially when data interactions is the end goal.
- Since the methods are designed to analyze thousands of chemical features there are often a large number of compounds that show statistical significance and investigatory resources are a limitation when there may be >100 unknowns to investigate. Compared to outputs of targeted methods that may indicate a few (<20 features) important areas of interest an untargeted model can have many (>100 features) that show statistical significance.
- Specifically for the proposed work in this thesis, these compounds are associated with aging chemistry and not explicitly flavor active and so identifying which of the statistically significant compounds should be emphasized first is challenging. Being able to differentiate between flavor active and significant chemical differences are two very different criteria. Algorithms only model chemical

changes, and are not able to determine if the chemical changes are flavor active or not. Working with a modeling infrastructure that is developed to identify chemistry differences and identify flavor active chemical changes is a challenge.

- Isolation, purification and further characterization of compounds presents challenges as untargeted methods are able to differentiate chemical differences in the parts per billion range in complex matrices. This leads to unique challenges for large scale isolation and purification of compounds that are present at very low concentrations. Instrumentation differences, scale challenges and detection limits all contribute to challenging method porting, especially scale changes from ultra-performance chromatography with sub 2 micron particles to semi-preparative high performance chromatography.
- There are two equally important aspects to characterization in this project, sensory and structural elucidation. In order to screen statistically relevant compounds a consensus panel will be used initially to evaluate individual compounds and then a trained descriptive analysis panel will validate the impact of purified compounds on recombination models. This panel will require extensive training and calibration in order to effectively characterize how added compounds change the perceived flavor. Once established as relevant, unknown compounds must be identified utilizing tandem mass spectrometry as well as nuclear magnetic resonance.
- Increasing actionable information from data driven experiments by identifying attributes of a data set that interact is also of value. Academic research is largely

driven by understanding relationships between two variables. By taking important variables from an untargeted investigation and understanding their relationship to each other can bring new value. Better insight into these relationships can identify targets for sensory investigation Being able to identify attributes of a data structure that demonstrate significant interactions during modeling can be a challenge but adds value to untargeted methods that are largely outcome sparse

1.5 Research Application: Applying chemometrics to understand age related compounds in model citrus systems.

Chemometrics applies statistical approaches to understand and address numerous aspects of chemical systems, while in this work it is applied in a flavoromic workflow. Although a departure from using flavoromics to correlate sensory and analytical data, its application in this work will focus on modeling aging and how age related compounds relate to flavor activity. For this work, model food systems were made through extraction of citrus products with food grade ethanol followed by subsequent aging of these extracts.

Freshness is critical in most foods, but terms like “fresh squeezed” are often associated with flavor and product quality in citrus systems. Consumer preference is moving more towards fresh citrus versus not, with freshness, flavor and appearance as drivers for consumer liking (Plotto, 2011). Freshness often comes with an association of increased product quality, and through understanding what aspects contribute to a fresh flavor profile actions to retain the fresh flavor are implementable. As freshness depends on the food system developing methods that allow food processors approaches that are

implementable across a number of food systems and food platforms is critical. A flexible approach to understanding freshness will bring value to food processors that are working on novel flavor compound discovery, better understanding manufacturing variance, understanding seasonal variation and much more.

Flavor research is a time and cost intensive research field (both instrumentally and sensory), and so developing methods that are able to supplement these traditional techniques is highly desirable. Untargeted chemometrics is one field that is often touted as acting as the bridge between chemical models and sensory outcomes, although this claim is still poorly supported by available literature. Developing untargeted methods that can lead to new discoveries in addition to supplementing already prominent targeted methods is highly desirable, as these data streams can lead to improved chemical understanding of flavor drivers in food systems which can lead to improved product quality and consumer acceptance. Applying a flavoromic approach beyond correlation of analytical and sensory results can lead to numerous advances and a change in how researchers understand this technology can lead to new advances and implementation in other areas of the food industry. The ability to identify specific chemical species given a change in a variable (time, composition, growing conditions, etc.) can lead to better understanding of the associated chemistry and an understanding of how that variable and chemistry can be manipulated to bring desirable flavor and consumer outcomes.

The current research focus is to establish analytical and data science principles that can lead to emphasis of flavor relevant information in a data structure. In this work the data structure is hyphenated data stream stemming from liquid chromatography and mass

spectrometry, which easily lends itself to chemometric applications. Applying this untargeted approach along with data handling approaches used to understand expressed phenotypes and second order data filtering can lead to age related compounds with flavor relevance. Citrus extracts (orange and lemon) of various ages were chemically analyzed through nonvolatile (Ultra Performance Liquid Chromatography-Mass Spectrometry). After chemical fingerprinting preprocessing extracted "all" chemical features for statistical modeling. Experimental design, data handling and variable selection methods are applied to support the identification of flavor active compounds since modeling is unable to differentiate compounds are flavor active from chemical changes. At the current time, to the best of the authors knowledge, there is no work that includes all of the considerations present in this work, or uses this approach of aging coupled with an untargeted flavoromic experimental design.

Chapter 2. Literature Review

Notes:

Summary: The second chapter works to present previous research that is considered relevant and critical to this work. This chapter focuses on areas of flavor perception, flavor analysis, and chemometrics. As the goal of this work to identify how flavor systems can be understood through analytics an emphasis is put on how data science and multivariate analysis can work with chemical data opposed to citrus flavor or mechanistic flavor perception.

2.1 Flavor Perception

2.1.1 Multimodal Perception

Extensive research has focused on understanding which chemical features contribute to consumer perception of food flavor. Through better understanding of what constitutes a food's flavor, food processors are better able to produce products that are liked, and so purchased, by consumers. A large bulk of research places the emphasis on aroma and analysis through Gas Chromatography Mass Spectrometry (GC-MS), frequently with olfaction (GC-MS-O). However, aroma is not the only contributor for flavor. Today it is widely recognized that flavor has numerous inputs and in order to understand the true character of flavor other aspects, such as taste and chemestetics, must be included. Beyond the chemical composition of food the physical aspects including appearance, texture, and even packaging play a role in consumer expectations and perception of flavor. Further, the emotional status of a consumer and their expectations are also critical to the success of an eating scenario. Ultimately it is all of these drivers combining which end up impacting the consumer experience. Chemists work to understand how product composition impact drivers of flavor liking and related mechanisms. Understanding how changes in processing, formulation, production and sourcing impact these drivers and so consumer liking is also a critical role of a chemist in the food industry. In order to fully understand and improve the quality of food, researchers must understand drivers for flavor perception and mechanisms of flavor release and perception. The following sections will discuss the areas of flavor perception and modulation.

2.1.2 Sensory drivers for flavor perception

Flavor is a complex combination of internal stimuli (aroma, taste, chemestetics, etc.) and extrinsic factors (consumer expectations, eating occasion, etc.). A consumer's product experience is improved through optimization of as many quality aspects as possible. Customer's perceived quality is a complex combination of product-oriented quality (ingredient composition, formulation, etc.), process oriented quality (organic, humane animal slaughter, etc.), quality control of each previous attribute (production and ingredient consistency) and finally user-oriented quality (subjective consumer opinion) (Brunsø, 2002). Researchers have also established that food flavor is one of the most important drivers of liking, this is largely considered an hedonic characteristic of food after a purchase event (Schultz and Wahl, 1981., Bower, 2000., Brunsø, 2002). The expectations of a consumer start when they purchase a food, and a food meeting these expectations at the time of consumption is critical for consumer liking. If a food validates the consumer's demands there is a greater likelihood for a repeat purchase event (Olson, 1972). Product success is a combination of product positioning and creating consumer expectations and delivering on these expectations. It is the goal of flavor chemists and food developers to create sensory profiles that repeatedly meet a consumer's expectations. Consumer demands are more easily met by understanding which specific compounds lead to sensory outcomes important to consumers. This is a driver for corporate investment into analytical instrumentation geared towards understanding flavor. From an analytical standpoint flavor can generally be broken down

into aroma and taste, and identifying which pieces lead to consumer preferences support long term product quality. Knowing what chemical make-up leads to desirable flavor profiles can help ensure product quality over the shelf life, help mitigate seasonal variation, and ensure consistency across processing locations by ensuring current product matches the required specifications (e.g. A gold standard).

Taste attributes are typically the non-volatile composition of food that contributes sweet, sour, salty, bitter, and umami on the tongue. Since there are 5 basic tastes it may initially appear simple compared to the numerous nuances in aroma. It is the complex combination of these tastes, each of which has varied attributes that lead to different taste characteristics. Taste perception is driven by compound interaction with taste receptor cells where the taste active compound, either directly or indirectly, depolarizes the cell activating a nerve response (Lindemann, 1996). Salt and sour tastes are direct transduction of the cell; while umami, sweet and bitter response comes from bonding of the taste compound with the surface G-protein receptor and a cascade triggering nerve response (Lindemann, 2001). These flavor compounds ranging from small molecules to large peptides and even small proteins represent an enormous amount of chemical diversity. These species range in taste thresholds from parts per million to parts per billion or lower in some cases. There is great challenge in understanding taste attributes as within each taste attribute, bitter for example, there may be a large range in chemistries that lead to a response (from ionic salts to massive proteins). Many of the taste attributes are also impacted by genetic diversity which will lead to differing perception as well. This is in part due to the number of taste receptors that exist for some attributes, in the

case of bitter there being over thirty classes of mediators identified (Chandrashekar, 2006).

2.1.3 Flavor Modulation

Research has established a number of instances of flavor modulation, with the research highlighting volatile and non-volatile modulation (Auvray & Spence, 2008). Along with this many other types of modulation exist including both receptor and cognitive level mechanisms (de Araujo, 2005., Shin, 2008). Early research indicated that sweetness was modulated by fruity aroma but the inverse was not true (Murphy & Cain, 1980; Murphy, Cain, & Bartoshuk, 1977). Other researchers have established the association of flavorants and how they can influence perception, as in enhancement between flavor materials that are commonly found together (Frank, Shaffer and Smith, 1991). This interaction seems to indicate a learned association, reinforcing the idea of pattern matching and cognition in flavor perception. The idea that humans associate specific attributes to specific foods, or their flavors are congruent, was proposed by Schifferstein and Verlegh (1996). It would make sense that if an aroma is commonly perceived with a sweet taste that the brain would try and match the aroma to the most commonly perceived flavor pattern, which may lead to an enhancement of commonly associated attributes even if the all attributes are not present.

Other mechanisms of flavor modulation exist beyond taste and aroma interaction such as taste receptor interaction and cognitive interaction. Zinc have been reported to act as a bitter and sweetness suppressor, which was suggested to be caused by reaction with the tastant or through reaction with the receptor protein causing change in

conformation which would explain why some bitterants are impacted while others are not (Keast, 2005, Keast, 2008.). Other examples of sugar (sweetness) modulation include salts which can lead to either suppression or enhancement depending on the type of salt, specifically carboxylates lead to enhancement (Van der Heijden, 1983). Ionic substances are unique since they are able to alter tastant solubility and impact the taste receptor protein structure leading to dual mechanisms of flavor modulation. Understanding the mechanism of modulation provides an understanding of how ingredients can be leveraged to accentuate and suppress other flavor attributes. Additional compounds, such as proteins and terpenoids, can lead to sweetness suppression (Kurihara, 1992). Umami is a well-established booster of savory flavor profiles, but there is also evidence that umami compounds (Mono Sodium Glutamate) as well as other compounds (adenosine mono-phosphate) can lead to bitterness suppression (Keast, 2002).

The complexity behind how we understand and evaluate food is challenging since flavor is multi-modal in nature. It is rare that individual compounds illicit a sensory response that fully encompasses a flavor profile. For example, there is no one compound that fully captures all of the attributes present in a food (taste and aroma). It is the complexity that allows nuanced differences to make large impacts on flavor profiles and consumer acceptance. Small chemical changes (ppb level) of off flavors can taint a consumer's experience. In order to effectively investigate and understand food flavor a wide variety of analytical tools is needed. A number of methods have been implemented to define the flavor chemistry of a food system. Often times this includes the extraction of a food system, and then evaluation using a chromatographic separation followed by

mass analysis. Other methods, like spectrometric methods, have also been applied in lieu of extraction and chromatographic methods.

2.2 Flavor Analysis

Flavor analysis has numerous aspects and considerations that are important to success, selecting methods that lead to an extract with a representative flavor profile increases this challenge. Flavor analysis can vary from sensory based approaches to methods using advanced analytical instrumentation, or a combination of both techniques. As this thesis largely relies on the information stemming from instrumental analysis, the flavor analysis section will focus on methods for flavor discovery. Critical to success is selection of starting materials that hold the sensory response to be analyzed, preparation methods that accurately retain the attributes of interest and then methods that accurately report the composition and provide a chemical basis for understanding drivers in the food. All of these steps should be done in a way that leads to an extract or analysis that closely mirrors how humans consume food. These methods will range greatly depending on the analytical approach and research outcome. The next section of the review will emphasize methods that are critical to this dissertation, and although there are numerous approaches to understand and address chemical complexity the discussion emphasizes the methods heavily utilized in this work.. This will principally involve liquid chromatography methods and sample preparation. Other methods are mentioned, but not emphasized, as they are relevant to certain aspects of flavor but not this specific project.

2.2.1 Sensory Analysis

Sensory analysis methods work to combine food science, physiological response,

mental perception of stimuli and statistics to identify characteristic perceptual differences in food. Using humans as "instruments" to identify what attributes of food differ and the drivers of food flavor is a challenging endeavor. Sensory analysis and flavor chemistry cooperation are critical as flavor chemistry provides food chemistry context to what sensory panels report. The chemical understanding of sensory drivers provides a critical understanding of how attributes can be attained and preserved in a large production setting. There are many different types of panels that can be run ranging from discrimination tests, descriptive methods, time intensity, preference and hedonic (Piggott, 1998.). There are some common forms of discrimination tests including: triangle test, paired comparison, A-not A, and duo-trio. Descriptive tests include Flavor Profile Method, Texture Profile Method, Quantitative Descriptive Analysis and Spectrum Method (Piggott, 1998). Ultimately, the goal with the method used is to measure and accurately report perceived flavor sensation; the utility of this information varies by the test and the panel size.

Previous work has attempted to associate panel perception with instrumental analysis. Panels are frequently time and resource intensive, so cost and time effective alternatives are desirable. Stemming beyond descriptive sensory work is coupling instrumental data and large scale consumer panels to attempt to predict consumer preference and chemical drivers of liking. Prediction of product quality through instrumental analysis would provide great value to food processors, as shortened time to action and reduction in cost help the product development pipeline. Correlation between sensory and instrumental analysis has a long history attempting to provide value with the

help of chemometrics (Aishima and Nakai, 1991). This untargeted approach was first emphasized as a way to better understand how multiple compounds could relate to sensory responses (Aishima and Nakai, 1991). Often times this area of research sets out to identify either the mechanism of flavor perception or through instrumental analysis prediction of the sensory response. These two approaches hope to use the chemical complexity to explain nuanced chemical differences between samples. There is extensive work done with Atmospheric Pressure Chemical Ionization Mass Spectrometry (APCI-MS) to try to understand how aroma release impacts sensory response (Taylor, 2003.). However there are challenges with these approaches, generally extensive correlation of variables is well handled by statistical projection methods but without appropriate scaling of redundant data any benefit from projection methods is largely lost due to the random enforcement of latent structure determination.

2.3 Chemometrics

In the realm of multivariate statistical analysis there is a wide breadth of applications, this thesis mainly draws on principles derived from the fields of chemometrics and metabolomics. Chemometrics is the application of multivariate analysis to chemical data where metabolomics is the application of multivariate statistics to the realm of metabolites (Beebe, 1998., Gates, 1978.). The term flavoromics stems from Reineccius and encompasses the use of multivariate approaches to flavor (Reineccius, 2008). Flavoromics can be considered a subset of chemometrics, and often includes sensory panel data in order to ground the modeling with human perception. The underlying goal of flavoromics is to take learnings from chemometrics and effectively

identify compounds for flavor in complex and multidimensional data sets. In literature there are often times when metabolomics methods are associated with flavor based omics, but often these approaches include both metabolic and chemically derived species and so chemometrics is often a more general and appropriate term (Lee, 2011., Tikunov, 2005., Ochi, 2012.). Within multivariate analysis there are a number of modeling classes including: exploratory analysis, regression and classification. The bulk of approaches fall into one of these categories, but there are other approaches that combine univariate operations within a multivariate framework, such as multiple regression and linear regression approaches. In order to key in on flavor active material both regression and classification modeling approaches have historically been used. Classification is exemplified in the work of Fisk et al. and the application of aroma profiling through Atmospheric Pressure Chemical Ionization (APCI) and the classification of geographical location of food stuffs (Fisk, 2014). Classification based models work to predict how a sample's class can be described based on the closest match within a data set. Andrade et al. showed use of partial least squares regression in understanding performance of spray dried model systems and flavor load for limonene and 2,5-Dimethylpyrazine (Andrade, 2008). In the case of PLS modeling, originally coined as partial least squares regression, a block of data (X block) is regressed (fit) to another identifier block of data (Y block). Regression based models work to correlate one data set with another, trying to relate multiple measurements to a desired property or outcome. As there is extensive use of PLS methods in literature, one could argue that regression is the most common method used, which is not surprising considering the thousands of citations on some of the early

PLS omic-based papers. Each approach, whether classification or regression, provides the ability to address and understand how classes of samples differ and identify statistical drivers for class differentiation. Classification based systems provide great utility in understanding how sample classes vary chemically and how new information fits into models. Delving further into the area of multivariate analysis shows there is significant depth in how researchers can approach statistical questions and outcomes. Ultimately in any chemometric application one of the first steps is converting the collected chemical data into a data frame, this is done via signal processing and is often referred to as preprocessing.

2.3.1 Chromatographic Preprocessing

Data coming from hyphenated separation methods is some of the most common data in flavoromics and methods to extract data from a chromatographic separation fall into the category of data preprocessing. Within this area there are numerous open source methods (e.g. XCMS, MZmine, ect) and many industrial options that are in essence "black boxes". No matter what option a researcher selects there are commonalities in the steps done, although the ways these steps are completed will differ algorithmically across implementation. Generally these steps include peak filtering/identification, peak matching across samples, retention time correction/alignment, fill missing peaks and then export data to a data frame that is populated with features (or variables). Once data is exported scaling and normalization is sometimes applied as well. Each of these steps can be completed by a number of different methods, but end goal is the same: convert chromatographic data into a data frame made up of sample identifiers, retention time and

mass to charge features, and the associated intensity. This is three-dimensional (time, mass to charge ratio and intensity) data which ends up being reduced to a two-dimensional data frame (mass to charge and retention time pairs with intensity). From the most basic applications the peak identification extracts data for a mass channel and identifies when there is significant increase above baseline, signifying the start of a peak. Once the mass channel has a rise above baseline the algorithm then identifies when the local maxima is reached which is the peak max and then the algorithm identifies where the mass channel converges with the baseline. The most relatable examples of these algorithms are the first derivative test (peak start and stop) and second derivative test (local maxima), but the complexity can be significantly higher than this using low and high pass filters or wavelet functions. Identified peaks are matched within sample grouping and scaling, warping, and shifting are done to produce more robust representations of the peaks (Smith, 2006). Depending on the analysis either peak height or peak area is used for the intensity of the feature, this is very dependent on the researcher's preference and mentality, the instrumentation used and the sample composition. Peak area relies more on the assumption that all peaks will share similar distributions and shapes, while peak height assumes that all compounds are retained similarly and there is equal analyte loading, no adsorption, and limited analyte co-elution while still following a Gaussian distribution (Weckworth, 2007). There are a number of dynamic controls associated with how parameters are set for these methods including signal to noise, absolute threshold, mass increment, retention time windows, and whether or not features should have their isotope signatures removed (Siuzdak, 2006). Once the

instrumental data is converted into a data frame the data can move to statistical modeling and as discussed earlier there are a number of modeling approaches ranging from general linear models to more complex machine learning approaches.

2.3.2 Modeling Mentalities

Within statistical methods there are a number of approaches to understanding and modeling data including explanatory modeling, predictive modeling and descriptive modeling (Shmueli, 2010). Explanatory modeling is most typically seen in the development of casual hypothesis testing with statistical models and has great utility for action based modeling since there is a description of drivers for the model, making interpretation easier. Predictive modeling is the use of modeling approaches to use an existing data set to define and understand where new observations fit within the context of an existing model, this modeling approach is commonly associated with the end goal of many flavoromic applications. Finally, descriptive modeling tries to communicate a complex data structure in more simplistic terms, differing from explanatory by relying less on the casual hypothesis (Shmueli, 2010). The reader should note that these approaches do not specify a modeling tool, as one could apply any number of statistical models to each mentality depending on the type of data. There is often a split between data scientists (those who commonly use machine learning) and statisticians in what methods are used and how the models are implemented but it should be noted that due to the staggering number of methods available saying that one mentality is superior is naïve.

2.3.3 Modeling Tools

The current variety of modeling tools and approaches is staggering and ranges from

simplistic univariate modeling (Linear Regression) to complex machine learning algorithms that utilize artificial intelligence (adaptive neural networks). To approach this wealth of methods in a digestible manner this work will briefly touch on some classes of modeling tools and then discuss the tools used in this work in depth. Firstly, a scientist must define the scope of their experiment and whether it should be supervised or unsupervised, which depends on the desired outcome, researcher mentality and data source. As new advances in unsupervised learning approaches show greater utility applications of algorithmic modeling appear to outperform those that work closer with stochastic data modeling (Breiman, 2003). The underlying assumption of stochastic methods is that the data is generated from a culmination of random noise, predictor variables and parameters (area of interest), where the algorithmic approach makes the assumption that there is too much complexity to use a simplistic analysis approach and so a complex association is needed to model responses (Breiman, 2003). In the context of food chemistry this algorithmic approach seems to hold additional promise over more traditional models since it relies on a complex set of associations to model a response. Since food chemistry is a complex and highly interrelated set of macro and small molecules the ability to identify data structure driven associations can be highly valuable. Though more complex modeling approaches may capture more nuanced associations they are often times more complex, and so to begin the discussion of modeling approaches an understanding of the more commonly applied methods is warranted. The assumption that a more complex model is better is rarely true, and use of a more complex model often makes the results difficult to understand. This along with the learning curve associated

with more complex methods make traditional multivariate tools the majority of modeling approaches used in literature.

2.3.3.1: Commonly used multivariate methods

Due to the ease of summarizing numerous dimensions of data at once projection methods see extensive use to visualize complex data sets and to create dimension reduction models. These projection based methods, such as principle component analysis (PCA) and projection to latent structures (PLS), create new variables that capture the majority of variation inherent to a data set allowing a summary of the data using a few statistically relevant vectors through data clouds (Trygg, 2007). In research PCA has seen extensive use in illustrating differences in multivariate space through simplified loading plots showing groupings of samples. PCA populates a data frame of X observations with N variables in a piecewise manner to generate a data cloud, with each new variable being added orthogonal to the prior variable (Wold, 1987). Once the entire data frame is populated vectors are drawn through the data cloud that capture the most variance, a variables position in this data cloud is now summarized relative to the new vectors (principle components). Since there is no added information such as a variable class or identity it is an unsupervised method. From an exploratory point of view PCA has the most utility in data exploration through untargeted data representation. Often times research will represent PCA as a model to differentiate sample classes, and although there may be some value in this it should never conclude an analysis, as this assumption relies on the principle components identified captured the attributes of the research at hand (Daszykowski, 2006). Often times literature will indicate that a PCA

model identifies samples as different due to a specific trait (e.g. geographical location, processing condition, ect.), but this is rarely a causative conclusion (Daszykowski, 2006). PCA has high utility in identification of highly leveraging outliers (variables or observations) and can even be used for dimensional reduction if large data sets want to be compressed into a number of meta-variables. PCA also suffers from the curse of dimensionality and is rapidly over fit with even small sized chemometric data. Application of PCA with a more supervised approach is Projection to Latent structures (PLS).

Projection to latent structures (PLS) couples the approach of PCA and regression methods to best fit a data frame X with another Y in order to maximize the captured variance (Wold, 2004). In other words, data matrix X is fit to data matrix Y to identify the underlying relationships in data sets that have many more variables than observations. Noisy data or high collinearity are generally well handled in PLS applications but data redundancy can be an issue depending on how the data structure is populated. In most uses within chemometrics a two block model is used, and commonly this is a discriminant analysis (PLS-DA) where the Y data matrix is categorical (Wold, 2004). There are many varieties of PLS models including some of the newest implementations such as orthogonal-PLS (OPLS) which makes models easier to interpret but does not increase the predictive power (Trygg and Wold, 2002). Since PLS fits matrices to each other it is used as a supervised method (sample classes are known) and has shown application is inclusion of sensory data as the Y matrix (Charve, 2011). Multivariate methods have a long and established place in data analysis, but with the dramatic increase

in data being generated (mostly electronic) new approaches were developed that can more effectively investigate massive data structures and better utilize computation resources.

2.3.3.2 Machine Learning

As data size increased and data structure grew more diverse and large, new methods came out to identify important information. Approaches like these that iterate and adapt over a data set are called machine learning, and have provided extensive value to those with the skillset to use them. Often times the first thought of machine learning is artificial intelligence, although this is only one approach and there are numerous uses of machine learning outside of artificial intelligence. This discussion will center on the use of machine learning to identify important data relationships and correlations within large data sets. The benefits of using machine learning in data mining is that large data sizes are easily evaluated and digested by machine learning algorithms. The efficiency of these algorithms dwarfs what a human can manually compute. Given thousands or millions of variables how would a human approach the investigation of such large data structure in an efficient, systematic way without using adaptive tools? Given the size of data present the likelihood that decision criteria and goals of the investigation would adapt as new information was discovered is large. The ability to systematically redesign and include new information into a model helps limit the amount of time a human would need to redesign models.

As there are numerous approaches to modeling complex data only the ones in this work will be discussed. The modeling of the data in this work will center largely on use

of supervised methods where sample class is used for a priori information. Within machine learning this information is called bias, a specific bias is provided to the algorithms which help target and increase efficiency in the machine learning algorithms. Given the size of the data systems we will be working with there are tens of thousands of inputs that can match with the outputs provided, so providing information to direct and target modeling makes the algorithm more specific and more selective to the research topic. That being said there are advantages to unsupervised learning, but with the already challenging interpretation of supervised methods is beyond the scope of the current work. In essence, adding in a sample identifier allows methods to achieve more rapid investigation of sample space as the narrowed focus provides significantly reduced space for hypothesis testing. If there were no sample class provided there would be a huge amount of iteration required to develop a hypothesis and this means increased analysis time. By providing a reduced hypothesis testing space the methods are better able to achieve efficient model generation. In the context of this experiment, some methods may be unsupervised (PCA) but most others will be fully supervised (decision trees, PLS). Machine learning is an ever-changing area of study with a foundation in computer science and a history in pattern recognition. Largely the goal is to utilize data in a manner that provides insight into a complex problem or issue and identify unique associations or combinations of data that would otherwise be obscured. As there are a number of different varieties, the application and approach taken can largely vary based on the needs of the analyst. There are approximately 17 types of machine learning that are commonly used for classification: discriminant analysis, Bayesian, neural networks,

support vector machines, decision trees, rule-based classifiers, boosting, bagging, stacking, random forests and other ensembles, generalized linear models, nearest-neighbors, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines (Fernandez-Delgado, 2014). Due to the number of considerations and coding challenges that may be present in each of these methods often times analysts will stick with a few familiar methods and only work with those, this is a common bias seen in machine learning which is often discussed but has limited publication (Fernández-Delgado, 2014). As each method can be challenging to understand and have complex intricacies that are beyond the scope of this work a few more common applications are touched on and should act as representative methods for how their classifier families work with data.

- Linear Discriminant Analysis (LDA)-

  Linear discriminant analysis is a great starting place as most, if not all, researchers have at one point conducted a Fischer's correlation when plotting data. This approach strives to understand which variables present in a dataset can create a good prediction of the observations, a higher degree of fit assumes a closer association of the variables. In essence LDA identifies variables present in a data set that lead to differentiation of two observations or classes. As the number of classes increases the number of observations used to define a fit increase as well. Success in linear discriminant analysis, as with many multivariate techniques, is closely linked with variable selection. Understanding how collinearity impacts data fit is very important, especially in LDA (Izenman, 2008). Often times in a

problem with C classes, there will be (C-1) classes fit to maximize the differences between samples. In a pool of thousands of compounds this can quickly lead to arbitrary selection of observations to select the vector for dimension reduction and does not preserve non-Gaussian distributions and so can lead to loss of complex data structure attributes.

- Quadratic Discriminant Analysis (QDA)-

  This method investigates whether or not variables have a Gaussian distribution and then using the Bayesian theory of posterior distribution develops a classification for a test point (Srivastava, 2008). These approaches work to generate curved classification boundaries in a dimensional reduction application, and may fit certain classification problems better than LDA. This approach works similarly to LDA but does not require a linear relationship, hence the potential for broader application. One of the larger points of differentiation between LDA and QDA is that QDA does not assume that the covariance between all points is similar, and in this sense works better when the data is highly heterogeneous.

- K-nearest neighbors (k-NN)-

  k-NN is a "lazy learning" method that develops sample boundaries in space using only voting of close proximity data. It is "lazy" in the sense that all of the decision making and associated computation is done during the classification step. This approach saves computation time, but also uses a very small amount of the data structure to drive classification which is in part why it is considered one of the most simplistic machine learning approaches. In essence the algorithm uses a

number of nearby observations (k) to define which sample group it should be associated with (Kuhn, 2013). This is often considered a method of model training, although it is not actually a training step. The number of nearby observations (k) is used to define which sample group the observation belongs to based on the majority vote in confidence interval (Weinberger, 2005). This method performs poorly when there is skewness in the dataset, creating bias in voting. Increasing the size of k, and so the voting pool, helps to alleviate bias but also makes the method more computationally intensive. There is also a challenge with whole model outliers and sample group outliers in introducing a voting bias, but this can be avoided through more complex sampling methods or larger k.

- Support Vector Machines (SVM)-

  SVMs are a set of mapped classification tools that take data and generate prediction criteria maximizing the distance between samples in a way to identify clear patterns leading to classification, this method stems from computational learning theory and gained traction by using non-linear methods (Kuhn, 2013). This is done using multiple or a single hyperplane in high dimensional space (similar to PCA/PLS data cloud). The hyperplane is generated to maximize the distance to a training data observation, and in doing so helps to minimize the generalization error, which is a metric of the performance. Development of many of these planes generate functions that help recognize structures in data that lead to classification. Any classification that is beyond binary needs special handling, and common methods include a one-vs-one scheme. The binary nature of SVMs

lead to their effective use for two sample group problems, but as more sample classes are added the statistical programming becomes more challenging and so the utility decreases.

- Tree methods-

  Tree methods are a broad category of methods that use a tree based system for either classification or for regression. From a programming standpoint these models often take form as a number of nested "if-then" statements, which as nodes are added become less powerful to the model (Kuhn, 2013). Through the combination of these statements, of differentiation cases, observations can either be classified or a regression model can effectively be built. These functions can be layered and the decisions can take into consideration multiple variables in a way that helps to contextualize variables populating in models. In the case of Boolean univariate binary decision trees it closely mirrors a two layer feed forward neural network, illustrating how decision trees are able to compete with newer approaches to modeling while maintaining ease of use. One of the most attractive parts of tree-systems is their robustness against data which may be detrimental to other modeling types, often sampling methods make them attractive when a data system is sparse or skewed, additionally magnitude differences in variables do not leverage models as much as projection methods (Kuhn, 2013., Fernández-Delgado, 2014). Of highlight within tree methods is Random Forest modeling, stemming from the work of Leo Breiman, which is a recursive and extremely robust implementation of tree based modeling

(Fernández-Delgado, 2014). Models are generated by randomly sampling a data frame with replacement (bootstrapping) to identify which variables continuously act as strong classifiers across numerous model generation instances (Breiman, 2001). As decision trees are prone to overfitting being able to model large and noisy data sets while still being robust to overfitting is a necessity, this combined with Breiman's claim that there is no need for cross validation in Random Forest models make it a highly attractive approach for data investigation.

- Adaptive boosting-

  Adaptive boosting uses a similar thought process to ensemble techniques where a number of weak learners are used to generate an adaptive system that eventually generates a strong learner (Chu, 2004). Adaptive boosting often uses ensemble decision trees as the basis for its weak classifier and slowly adjusts data presented to these learners to optimize a decision network (Freund, 1999). This system is successful since over the many iterations of data models that are marginally better than guessing are formed. This identifies clear data trends that would only stem from iteration and emphasized by added weighting of marginal improvements. Some researchers suggest that adaptive boosting is one of the best "out of the box" classifiers, as it often performs well on a number of varied data sets (Fernández-Delgado, 2014., Mayr, 2014). This claim is not always supported in literature since comparing modeling approaches is difficult as there are so many parameters to tune and adjust, but it does frequently edge towards higher quality classification even in comparisons (Fernández-Delgado, 2014).

- Naïve Bayes Classifier (NBC)-

  NBC is an approach that uses methods assuming complete independence of variables and their relationship to an observation. These assumptions completely ignore multiple trends that may contribute and compound to create strong data trends (Kuhn, 2013). This method uses a probabilistic approach to classification and provides classification and an associated degree of certainty. Some of the challenges with NBC stem from its consideration of independence, but the probabilistic approach also is an issue when variables manifest at numerous levels as developing probability tables off diverse populations and numerous variables is challenging. Despite these considerations NBCs have shown a high utility to produce simplistic results that are rapidly achieved using real world data sets, with some proponents of the model indicating it as one of the superior methods of analysis (Zhang, 2004).

2.3.4 Data Reduction and Variable Selection:

In the context of the data set one of the important considerations for researchers is: which of the many statistically significant variables are of interest. This question is at the core of many comprehensive research approaches and numerous methods for variable selection have been created and used. In essence the goal is to reduce the initial data set size into a more targeted and directed, making a subset that hopefully contains the area of interest for the researcher. The area of feature selection is relatively new as papers as early as 1997 were dealing with fewer than 100 variables, a much more manageable dataset than the thousands that are more common now. There has been a massive

increase in available data to address research questions and support data driven research, but much of this data is considered outcome sparse. In a sea of tens to hundreds of thousands of variables only a few variables are realistically investigable. From an optimistic point of view a method would remove data that was noisy, poorly reported, incomplete, redundant, or not relevant. By doing so one would increase the modeling efficiency, decrease the analysis time and with fewer dimensions produce models that were less likely to be over fit and more likely to focus on the research question at hand. The challenge lies in developing an infrastructure that identifies data that is relevant to an expressed output with little guiding form the researcher. This is not the norm when multiple data streams are used and one wishes to use a truly comprehensive approach. One approach is to filter the data from a strictly reliability approach, if compounds are not reliably reported within a sample group they should be eliminated (e.g. precision). The caveat to this is that without agreement for other sample groupings one increases data sparsity and does not provide adequate statistical power, so a variable with a high variance should only be eliminated if it is also highly variable in all other sample groupings. This idea can be expanded to cover any number of data metrics like skew or kurtosis. Expanding on these approaches there are also methods called variable ranking like Fisher's linear discriminate which are more akin to a preprocessing methods and help to reduce data size before modeling (Duda, 2001). One of the main benefits of this approach is that it computationally only evaluates the number of variables so it is considered efficient from a time perspective. Another approach is linear correlation such as Pearson's correlation, which many are familiar with. This allows for investigation of

the amount of variance that is explained by variables, and highly correlated variables may indicate redundancy and warrant combination of variables. Single variable classification investigates error rates when individual variables are used for sample group classification, and can help identify novel features as well as poorly performing variables. Although these methods might perform well one must account for the fact that if a variable is not important in its own right, it can provide leverage and context for other variables and this is a consideration important in food systems as the chemistry is complex and intertwined. It is also worth noting that there are multiple steps in a multi-variate analysis that warrant variable selection and include before modeling (increasing model efficiency) and post modeling (identifying which variables to focus on initially).

One such approach of note is a meta-analysis, where statistical models are built on statistical models or previously mined data. The first model captures the major variances and major drivers for differentiation. Then a metric is established and data is output for a second set of modeling or investigation, this is also referred to as a second order analysis. In this approach the initial model is used as a way to roughly approximate the area of interest that the model should capture, then from there smaller data sizes and a second set of models which will provide more directed and targeted investigation into the research question at hand. This approach was exemplified in the work of Tautenhahn et al. and the development of metaXCMS, which has generated a platform for users to identify areas of interest through sample overlap in the goal to identify a specific phenotype. The underlying idea behind this approach to metabolomics is that there may be significantly different sample composition but the overlap allows researchers to better understand the

commonalities of a phenotype of interest, this idea strongly benefits from having the area of interest present in many samples of a diverse population. From an experimental point of view this treats non-relevant biological variation as "biological noise" and allows for data reduction through identification of commonalities. By using a layered method such as this contextual data impacts are preserved, and so one of the underlying drivers for a multivariate analysis is kept as an investigation moves forward-an aspect that is lost with other preprocessing conditions (Duda, 2001). Preserving this inter-data frame relevance is an important aspect that can lead to better downstream identification of compounds of interest that would otherwise not have strong model significance. Ultimately, generated models are creating constructs to fit the data and not the research question at hand, so it is important for researchers to understand how their data manipulation impacts both modeling but also the model's ability to associate to the research question at hand.

2.4 Analytical Instrumentation and Theory

There are numerous approaches to comprehensive research, in the truest sense the entire product is characterized natively but this goal is rarely achievable. A true comprehensive study would entail data streams all the way from seed to the consumer and impactors at every point (storage, weather, etc.), as all of these are known to impact a human's perception of a food. As this is not possible the term comprehensive often takes on a new meaning in "omics" studies. In essence a comprehensive study entails the inclusion of as much data as possible, ideally prepared in a manner that is systematic, logical and is collected by a sensitive, stable and reliable analytical platform. As whole food analysis goes there are very few methods that analyze the chemical composition as

well as the undisturbed physical structure of a food, the exception being solid state nuclear magnetic resonance, so most of the time an isolate is taken for analysis. In the realm of food flavor chromatographic separations have provided a massive wealth of information and learnings. When chromatographic instruments are coupled to informative detectors like mass spectrometers the quality of information is quite high. Although an established and informative analytical platform samples need to undergo rapid, representative and unbiased sample preparation before they are able to be analyzed in an efficient manner. For a good expression of natural variance numerous samples must be prepared and so methods must balance the amount of information gained with the amount of time spent on generating the data. When working in this space it must be acknowledged that there is always a tradeoff between improvement of results, introduced bias and variation and time spent on the analytical step.

2.4.1 Sample Preparation and Clean up

In the space of chromatographic separations sample preparations aim to enhance the analytical signal, while limiting noise and remove analytes that will foul columns or interfere with ionizations. This is an incredibly varied step as it depends on the food system and can include removal of large macro-molecules, de-fatting, sample concentration, and removal of unwanted analytes. With each of the steps there is an associated improvement but added time and a bias introduced into the experiment, so the number of steps should be kept minimal. Often times Solid Phase Extraction is employed to concentrate the sample, remove sugars and compounds that would foul the analytical column (Simpson, 2000). Ultra-filtration is an approach to remove large macro-

molecules from complex mixtures in order achieve better chromatographic performance (more analytes on column with same injection volume) as well as reduce ion suppression stemming from these compounds.

2.4.2- Liquid Chromatography Mass Spectrometry

Chromatographic separation has a long history of application to flavor science including both liquid and gas to better understand chemical divers of liking. Liquid chromatography is the application of a pump system to adhere compounds to an analytical column and then selectively remove analytes by applying a selective pressure (organic solvent, salt, pH). There are multiple types, but reversed phase (RP) and normal phase (NP) are the most common approaches, with hydrophilic interaction liquid chromatography (HILIC) often times more selective than normal phase application (Snyder, 2011). Food chemistry analysis is quite amenable with reversed phase separations due to the increased selectivity of non-polar compounds and their ubiquity in food systems. Foods hold a large library of compounds that are well separated on reversed phase applications, and with advancing technology came increased ability to understand chemical drivers of food liking. As analytical technology progresses and new advances create more sensitive instruments more concrete information can stem from preliminary runs, allowing for increased insight with less development. Advances like ultra-performance liquid chromatography (UPLC) provide high separation efficiency in very short run times and when coupled with high mass accuracy mass spectrometers like Quadrupole Time of Flight (QTOF) instruments large amounts of data is rapidly collected.

Mass spectrometry (MS) is also a well-established method of food investigation, and brings critical information about a molecules mass, through the mass to charge ratio. Mass analyzers have long been used in aroma analysis which is supported by the extensive library for GC-MS flavor compounds. Mass spectrometry based metabolomics has grown massively with the improvement of mass spectrometry platforms (Dettmer, 2007). Although there are approaches that use direct infusion for MS analysis, saving time and dramatically enhancing the throughput of samples, this approach suffers from extensive ion suppression and so may not realistically represent the sample composition (Dettmer, 2007). The addition of a chromatographic separation on the front end turns mass spectrometry detectors into highly robust analyzers that provide highly informative information. In the case of accurate mass instruments that are able to produce highly resolved spectra with low mass error elemental composition is possible as a way to better understand and identify unknown compounds. For the application presented here that is highly valuable and can lead to a more efficient workflow for nuclear magnetic resonance structural elucidation.

2.4.3- Nuclear Magnetic Resonance

Nuclear Magnetic Resonance has a long and extensive history, with some of the earliest publications arriving in the late 1930s (Rabi, 1939). The principle behind nuclear magnetic resonance applies the idea that given the application of an oscillating magnetic field, in the presence of a static field, atoms with a property called spin absorb and re-emit the applied field (Chizhik, 2014). Spin is present in isotopes with un-equal number of protons or neutrons and this is referred to as having an intrinsic magnetic moment.

This can further be broken into fractional spins (odd protons, odd number nucleons), integral spins (even protons, odd number of protons and nucleons), and zero spin (even protons and even nucleons) (Chizhik, 2014). Changes in the structure of an analyte lead to different field impact on the nucleus, this is due to shielding and provides a significant amount of structural information. There are numerous concepts that lead to the power of NMR for structural elucidation, but that is beyond the scope of this work. In this work, NMR is used as a method for structural elucidation and this process usually involves three major aspects: determination of skeletal connectivity, determining of relative stereochemistry, and verifying proposed structure. Developing the skeletal connectivity entails a number of 1D and 2D runs. Initial count of the 1D 1H NMR spectrum should match the information coming from the accurate mass elemental composition, the number of protons allows for the determination of the degree of unsaturation. One bond heteronuclear correlation (HSQC) generates a spectra that identifies the proton and the directly connecting carbon. Homonuclear approaches can also be applied to understand geminal, vicinal and some longer range coupling and piece together separate segments that proton and heteronuclear techniques identified. Some of the last scans run, due to the limited sensitivity, are multibond heteronuclear approaches. This allows further piecing together of skeletal sections of identified functional groups through understanding long range heteronuclear couplings (Ernst, 2001).

Initial interpretation of proton NMR starts with a molecular formula, which is often provided by a high mass accuracy mass spectrometry analysis for elemental composition. The number of protons in the formula and the NMR spectra should match. Understanding

of the chemical shift, the splitting and the integration values provide additional information on the type of group and associated molecules in the structure. Hetero Nuclear Single Quantum Coherence (HSQC) is a technique that allows for resolution of both the proton and the carbon spectra and aids in showing if any of the proton NMR signals were overlapped. This technique eliminates carbon signals that are not bound to a hydrogen, which helps in many of the carbon assignments. This coupled with Heteronuclear Multiple bond Correlation NMR provide significant information that can piece together un-assigned structures. Further piecing all of these pieces of information together come from the use and understanding of how multiple bond NMR scans (TOCSY- Total Correlation Spectroscopy) fit the pieces together.

2.5 Citrus Flavor

Due to the numerous value streams associated with citrus products there is extensive background on the volatile profile for citrus fruits and products, not surprising given the complexity and benefit of preserving the native fruit flavor profile (Rouseff, 2009). Orange Juice alone is one of the most consumed juices in the world, with the flavor quality being one of the main drivers for consumer liking. Similarity to freshly squeezed juice being a major consumer driver, recreating this perception in commercial juice can have a large financial impact (Rousseff, 2009). By better understanding the chemical drivers of flavor quality in juices, flavor improvements can lead to new and improved sensory profiles which could help the struggling citrus markets. The United States Department of Agriculture (USDA) Foreign Agriculture Service (FAS) indicate that there is expected to be a 7% decline in global orange production, and the United

States shows near a 350,000 ton reduction to 5.8 million tons. A major player in reduced orange yields worldwide is the impact of citrus greening. Although the orange production and consumption is declining, the lemon crop is estimated to grow by 10% (FAS). As with any food system there are multiple routes commodities can take depending on their initial quality and understanding how this system can be better utilized can lead to improved margins and better utilization of lower value raw materials. Over time citrus production has developed numerous value streams stemming from the processing of raw fruit ranging from direct consumption of fruit and fruit juices, use of essential oils for flavors, and the use of polysaccharides as functional ingredients. The key for each of these value streams is high quality flavor profiles, which are hard to understand given the aroma complexity and diversity of non-volatiles. There are numerous different approaches one can take to investigate citrus flavor, but the bulk of the research has long focused on the volatile fingerprint. In fact, even relatively recent reviews indicate that attempts to understand nonvolatile quality predictors have been largely unsuccessful, and only briefly touch on the contribution of limonin to the flavor profile of juices (Rouseff, 2009). One simply has to look at the prevalence of brix to acid ratio in citrus production to get an appreciation of how heavily the research has emphasized the volatile half of flavor and rarely looks outside of a few simple taste aspects. Although taste may be comprised of just 5 basic tastes it should not be relegated to the side, or summarized by a metric as simple as the brix to acid ratio.

The bulk of previous citrus flavor work has emphasized aroma and so this section will discuss the contribution of such and will tie in any available discussion of

nonvolatiles to the flavor profile. An area that does investigate the non-volatile flavor of citrus juice is the investigation of pulp on flavor release. Guichard et al. illustrated that there were significant impacts on flavor perception when different filtration methods were conducted on orange juice. When large particulate was retained there were significant amounts of terpenes and aldehydes while the non-retained particulate showed retention of ethyl butanoate and hexenal, compounds that have shown strong importance on orange juice flavor (Rega, 2004). This result was long supported by the different association of volatiles with different pulp sections (Radford, 1974). As there are often numerous and diverse macromolecules in a food system along with a diverse composition of tastants and aroma compounds, getting a better understanding of association phenomena is highly valuable. Beyond pulp there are flavonoids, carotenoids and liminoids, and all contain a significant amount of chemical variety. Flavones are a predominant flavonoid present in citrus products, often found as a glycoside and this conjugation will often dictate the flavor impact of the compound.

The bulk of nonvolatile work has focused on the aspect of delayed bitterness in citrus products, and many of these compounds stem from nomilin. When these liminoids are over abundant the consumer may perceive reduced quality and value through increased bitterness (Hasegawa, 1996., Rouseff, 1994). Liminoids have long been reported to contribute to the delayed bitterness in citrus juices and a number of approaches to limit their flavor contribution have been developed but are often expensive from an ingredient and processing perspective. Beyond delayed bitterness there is also bitterness contributed from flavonone neohesperidoses, which are perceived as

immediately bitter (Rouseff, 1994). This type of bitterness comes from a flavonone linked to a rhamnose and glucose, when linked 1-2 are bitter and when linked 1-6 are not bitter (Rouseff, 1994). Even with this understanding of flavor profiles it is often rare to specifically breed cultivars for flavor quality as aspects like horticultural vigor, disease resistance, fruit size and shape, seediness, peel thickness, color, and seasonal maturity often dominate (Rouseff, 1994). This indicates that the flavor is considered acceptable for market, but not the main driving force when selecting varietals. This is further supported by the lack of formal sensory panels run on new cultivars, and likely some of this is due to the amount of resources needed to develop and cultivate a large number of new crosses. Rather than running large consumer panels for liking, or smaller descriptive panels there is likely a single taster, often the breeder, or a small number of tasters that use consensus methods (Rouseff, 1994).

Outside of citrus research the investigation of non-volatiles on flavor profile is more established and works to understand phenomena beyond direct flavor activity. Understanding other phenomena is more challenging, as the presence of nonvolatiles shows an impact on the headspace composition of aromas in wines and model wine systems (Aronson, 2004., Mitropoulou, 2011, Rodríguez-Bencomo, 2011). It may be an over simplification to denote changes in perception to direct interaction, but this certainly has a contribution, and given the complexity of flavonols in citrus products there is very likely to be an impact on the aroma profile given changes in the non-volatile composition. One of these studies specifically targeted complex flavonol extracts of skin and seeds of grapes to present a more realistic flavonol composition to better understand

the role of phenolics on head space aroma concentration and showed significant differences in the headspace concentration of characteristic wine volatiles (Mitropoulou, 2011). Dufour and Bayonove (1999), showed the impact of catechin interaction with aroma release and the potential for hydrophobic interaction by investigation via proton NMR. Perez-Cacho and Rouseff (2008) discuss the importance of aldehydes, esters, terpenes, alcohols and ketones to the aroma quality of orange juice. As carbonyl compounds are critical to the aroma quality, changes in phenolic composition and concentration have large potential to impact the headspace concentration and therefore the sensory profile of citrus products, via the mechanisms discussed and through carbonyl trapping of aldehydes and ketones by phenolics (Totlani, 2006).

Expanding the discussion beyond orange products reduces the amount of literature present, there is again a larger amount of research conducted on the aroma aspect of lemon products, and very little about the taste. Researchers have long established the role of citral in citrus flavor profiles and how degradation of citral can lead to altered sensory profiles (Kimura, 1983). Researchers have just recently reported more comprehensive approaches to understanding lemon oil volatile and non-volatile differences between growing regions (Mehl, 2014). It should be pointed out that much of the data presented in the Mehl et al. work showed samples falling outside of the confidence interval of the models that were being used, ultimately these models may be capturing sources of variation outside of the researcher's hypothesis. The team also went to identify compounds relating to origin on citrus oils by use of metabolomics approaches to better understand how non-targeted methods can fit into the quality infrastructure of quality

control organizations (Marti, 2015). But these methods do not emphasize a sensory validation step, ensuring that the drivers for differentiation are the key reasons flavor quality differs.

The work done in the following chapters takes many of the presented ideas and develops a better understanding of how large chemical data sets can be mined for flavor active information. The following chapter will illustrate how preprocessing methods are optimized, and mined for minute chemical changes. Once optimized these methods are then applied to better understand aging in citrus systems and how flavor active information can be identified, and how value may be added through variable interactions.

Chapter 3. Development of analytical fingerprinting for untargeted analysis of non-volatiles.

Summary: This chapter highlights how chromatographic separations are investigated using multivariate modeling approaches (Projection to Latent Structures and Principle Component Analysis) to identify chemical changes at minute levels (i.e. part per billion). Preprocessing methods were optimized to produce highly selective and effective models. Altered data preprocessing methods were demonstrated to produce to different data structure size, utilization and ultimately to differing model quality and fit. Ultra Performance Liquid Chromatography-Mass Spectrometry was used to chemically characterize a citrus extract, while multivariate statistics were implemented to identify compounds that were added in at 5 and 10 parts per billion. Included is how preprocessing methods are adapted in order to generate informative but concise data structures that support effective multivariate investigation. Through optimization of preprocessing parts per billion levels of differentiation are possible, even when in the presence of complex comparisons.

Note: Parts of this chapter appear in: Abstracts of Papers of The American Chemical Society, vol. 247 (2015) titled “Application of untargeted LC/MS techniques (flavoromics) to investigate freshness of oranges.”

3.1 Introduction

As this dissertation reports the use of untargeted methods to understand food flavor, the foundation is converting the chemical fingerprint to a data frame. The conversion chromatographic-mass spectrometry techniques to a data matrix is called preprocessing. Preprocessing is a set of algorithms that takes a three-dimensional data structure (retention time, mass to charge ratio, intensity) and converts it to a two dimensional structure (retention time/mass to charge pair, and intensity). This chapter is focused on developing sensitive and selective preprocessing methods to characterize chemical changes at a minute level. More accurate and efficient preprocessing can lead to faster data analysis times, data structures with reduced noise, and data that better represent the true composition of the food. A number of methods exist to analytically characterize food, but liquid chromatography mass spectrometry methods are uniquely suited to produce large data quickly.

In this thesis untargeted methods were used to identify chemical changes related to aging of orange extracts, with the goal of flavor discovery. Ultra performance liquid chromatography mass spectrometry (UPLC-MS) is utilized to characterize chemical changes in citrus extracts while multivariate methods allow for statistical investigation of the chromatographic data. This platform is selected since it is able to rapidly characterize compounds that range from volatiles to non-volatile. UPLC-MS is not without limitation but produces large, dense and accurate data with a wealth of compounds for investigation and characterization. Although not a fully comprehensive study analytically, the collected data was investigated comprehensively. While attractive, adding additional data

streams, via a multi-block experiment increases data redundancy (e.g. positive and negative electrospray ionization) and will exacerbate the short comings of some modeling methods while reducing the statistical power of many modeling approaches. The researchers acknowledge that there may be limitations to the current study as far as scope, but by taking this approach more in-depth analysis of the collected data stream is possible. UPLC-MS methods are able to generate significant amounts of data and coupled to highly sensitive and fast scanning mass spectrometers can produce very large and information rich data sets. Instruments, like quadrupole Time of Flight (QTOF) mass spectrometers, are able to generate spectra with high mass accuracy (<3 ppm) while maintaining high mass resolution. These fast scanning instruments couple well with the short separations provided by UPLC and keep a large dynamic range, helping researchers reach new levels of sensitivity and consistency, providing a strong platform for characterizing the chemical makeup.

The optimization of preprocessing conditions for untargeted chemical fingerprinting methods is a challenging, since there is no variable to optimize. Often times this leads to "black-box" applications of commercially available preprocessing methods. Data quality is evaluated using a number of differing metrics, but are largely based on the speed of modeling and less on the quality of the model generated. A more controlled approach is to understand how the data leads to powerful models, and these metrics are: amount of data, completeness, concise representation, ease of manipulation, and relevancy. Other metrics investigate the time needed to analyze a data set and include accessibility, interpretability, timeliness and security (Pipino, 2002) however do not fit in the context of

the current thesis project. Given the data size and implementation of multivariate modeling at hand the speed of analysis is not as relevant as the data quality. This chapter will emphasize the amount of data, completeness, concise representation and relevancy of the collected data. In chemometrics the generation of more data is very simple, but the generation of relevant and concise data is much more obtuse. Optimizing data streams for an untargeted data driven experiment is a challenge since there are no outcomes to optimize to. Yet optimization of the data structure is critical to being able to effectively generate a data driven hypothesis. To achieve preprocessing methods that generate enough data to express the chemical complexity of the food system while doing so in a concise manner requires tuning. For this we can turn to unbiased and untargeted methods of investigation, and generate models using the data generated by various preprocessing conditions to identify the parameters that lead to high quality unsupervised models. If the data set generates a model that can achieve good data utilization, good model fit and prediction then the researcher can likely infer that the preprocessing achieved a good outcome.

The goal of the current study is to develop instrumental techniques that are able to monitor chemistry in the parts per billion ranges while reliably and robustly generating data and further mined for chemical changes. Once chromatographic data is generated it must undergo preprocessing before further investigation. Preprocessing converts spectral information into retention time and mass to charge ratio pairs (RT_m/z) with an associated intensity. Preprocessing is a critical step and is a step known for introducing bias (Bijlsma, 2006). Flavor compounds range in flavor activity and so analytics must be able to detect trace analytes and still reliably report the analytes. This chapter will

highlight how tolerances were established for the instrumentation, and how a statistical platform was generated around this data in order to optimize data preprocessing and create a workflow that could reliably detect minute chemical changes in the parts per billion range even in the presence of complex samples present.

3.2 Materials and Methods

Chemicals and Reagents. The following chemicals were obtained from the sources given: 200 Proof USP Ethanol (Fischer Scientific), >99% reserpine (Sigma Aldrich), UPLC Acetonitrile (JT Baker), Mass spectrometry grade Formic Acid (Fluka), NanoPure Water (Barnsted), 6 mL 1 g C-18 SPE tubes (Supelco), Methyl parabens (Sigma Aldrich), Ethyl Parabens (Sigma Aldrich), propyl Parabens (Sigma Aldrich), Butyl Parabens (Sigma Aldrich), #4 paper filers (Whatman), 3 kDa ultrafiltration membranes (Millipore). Fruit was purchased from local markets.

Model System:

Oranges were washed, rinsed, and then cut <5 mm thick. 500 g of citrus was extracted with 200 g ethanol for 24 hours, protected from light and purged with nitrogen. Ethanol was selected since it recreated a realistic fresh citrus character and is food grade. To this extract four compounds (methyl through butyl parabens) were added at five and ten parts per billion, the compounds selected are not native to food products. These were selected because the increasing hydrocarbon chain produces a range of polarity, providing us with an understanding of how well methods work for cleanup and concentration on a range of compound polarity.

UPLC-MS Analysis:

Sample Preparation: Samples were passed through a 3kDa ultrafiltration membrane (Millipore, MA) for improvement of chromatographic performance through removal of large molecular weight compounds. Solid phase extraction was performed using a 6 mL tube, 1 g packing C18 phase cartridges. Prior to loading the cartridge, samples were diluted to 10% ethanol. Cartridges were conditioned with acetonitrile (0.1% formic acid) and re-equilibrated with 5% acetonitrile/ 95%water before sample loading. Samples were eluted with 600 µL UPLC grade acetonitrile (0.1% formic acid) (JT Baker) and to this 400 µL Nanopure water (Barnsted, Waltham, MA) was added. The total sample concentration was 5 fold. Samples were filtered through 0.2 µm nylon filters (Millipore, MA) into 2mL auto sampler vials.

Chromatographic Analysis of Orange Extract Doping Experiment: A Waters I-class sample manager and binary solvent manager were coupled to a Waters Xevo G2 Q-TOF. A Waters BEH C18 (2.1 x 50 mm) was kept at 45°C in a Waters Column Manager with flow of 0.55 ml/min with linear gradient conditions (solvent A, Nanopure water 0.1% formic acid; Solvent B, Acetonitrile): solvent B at 3% (0-0.5 minutes), increased to 30% (0.5-8 minutes), increased to 60% (8-16 min), followed by column wash (100% acetonitrile) and re-equilibration.

Sample Preprocessing Doping Experiment: Sample processing was done in a manner to ensure sensitivity for minute chemical differences. Optimization evaluated five different levels of thresholding (minimum intensity for a peak) and three levels of noise elimination. These processing methods produced a range of chemical features for modeling between 917 and 21,434 compounds. The quality of the data set is entirely driven by the

preprocessing applied. Chromatograms were processed from 0.5 minutes to 16 minutes using waters MarkerLynx software. Mass range was set to 100-1000 m/z and the mass step was 0.01 dalton. Noise elimination was evaluated at 1, 10 and 50, while intensity count threshold holding was evaluated at 100, 500, 1000, and 5,000. Signals were included if they were in >75% of samples analyzed. Normalization was done to peak area.

Chromatographic Analysis of Orange Extract Doping and Aging Experiment: Although there was a high quality separation in the initial separation reducing the length of runs dramatically decreases the time needed to run longer sample lists and so the method was adapted to produce shorter analysis times. A flow rate of 0.55 mL/min was used with initial gradient conditions of 3% acetonitrile (ACN) and 97% Water (0.1% Formic Acid), which was held for 0.5 min and raised ACN content to 15% at 1.5 min, 45% ACN at 8 min followed by a 1 min column wash (100% ACN) and re-equilibration. Electrospray Spray Ionization (ESI) was run in negative mode with source temperature of 120°C, the desolvation gas was run with a flow of 600 L/hour at a temperature of 400°. The reference compound was reserpine and 6 traces were used for mass correction per injection, each reference scan was collected for 0.5 seconds and with a capillary voltage of 3 kV. Instrument was checked and validated to operate at a mass accuracy lower than 2 ppm. Each sample was injected 5 times in randomized blocks. Sample preprocessing was done from 0.5 to 8 minutes.

Sample Preprocessing Doping and Aging: Sample processing was done in a manner to ensure sensitivity for minute chemical differences. Preprocessing optimization was done using a two factor uneven level design, with 4 levels for intensity threshold and three levels

for noise elimination. These parameters were identified as the most impactful on preprocessing quality and also the most dynamic conditions. These processing methods produced a range of chemical features for modeling between 55,131 to 515 compounds. The importance of data preprocessing is clear. The quality and depth of the data set is largely driven by the preprocessing applied. Chromatograms were processed for the respective times listed above for each experiment. Mass range was set to 100-1200 m/z and the mass step was 0.01 dalton. Noise elimination was set to 10, intensity count thresholding was set to 500, the mass window match was set to 0.1 dalton and the retention time window was set to 0.1. Signals were included if they were in >75% of samples analyzed. Normalization was done to peak area.

Development of preprocessing conditions:

Preprocessing optimization was done using a two factor uneven level design, with 5 levels for intensity threshold and three levels for noise elimination. These parameters were identified as the most impactful on preprocessing quality and also the most dynamic conditions. Within the MarkerLynx (Waters, Millford MA) there are a number of method parameters that are adjustable, and many that are set by the instrumental parameters. In this experiment the retention time window (0.1), mass step (0.01), mass window match (0.1), and retention time window (0.1) were all fixed at the respective values. The intensity threshold was evaluated at the levels 50, 100, 1000, 5000, while the noise elimination was evaluated at 1, 10 and 50. This produced data sets ranging from 55,131 compounds to 515. Table 3.2 showed the number of compounds identified given each set of preprocessing conditions.

3.3 Results and Discussion

After the data collection preprocessing of that was done to convert the data to a matrix for multivariate modeling. In this work the tunable parameters of threshold (minimum intensity count) and signal to noise (level of identified signal relative to surrounding noise) were adjusted to create data of differing quality and size. Figure 3.1 reported the impact of threshold value and signal to noise ratio on the amount of variables

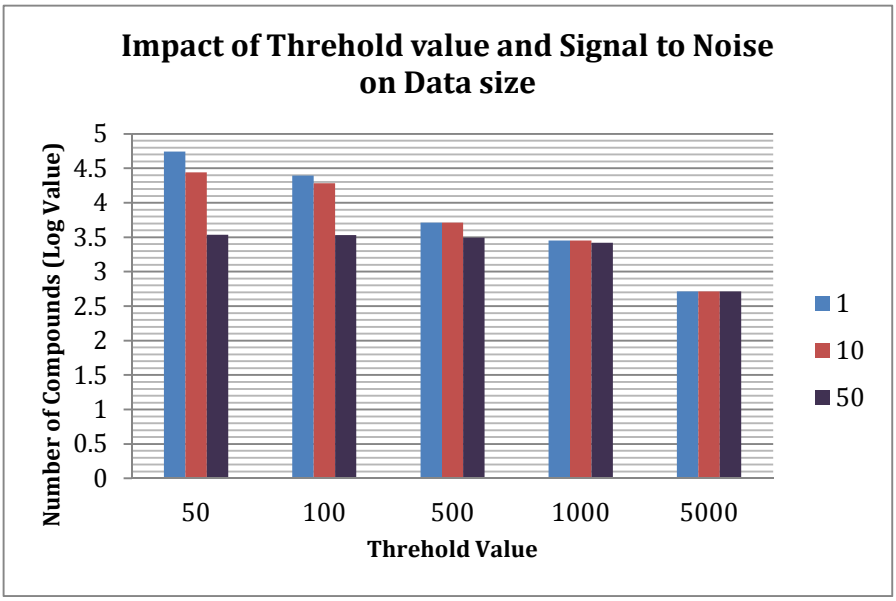


Figure 3.1: Impact of Threshold value and Signal to Noise ratio in variables generated frame

collected, a log scale is used to make the plot more interpretable. A change in threshold value leads to an impact based on the signal to noise value until the threshold value reaches above 1,000 then only the higher signal to noise ratios reduce the number of compounds. Once the threshold value reaches 5000, signal to noise has minimal impact to the data generated. This is indicative of two things, at a certain intensity the algorithm identified signals above background noise. Knowing this value can helps to understand the

approximate signal to noise average across a data set. Secondly, signals at very low thresholds can be introduced to the data set as long as the signal to noise ratio is at a level ensuring signal is included while noise is kept to a minimum. Lower preprocessing thresholds leads to more noise generation but also include lower intensity analytical signals. Including more trace analytes can lead to new discoveries and so is desirable in untargeted analyses.

Ensuring that the data generated by an UPLC-MS system can produce signal for trace analytes while limiting noise is a challenge. Separations must balance information

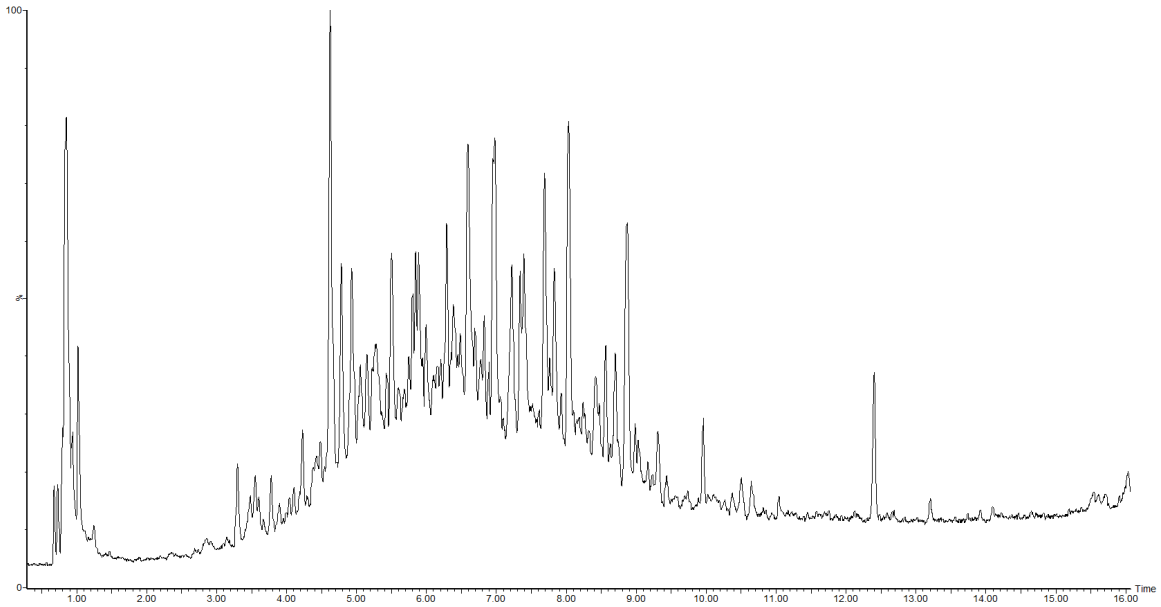


Figure 3.2: Total Ion Chromatogram of the orange model system showing the time scale used for preprocessing. The chromatogram shows a number of sharp Gaussian peaks from 3-10 minutes. 2x2 smoothing applied.

generated with the analysis time. Since longer analysis times are compounded over numerous injections and lead to increased sample preprocessing time along with increased cost more rapid analysis is desirable. Shorter analysis times also lead to sharper peaks, which allows for better filtering of poorly performing peaks (broad or non-

Gaussian), but longer separations leads to less co-elution. To better understand how a workflow is performing being able to model chemical changes (at minute levels) is desirable, but often challenging since not all compounds are known. To understand how models can identify changes in complex sample, known compounds were spike into the citrus extract at 5 and 10 parts per billion (µg/L). Evaluation of this chromatographic data establishes how well the workflow performed and whether the UPLC-MS data could effectively be classified based on the spiked compounds using multivariate methods. The initial gradient was generated and can be seen in figure 3.2. Initial investigation was done with Principle Component Analysis (PCA) and depicted tight grouping and statistically distinct clusters of samples (Figure 3.3). Principle component analysis allows for an unsupervised investigation and identification of any observation outliers as well as a better understanding of the major chemical diversity of the data set.

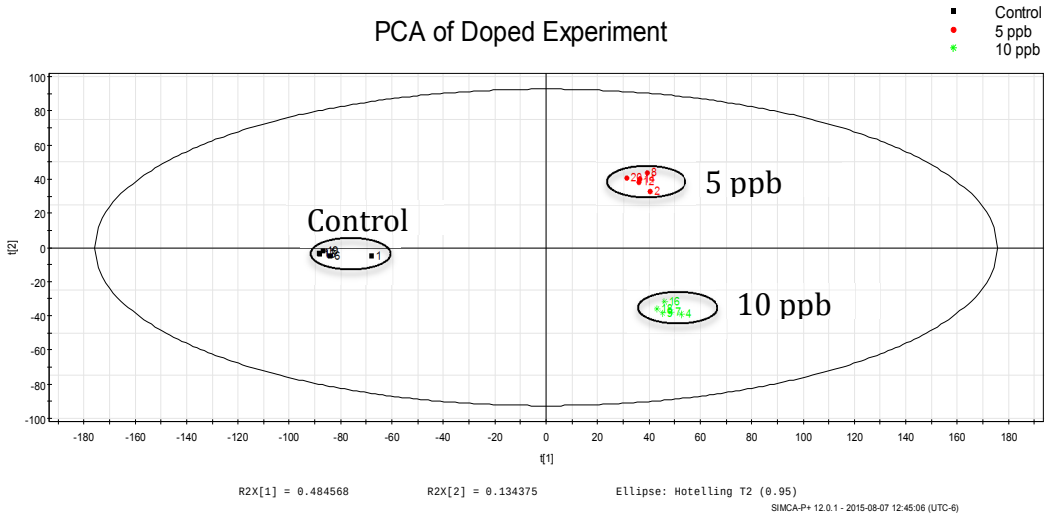


Figure 3.3 Principle Component analysis of the initial preprocessing of the doping experiment, showing excellent separation of the doped and control samples on the first principle component and separation of the doping level on the second principle component. Illustrated is sample preprocessing conditions with a 500 threshold and a S/N ratio of 10.

Figure 3.3 clearly suggested both selective and sensitive instrumental analysis, data preprocessing and statistical modeling. There is clear distinction between doped samples (5 &10 ppb) and the control. Presence of doped compound is captured along the initial principle component of the model, which is the X axis in this plot. There is also further statistical separation between the 5 and the 10 parts per billion samples via the second component, indicating the level of addition lead to the second principle component being identified. This initial investigation using an untargeted PCA on the X block of data allowed for investigation of chemical changes at a minute (bbp) level and showed differentiation of presence of added compounds along the first principle component, and the level of addition along the second principle component. The model metrics suggested that the model is over fit (R^2X=0.452, Q^2=0.265) and so PCA is not the only warranted investigation. With significant overlap in the composition of the samples, the model is fitting noise to drive classification. Any statistically significant features beyond the added compounds arise from errors in instrumental analysis, variance during sample preparation, or random noise. Alternatively, PCA will often show significant differences even using randomly generated numbers as data, so it should not be the final model of any analysis (Daszykowski, 2006). To better investigate what variation the model is capturing Projection to Latent Structures models were fit to understand the systemic variance between the sample class and chromatographic data (Model Fit Metrics: R^2X= 0.599, R^2Y= 0.993). To delve deeper into the data the Variable of Importance (VIP) for the projection to latent structures is extracted in order to identify the

top contributing compounds. Table 3.1 illustrated that the four compounds added in all had a statistically significant VIP, and were all in the top variables for VIP. The expected

Table 3.1: Top PLS compounds stemming from the Variable of Importance metric show the added compounds were the most statistically powerful reasons differentiation was achieved

Expected Retention Time	Expected m/z	Compound ID	Retention Time	Mass	VIP
10.37	179.0009	10.37_178.9949	10.37	178.9949	1.59713
8.66	165.0028	8.66_164.9838	8.66	164.9838	1.59426
11.87	193.0014	11.87_193.0104	11.87	193.0104	1.5915
6.91	151.0156	6.91_150.9717	6.91	150.9717	1.58538

retention time and m/z were generated from injection of the mixture of paraben compounds directly injected with the same gradient.

Consequently the established analytical platform was demonstrated to model true chemical differences with a low expression of systemic variance, in the complex food system. This suggested an expression of system stability and post data collection performance as the models generated were high performing as far as identifying true chemical differences even with a high amount of chemical complexity. Knowing the range of preprocessing conditions that led to appropriate data structures and supports the utilization of this analytical platform for comparisons.

Introduction of new samples and new varietals will require new tuning of the preprocessing conditions in order to address the new chemical complexity, so the preprocessing methods were validated with new samples. Since changes in sample composition will always require new preprocessing conditions it can be assumed that within similar food matrices and extracts preprocessing are likely to have similar parameters. The analytical gradient was shortened to increase sample throughput and reduce solvent usage (37.5% increase in throughput), and so new preprocessing conditions were generated. The next section illustrates how a preprocessing method was further optimized with a shorter gradient (Figure 4.3).

Analysis of Orange Extract Doping and Aging Experiment:

Although the previous experiment demonstrated accurate classification of observations using a small number of compounds of interest (Figure 3.3), the amount of systemic variation seen in more diverse experiments can easily over leverage this minute chemical variance. In order to better understand how the analytical framework will operate with added chemical complexity, samples with more chemical diversity must be built in. Since this dissertation has the goal of understanding flavor change with time, a sample aged for 24 hours was introduced. The experiment was re-run with the faster gradient (Figure 3.4) and the aged samples were analyzed. The aged samples were expected to maintain much of the chemistry present in the fresh samples, but add chemical complexity stemming from changes that have happened during aging. By understanding how preprocessing conditions generated different performing data sets

allowed for rapid generation of conditions that produced selective models that were sensitive and selective.

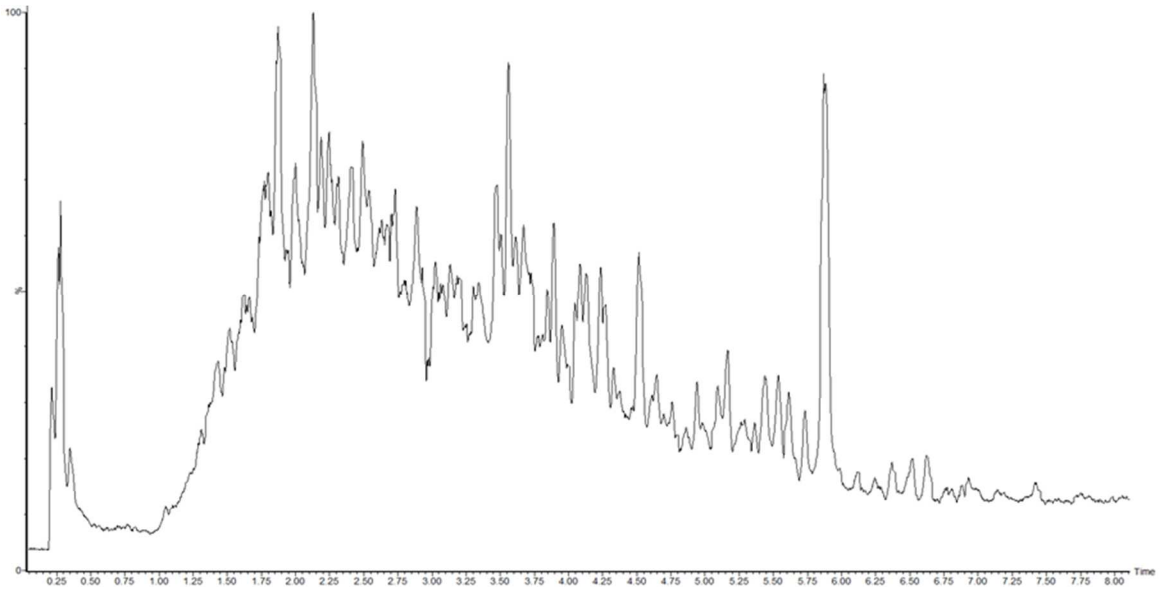


Figure 3.4: Total ion chromatogram from the shortened gradient for the selected preprocessing range. 2x2 smoothing applied. Shows high information density and good peak shape across the majority of the separation, but co-elution is present.

The previous section reported that the preprocessing conditions produced models that effectively modeled and identified compounds at low parts per billion levels the preprocessing was revalidated since there were new chromatographic. Signal to noise (S/N) at levels: 1, 10, 50 and threshold at levels: 50, 500, 1000, 5000 were varied to generate data frames of differing size. Turning chromatographic signals into data frames is not as simple as generating large data structures, the data should be highly utilized and generate quality models. Preprocessing methods may identify a signal, but populate it with an intensity factor of zero if the signal is not reported consistently within the sample group. This allows for an approximation of the noise level identified by the preprocessing conditions as signal, and further the sparsity of a data set. Illustrated in table 3.2 is the data density and utilization for selected data frames in a PCA application.

Table 3.2. Impact of preprocessing conditions on data frame sparsity and data usability, varied pre-processing conditions led to dramatic changes in data frame size and density

Selected Preprocessing Conditions (Threshold-Signal to noise)	Number of variables generated	Variables usable for model	Percentage of data unusable
50-1	55153	24229	56.06%
500-10	5180	3793	26.77%
1000-10	2839	2171	23.52%
5000-50	515	407	20.97%

Unused data occurs when preprocessing identifies a signal that is reliably replicated in the sample grouping, and is populated with a null value or has no change between observations. Either of these outcomes contribute noise in projection methods. Ultimately it is not the number of data points used that dictate model quality, it is how explanative and

predictive the data frame is. Generating unsupervised models for each of the varied conditions helps to better illustrate how informative a data set is, by having high data use and well fitted PCA models there is increased likelihood that the preprocessing conditions are well preforming. This is illustrated by the PCA generated for data frames; threshold 50- S/N1, threshold 500- S/N 10, threshold 5000- S/N 50 and the associated model quality metrics$R^2(X)$ and Q^2. $R^2(X)$ is the quality of the model at explaining the X block of data, while the Q^2 is the metric of how well the model predicts the X-block of data. These analytics provide insight into how well a model is at representing the data, independent of normal loading plots. Each of the PCAs is generated with multiple iterations (permutation) to ensure that there is no order bias when model generation takes place, and further validated with leave one out cross validation.

Figure 3.5 reported the loading plot from the Threshold of 1 and signal to noise ratio of 1, this data set generated over 55,000 variables but used less than half of them (Table 3.2). This suggested that the model is identifying noise as reasons for differentiation of the samples. Investigating the fit coefficients the R^2 is 0.525 and the Q^2 is 0.298. With a difference of 0.227 between the R^2 and the Q^2 the model indicated that it is somewhat over fit. When a difference between 0.25-0.3 between the Q^2 and R^2 is observed, this indicated an over fit model especially coupled with a low Q^2. Furthermore, an investigation of the variable contribution to the model reported that 4,304 variables negatively contribute to the Q^2 value, which illustrated that 17% of used compounds are detrimental to the model fit. So although the loading plot reported, what appeared to be a highly performing model, review of the modeling metrics show it is not performing in the way a researcher can get

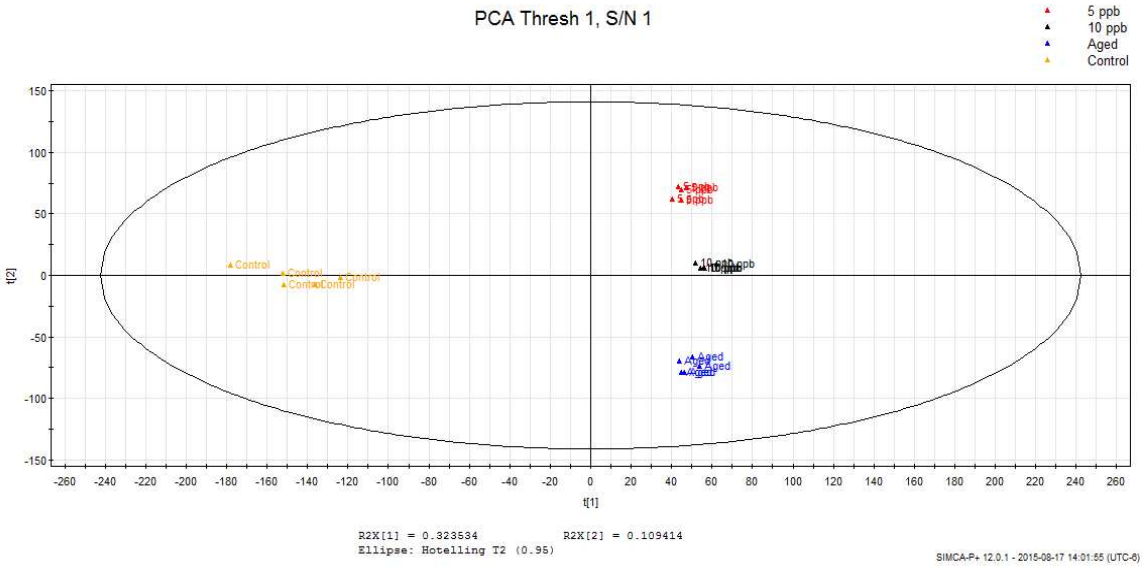


Figure 3.5: Principle Component Analysis of data structure generated with a S/N of 1 and threshold value of 1, loading plot shows good visual separation of samples.

usable information. There are thousands of compounds contributing against model quality and at the same time the model showed a low level of prediction (Q^2).

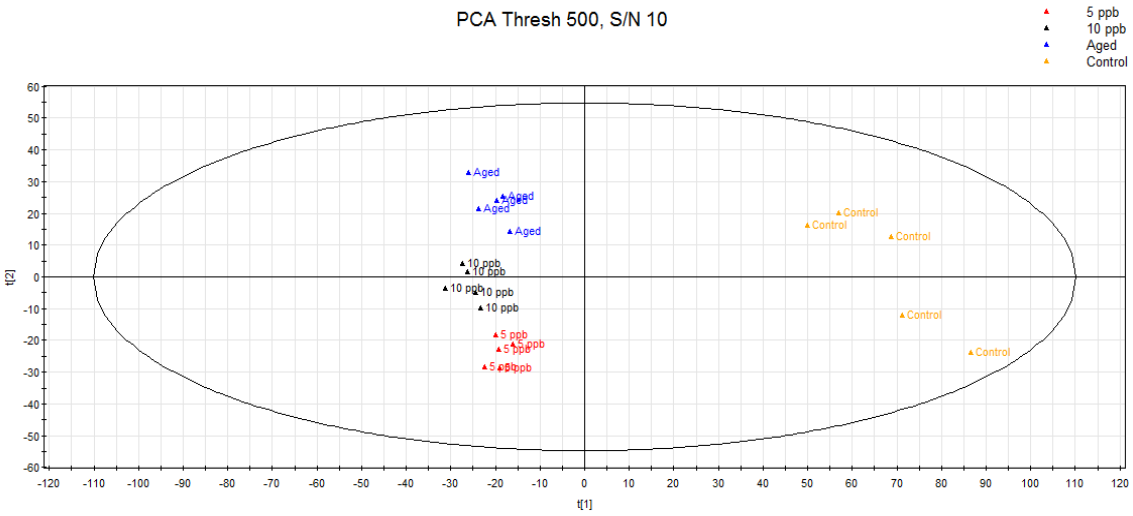


Figure 3.6: Principle Component Analysis for data generated from the a threshold of 500 and Signal to Noise of 10, reported classification but is suggested to be picking up sample or instrumental variance in principle component 2.

When utilizing a threshold of 500 and signal to noise of 10 graphically produced the model (figure 3.6) that has 'looser' clustering than the PCA from a threshold 1-S/N 1 preprocessing (Figure 3.5), but reported a better model quality. With an R^2 of 0.61 and a Q^2 of 0.39, the data set is producing a model that was better at representing the data structure and therefore better at prediction. Coupled with the fact that the data utilization is much higher, with only 26.7% of the data populating as noise, indicated the model has a much higher quality than the threshold 1- 1 S/N model (Figure 3.5). Although the difference between the R^2 and Q^2 was 0.222 which is lower than the previous model, it is

only slightly lower but did have a better Q^2 than the 1:1 model. Additionally, the 500:10 model has a number of attributes that detract from the Q^2, with fewer than 600 variables detracting from the prediction metric. With an increase in data utilization and improved quality metrics the model is more informative of the true differences in the data stream. The looser grouping of the samples likely stem from more accurate representation of systemic variation, and it is important to rely on the model metrics rather than the geometric representation of the model.

Figure 3.7 reported the PCA from preprocessing of threshold 5000-S/N50, the loading plot indicated that the model can classify the difference between the control and the other samples, but the model is unable to capture the chemical variance that allows

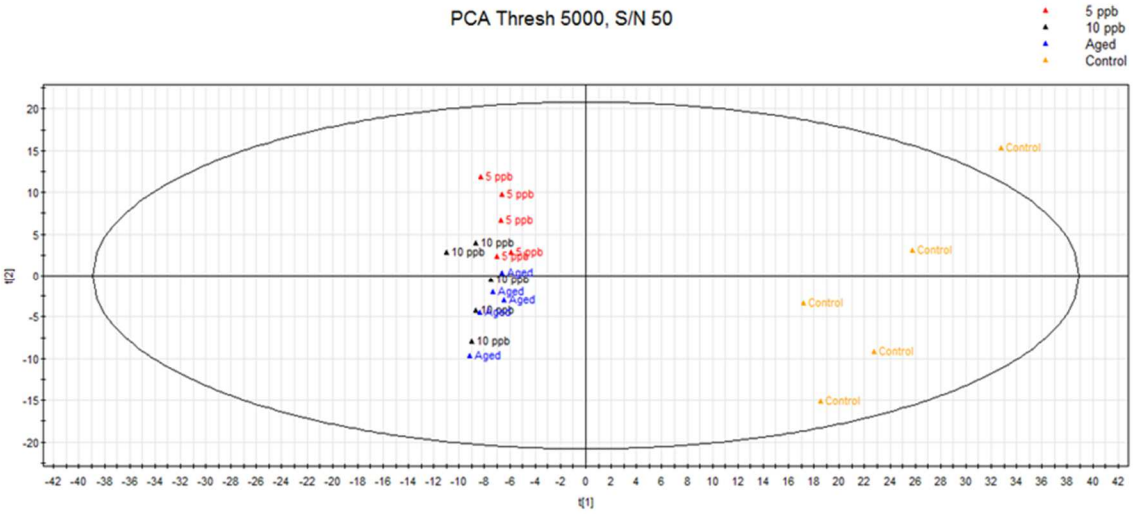


Figure 3.7: Principle Component Analysis for data generated from the threshold of 5000 and Signal to Noise of 50, reported poor classification and is picking up noise so has filtered out chemistry that related to chemical differentiation.

for differentiation between the aged and 10 parts per billion samples in the second principle component. The model has the highest R^2 of 0.638 and Q^2 of 0.508, so although the model accurately represents the data frame, it is unable to model the

research question at hand. The first principle component captures the difference between the endogenous extract and the treatments. The preprocessing conditions used likely eliminate the analytical signal for the doped compounds, and may obscure some of the minute chemistry changes during the 24 hours of aging. As the preprocessing threshold of 500 and S/N of 10 provided a data structure that had an appropriate depth (5180 variables) and captured the research question at hand (0% class error), while still effectively filtering out noise this data frame was used to generate a projection to latent structures model. These conditions should translate well to supervised modeling. PCA provides insight into how pre-processing can lead to different modeling qualities, but it is

a preliminary investigation technique and a PLS model will produce models with more

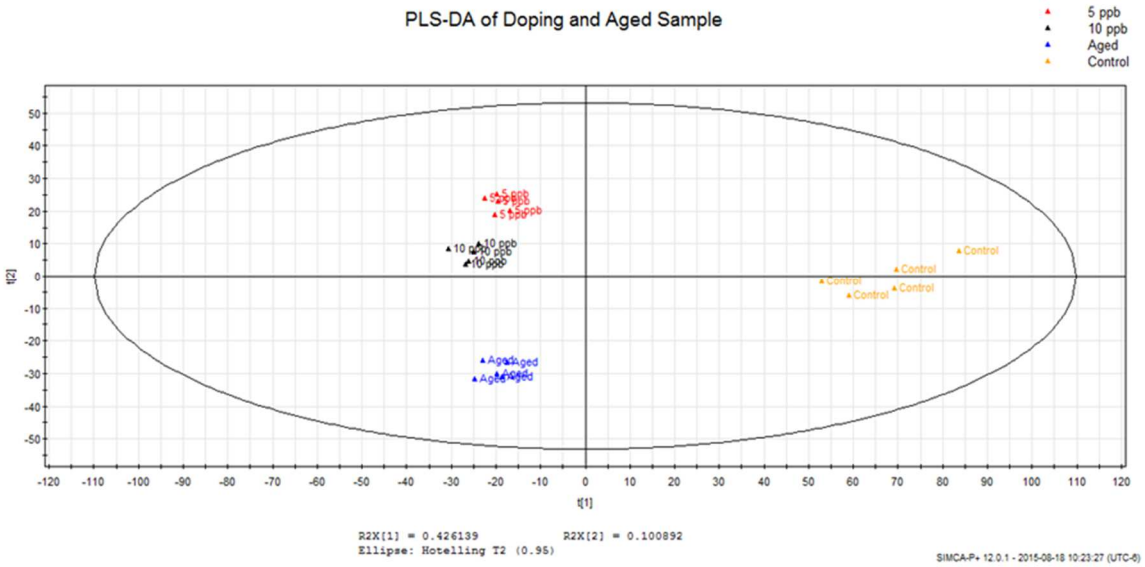


Figure 3.8: Reported the loading plot from the Projection to Latent Structures model that is generated using preprocessing conditions with a 500 threshold and 10 signal to noise level.

discriminatory power.

The data was fitted with a PLS model and showed distinct classification of the control from the other sample groupings, with this being captured along the first principle component (Figure 3.8). This suggested the model is producing explanative models on presence of chemical differences versus the control as the doped samples and the aged sample are both differentiated along the first principle component, with this principle component capturing 46% of the X block variation and 33% of the Y block variation. Further the second principle component differentiates the doped samples from the aged sample indicating that the model is further differentiating the aspects those samples are different from the control, this component captures 10% of the X block variation and 32% of the Y block variation. The model has an RX^2 of 0.671, an RY^2 of 0.993 and a Q^2

of 0.944, indicating high model quality with large explanative and predictive power. The RY^2 supports the graphical representation of the model, and how visual classification translates to statistical validity. To further represent this, a classification table is generated to illustrate how well excluded observations classify into the existing model, class error reported 0% in table 3.3.

Table 3.3: Classification table from PLS model of aging and doping experiment (Figure 3.8), which validates model classification.

Sample Class	# of Observations	Number Correctly Classified	Number Incorrectly Classified	Unknown	Classification Error
Control	5	5	0	0	0%
5 ppb	5	5	0	0	0%
10 ppb	5	5	0	0	0%
Aged	5	5	0	0	0%

This confusion matrix reported the results of leave one out cross validation, indicating that the 5 and 10 parts per billion samples are still classifying as statistically different. This experiment indicated using the faster elution gradient (Figure 3.4), more complex sample group models that differentiate between 0, 5 and 10 parts per billion samples were possible while still explaining the more complex chemical diversity of the sample aged for 24 hours. This indicated that the models generated are both sensitive to complex chemical changes while maintaining selectivity base on small changes in the chemical composition. In order to generate applicable data-driven hypothesis models must be sensitive to diverse sample composition but also to minute chemical changes.

This is especially critical to the understanding of flavor systems, due to the range in concentration and sensory activity. Each constituent of a flavor does not have the same activity and so concentration does not dictate the importance or contribution to the sensory profile. This is also complicated by the fact that when researchers establish flavor activity values for compounds there is a wide range of methodology and the flavor attribute is contextual to the food system it is being tasted in.

3.4 Conclusion

This chapter reported preprocessing tools were used for generating data frames from chromatographic data in a way that produces sensitive, selective and powerful models. In order to appropriately model chemical data and identify minute changes in chemical information, preprocessing plays a very important role in this process. Often times preprocessing software packages are treated like black boxes, and in much of the literature the optimization is minimal, or not discussed. Developing the preprocessing parameters and knowing their impact on the data frame and model quality leads to a higher quality data driven hypothesis. This leads to a stronger statistical foundation for variable selection and elimination, which is critical in data driven experiments. In this chapter different preprocessing conditions were manipulated to illustrate how changes in parameters lead to changes in data size, data density, data utilization and help to show how all of these aspects impact model quality and further usability. Simply because a preprocessing program outputs a data frame does not mean that it is a high quality data frame addressing the research question. Experiments that identify how minute chemical differences are seen in data help to develop a untargeted workflow. Optimization of this

work flow helps researchers understand and identify minute chemical differences within complex and diverse experiments. In this experiment the threshold of 500 and S/N of 10 produced a well fit model (R2Y=0.993, Q^2= 0.944) with good data depth (n=5180), data usage (usage = 73.23%) and so was selected as the preprocessing conditions for future experimentation.

Chapter 4: Application of untargeted methods to identify compounds relating to aging in citrus extracts

Summary: The objective of this chapter was to model chemical aging in ethanol extracts of three varietals of orange fruit using untargeted chemometrics and to further screen the chemical changes for flavor activity. Employing optimized sample analytics as detailed in Chapter 3, and implementing both multivariate and machine learning approaches, compounds of interest were identified, isolated, purified and evaluated for their sensory impact in a model orange juice system. Statistically significant compounds showed a range of flavor impact on the tasting medium using a descriptive analysis panel, three of the evaluated compounds were positively identified. Nomilin glucoside and two novel compounds were identified through nuclear magnetic resonance and mass spectrometry. The identified compounds all positively correlated with age. Nomilin glucoside showed suppression of orange character. Compound 383 showed suppression of orange character, enhancement of cooked and green bean character, and suppression of floral character. Compound 661 showed suppression of floral character.

4.1 Introduction

Flavor analysis has long used targeted methods to analytically characterize flavor compounds within a food system. Targeted methods largely focus on identifying compounds based on their individual contributions, largely focusing on aroma attributes while overlooking the non-volatile portion (Nisperos-Carriedo, 1990, Rouseff, 2008). When the non-volatile composition is evaluated it often only employs a basic investigation rarely extends beyond identification of compounds that directly activate the basic tastes (e.g. Overlooks modulators). As there are thousands of chemical species in food systems this reductionist approach is logical, but has limitations. New methods that address prior analytical challenges are desirable and can expand current knowledge. Due to the chemical complexity surrounding food flavor Reineccius (2008) discussed the approach coined "Flavoromics" as a new frontier in flavor research by using some of the proven workflows seen in metabolomics and other "omics" fields. Although this analytical technique was introduced a number of years ago, it has only been used by a small number of prior studies to identify flavor compounds (Charve, 2011). Of the available literature none have conducted sensory recombination studies, as flavor holds unique challenges. To date, flavoromic studies have been limited in the lack of sensory validation of identified compounds as well as the development of data handling (Gracka, 2015).

The current study further examines the application of untargeted methods to identify compounds that contribute to the changes seen during the aging of orange extracts. The workflow utilizes both multivariate and machine learning techniques with

further isolation and sensory evaluation of statistical compounds. As far as these researchers understand there has been no prior works that established causation through isolation, purification and recombination based sensory validation. This chapter will illustrate a more comprehensive Flavoromic workflow that includes analytical fingerprinting, isolation, purification and recombination validation in order to establish a causative flavor impact of the compounds that illustrated a statistical significance.

4.2 Materials and methods

Matrix Preparation:

Navel Oranges, Mineola Tangerines, and Valencia Oranges were washed, rinsed, and then cut to be <5 mm thick. 500 g of citrus was extracted with 200 g ethanol for 24 hours, protected from light and purged with nitrogen. Ethanol was selected since it reduced biological and enzyme activity, recreated a realistic fresh citrus character and is food grade. The extract was passed through a whatman #4 filter and was either frozen (-80°C, time=0) or further aged (time=2, 4 and 6 days) to yield appropriate treatments. Samples were passed through a 3kDa ultrafiltration membrane (Millipore) for improvement of chromatographic performance through removal of large molecular weight compounds. Solid phase extraction was performed using a 6 mL tube, 1 g packing C18 phase (Sigma Aldrich), 5 mL of sample was diluted to 10% ethanol before loading onto the cartridge. Samples were eluted with 600 µL UPLC grade acetonitrile (JT Baker) and to this 400 µL Nanopure water (Barnsted, Waltham, MA) was added.

UPLC-MS Conditions:

A Waters I-class FTN sample manager and flow binary solvent manager were coupled to a Waters Xevo G2 Q-TOF. A Waters BEH C18 (2.1 x 50 mm) was kept at 45°C in a Waters Column Manager. A flow rate of 0.55 mL/min was used with initial gradient conditions of 3% acetonitrile (ACN) and 97% Water (0.1% Formic Acid), which was held for 0.5 min. A linear gradient raised ACN content to 15% at 1.5 min, 45% ACN at 8 min followed by a 1 min column wash (100% ACN) and re-equilibration. Electrospray Spray ionization was run in negative mode with source temperature of 120°C, desolvation temperature of 350°C, capillary set to 1.75 kV, sample cone of 25 V, TOF scan range was 100-1200 m/z, with lock mass corrected automatically. The reference compound was reserpine and 6 traces were used for correction per injection. Each sample was injected 5 times in a randomized block design. Injection volume was 1 µL, with randomized blank and standard injections added to each randomized replicate block. Chromatographic drift over the entirety of the run was found to be less than 0.2% for each of the standard peaks, and less than 5% for peak area for each standard.

Preprocessing:

Preprocessing of the UPLC-MS data was done using Markerlynx software (Waters, Milford, CT). Preprocessing conditions started with methods identified in previous chapters and developed to produce good data frames (Chapter 3). Peak detection was performed between 0.2-8 min, to exclude column wash and dead volume, m/z range was 100-1200 m/z using a m/z step of 0.01, noise elimination was set at 10 and a threshold of 500, and spectra smoothing was applied. Peaks were matched if the retention time was within 0.1 min, established by the variation observed in elution of

standard runs across analysis, the mass within 0.03 m/z and the peak detected in four of five injections. Peak lists were exported in .csv format for further analysis. Unit variance scaling was applied. Variables with a coefficient of variation of zero (no significant change) were eliminated, to reduce model over fitting and reduce time required to generate models.

Model Generation

Principle component analysis and projection to latent structures (PLS) models were generated using SIMCA-P+ 12 (Umetrics, Umeå, Sweden), random forest (RF) models were generated using the R (R v.3.0.1 "Good Sport", University of Aukland) package "randomForest". Data was divided into training and test sets (70% training, and 30% test), and was sampled randomly. Sample groups were identified by age of extract and binning varieties across an age point. Before analysis PCA was used to screen for outliers in data sets to prevent over leveraging. PCA and PLS models went through permutation testing and Leave One Out Cross Validation (LOOCV) to ensure model quality. For random forest generation, the forest depth was optimized based on minimization of the classification error. Model optimization produced a model with 2.38% out of bag error using 110 trees and 110 variables tried at each split. Forest depth and variable tried at each split was piecewise optimized for model quality and speed of analysis. An importance plot provides insight into the variables that provided the most powerful leverage into classification of the samples.

First Dimension LC-MS Directed Fractionation

A Shimadzu 10ADVP LC system was coupled to a Waters QuattroMicro triple quadrupole mass spectrometer with the flow split to a Waters Fraction Collector III. Compounds were isolated using a m/z trigger in Single Ion Monitoring (SIM) mode (383, 695, 191, 337, 457, 413, 563, 661, 915, 148, 295 m/z), Multiple-Reaction-monitoring, retention time on the column and ion intensity. First Dimension separation was conducted on a Waters Xbridge Prep Shield RP18 Column (10X250 mm, 5 μm particle size) held at 40°C injections were 350 μL. Initial gradient started at 5% Methanol (0.1% Formic acid) and 95% water (0.1% formic acid) and held for 3 minutes, ramped to 30% methanol at 10 minutes and 100% Methanol at 38 minutes, 100% methanol was held until 41 minutes and then re-equilibrated to 5% until 45 minutes. After isolation, methanol was removed using a rotary evaporator and further lyophilized twice to ensure sub parts per billion levels of residual methanol. Isolates were re-injected into UPLC-MS to ensure retention time and m/z match.

Second dimension purification:

Second dimension was done with a Phenomenex Phenyl-Hexyl Luna (10x250 mm, 5 μm particle size). Gradient varied based on the compound of interest.

Compound 383.114. Initial solvent make up was set to 5% methanol and linearly ramped to be 25% methanol at 3 minutes, 70% at 29 minutes, 100% at 31 minutes, followed by column wash and equilibration at 5% methanol for a total run of 36 minutes.

Compound 695.283. Initial solvent make up was set to 5% methanol and linearly ramped to be 15% methanol at 3 minutes, 40% at 10 minutes and 100% at 30 minutes, followed by column wash and equilibration at 5% methanol for a total run of 36 minutes.

Compounds 413.121E1 & E2 and 661.265. Initial solvent make up was set to 5% and linearly ramped to 30% at 3 minutes and 80% at 29 minutes, followed by column wash and equilibration at 5% methanol for a total run of 36 minutes.

Compound 457.256. Initial solvent makeup was 10% methanol and linearly ramped to 60% at 3 minutes, 70% methanol at 29 minutes, followed by column wash and equilibration at 5% methanol for a total run of 36 minutes.

Third Dimension Separation. As some separations were not able to fully separate isolated fractions with adequate purity a Waters Charged Surface Hybrid flouro-phenyl column was used.

Descriptive analysis panel:

Panelists were recruited from the Department of Food Science and Nutrition at the University of Minnesota, panelists had previous experience on descriptive analysis panels. Panelists evaluated taste attributes and retronasal character of the provided coded samples. A citric acid scale was provided for cross modal matching. Samples were evaluated using a 10 point scale, anchored on the left with "Low" and "High" on the right. Samples were evaluated in duplicate for each session. Panel run under the University of Minnesota Internal Review Board approval number 1505E70948.

Two mediums were used for the tasting base:

- One tasting solution was a Solvent Assisted Flavor Evaporation (SAFE) (Schieberle, 1999) aroma isolate of a commercial orange juice diluted with 5% food grade 200 proof ethanol. This method recovers the volatile portion of the juice and the tasting solution is made using this isolate with sucrose at 8% and

citric acid at 0.05%. This panel evaluated 5in this medium, by 8 panelists who went through 8 hour long training sessions.

- The second tasting medium was a commercial water soluble volatile orange flavor (VOF) with sucrose added at 8% and citric acid added at 0.05%, 7 compounds were evaluated with this medium by 9 panelists who had completed 10 hour long training sessions.

The panelists were asked to evaluate samples for the taste attributes: sweet, sour, bitter and the feeling of astringency. The panel also evaluated the retronasal character of the sample for the attributes: orange character, orange peel, cooked, floral and green bean. The panel was provided with a citric acid scale for cross modal matching of the intensity. Table 4.1 illustrated the references provided to the panel during training and testing. Compounds levels were as follows: Compound 383 1.2 mg/L, Compound 413E1 13 mg/L, Compound 413E2 12 mg/L, Compound 661 .35 mg/L, Compound 693 46 mg/L, Compound 457 0.40 mg/L. These were levels that were estimated to be found in food systems. A calculation was done based on the amount of food matrix used for fractionation and accounts for an estimated 20% losses in transfers and fractionation inefficiencies. Descriptive testing was done using Compusense Cloud, and samples were presented in a randomized order in duplicate, over two sessions.

Data was collected with Compusense Cloud Software (Compusense Inc., Guelph, Ontario, Canada) and exported in .csv for analysis through R. Analysis was conducted

using an initial two way-ANOVA and further post-hoc analysis via Dunnett's test with the sample blank (tasting base) as the control sample (Appendix II).

Table 4.1: Attribute and provided references for the descriptive panel

Attribute	Provided Reference
Sweet	Sugar Solution 8%
Sour	Citric Acid Solution
Bitter	
Astringent	
Orange Character	Not-from concentrate orange juice
Orange Peel	Freshly cut navel orange peel
Cooked	From concentrate orange juice
Green (grassy, herbal like)	Freshly minced grass
Green Bean	Canned green beans

Compound Library Searches:

For each compound evaluated the accurate mass (reported to a sub 2 ppm accuracy) was used to generate an elemental composition and potential chemical formula). Both accurate mass and chemical formula were used to search a number of databases including; Metlin (Scripps Institute), Dictionary of Natural Products (Taylor & Francis Group), The Human Metabalome Database (The Metabolomics Innovation Center), FooDB (The Metabolomics Innovation Center), ChemSpider (Royal Society of Chemistry), Universal Natural Products Database (Peking Database).

NMR Analysis:

NMR analysis was conducted using either a Bruker Avance III 750 MHz and Bruker 900 US^2. The Bruker Ultrashield 700 was equipped with a 1.7mm TCI probe

while the Bruker 900 US^2 was equipped with a 5mm TCI probe. Compounds 383, 413, and 661 were run in water, compounds 693 was run in methanol.

4.3 Results and Discussion

The citrus extracts were aged and a 5 person consensus panel established that the samples showed a flavor change at each aging point (0, 2, 4 and 6 days). At this point samples were chemically fingerprinted to understand how the systems chemically change with time. Since fresh flavor is an attribute that is lost with time, if the chemistry of aging is effectively modeled then freshness is also modeled. UPLC-MS was used to chemically define the samples followed by untargeted multivariate investigation through principle component analysis (Figure 4.1). Figure 4.1 reported that initial model generation identified that the selected fruit varietals are different, which is not the focus of this study. The PCA in figure 4.1 indicated that the biggest chemical variance in the model stems from the variety of fruit being different, not the aging of the samples.

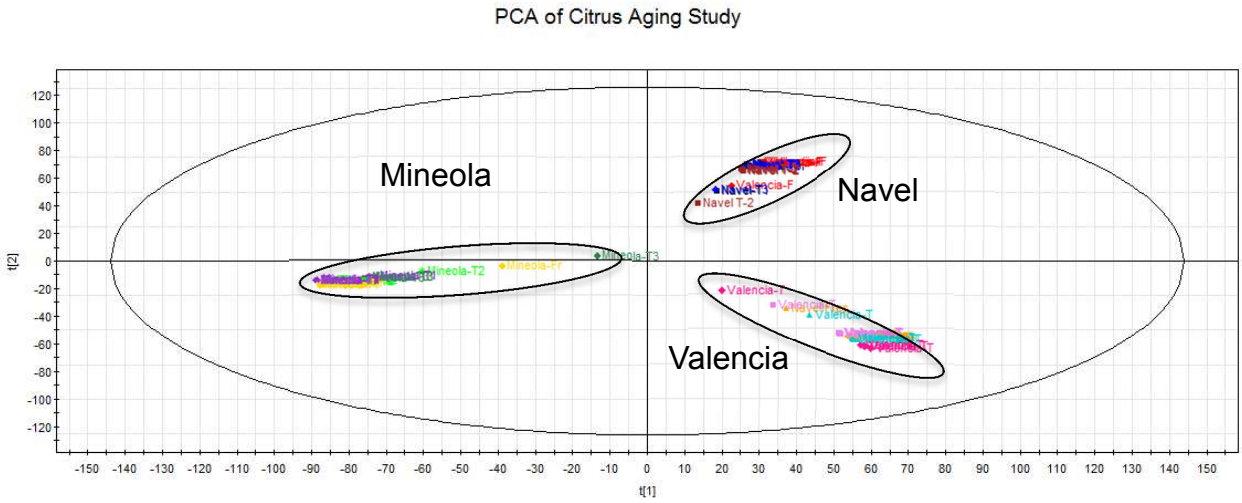


Figure 4.1: Principle Component Analysis of collected data set, showing strong classification based on the varietal of citrus.

As PCA is an unsupervised multivariate method the model will capture and report the largest variation between samples, so the classification by varietal should not be surprising. The PCA in figure 4.1 reported that Navel oranges and Valencia oranges are more similar than the Mineola, supported by both orange varietals grouping in the same region of principle component 1 (PC1). Navel and Valencia samples are then further separated by the second principle component (PC2). Mineola falls on the opposite end of PC1, indicating that the main chemical differences among the samples are between Mineola and the other two orange varietals, which is logical from a taxonomical standpoint. Modeling the origin or differences in varietals is a very common application of multivariate methods in food (Aishima, 1987., Tewari, 2008., Cano, 2008., Liu, 2010). The matrix composition was shown to be the largest source of variance (Figure 4.1) but is not the specific area of interest for this chapter, so the modeling approach is adjusted to emphasize sample age. It is important to understand if modeling age is possible within

each of the varietals. The changes in chemistry associated with each varietal aging (freshness) were also modeled and shown Figure 4.2(a, b, c). This established that models could capture aging, since it is the chemistry associated with time that is leading to the different groupings.

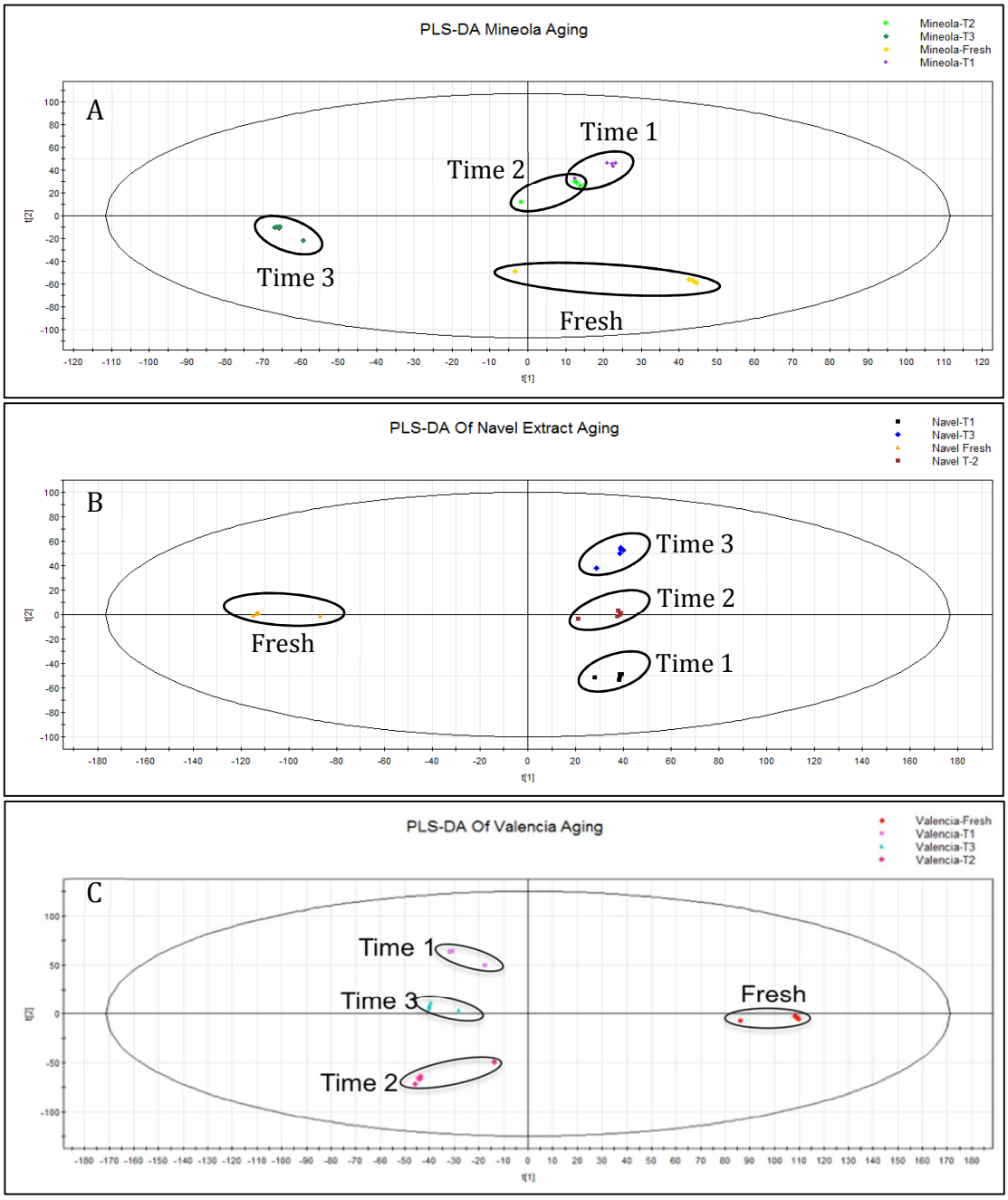


Figure 4.2: PLS models of aging chemistry for data subsets of each of the varietals. All models show clear differentiation. A. Mineola. B. Navel. C. Valencia.

PLS models were generated for each varietal to understand the ability to chemically characterize each sample by age. Figure 4.2a reported that the model is able to differentiate the samples within the Mineola varietal by time. The 144 hour sample

(T3) was classified distinctly different than the other ages by PC1 whereas sample time points (T1 and T2) were differentiated along PC2. Model quality metrics show a R^2X= 0.571, R^2Y=0.988 and a Q^2=0.923. This plot depicted that the final aging point (Mineola-T3) is the most different from the other sample time points, with separation along the first principle component. This is a unique trend in this set of models, and may indicate that the Mineola sample initially undergoes aging slower than the Valencia or Navel systems (Figure 4.2 b,c). Since the largest chemical difference is seen in the last 48 hours of the aging experiment rather than the first 48 hours like the other varietals (Figure 4.2b,c). Figure 4.2a, seems to show that Time 2 and Time 1 are clustering together but this is due to a multi-dimensional space being represented by two dimensions. A three dimensional view illustrated that samples separate on a dimension that is not represented in the two dimensional space (PC3), further confirmed by the confusion matrix which has a 0% class error. The second principle component captures 13% of the X block variance and 60% of the Y block variance, which indicated that there is still a large chemical diversity driving classification in the second principle component. One sample from the fresh grouping and one from the time point 2 appear to deviate from their respective clusters this is in part due to the two dimensional representation a multidimensional system(figure 4.2a). This is likely due to variation in either sample preparation or during analytical fingerprinting, to validate these samples the residuals are investigated to determine if they are outliers. The distance to the plane in X matrix space (DModX) were less than 2.5 times the overall residuals, indicating that these points are not outliers.

Figure 4.2b reported the navel model has very clear differentiation graphically, and model performance metrics of: $R^2X=0.631$, $R^2Y=0.988$, $Q^2=0.915$. The navel model reported that the fresh sample is differentiated from the other ages by principle component 1, indicating the most chemical differences in the first 48 hours of aging (Figure 4.2b). The other ages (48, 96 and 144 hours) were all further differentiated by principle component 2, clearly differentiating between T1(48 hours), T2 (96 hours), and T3 (144 hours) (Figure 4.2b). Principle component 2 also represents that samples in the logical time progression with T2 falling between T1 and T3. This trend was somewhat present in Figure 4.2a with the Mineola samples but does not have as clear separation of time 2 and time 3 as the navel model (Figure 4.2b), and the Valencia (Figure 4.2c) does not have this same time progression which is due to different model rotation.

The Valencia model noted a very similar trend as the Navel model (Figure 4.2c). The Valencia PLS-DA model has the quality metrics of: $R^2X=0.632$, $R^2Y=0.989$, $Q^2=0.923$. These again show good relation of the model to addressing the chemical diversity present, which is visualized by the loading plot as in Figure 4.2c. The fresh (age 0 hours) and aged samples (age 48, 96, 144 hours) were clearly separated along the first principle component with further differentiation of the aged samples across the second principle component. Unlike the other two varietal models (Mineola and Navel) there is different rotation and projection that resulted in T3 being between T1 and T2. The difference in projection suggested that the Valencia model had different relationships between the latent structures beyond PC2 and resulted in different sample projection. This unique data trend for Valencia, was similar to that noted for Mineola where the

grouping for T3 was the most different among the different time points (Figure 4.2a). Navel and Valencia also showed the Fresh sample as distinct from all of the other time points, where the Mineola has the oldest sample (T3) as the most different, with differentiation along PC1. This data trend may hint at which point in aging indicated the biggest change in flavor quality.

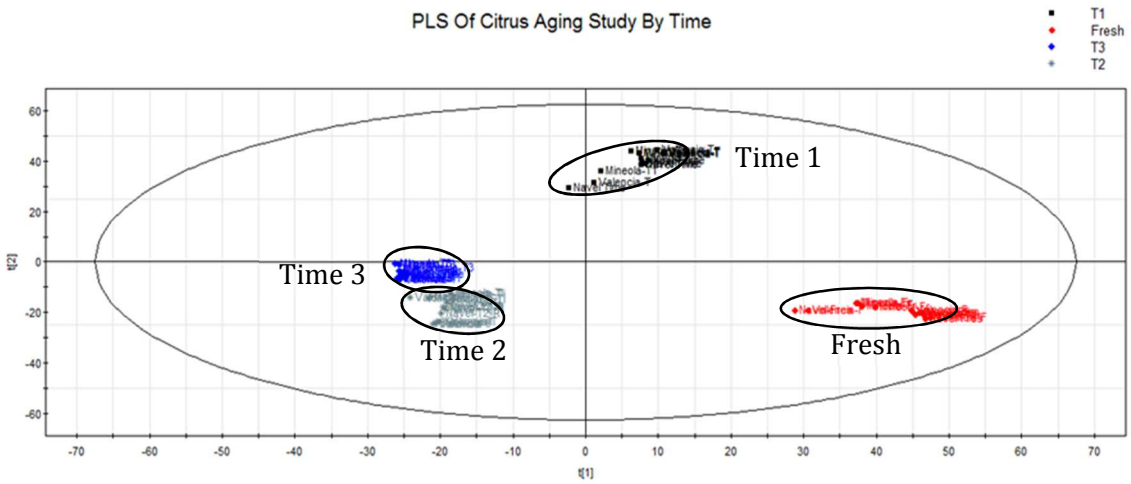


Figure 4.3: Projection to Latent Structures model generated to model the aging chemistry associated with the entire citrus platform, indicating there is common chemistry within all of the varietals related to sample age

In addition to modeling aging for each individual varietal, a PLS model consisting of all varieties by age was developed (Figure 4.3). While investigating individual varieties does examine the chemical varietal diversity, a model combining all the varietals provides a large data pool and filters out the varietal differences. Figure 4.3 depicted how a model is able to handle chemistry differences on a more diverse data while eliminating varietal effects. This modeling approach would evaluate changes

related to citrus, in general. The aim of this dissertation is to understand how these systems age similarly and assumes that within this large data set there are a number of universal chemistries and trends that can be teased out. In this context we are defining a food platform as a group of similar food systems, the citrus extracts in this case. Each of the samples selected are different species and varietals, but fall close to each other from a taxonomical standpoint. Through understanding this and only investigating how the food platform ages provides unique modeling opportunities. This experimental design provides unique chances for variable selection and data filtering.

Since the modeling is driven by time rather than the varietal, the chemistry stemming from the varietal differences is suppressed. This filtering allows for elimination of sample makeup and emphasizes commonalities in aging (Tautenhahn, 2010). This model preserves the preprocessing within time and varietal, and identified ubiquitous data trends while minimizing varietal contribution. A high quality multiple correlation coefficient ($R^2Y= 0.95$) and a Q^2 of 0.981 was reported for model. This modeling approach ensures that the model identified chemistry that is associated with aging, and this approach has previously shown utility in identifying common impacts across a number of different phenotypes (Tautenhahn, 2010). Thus the model effectively differentiated samples using aging chemistry which is common for all the varietals. By identifying common chemistry changes across the food platforms freshness character is included in this chemistry. The essence of this is to develop an approach that can filter chemical data in large data sets that do not pertain to the research question at hand,

making the pool of statistical variables that are contributing to differentiation more generically related to product aging

From the model in Figure 4.3, differential analysis (PLS-DA) was used to compare two sample groups and further meta-analysis was conducted. Pairwise models were generated comparing time 0 hours to 48 hours, 48 hours versus 96 hours, and 96 hours versus 144 hours (data not shown). This produced three PLS-DA models that were individually mined for statistical features a first order analysis was used. Model metrics were as follows: Time 0 versus T2 (48 hours) with model metrics $R^2Y=0.982$ and $Q^2=0.942$, T2 (48 hours) versus T4 (96 hours) with model metrics $R^2Y=0.987$ and $Q^2=0.941$, T4 (96 hours) versus T6 (44 hours) with model metrics $R^2Y=0.985$ and $Q^2=0.963$. Overall these models were able to successfully differentiate the food platform based on the product age. From each of the three models, the top 1200 compounds were selected based on the variable of importance (VIP) metric. This cutoff was selected as this was approximately the point when the VIP for each feature had a low contribution to the

Table 4.2. Top features from modeling age with PLS after variable selection

Retention Time	M/z	VIP From Overall Analysis
2.24	473.2481	1.70521
2.00	877.3165	1.69654
1.50	541.1749	1.66511
5.48	457.2563	1.65814
1.75	213.0026	1.64311
6.93	563.2393	1.63663
2.38	661.2653	1.60764
3.50	914.3524	1.58976
3.07	684.3070	1.57854
2.72	901.3182	1.57389

model (VIP <1.0). These three comparison sub-data sets were further evaluated for features in common (using a second order analysis).

Overall, 97 common chemical features during the aging process were identified that were present in each comparison model. This list was further ranked by the Variable of Importance (VIP) and the top 10 features are reported in Table 4.2, the ranking was done based on the VIP of each feature from the model in figure 4.3. All 97 compounds had a variable of importance of at least 1.2, which indicated that each selected feature is strongly important to classifying the time points. VIP statistics above 1.0 are considered strongly contributing towards the model. Although PLS is a well performing method, it is rooted in regression and so an alternative modeling approaches may capture alternative chemistry. Machine learning is an attractive approach to data modeling, and provides differing insight into the data structure.

Random Forest is a machine learning algorithm that uses a combination of nested decision trees and bootstrapping (a method of randomly sampling with replacement) to identify variables of interest. Random Forest was used as it provides orthogonal non-regression based modeling approach for data analysis that is not as focused on strictly linear data trends. PLS models typically favor data trends whereas Random Forest does not necessarily emphasize linear trends, and can identify statistical features that may have maxima or minima inflections at an intermediate time point. Numerous collinear variables can increase the challenge of interpretation and variable selection in PLS, combined with very large and noisy data sets can lead to over emphasis of variables with high leverage (Wold, 2004). Additional benefits to random forest include explicit noise

elimination, not present in PLS, and Random Forest does not need cross validation since it is already a part of the modeling design (Menze, 2009. Breiman, 2001). A classification model was generated to avoid linear regression trends. Random Forest model fit is evaluated using a metric called out of bag error, and is a metric of classification error. The model testing error generated for this aging experiment was 2.38%. The out of bag error is similar to leave one out cross validation where the unused samples are run through the generated tree, and this information is summed across forest generation for an unbiased estimate of classification power (Breiman, 2001). The variables of importance are shown in figure 4.4, within the topmost compounds (30) there was no overlap with variables identified in PLS. The PLS model however, did not have as many matching m/z as random forest did (eg, 413.121 at retention times 2.193 and 2.454). This indicated that a number of closely related chemical constituents were

present and important to aging (eg. Features 2.454_413.121 and 2.193_413.121). As these similar compounds have differing retention times, they are likely isomers or contain a similar chemical moiety. Once compounds of interest were statistically identified the features were isolated from a food system for preliminary sensory analysis and screening.

Initial screening of the isolated compounds was conducted at an estimated purity of 50-60% (by UPLC-mass spectrometry). The identified compounds were isolated and purified through mass spectrometry guided fractionation. An initial screening of the

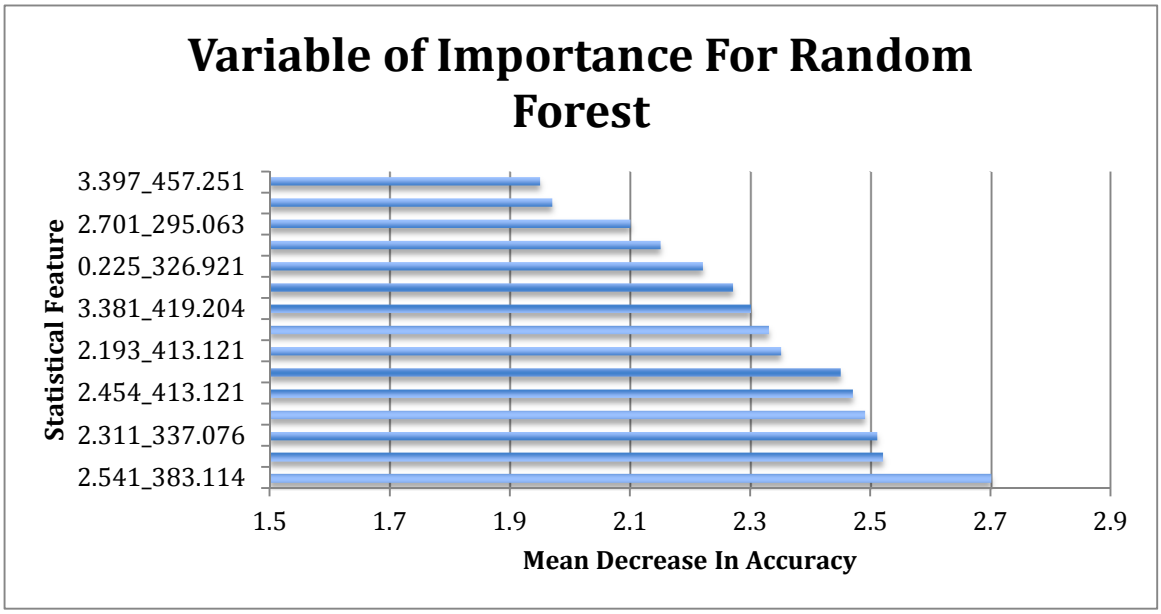


Figure 4.4: Random Forest Variable of Importance based on the increase of out-of-bag error

isolated compounds at moderate purity was conducted using a 6 person panel. Panelists reported if there was any change over a tasting blank, and what the flavor difference was, any reported difference was described by half of the panel. Table 4.3 Provides the statistical feature, the sensory attribute and further whether it was identified by Random

Forest or by PLS. Sensory evaluation at low purity is inconclusive, but establishes whether or not a compound should be further pursued.

Table 4.3 Initial screening of isolated compounds including feature ID, reported attribute and the which modeling approach identified the feature

Feature (Retention Time_M/z)	*Sensory Attribute*	*Modeling Source*
2.541_383.114	Reduced sweetness, bitterness	Random Forest
3.145_695.283	Bitterness	Random Forest
4.909_191.091	Increase in sweetness	Random Forest
1.662_148.972	Slight increased sweetness	Random Forest
2.208_295.063	Increased sweetness, floral, fruity	Random Forest
2.454_413.121	Oxidized, astringent	Random Forest
2.193_413.121	Cardboard, astringent	Random Forest
2.24_473.2481	Slight increase in sweetness	PLS Analysis
1.50_541.1749	More bitter, late onset acidity	PLS Analysis
1.75_213.0026	Slight astringency	PLS Analysis
6.93_563.2393	Mouth numbing, slight bitter	PLS Analysis
2.38_ 661.2653	Bitter, slight sour	PLS Analysis
5.48_457.2563	Increased sweetness	PLS Analysis
3.07_684.3070	Increased bitterness	PLS Analysis

Once these compounds showed initial sensory contribution more time intensive isolation and purification took place for a larger descriptive analysis. As mentioned in the materials and methods section some of these features required three dimensions of separation to achieve acceptable purities. Since flavor compounds have a wide range of flavor activities, high purity is critical to establishing causation during sensory recombination. All of the descriptive analysis results are done with compounds that have a purity >97% as estimated by both proton nuclear magnetic resonance and UPLC mass

spectrometry. Once purity was established, a descriptive analysis panel is able to evaluate selected compounds to better understand how they relate to the flavor profile.

Table 4.4- Each of the SAFE recombination models and the associated descriptive analysis ratings for the attributes, value in parenthesis is the p-value for Dunnett's test comparison to the sample blank

Sample	Orange Character	Cooked	Green Bean	Floral
Sample Blank	5.25	1.92	0.85	2.1
383 Recombination	3.71 (0.0012)	2.82 (0.22)	1.82 (0.029)	1.37 (0.054)
413E1* Recombination	3.82 (0.0029)	3.93 (<0.001)	1.85 (0.022)	1.14 (0.006)
191 Recombination	4.03 (0.013)	2.68 (0.36)	1.64 (0.096)	1.46 (0.11)
693 Recombination	4 (0.011)	2.89 (0.17)	1.5 (0.21)	1.57 (0.24)
*Marker 413E1 relates to marker 2.193_413.121, which is first in elution order				

A SAFE extract of orange juice was selected as it represents a complex aroma profile that effectively conveys a number of fresh and aged attributes. Table 4.4 illustrates the sensory changes seen for significant attributes and the associated p-value for Dunnett's test. All of the isolated compounds showed a suppression of orange character over the sample blank with compound 383 showing the largest suppression (Table 4.4). Compound 413E showed almost a doubling of cooked character over the sample blank. Compounds 383 and 413E1 also showed increase in green bean character over the sample blank (Table 4.4). Compound 413 reported suppression of floral character, and with a relaxed significance level ($p = 0.054$) indicated that compound 383 also suppressed floral character. Review of the sensory results supported the features identified by untargeted methods were successful in changing the flavor of the sample

blank to a profile closer to that of the aged sample used in term generation. The nebulous understanding of freshness is hard to capture in a sensory panel, but certain attributes were associated with aged samples during term generation (high green bean, cooked). With compound 383 showing a suppression of orange character and floral coupled with an increase in green bean character, these noted changes support that this compound could negatively impact the flavor quality of citrus systems as they age. Terms like green bean, cooked were closely associated with aged samples during term generation as was decrease in orange character. Compound 413E1 reported suppression in orange character and floral as well as an increase in cooked character, which again would negatively be associated with flavor quality of a citrus system. It is also important to note that these compounds did not have aroma character, and so there is likelihood for impact either through modulation or a mechanism that impacts the aroma profile. The mechanism is beyond the scope of this dissertation, but should be included in future work. For the SAFE model tasting system, ethanol is critical for the mechanism of SAFE to reproduce the endogenous aroma. The concentration of ethanol in the tasting samples was 5% (v/v) and therefore was detectable by the panelists and may have altered the perception or release of some compounds, however this isolate did provide a 'fresh' orange aroma for evaluation.

To further remove any potential impact of alcohol in the tasting medium, a commercial volatile orange flavor (VOF) system was selected after small consensus screening with experienced panelists to have representative orange aroma. In this tasting medium there was an impact on sweetness for a few of the compounds, namely

compound 413E2 and 457 showing an increase over the sample blank, and compound 383 also showing an increase over the sample blank although with a reduced significance level(p=0.06). Compound 383 showed a decrease in cooked over sample blank, which differs from the SAFE extract system. This change in sensory response is not unexpected given that the VOF sample would not have the same chemical composition as the SAFE model, and is likely less complex. Often times formulated flavors are not as complex as their natural counterparts, and this lack of complexity may be part of the reason for the differing sensory results. Compound 413E2 also reported an impact on the cooked character with a reduced significance level (p=0.08). Compound 661 showed a suppression of floral character compared to the sample blank, with none of the other compounds showing an impact which also differs from the SAFE as 413E1 showed an impact in the SAFE extract but not the VOF system. This impact is interesting as compound 413E1 showed an impact in the SAFE extract, as compound 383 with the alpha reduced to 0.1. Compound 413E2 suggested a suppression of green bean character

Table 4.5: for the VOF recombination models and the associated descriptive analysis ratings for the attributes, value in parenthesis is the p-value for Dunnett's test comparison to the sample blank

Sample	Sweetness	Cooked	Green Bean	Floral
Sample Blank	4.78	1.89	1.0	1.88
383 Recombination	5.27 (0.060)	1.11 (0.029)	0.77 (0.84)	1.66 (0.81)
413E1* Recombination	5.22 (0.11)	1.56 (0.68)	0.77 (0.84)	1.66 (0.81)
413E2** Recombination	5.33 (0.029)	1.22 (0.08))	0.33 (0.021)	1.66 (0.81)
457 Recombination	5.33 (0.029)	1.72 (0.40)	0.83 (0.95)	1.5 (0.30)
661 Recombination	4.88 (0.98)	1.44 (.40)	0.88 (0.99)	1.27 (0.030)
693 Recombination	4.83 (0.99)	1.61 (0.82)	0.88 (0.99)	1.5 (0.298)

*Marker 413E1 relates to marker 2.193_413.121, which is first in elution order
**Marker 413E2 relates to marker 2.454_413.121 which is second in elution order

compared to the tasting solution as well. To better characterize how these compounds impacted the flavor profile, it is important to identify their structures.

The isolated compounds showed sensory significance in a descriptive analysis panel, and so a structural understanding would be of interest. Chemical libraries are often employed to identify unknowns. In order to better understand the isolated compounds library searches were conducted to identify unknowns.

As the above searches provide challenging the MS-MS spectra was used to search as well. When possible the same databases were used with more selective MS-MS searches, these searches only yielded a positive response with compound 693. This compound was shown to be nomilin glucoside, which was further confirmed but Nuclear Magnetic Resonance (NMR). To the above databases keyword searches were conducted using the MS-MS fragments in both Google Scholar and Scifinder. These searchers did not produce acceptable matches and so NMR was used for structural elucidation.

Understanding the chemical structure of these compounds is highly desirable, by understanding the chemistries of each and their structure provides information about the source of each compound. To characterize the compounds of interest MS-MS and NMR were employed for structural elucidation. For the case of MS-MS profiling library searches were conducted using elemental composition from accurate mass data and further fragmentation data was also searched in available libraries. This was found to be highly unsuccessful and so elucidating the identity of the compounds was largely done through NMR.

NMR is a powerful tool to determine the structure of unknown compounds. Coupled with MS-MS a significant amount of information about unknowns can be generated. For each of the isolated compound Heteronuclear Multiple Bond Correlation (HMBC) is shown, and the other NMR spectra can be found in the appendix II. Each NMR spectra is accompanied by the bond correlations for the HMBC spectra, which comprise a two dimensional carbon and proton plot.

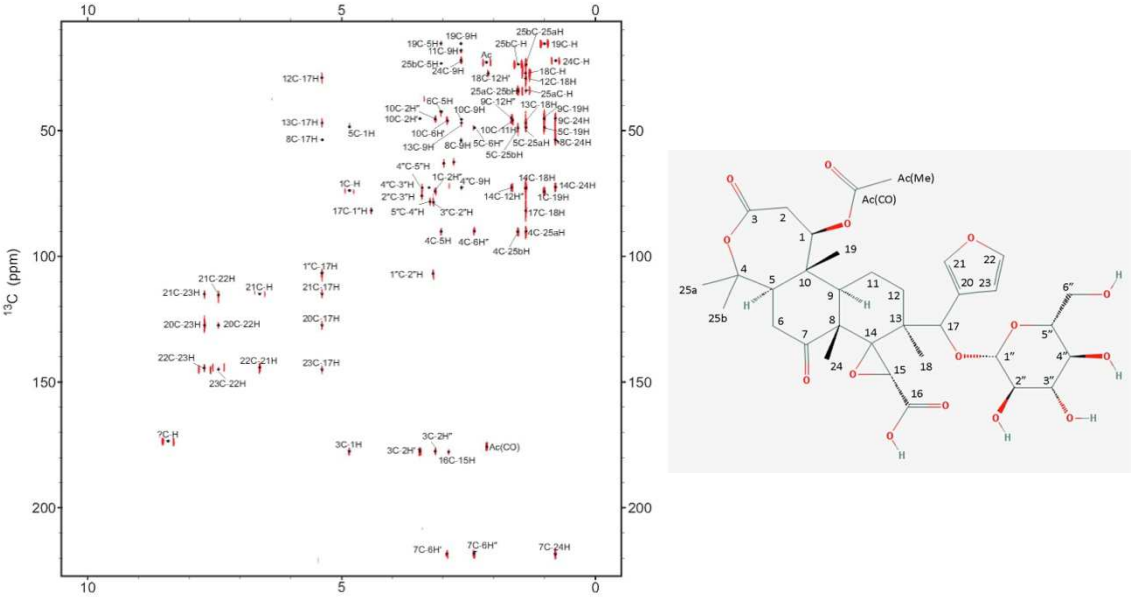


Figure 4.5: HMBC NMR and the associated assignments for Nomilin Glucoside. Nomilin glucoside is a known flavor agent in citrus and orange juice and showed a suppression of orange character in the descriptive analysis panel.

Compound 693, Nomilin 17-O-beta-D-glucopyranoside (Figure 4.5): MS-TOF: m/z 693.2770 ([M-H]-, 2 ppm), 179.0564 ([Glc-], 2 ppm), 161.0446([Glc-], 2 ppm). 1H NMR (900 MHz, CD_3OD): δ 0.79 [s, 3H, H-C(5C, 7C, 9C, 14C)], 1.01 [s, 3H, H-C(1C, 5C, 9C)], 1.37 [s, 6H, H-C(4C, 5C, 25bC)], 1.37 [s, 6H, H-C(12C, 13C, 14C, 17C)], 1.52 [s, 3H, H-C(4C, 5C, 25aC)], 1.64 [m, 2H, H-C(9C)], 1.40 [m, 1H, C-H(9C)], 2.11 [m, 1H,

H-C (9C, 14C, 18C)], 1.65 [m, 2H, H-C (9C, 14C, 18C)], 2.14 [s, 3H, H-C[OAc(CO)], 2.64 [dd, J= 12.0, 7.7 Hz, 1H, H-C (8C, 11C, 13C, 19C, 24C)], 2.89 [s, 1H, H-C(16C)], 2.92 [dd, J= 20.3, 7.4 Hz, 1H, H-C (4C, 5C, 7C, 10C)], 2.38 [dd, J=20.3, 12.0 Hz, 1H, H-C(4C, 5C, 7C, 10C)], 3.04 [m, 1H, H-C(4C, 6C, 19C, 25bC)], 3.15 [dd, J= 15.9, 7.7 Hz, 1H, H-C (1C, 3C, 10C)], 3.19 [d, J=8.5 Hz, 1H, H-C (1"C, 3"C)], 3.26 [t, J= 9.3 Hz, 1H, H-C (5"C)], 3.28 [m, 1H, H-C(4"C)], 3.41 [t, J=9.3 Hz, 1H, H-C (2"C, 4"C)], 3.46 [d, J=15.9 Hz, 1H, H-C (1C, 3C, 10C)], 3.54 [dd, J= 12.3, 5.6 Hz, 1H], 3.70 [dd, J= 12.3, 2.0 Hz, 1H], 4.41 [d, J=8 Hz, 1H, H-C(17C)], 4.84 [d, J=7.7 Hz, 1H, H-C (3C, 5C)], 5.39 [s, 1H, H-C (8C, 12C, 13C, 21C, 23C, 1")], 6.62 [s, 1H, H-C(22C)], 7.43 [s, 1H, H-C (20C, 21C, 23C)], 7.7 [s, 1H, H-C(20C, 21C, 23C)]. 13C NMR (175 MHz): δ 15.2 [C-19], 17.8 [C-11], 22.1 [C-24], 22.8 [OAc(Me)], 23.4 [C-25b], 27.2[C-18], 29.3 [C-12], 34.1 [C-25a], 37.5[C-2], 42.5 [C-6], 44.7 [C-9], 45.5 [C-10], 46.7 [C-13], 48.6 [C-5], 53.8 [C-8], 62.6 [C-15], 63.2 [C-6"], 72.36 [C-4"], 72.44 [C-14], 74[C-1], 76.3 [C-2"], 78.2 [C-5"], 78.5 [C-3"], 81.6 [C-17], 90.1 [C-4], 106.8 [C-1"], 114.9 [C-21], 127.4 [C-20], 144.2 [C-22], 145 [C-23], 175.7 [OAc(CO)], 177.4 [C-3], 177.5 [C-16], 218.4 [C-7].

Nomilin is a known compound of interest in citrus juices, the terpenoid by itself has long been a target for intervention and mitigation of bitterness (Shaw, 1984).

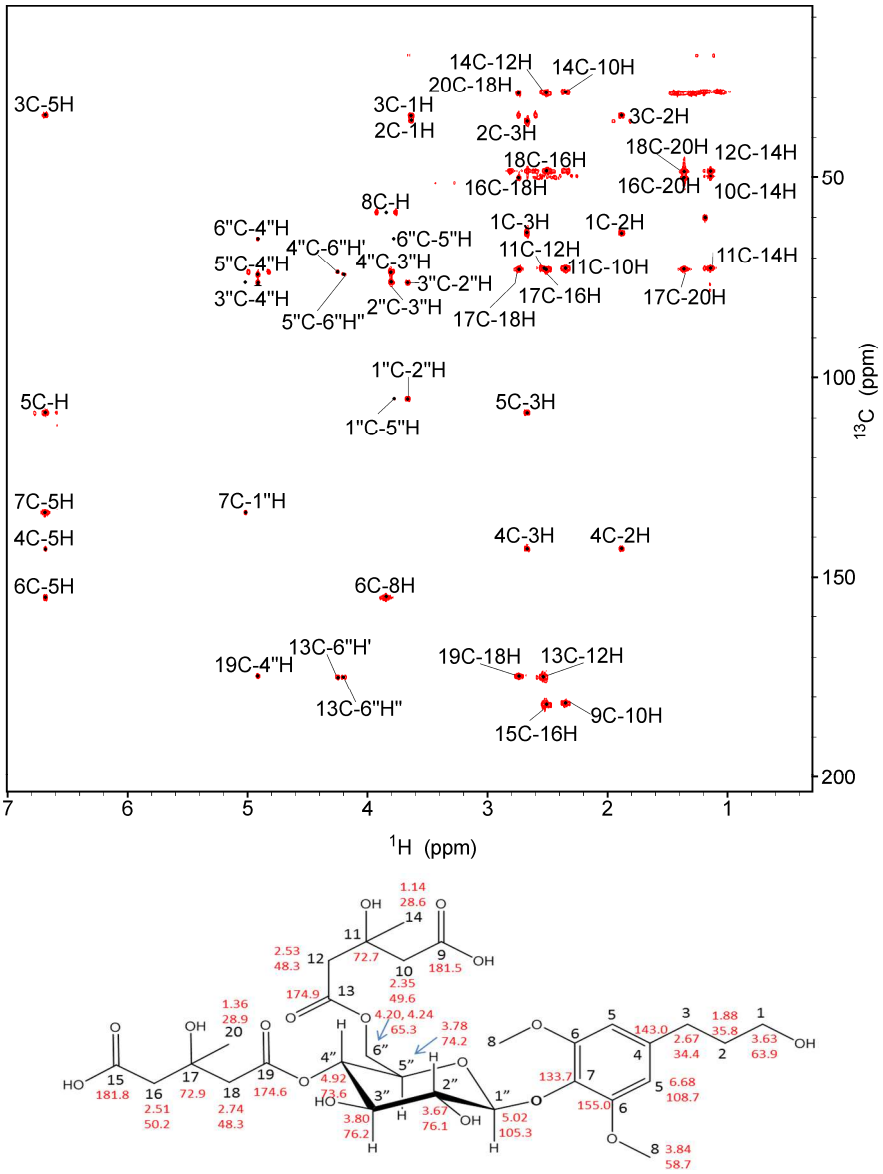


Figure 4.6. HMBC and associated assignments for novel compound 661. Compound associated with suppression of floral character

Compound 661 (Figure 4.6), systematic name: (5-(((2R,3S,4R,5R)-4,5-dihydroxy-3-((3-hydroxy-3-methyl-5-oxohexanoyl)oxy)-6-(4-(3-hydroxypropyl)-2,6-dimethoxyphenoxy)tetrahydro-2H-pyran-2-yl)methoxy)-3,3-dimethyl-5-oxopentanoic acid): MS-TOF: m/z 661.2349 ([M-H]-, 2 ppm]), 517.1909 ([M-162-18]-, 211([M-

$C_{18}H_{27}O_{13}$]-, 2 ppm). ^{1}H NMR (750 MHz, D_2O): 1H NMR (750 MHz, D2O): δ 1.14 [s, 3H, H-C(10C, 11C, 12C)], 1.36 [s. 3H, H-C(16C, 17C, 18C)], 1.88 [q, J= 6.6, 7.8 Hz, 2H, H-C(1C, 3C, 4C)], 2.35 [q, J= 12.6, 15.0, 2H, H-C(9C, 11C, 14C)], 2.51 [q, J= 16.7, 14.6 Hz, 2H, H-C(15C, 17C, 18C)], 2.53 [q, J= 14.3, 5.1 Hz, 2H, H-C(11C, 13C, 14C)], 2.67 [t, J=7.8 Hz, 2H, H-C(1C, 2C, 4C, 5C)], 2.74 [s, 2H, H-C(16C, 17C, 19C, 20C)], 3.63[t, J= 6.6 Hz, 2H, H-C (2C, 3C)], 3.67 [m, 1H, H-C(1”C, 3”C)], 3.78 [m, 1H, H-C(1”C, 6”C), 3.80 [t, J=9.5 Hz, 1H, H-C(2”C, 4”C)], 3.84 (s, 6H, H-C(6C)], 4.20 [dd, J= 12.3, 2.0 Hz, 1H, H-C(4”C, 5”C, 13C)], 4.24 [dd, J= 12.3, 6.6 Hz, 1H, H-C(4”C, 5”C, 13C)], 4.92 [t, H=9.8 Hz, 1H, H-C(3”C, 5”C, 6”C, 19C)], 5.02 [d, J=8.1 Hz, 1H, H-C(7C)], 6.68 [s, 2H, H-C(3C, 4C, 5C, 6C, 7C). 13C NMR (175MHz): δ 28.6 [C-14], 28.9 [C-20], 34.4 [C-3], 35.8 [C-2], 48.3 [C-12], 48.3 [C-18], 49.6 [C-10], 50.2 [C-16], 58.7 [C-8], 63.9 [C-1], 65.3 [C-6”], 72.7 [C-11], 72.9 [C-17], 73.6 [C-4”], 74.2 [C-5”], 76.1 [C-2”], 76.2 [C-3”], 105.3 [C-1”], 108.7 [C-5], 133.7 [C-7], 143.0 [C-4], 155.0 [C-6], 174.6 [C-19], 174.9 [C-13], 181.5 [C-9], 181.8 [C-15].

Compound 661 is a multi-substituted sugar moiety, with 3-Hydroxy-3-methylglutaric acid bonded to the 4th and 6th position of the sugar backbone. Bonded to the anomeric carbon is a dihydrosinapyl alcohol moiety.

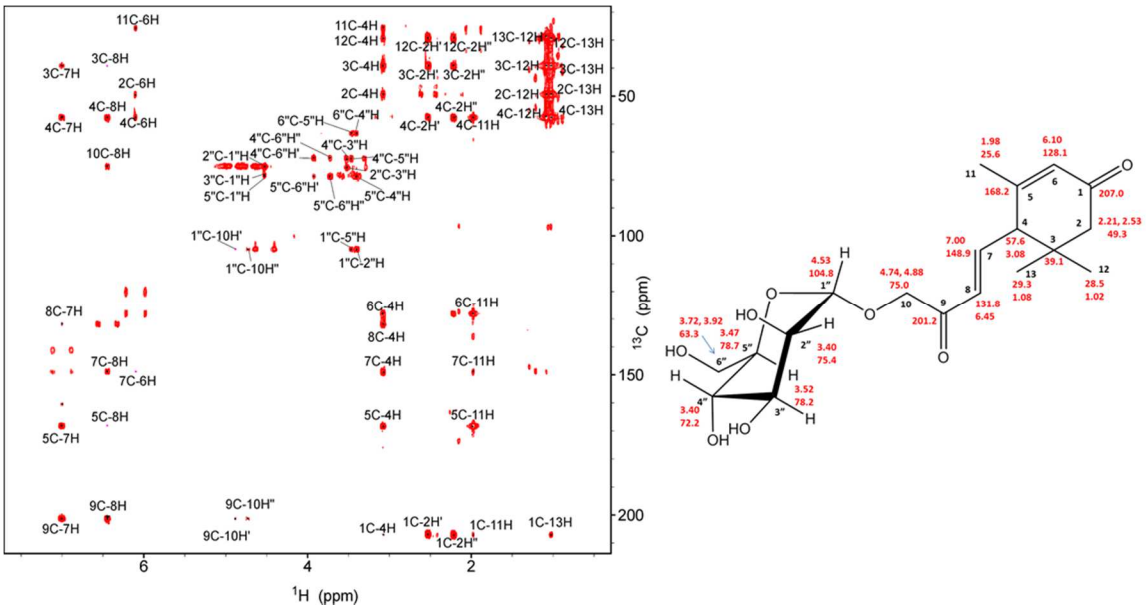


Figure 4.7 HMBC and associated assignments for novel compound 383. Compound is associated with suppression of orange character and increased green bean character

Compound 383 (Figure 4.7), systematic name: (3,5,5-trimethyl-4-((E)-3-oxo-4-(((2S,3S,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)oxy)but-1-en-1-yl)cyclohex-2-en-1-one). MS-TOF: m/z 383.1727 ([M-H]-, 2 ppm), 365.1620 ([M-H_2O]-, 2ppm), 221.1178 ([M-$C_6H_{11}O_5$]$^{\bullet}$, 2ppm), 163.1128 ([$C_6H_{11}O_5$]$^{\bullet}$, 2 ppm). 1H NMR (700 MHz): δ 1.02[s,3H, H-C(2C, 3C, 4C, 12C)], 1.08 [s, 3H, H-C(2C, 3C, 4C, 13C)], 1.98 [s, 3H, H-C(1C, 4C, 5C, 6C, 7C)], 3.08[d J=9.3 Hz, 1H, H-C(1C ,2C, 3C, 5C, 6C, 7C, 8C, 11C, 12C)], 3.40 [m, 1H, H-C(1"C, 3"C)], 3.41 [t, J=8.5 Hz, 1H, H-C(3"C, 5"C, 6"C)], 3.47 [t, J=8.8 Hz, 1H, H-C(4"C, 6"C)], 3.52 (t, J=8.5 Hz, 1H, H-C(2"C, 4"C)], 4.53 [d, J=6.9 Hz, 1H, H-C(2"C, 3"C, 5"C)], 6.10 [s, 1H, H-C(2C, 4C, 7C, 11C), 6.45 (d, J=15.8 Hz, 1H, H-C(3C, 4C, 5C, 7C, 9C, 10C)], 7.00 [dd, J=15.8, 9.3 Hz, 1H, H-C(3C, 4C, 5C, 8C, 9C)], 2.21 [d, J= 16.8 Hz, 1H, H-C(1C, 3C, 4C, 12C, 13C)], 2.53 [d, J=16.8 Hz, 1H, H-C(1C, 3C, 4C, 12C, 13C)], 3.72[m, 1H, H-C(4"C,

5"C)], 3.92[d, J=12.7 Hz, 1H, H-C(4"C, 5"C)]. 4.73 [d, J=8.0 Hz, H-C(1"C, 9C)], 4.88 [d, J= 8.0 Hz, 1H, H-C(1"C, 9C)]. 13C NMR (175MHz):δ 25.6 [C-11], 28.5 [C-13], 29.3 [C-12], 39.1 [C-3], 49.3 [C-2], 57.6 [C-4], 63.3 [C-6"], 72.2 [C-4"], 75.0 [C-10], 75.4 [C-2"], 78.2 [C-3"], 78.7[5"], 104.8 [1"], 128.1 [C-6], 131.8 [C-8], 148.9 [C-7], 168.2 [C-5], 201.3 [C-9], 207.0 [C-1].

4.4 Conclusions

This chapter illustrates how an untargeted method identified flavor active materials that were associated with age. Sensory validation was shown through descriptive analysis of recombination testing. Flavor characterization studies have long utilized targeted methods, this work established how untargeted methods can be utilized for flavor compound discovery. More specifically this work illustrated an untargeted workflow that could be applied to any number of food systems to better understand the non-volatile impact of flavor. This dissertation reinforces how modeling aging chemistry can lead to understanding how a food moves away from its 'optimum fresh' flavor attributes. The reported impact on flavor profile supports that the compounds moved the sample blank away from its initial flavor quality, which may provide insight into understanding attributes like freshness in foods. For the isolated compounds the mechanism of impact is not understood and there are a number of mechanisms that lead to the perceived differences, including modulation of the headspace aroma or aroma release from the matrix (Rodríguez-Bencomo, 2011), trapping of chemical constituents through a number of binding mechanisms (Mitropoulou, 2011), or impact perception via cognitive effects.

This chapter illustrates the utility of flavoromics for identifying data trends that can lead to a biological outcomes, and how those relationships can be established through modeling, and sensory validation with recombination models and descriptive sensory evaluation. This chapter contributes to advances in analytical flavor chemistry by further illustrating the contribution of non-volatile compounds to complex flavor perception. Understanding drivers of consumer liking is extremely challenging and chemically complex, so in order to understand and evaluate this complexity new innovation on how food is understood is critical to better understanding how food relates back to the consumer. Flavoromics is one of the few methodologies that is able to handle the chemical complexity that would be needed to holistically define chemical drivers and their relation to the perception of flavor and a food products performance.

Chapter 5: Optimization of an untargeted machine learning workflow to model aging and identify contextual data interactions

---

Summary: A lemon extract aging study illustrated methods to derive additional value from supervised machine learning. As data driven research is still largely outcome sparse developing methods to add supplementary value while increasing the utility of data collection can lead to wider use. Machine learning methods were screened for model quality as Random Forest produced the best fit model it underwent tuning (max features, number of trees, and samples per terminal node) and produced a final model with high fit (training = 0.951, test = 0.928). Statistically important features were then mined for additional contextual relationships using bivariate scatterplot matrices. The goal of this chapter is to identify bivariate interactions that establish relationships and support additional outcomes. The bivariate scatterplot approach found contextual relationships between many of the identified features including; co-degradation, curve linear relationships, positive split correlations. The identified contextual relationships support additional outcomes from untargeted research by better contextualizing the statistical features.

Notes:

5.1 Introduction

Untargeted chemical finger printing and statistical modeling provide robust methods to address chemically complex systems from a more holistic perspective. Chapter 4 illustrated some multivariate approaches (second order analysis) and machine learning methods (RandomForest) to identify areas of investigation for increased emphasis on chemistry that relates to a flavor change over time. While multivariate analysis has provided value in numerous areas of research, increased insight or context through the use of machine learning provides additional opportunity to advance the experimental outcomes. Machine learning is a dynamic modeling approach that allows algorithms to learn without explicit programming towards a problem (Witten, 2005). Machine learning can often utilize more data and provide orthogonal insight to traditional multivariate methods (Brieman, 2001). Being able to piece together more variables and contextualize them to a food system provides more utility to an untargeted methods of analysis. Generating matrix scatterplots is an established method to understand how variables exist with conditional relationships to one another (Becker, 1987., Davison, 2000., Murdoch, 1996., Wong, 1994). These scatterplots provide contextual inference and the ability to handle more variables visually than other more time intensive methods (e.g. Pearson correlation). By knowing the context behind how variables relate to one another allows for scientists to better understand complex problems. Being able to piece together singular variables by understanding relationships can dramatically increase value of untargeted methods, and may lead to wider application in the food industry.

For example, being able to understand how variables relate to each other using

machine learning can help flavor researchers better understand how untargeted methods supplement more traditional methods of analysis. Identifying how unknown compounds relate to known flavor materials can lead better control of food flavor during processing or aging. Ultimately, the goal of this is to better understand and predict how changes in food systems impact consumer acceptability and to develop strategies to mitigate any deleterious flavor changes. Current approaches to understand and evaluate changes in flavor are costly, time consuming and are limited in scope. Having more contextual information helps flavor researchers better understand flavor formation and degradation and what chemistries are involved. Understanding how flavor materials interact and which compounds of interest are altered when a food system is changed can help support understanding flavor drivers. At the same time identifying contextual interactions may lead to understanding flavor modulation and interaction. A systematic method to either of these aspects would be highly desirable. One of the challenges in addressing these complex interactions is the sensory validation.

. Identifying data based interactions can create a supportive infrastructure to reduce the number samples (or select probable interactions) investigated in a sensory validation. This still uses systematic decision criteria while identifying what compounds to target for investigating flavor generation or perception. Using data trends to identify contextual interactions is a method to understand what features from an untargeted analysis might provide the most value to a food processor. Beyond sensory panels cell cultures or employed to understand flavor interaction. Cell culture assays are high throughput techniques to identify which compounds may activate certain taste receptors, but require

sensory panel validation. Cell culture screening of statistically significant compounds is hindered by available compound libraries, especially when the number of samples increase (n^2). Modern sensory approaches are ill equipped to handle this number of samples, given the associated cost and time. The ability to further screen and select interactions to present to a panel would support this research field of interest. In order to systematically identify compounds to evaluate, a machine learning workflow can help identify which species are of interest for further evaluation. Through understanding data trends there is a chance to identify compounds that contextually related with a concentration relationship or time correlated which can lead to further understanding of food chemistry.

There are a number of benefits to approaching sensory interactions from a data driven mentality. Data driven methods produce targets that are contextually based on the rest of the data set, show statistical significance and are systematically identified. Some of the identifiable variable interactions include; concentration (e.g. correlation), time contextual relation (e.g. important in context of other variables), absence and presence. Investigating these bivariate relationships provides added value to the identified compounds and their relation to known chemical species (e.g. flavor compounds). Knowing the chemistry that influences generation or degradation of a flavor material can help researchers control the flavor profile. Data interactions can help to identify how unknowns fit into the experiment and better understand their role in a food's flavor. This contextualization can greatly enhance the value of untargeted research platforms.

The ability to contextualize unknowns will help advance existing courses of

research and better identify how they relate to the food system. Providing more value to the chain of information that is already being collected, utilizing previously unusable data supports new discoveries and can help drive innovation. In this work an aging experiment to monitor freshness is presented and adds an extra dimension to understand data interactions. Unlike other omics based fields (e.g. metabolomics, systems biology, proteomics) the tools within food omics currently lack the software packages and libraries to get a more in depth understanding of how data fits into complex reaction pathways. This reinforces the need to increase the number of outcomes from untargeted methods due to the resources involved in understanding unknowns. This chapter will illustrate how data based interactions are identified using an optimized machine learning approach, the end goal being identifying interacting variables for further investigation and characterization.

Within the realm of data analytics there are a number of highly powerful methods for data analysis, ranging from untargeted exploratory investigation to explanative classification and robust predictive regression methods. Within machine learning the most popular are often considered to be K-nearest neighbors, support vector machines (Linear and Radial Basis Function), Decision Trees (random and structured), Adaptive boosting, Naïve Bayes Classification (probabilistic classifiers), Linear Discriminate Analysis, and Quadratic Classifiers. Each of these methods use different decision criteria, approaches and assumptions to work with a data set leading to different learning outcomes.

Multiple machine learning approaches are often implemented on a data set to

understand which has the best "out of box" performance. Out of box refers to the best model fit with limited or no tuning. In this chapter, numerous machine learning methods are evaluated for their fit on an aging experiment. Once an algorithm shows high performance for this aging data it can be further tuned and evaluated for performance.

5.2 Materials and Methods:

Chemicals and Reagents. The following chemicals were obtained from the sources given: 200 Proof USP Ethanol (Fischer Scientific), 95% Leucine Enkaphaline Acetate Salt, UPLC Acetonitrile (JT Baker), NanoPure Water (Barnsted), 6 mL 1g C-18 SPE tubes (Supelco), Methyl parabens (Sigma Aldrich), Ethyl Parabens (Sigma Aldrich), propyl Parabens (Sigma Aldrich), Butyl Parabens (Sigma Aldrich). #4 paper filers (Whatman), 3 kDa ultrafiltration membranes (Millipore). Fruit was purchased from local markets.

Lemon: Meyer Lemons, Sweet Lemons, and Eureka (California grown, Organic California Grown, Mexico Grown) were sliced <5mm thick extracted at a ratio of 500 g fruit to 200 g ethanol for 24 hours. The lemon extracts were each aged for 0, 2, 4, or 6 days protected from light and with the headspace purged with nitrogen.

Sample Preparation: Samples were passed through an ultrafiltration membrane for improvement of chromatographic performance through removal of large molecular weight compounds. Extract was diluted to 10% ethanol before solid phase extraction. Solid phase extraction was performed using a 6 mL tube, 1 g packing C18 phase Samples were eluted with 600 µL UPLC grade acetonitrile (JT Baker) and to this 400 µL Nanopure water (Barnsted, Waltham, MA) was added.

Chromatographic Analysis of Lemon: Analysis was run using a Waters CORTECS™ C18+ column (2.1x100 mm) with corresponding guard column held at 40°C with a flow of 0.4 ml/min with linear gradient conditions (solvent A, Nanopure water 0.1% formic acid; Solvent B, Acetonitrile 0.1% Formic Acid):Solvent B 3% (0-0.75 min), increased to 50% (0.75-6 min), then to 90% (6-9 min) followed by column wash and re-equilibration. Mass spectra were both collected using sensitivity mode on a Waters Xevo G2 TOF instrument scanning from 100-1600 m/z with a scan speed of 0.3 sec, real time lock mass correction was done using leucine enkaphalin using 6 scans across a chromatogram with each scan 0.2 seconds, instrument was calibrated to <1ppm mass accuracy. Ionization conditions had desolvation temperature set to 450°C with a gas flow of 800 L/hr and the capillary set to 2.5 kV. The cone was set to 35 V.

Data Preprocessing: Data was analyzed using the Progenesis QI 2.0 software platform. The .RAW waters data was imported as Centroided data with a resolution set to 10,000, the operating resolution for sensitivity mode on the instrument. Data importation and alignment were done with the stock conditions, and all runs were evaluated for the suitability for alignment. While computationally expensive, this produces an alignment reference that is optimal for the experiment. Further peak picking was done with the inclusion criteria relaxed, which will include more peaks but with the chance to additionally include spectral noise. After the peak picking and review of the spectral de-convolution the identified peaks and data was exported as a CSV for additional analysis via Python and R.

Data Filtering and Modeling:

Python was used as a data manipulation and filtering tool. Initial filtering is completed using the coefficient of variation for a variable within a sample and age. For each variable the CV is calculated within a sample group and time, this is then compared to the other CV's present for that variable across observations. This script was developed to eliminate highly variable data unless the variables are reliably reported in another sample set. This ensures that noisy data is eliminated across a dataset, while the integrity, distribution and variance of other data is kept and reported.

Machine learning was conducted via the Scikit-learn package and was done using a 33% test, 67% training data split (Pedregosa, 2011). Interaction plotting was done using python package seaborn (0.6.0) and matplotlib (1.5.1).

5.3 Results and discussion-

Ethanol extracts of six lemons (Lisbon 1, Citrus Meyeri, Eureka 1, Citrus Limetta, Eureka 2, and Organic Eureka) were aged for 0, 2, 4 and 6 days and then chemically fingerprinted with UPLC-MS. The lemon data preprocessing produced two data streams, one from negative ionization with 27,872 variables and one from positive ionization with 38,653 variables. The size of the data streams indicated that preprocessing used low filter to low level compounds were included. Logical data reduction as described in materials and methods used a 15% coefficient of variation cutoff, was used to reduce the data structures to 11,746 variables in positive ionization mode (69.6% decrease) and 3,003 (89.2% decrease) variables in negative ionization mode. The method looked at features above the cutoff (>15% CV) and referenced that variable in all other observations, the feature is eliminated unless it is well reported (<15% CV) in another observation. This

filtering approach preserves data density through not promoting sparsity or null representation. This sort of logic filtering is applied in a manner that any metric could be used as a filter (e.g. data skew and kurtosis). This filtering may even build off earlier models that suited for outlier detection (e.g. PCA) and eliminate lowly contributing data (noise) and highly leveraging observations/variables (outliers). These two data structures were merged based on observation, for a multi-block experiment. Two scaling methods and a number of machine learning algorithms were investigated for their ability to fit the data. Minimum-maximum scaling was done using a 0-1 scalar length and so was a unit variance method. The methods selected were all of interest as they are commonly used for complex modeling situations and historically perform well on a number of diverse datasets. For tunable models (random forest, ensemble trees and SVC) baseline methods were used to not bias selection of an already tuned model. Each machine learning approach was investigated using Python and the associated package in scikit-learn, and results for both scaling methods are illustrated in figure 5.1.

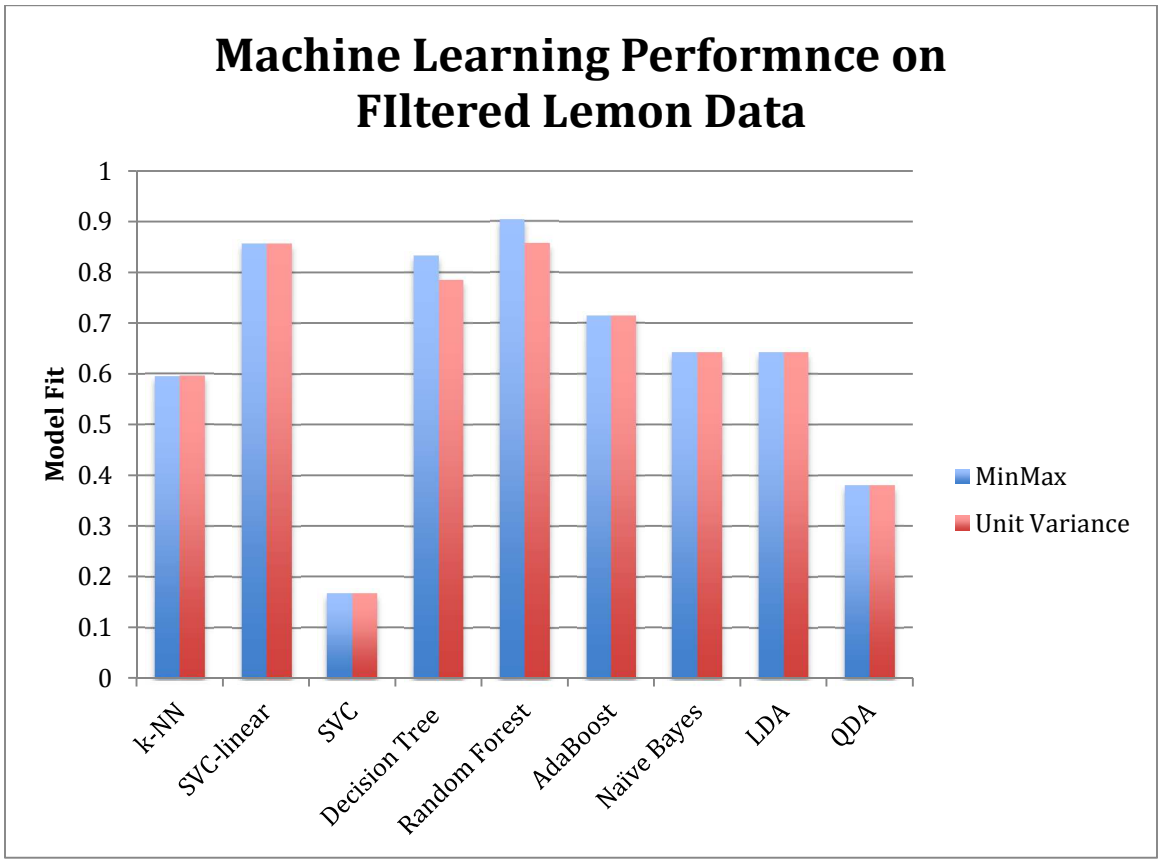


Figure 5.1 Performance of various machine learning approaches with two different scaling approaches, the two ensemble techniques and the linear kernel support vector machine produced models with the best fits

Model performance was generated using a training/test split of 0.33, meaning 33% of data was kept out of the training dataset and used to evaluate model performance. Model fit indicates how well the algorithm models the data and a value closer to 1.0 indicates a better fit. Each machine learning approach was evaluated with the two scaling methods (Unit variance and MinMax), limited impact on model performance was observed. The exceptions to this are the ensemble tree methods (Decision trees and RandomForest), which show MinMax scaling to outperform unit variance scaling (Figure 5.1). Support vector machine with a linear kernel, and the two ensemble methods

generated the models with the best fit. It is interesting that the linear module of the support vector machine (SVM-Linear) outperformed the radial basis function kernel model (SVM) which may indicate the screened model was over fit, as increased flexibility (in RBF model) in hyper plane creation should improve upon the linear model if not match it. This trend was also shown in the fit of the Linear Discriminant Analysis compared to the Quadratic Discriminant Analysis. This trend in both sets of models indicated there is likely a hyper plane that is well modeled by linear methods. Overall random forest was selected as the best modeling approach, therefore was selected to further investigate tune parameters. In addition to the quality of fit, random forest models perform well with noisy within data sets and keep their performance when variables outnumber observations (Brieman, 2001), which is true for the current dataset. The random forest model benefitted from MinMax scaling, which is advantageous as some of the assumptions of random forest are supported from all of the variables being the same length.

Within Random Forest there are a number of tunable parameters. These include n_estimators, max_features, and min_samples_leaf. Each of these methods is used to optimize and decrease the out of bag error (oob) which is a metric illustrating model quality.

n_estimators: This parameter is the number of trees in the forest and dictates the number of iterations used to generate a model. Generally the model reaches a point of decreasing returns where the cost (computing time) to train the forests is not outweighed by the improvement in model quality.

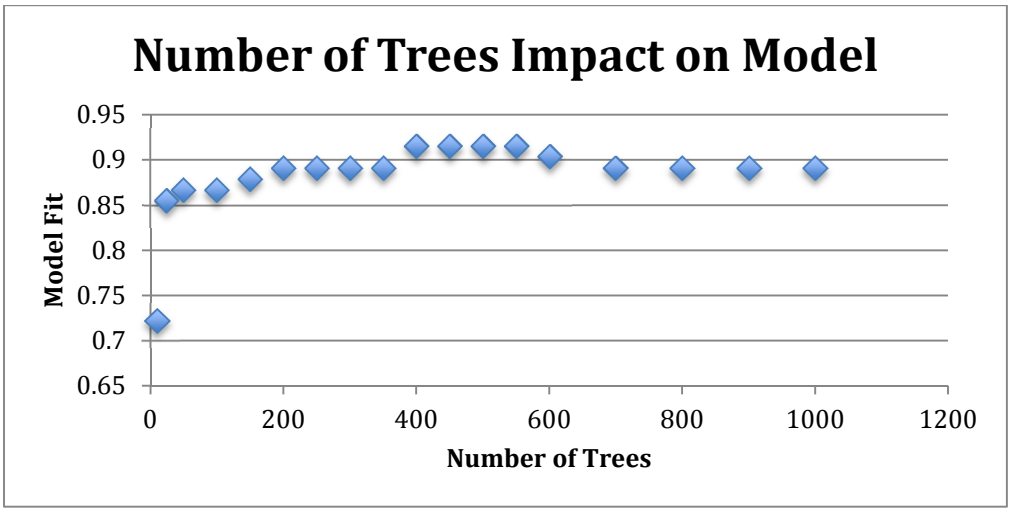


Figure 5.2 The impact on the number of trees used in a random forest model and the model fit

Figure 5.2 illustrated how with an increase in trees, there is an increase in model quality up to a certain point. This is well illustrated moving from a model with 10 trees (oob=0.722) to 25 trees (oob=0.855), but diminishes as trees are added with 50 trees having an out of bag error of 0.867. The optimal selection of trees was found to be between 400 and 550 trees. Random forest models are not always stable in repeat generation, but it was found over multiple model generations that 450 trees was the optimal with an oob score of 0.915 and a model fit on the test data of 0.915.

max_features: This value is the number of variables tried at each node when a forest is generated. During node generation a population(k) of the training data is sampled and used for node generation. From population k, a selection of variables(n) is sampled to determine the best variable to use as a decision criteria. The max_features determines the n, and as n increases so does the computation time as there are more variables tried every time a node is generated. At each node of each forest n variables are randomly selected from the training data in order to identify a variable that a

classification decision can be made from. This is often considered a parameter that is computationally expensive. As max features approaches inclusivity the likelihood to over fit the model and lose the benefit of random bootstrapping is almost certain, and only sees application in regression applications (Geurts, 2006). Often times a starting point of the square root of the number of variables is used for initial generation and then further tuned for model performance and stability.

In this case the types include; Auto: the square root of the number of variables are used, None: all of the features are used, Log_2: the logarithm base 2 of the number of variables are used, 0.9: 90% of the variables are used, 0.2: 20% of the variables are used. SqRt: same as Auto, but added to look into forest variability.

Evaluation of these parameters suggested there is improvement of model performance by using a 0.9 or 0.2 as the parameter (Figure 5.3). The differences observed between Auto, None, 0.9 and 0.2 likely stem from different forest manifestations. Parameters of SqRt and Auto use the same metric for determination of features so it is likely the model stability is acceptable. There is a clear improvement over the use of a base two logarithmic, which is logical since this metric tries fewer than 14 variables(<0.1% of data). Even with the small number of variables used by the Log_2 setting it still generated a model with a fit above 0.84 which is a well performing model. As the number of variables tried increases there is an increase in model performance, but diminishing returns above the Auto function, as the model fit is only slightly better in the 0.9 function and this is associated with a >1,500% increase in processing time. Understanding the number of features that create the highest quality model in the least

amount of time helps ensure that resources are utilized in an effective manner.

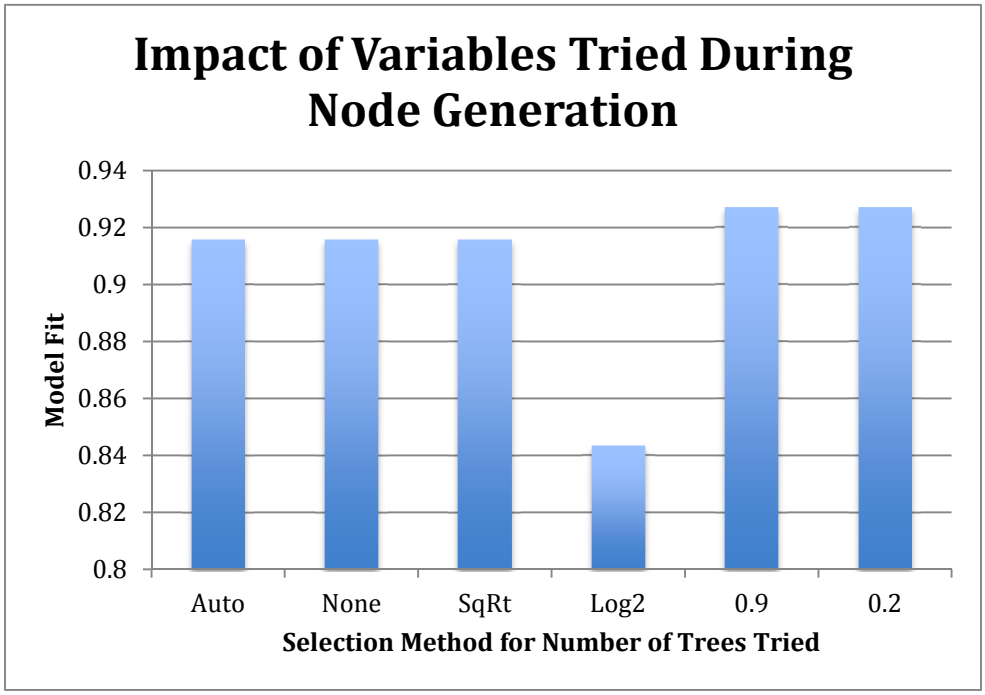


Figure 5.3 Impact of variables tried during node generation on RandomForest model fit quality

min_samples_leaf: This parameter dictates how many samples can end at a terminal node. After a decision at a node is made the min_samples_leaf establishes how many samples can preside in a terminal leaf before that decision fork ends. For example, if this attribute is set to one then each terminal node must contain one or more observations. If this metric was set to 10 then the terminal node must contain at least 10 samples. This metric is important to understand as it helps to dictate how the algorithm ends the decision making process. If 10 samples must be grouped together at the end of a node, this captures less interclass variability than a value of one (one sample per terminal

node). This helps to tune the model sensitivity and preserve inner class discrimination.

Figure 5.4 illustrated how the minimum number of variables in each leaf impacts model quality, with a lower of samples per leaf the model is better able to better differentiate samples. With an increasing min_samples_leaf the model fit decreases, but at a certain point model quality increases. This illustrated how with changing model flexibility interesting modeling outcomes arise, and some research questions may be better answered by adjusting how sensitive the model is to interclass variability. In addition to the settings tried in the above figure a value of 30 was also tried, which led to a model quality of 0.277 and was not included due to poor fit. A min_samples_leaf of 30 was tried since there were 30 observations of the same time scale (age).

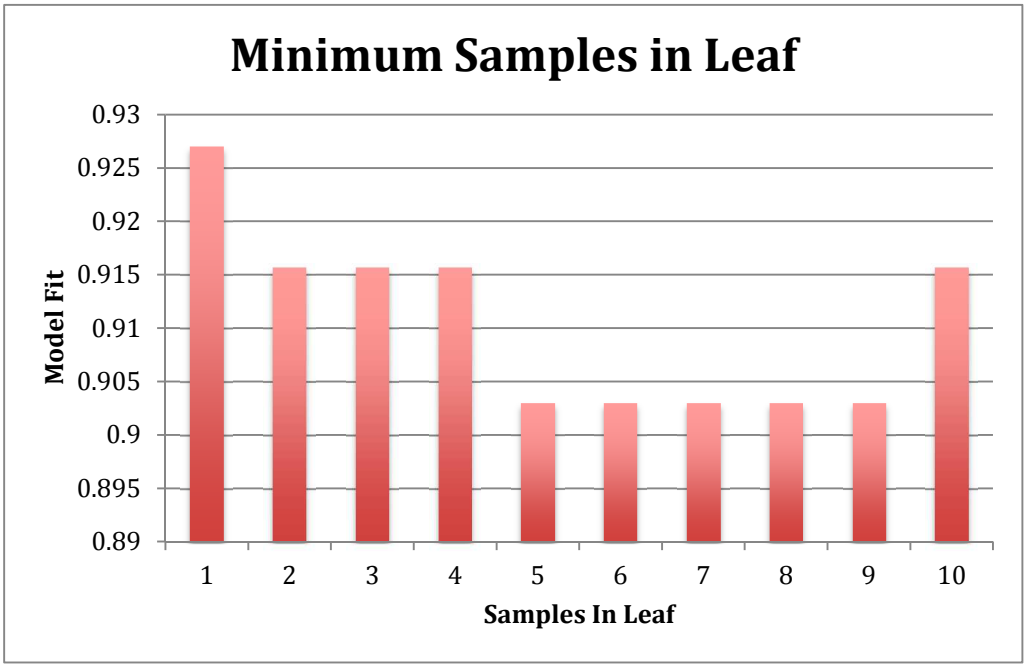


Figure 5.4 Impact of the minimum samples at a terminal point in the random forest on the model fit

Integrating all of this information allows us to understand how to best tune the model in order to produce a robust output. A model was generated with the optimized conditions with both importance measures (Gini and Entropy). In the context of machine learning these are indicators of model quality, entropy indicating the gain in information a model gets before and after a decision event including that variable and gini is a calculation of whether or not an observation would be randomly labeled incorrectly given the labels present in the subset. Provided the same data structure these settings contribute changes due to the difference in decision criteria, and each should be evaluated for a data stream.

Table. 5.1: Training and Test model fit using each Gini and Entropy important measure.

Decision Criteria	Training Score	Test Score (+/-)
Gini	0.927	0.904(0.0059)
Entropy	0.951	0.928(0.0049)

The two important measures lead to well performing models, but Entropy provides a reported improvement over Gini (Table 5.1). The table 5.1 suggested that there is a an improvement in model quality using the Entropy decision basis for both the training and test validation including the variability of model stability over five independent model generations. The features of importance are extracted and seen in Table 5.2, which indicated the contribution of feature score in both Gini and Entropy.

It should be noted that the top compounds are largely the same between the two methods, the order of importance is what is impacted by the decision metric (Table. 5.2).

This piece of information may be relevant if there was a single variable that was to be investigated however multiple compounds may also be of interest. Of the compounds that are present in the top ten for each method there is a wide range of polarities of compounds, as indicated by the retention time on a reversed phase LC column (Table 5.2). Interestingly, as chapter 4 mentioned there were similar features identified at varied retention times which is seen in markers 3.88neg383.0985m/z and 3.60neg383.0979m/z. The difference in retention time and a m/z difference of 0.004 suggests similar chemistry formula of different isomers.

Table 5.2 the Gini Score and Entropy Variable of Importance score is presented for the top variables for each decision criteria. Variables take the format: retention time, ionization polarity, mass to charge ratio. The underlined variable indicates which decision parameter gave the highest model contribution.

Feature	Gini Score	Entropy Score
4.32pos297.0615m/z	0.12915	<u>0.15112</u>
3.59pos212.0929m/z	0.04271	<u>0.06046</u>
3.99neg239.1282m/z	<u>0.03976</u>	0.03748
6.47pos277.1079m/z	0.03515	<u>0.03624</u>
6.54pos309.0743m/z	0.01272	<u>0.03492</u>
3.69pos376.1398m/z	0.00959	<u>0.03339</u>
3.88neg383.0985m/z	<u>0.04664</u>	0.03326
3.67pos210.0406m/z	0.02187	<u>0.02694</u>
0.61neg209.0657m/z	0.01594	<u>0.02341</u>
5.88neg239.1281m/z	<u>0.03333</u>	0.02233
0.56neg195.0506m/z	0.00399	<u>0.01939</u>
3.78neg402.1157m/z	<u>0.02834</u>	0.01674
3.60neg383.0979m/z	<u>0.02123</u>	0.01364
5.24pos180.1031m/z	<u>0.02264</u>	0.01329
5.41pos129.0560m/z	0.00615	<u>0.01216</u>

A goal of this chapter was also to investigate data trends related to potential flavor

interactions. Many researchers initially start with the Pearson's coefficient to understand how compounds relate. However Pearson's is ill suited for discovery work, as it ignores the presence of some time scale (present then not) specific trends and would ignore intermediate compounds. Machine learning was selected to model the data in order to move beyond corollaries. Matrix scatterplots based on Random Forrest modeling are reported in Figure 5.5. This approach can illuminate more information than a normal correlation, as scatterplot densities (diagonal in scatter matrix) can illuminate information that a correlation misses. Contextual variable interactions go beyond corollaries and can identify a number of data trends that are discussed in subsequent sections. A bivariate scatterplot matrix of the top statistically identified compounds from the Random Forest model (Table 5.2) is illustrated by the scatterplot matrix in figure 5.5.

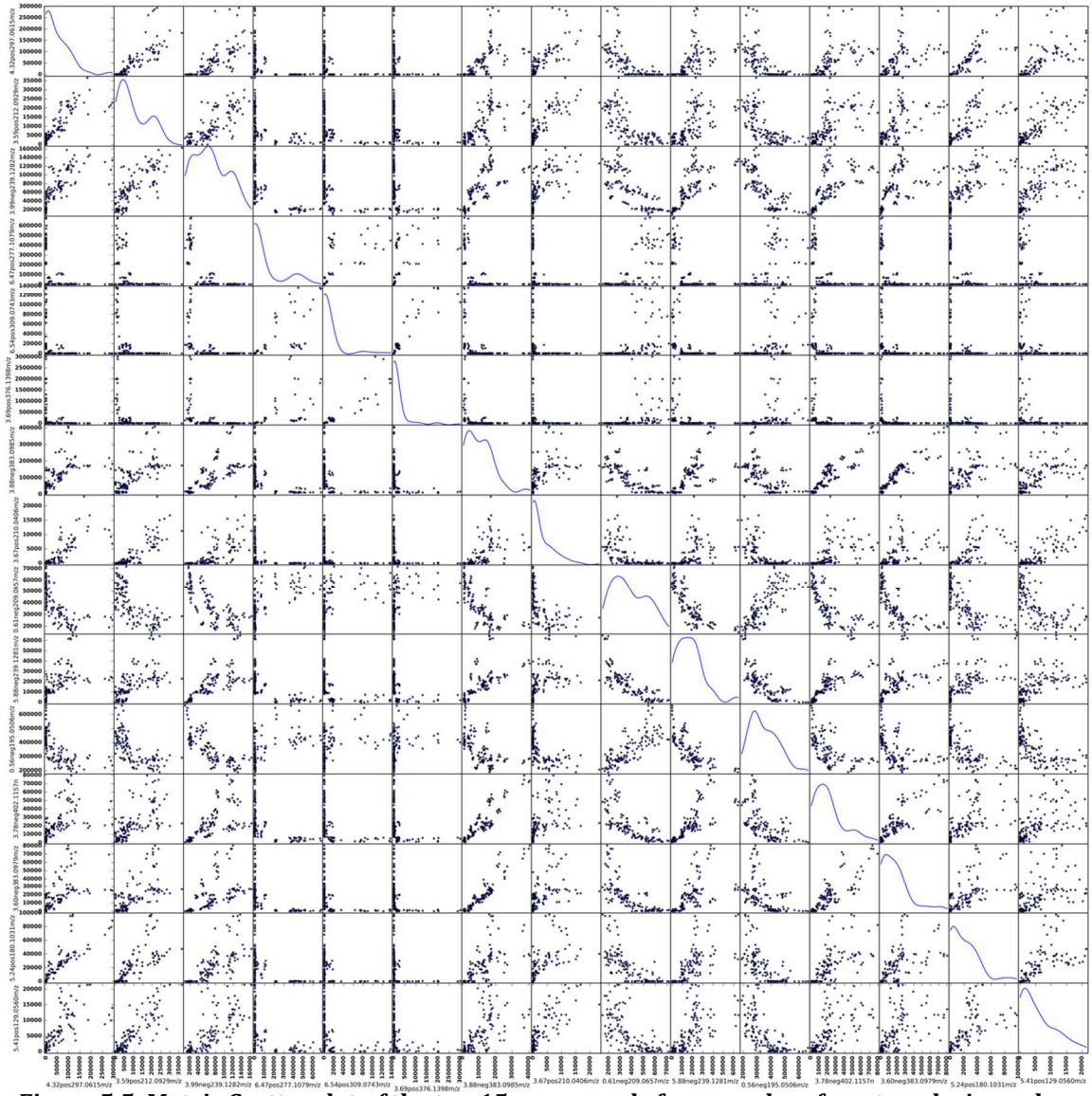


Figure 5.5. Matrix Scatterplot of the top 15 compounds from random forest analysis, each plot is the bi-plot of the compounds concentration, the middle diagonal line is the frequency histogram for the population.

Although there is a lot going on in figure 5.5, trends are clearly observed. A density distribution is presented along the diagonal of the scatter plot. In order to make this plot more interpretable, smaller subsets of these variables are extracted and plotted for investigation and illustration of specific trends, included in this plot is a histogram

distribution, color coded for age. Figure 5.5 was utilized to initially identify what variables have existing concentration dependent relationships. Once the pairs of compounds are identified the date can be remodel in a manner that is more interpretable. This scatterplot matrix holds the same information as plotting each of the pairs individually and acts as an information dense screening approach. In all future variable scatterplots the fresh samples (T0) are colored green, samples aged for 2 days (T2) are colored maroon, samples aged for 4 days (T4) are colored purple, and samples aged for 6 days (T6) are colored dark yellow, while the mix sample (M) is dark blue.

Moving beyond linear relationships there are a number of aspects to these graphs that are of specific interest in this application, namely: binary absence/ presence, negative curve linear relationships, positive split correlations and clusters. Each of these trends relate to concentration dependence, which can translate to a number of important applications. One of the key thoughts behind applying this approach is how to add additional outcomes to normally outcome sparse big data experiments. By identify which variables have unique data trends a chemist can gain value by understanding what chemistries add context. For example, if a known flavor agent wants to be better understood then knowing what data associates with this compound can lead to novel discoveries. Since each of the identified concentration relationships stem from statistically powerful variables and are still relative to a time scale, these relationships provide an additional scientific approach to investigate reaction chemistry or sensory compound interaction relationships.

Figure 5.6 depicted the relationships between variables 6.47pos277.1079, 6.54pos309.0743 and 3.69pos376.1398. Which are all examples of compounds that were reported in the fresh samples (T0), and degraded with time. The histogram showed that

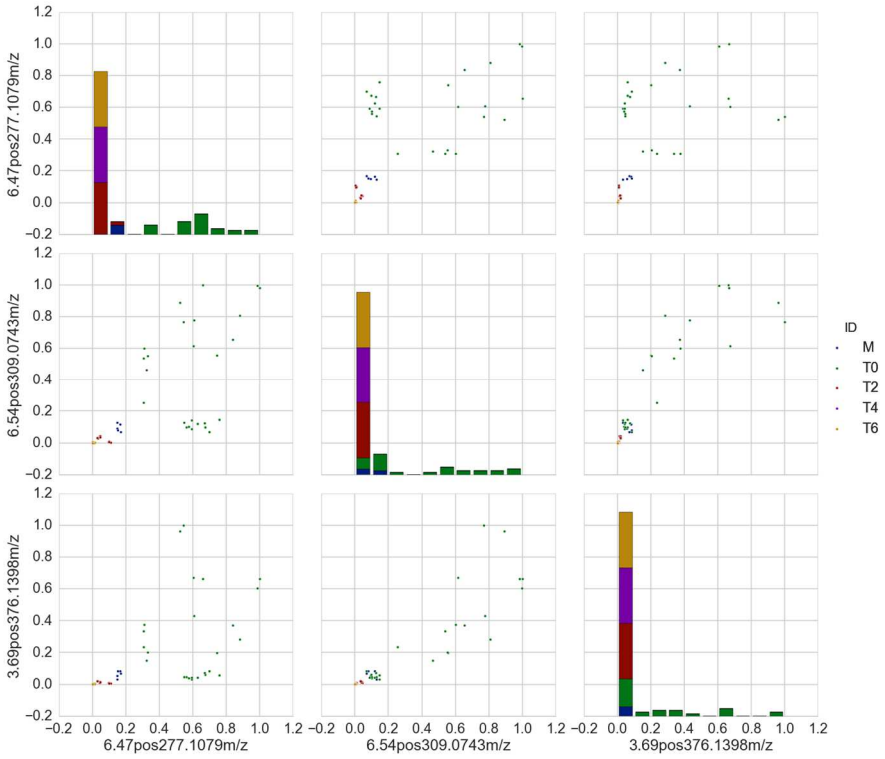


Figure 5.6. This bivariate plot illustrated sets of markers that are initially present in the fresh samples but are degraded after the initial aging period

all of the features have minimal presence in non-fresh samples. This could indicate a number of phenomena but of major interest is the possibility that these reactants degrade after the initial aging period. Reported data trends may relate to compounds that share similar degradation kinetics, or pathways.

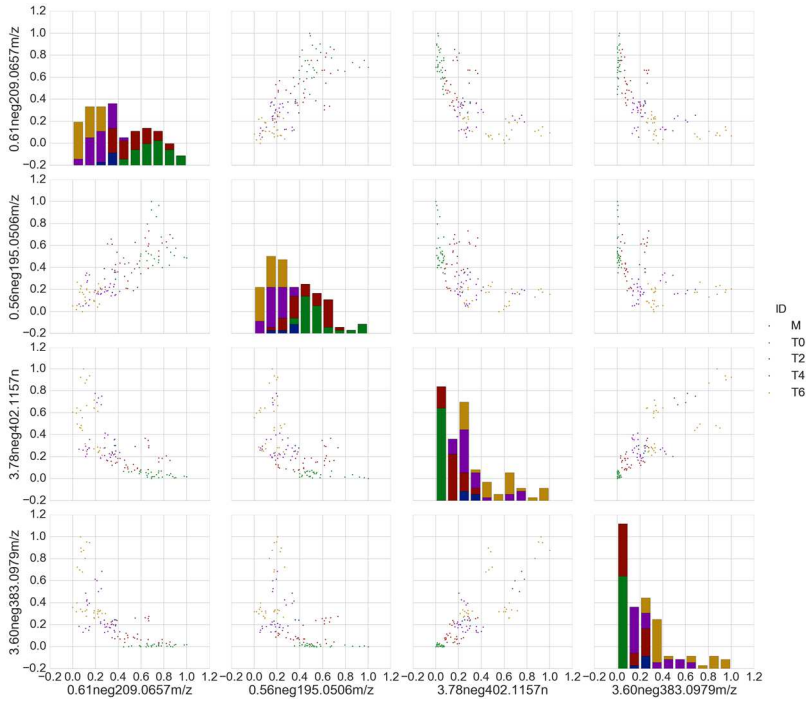


Figure 5.7: Illustrates curve linear bivariate relationships for samples. Aging is further color coded and illustrated a number of compounds that form over time and may help link products to reactants

A negative curve linear relationship is a seen in the plots of 3.78neg402.1157 and 3.60neg383.0979 (Figure 5.7) and similar trends are seen with features relating to 0.61neg209.1281 and 0.56neg195.0506 (Figure 5.7). These relationships show variables that might have a partial concentration dependence or stem from a reaction pathway that has slower kinetics than the trends in Figure 5.6. Within some of these plots there are additional aspects that show interesting behavior, like the bi-plot of 0.56neg195.0506m/z and 3.60neg383.0979m/z with high data density close to the x-axis, showing how much of the population has an intermediate concentration of compound 0.56neg195.0506m/z when there is a baseline concentration of compound 3.60neg383.0979m/z. As compound

3.60neg383.0979m/z increases in concentration compound 0.56neg195.0506m/z reported a decrease in concentration. These negative curve-linear trends are also seen in connection with compounds 3.78neg402.1157m/z and 3.60neg383.0979m/z and the association with compound 0.61neg209.0657m/z. Within each of these interactions there seems to be unique groupings in these larger trends. It is likely these groupings stem from the different varietals. Knowing how these compounds related to a food platform (e.g. citrus) is important, but there might still be interest in how these different compounds are generated based on varietal differences.

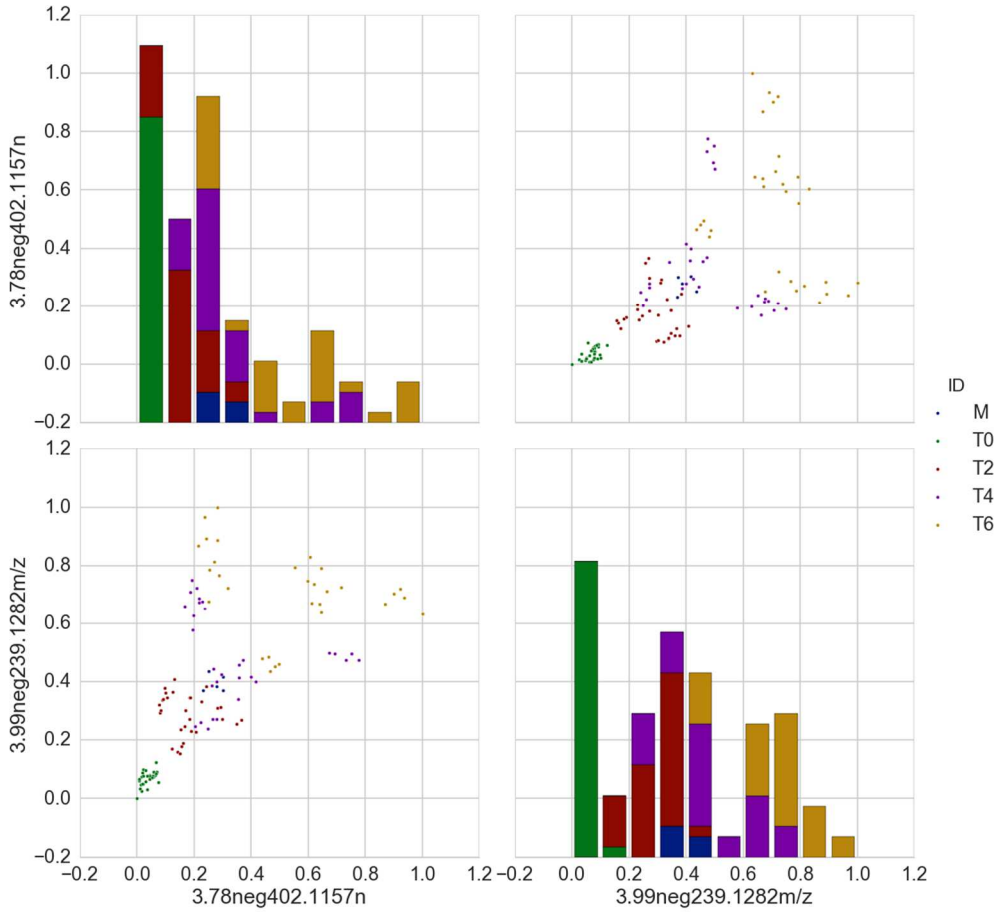


Figure 5.8: Positive correlation with data splitting is seen where compounds are generated as aging occurs but digress at a certain point, as shown by clustering of green samples near the origin and yellow samples showing high concentration of each variable

A positive correlation with a data split seen in compounds 3.78neg402.1157m/z and compound 3.99neg239.1282m/z was observed and illustrated how two variables can follow a linear trend and then fork at a certain concentration (Figure 5.8). This change in the data trend can illustrate how certain varietals may undergo similar reactions at different rates. Interestingly, these compounds also show a trend of formation as all of the fresh samples group near the origin and as the aging progresses the identified features are

generated. Understanding how different varietals age can help researchers identify how differences in food composition leads to differences during aging. This trend is also seen in compounds 5.41pos129.0560m/z and 5.24pos180.1031/z but with a later split in the data (Figure 5.9). This pair of compounds was also reported to generate over time, with many of the fresh, and T2 samples around the plot origin. Other trends were also reported, like those seen between 3.60neg383.0979m/z and 3.99neg239.1282m/z (figure 5.5), that share an initial low concentration and fork at lower concentration than the previous examples.

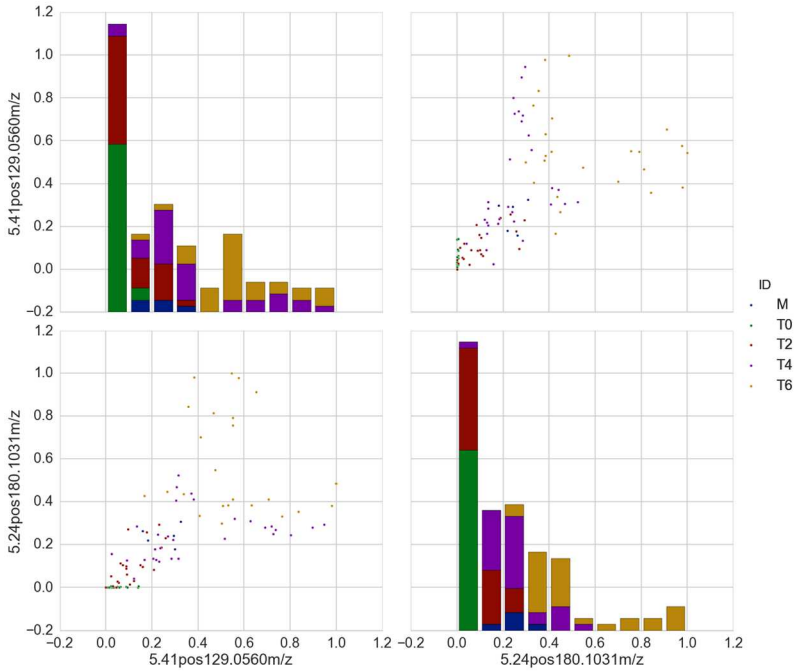


Figure 5.9: Positive correlation with data splitting is seen where compounds are generated as aging occurs, but at differing rates. Showed a stronger skew than the pair depicted in Figure 5.8.

Compound generation over time as influenced by orange variety is shown in Figure

5.10. This plot depicted how varietal impact is present in some bivariate relationships like those seen between 5.88neg239.1281m/z and 3.99neg239.1282m/z (Figure 5.10A). What is particularly interesting about these two compounds is that they share very similar mass to charge ratios, which may strongly indicate degradation of a compound that leads to two similar isomeric moieties with differed polarity. As one compound has a dramatically longer retention on a reversed phase separation. As sample aging progressed there are clear groupings within the plot, which are likely stemming from

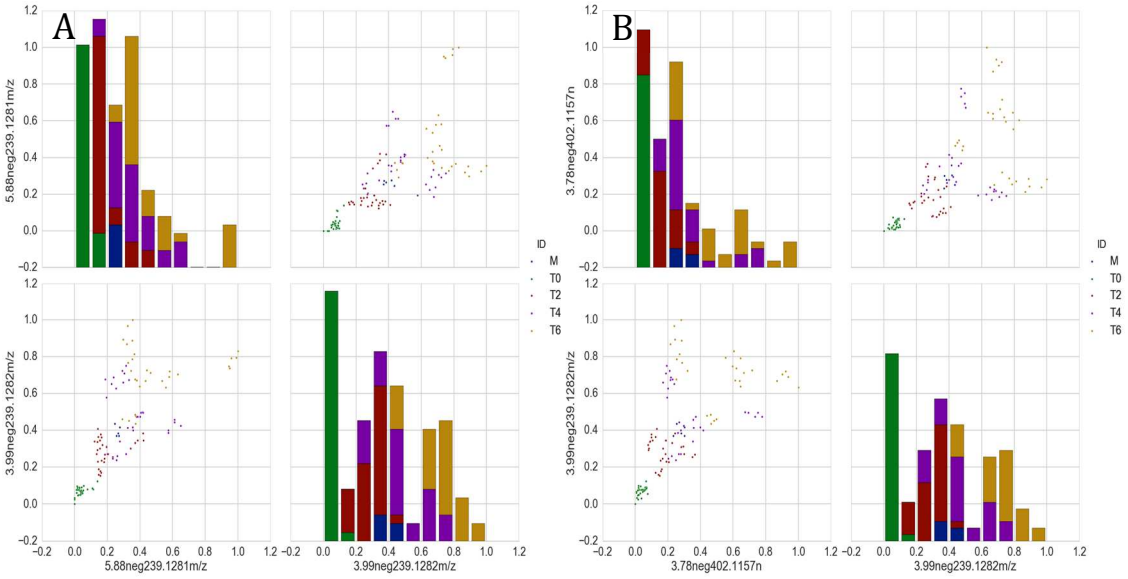


Figure 5.10: Bivariate plots reported formation over time, but has groupings likely stemming from the varietals present in the experiment. This can be an interesting tool to understand varietals can differ in aging.

different varietal chemistry. These can also be visualized with more traditional methods like bar plots, but the unique aspect of using scatter bi-plots is that the clusters are present in the context of two variables. This trend is also present in the variables 3.78neg402.1157m/z and variable 3.99neg239.1282m/z (Figure 5.10B).

This modeling approach identified a number of variables that are contextually

related which can be further targeted for investigative analysis. These approaches can help researchers better establish contextual relationships for unknown compounds. This dissertation has shown the value of working to isolate and evaluate compounds identified through flavoromics (Chapter 3 & 4). However, additional modeling can provide further understanding of how unknown compounds relate to the chemistry of the food. Since isolation, purification and structural elucidation is very expensive, anything that adds value to this workflow is highly desirable. Understanding the contextual relationships may help researchers better understand their relation to a flavor profile or the endogenous food platform.

5.4 Conclusions

As research constantly increases the ability to collect data, a major opportunity stems from effectively characterizing the data, and how conclusions are drawn. Informatics experiments typically are only able to investigate very few compounds, relative to the number analyzed. Likewise being able to draw more from the experimental output is highly desirable. While libraries will continue to improve, they are only as good as the data behind them and will always lack unknowns. This chapter illustrated a workflow to develop a robust machine learning platform, optimize a machine learning method, statistically identify compounds relating to age, and further identify data dependent interactions. Through adding context around variables of importance their relation to a food system may be easier to understand. Being able to rapidly contextualize identified variables can help researchers expand their breadth of

investigation. This adds to the value stream of industry and academic researchers that are already spending instrumental analysis time on informatics platforms.

Chapter 6. Discussion and Conclusion

Investigating the flavor chemistry of food can help address some of the major challenges in food security and support consumer wellbeing through ensuring positive eating experiences of healthful products. Untargeted analytical methods provide a comprehensive analytical approach to support new discoveries. This dissertation has established how untargeted methods are able to identify non-volatile compounds that impact a flavor profile. Increasing the amount of data food researchers investigate helps to better characterize the food systems and reactions while increasing the understanding of complex phenomena. This thesis has established a causative relationship for statistically important compounds through recombination validation, which is a first for the area of flavoromics.

6.1 Flavoromics Value and Application

Flavoromics holds a promising future for flavor analysis as an approach that can supplement existing research paths by identifying areas that may otherwise be missed. As flavoromics uses a high amount of the chemical composition there is a lot of potential to gain new information from the data, where targeted methods are limited to a narrowed focus. Ideally, these two methods would be used together as a way to increase the efficiencies of each while avoiding pitfalls. By better understanding chemical trends and identifying new unknowns (by chemometrics) targeted methods can provide more value to researchers and food processors. In this thesis flavoromics was applied to understand how aging chemistry related to flavor quality, and showed that a number of statistically important compounds led to altered flavor perception. As machine learning and data

science continue to show value through identifying novel information with immediate and real world applications their use in understanding chemistry as it relates to health and consumer acceptance is an exciting space to move towards. When the chemistry of a food is understood obstacles that prevent palatable, healthful and high quality food can be developed. Innovation is needed for the development of healthful, profitable, and likeable food that is currently hindered by the lack of information surrounding food and flavor chemistry.

6.2 Limitations

This dissertation highlights some of the benefits to utilizing flavoromic applications for flavor discovery, but further developing of this method is needed. Untargeted food research is limited by the identification of unknowns, this leads to time intensive isolation, purification and characterization of unknown compounds. The lack of available libraries for non-volatile food constituents makes rapid identification a challenge. This means that a skill set in tandem mass spectrometry and nuclear magnetic resonance is required to identify unknown compounds. In order for flavoromics to achieve wider application this hurdle must be overcome.

The advancement of analytical techniques would benefit flavoromics. As LC-MS based methods rely on separation for accurate reporting, and are not yet able to achieve a "single method that does it all". Even with some of the newest technology achieving optimal separation would lead to extremely time and solvent intensive runs. Newer method like comprehensive two dimensional LC-MS can help address some challenges but are still in their infancy and still suffer from shortcomings. To compromise methods

that shorten separation time are used, which leads to more ion suppression and skew reporting of compounds intensities. Although this facilitates sample throughput, leading to more observations however variable reporting can be skewed. Additional challenges surround multi-block experiments with multiple ionization techniques, as data redundancies can lead to problems in some machine learning applications. There is no singular method that allows for accurate, sensitive and complete reporting of chemical species but with new analytical developments there is likely to be analytical methods that allow for more comprehensive analysis.

Data preprocessing and handling is also showing new advances that address some of the issues that arise from older algorithms. The newer methods and platforms (eg. Progenesis QI) are able to use ionic ratios to group ionic peaks as one compound, this information is further used to aide in compound identification and for elemental composition. These algorithms are able to handle data in a more intuitive way, with added quality control along the workflow and should help researcher produce higher quality data structures. One of the challenges with data processing, even with higher end computing resources, is these systems do not effectively use computer resources and so running preprocessing on a few hundred samples can take a significant amount of time. Again, a balance between the quality, accuracy and depth of data must be struck.

As data science continues to develop alongside statistics new methods and approaches appear that innovate and help researchers solve more specific questions. New methods are able to pull more information out of data structures, with less input from the researcher, driven by the area of Deep Learning (artificial intelligence and neural

networks). Overall these methods can help address some of the above concerns through more efficient analysis and modeling of data, to identifying compounds of interest.

6.3 Future Work

In order to further innovate, food informatics methods need to cross borders and help understand how ingredients and food perform from the seed all the way to the consumer. Understanding how to make ingredients and food profitable, palatable and healthy across the entire processing chain is one way to facilitate change in complex agricultural environments. Ensuring that value is derived across the supply chain, while maintaining flavor attributes is one approach to produce better and more healthful food. Machine learning provides a toolbox to understand immensely complex problems and is an ideal starting place to understand steps along food processing are linked. This work provides a small contribution to understanding a vastly complex food system.

Citations:

1. Aishima, Tetsuo, and Shuryo Nakai. "Chemometrics in flavor research." Food reviews international 7.1 (1991): 33-101.
2. Aishima, T., and S. Nakai. "Pattern recognition of GC profiles for classification of cheese variety." Journal of Food Science 52.4 (1987): 939-942.
3. Andrade, L., Farhat, I. A., Aeberhardt, K., Normand, V., & Engelsen, S. B. (2008). Characterization of encapsulated flavor systems by NIR and low-field TD-NMR: A chemometric approach. *Food Biophysics*, *3*(1), 33-47.
4. Auvray, Malika, and Charles Spence. "The multisensory perception of flavor."*Consciousness and cognition* 17.3 (2008): 1016-1031.
5. Becker, Richard A., and William S. Cleveland. "Brushing scatterplots." Technometrics 29.2 (1987): 127-142.
6. Beebe, K. R., Pell, R. J., & Seasholtz, M. B. (1998). Chemometrics: a practical guide (Vol. 4). Wiley-Interscience.
7. Bijlsma, S., Bobeldijk, I., Verheij, E. R., Ramaker, R., Kochhar, S., Macdonald, I. A., ... & Smilde, A. K. (2006). Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Analytical chemistry, 78(2), 567-574.
8. Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial intelligence, 97(1), 245-271.
9. Breiman, Leo. (2001). Random forests. *Machine learning* 45, no. 1, 5-32.
10. Breiman, L. (2003). Statistical modeling: The two cultures. *Quality control and*

applied statistics, *48*(1), 81-82.

11. Bower, John A., and Robert Whitten. "Sensory characteristics and consumer liking for cereal bar snack foods." Journal of Sensory Studies 15.3 (2000): 327-345.
12. Brunsø, K., Fjord, T. A., Grunert K. G. (2002) *Consumers' food choice and quality perception*. Aarhus School of Business, MAPP-Centre for Research on Customer Relations in the Food Sector.
13. Buettner, Andrea,. Schieberle, Peter. "Evaluation of aroma differences between hand-squeezed juices from Valencia late and Navel oranges by quantitation of key odorants and flavor reconstitution experiments." Journal of Agricultural and Food Chemistry 49.5 (2001): 2387-2394.
14. Cano, Antonio, Alejandro Medina, and Almudena Bermejo. "Bioactive compounds in different citrus varieties. Discrimination among cultivars."*Journal of Food Composition and Analysis* 21.5 (2008): 377-381.
15. Chandrashekar, J., Hoon, M. A., Ryba, N. J., & Zuker, C. S. (2006). The receptors and cells for mammalian taste. Nature, 444(7117), 288-294.
16. Charve, J. I. M. (2011). Prediction of mandarin juice flavor: a flavoromic approach (Doctoral dissertation, University of Minnesota).
17. Chizhik, Vladimir I., Yuri S. Chernyshev, V. V. Frolov, A. V. Komolkin, and M. A. Shelyapina. *Magnetic Resonance and Its Applications*. Springer, 2014.
18. Chu, Fang, and Carlo Zaniolo. "Fast and light boosting for adaptive mining of data streams." In *Advances in knowledge discovery and data mining*, pp. 282-292.

Springer Berlin Heidelberg, 2004.

19. Daszykowski, M., & Walczak, B. (2006). Use and abuse of chemometrics in chromatography. *TrAC Trends in Analytical Chemistry*, *25*(11), 1081-1096.
20. Davison, A. C., and S. Sardy. "The partial scatterplot matrix." Journal of Computational and Graphical Statistics 9.4 (2000): 750-758.
21. de Araujo, Ivan E., Edmund T. Rolls, Maria Inés Velazco, Christian Margot, and Isabelle Cayeux. "Cognitive modulation of olfactory processing." Neuron 46, no. 4 (2005): 671-679.
22. Dettmer, Katja, Pavel A. Aronov, and Bruce D. Hammock. "Mass spectrometry-based metabolomics." *Mass spectrometry reviews* 26, no. 1 (2007): 51-78.
23. Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.
24. Dufour, C., & Bayonove, C. L. (1999). Interactions between wine polyphenols and aroma substances. An insight at the molecular level. Journal of agricultural and food chemistry, 47(2), 678-684.
25. Ernst, Richard R., Geoffrey Bodenhausen, and Alexander Wokaun. *Principles of nuclear magnetic resonance in one and two dimensions*. Vol. 14. Oxford: Clarendon Press, 1987.
26. Frank, R. A., Shaffer, G., & Smith, D. V. (1991). Taste–odor similarities predict taste enhancement and suppression in taste–odor mixture. Chemical Senses, 16, 523.
27. Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we

need hundreds of classifiers to solve real world classification problems?. The Journal of Machine Learning Research, 15(1), 3133-3181.

28. Freund, Yoav, Robert Schapire, and N. Abe. "A short introduction to boosting."*Journal-Japanese Society For Artificial Intelligence* 14, no. 771-780 (1999): 1612.
29. Gates, Sweeley; Sweeley, CC (1978). "Quantitative metabolic profiling based on gas chromatography". Clin Chem 24 (10): 1663–73.
30. Gan, H. H., Soukoulis, C., & Fisk, I. (2014). Atmospheric pressure chemical ionisation mass spectrometry analysis linked with chemometrics for food classification–A case study: Geographical provenance and cultivar classification of monovarietal clarified apple juices. *Food chemistry*, *146*, 149-156.
31. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42.
32. Hasegawa, S., & Miyake, M. (1996). Biochemistry and biological functions of citrus limonoids. Food Reviews International, 12(4), 413-435.
33. Hough, G., Langohr, K., Gómez, G., & Curia, A. (2003). Survival analysis applied to sensory shelf life of foods. Journal of Food, 68(1), 359-362.
34. Izenman, A. J. (2008). Linear discriminant analysis. In Modern Multivariate Statistical Techniques (pp. 237-280).
35. Kwan, E. E., & Huang, S. G. (2008). Structural elucidation with NMR spectroscopy: practical strategies for organic chemists. European journal of organic chemistry, 2008(16), 2671-2688.

36. Keast, R. S. J., & Breslin, P. A. S. (2002). Modifying the bitterness of selected oral pharmaceuticals with cation and anion series of salts. Pharmaceutical Research, 19, 1019-1026.

37. Keast, Russell SJ, and Paul AS Breslin. "Bitterness suppression with zinc sulfate and Na-cyclamate: a model of combined peripheral and central neural approaches to flavor modification." *Pharmaceutical research* 22, no. 11 (2005): 1970-1977.

38. Keast, Russell SJ. "Modification of the bitterness of caffeine." *Food quality and preference* 19, no. 5 (2008): 465-472.

39. Kell, Douglas B., and Stephen G. Oliver. "Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era." Bioessays 26.1 (2004): 99-105

40. Kimura, K., Nishimura, H., Iwata, I., & Mizutani, J. (1983). Deterioration mechanism of lemon flavor. 2. Formation mechanism of off-odor substances arising from citral. Journal of Agricultural and Food Chemistry, 31(4), 801-804.

41. Kuhn, Max, and Kjell Johnson. “Nonlinear Regression Methods: K-Nearest Neighbors” *Applied predictive modeling*. New York: Springer, 2013.

42. Kurihara, Yoshie. "Characteristics of antisweet substances, sweet proteins, and sweetness-inducing proteins." *Critical Reviews in Food Science & Nutrition* 32, no. 3 (1992): 231-252.Lindemann, B. (1996). Chemoreception: tasting the sweet and the bitter. *Current Biology*, *6*(10), 1234-1237.

43. Lee, S. M., Kwon, G. Y., Kim, K. O., & Kim, Y. S. (2011). Metabolomic approach for determination of key volatile compounds related to beef flavor in

glutathione-Maillard reaction products. Analytica chimica acta, 703(2), 204-211.

44. Lindemann, Bernd. "Receptors and transduction in taste." Nature 413, no. 6852 (2001): 219-225.
45. Liu, Yande, Xudong Sun, and Aiguo Ouyang. "Nondestructive measurement of soluble solid content of navel orange fruit by visible–NIR spectrometric technique with PLSR and PCA-BPNN." *LWT-Food Science and Technology*43, no. 4 (2010): 602-607.
46. Madrera, R. Rodriguez, D. Blanco Gomis, and J. J. Alonso. "Characterization of cider brandy on the basis of aging time." Journal of food science 68.6 (2003): 1958-1961.
47. Marti, G., Boccard, J., Mehl, F., Debrus, B., Marcourt, L., Merle, P., ... & Wolfender, J. L. (2014). Comprehensive profiling and compound identification in non-volatile citrus oil residues by mass spectrometry and nuclear magnetic resonance. Food chemistry, 150, 235-245.
48. Mayr, Andreas, Harald Binder, Olaf Gefeller, and Matthias Schmid. "The evolution of boosting algorithms." *Methods Inf Med* 53, no. 6 (2014): 419-427.
49. Mehl, F., Marti, G., Boccard, J., Debrus, B., Merle, P., Delort, E., ... & Rudaz, S. (2014). Differentiation of lemon essential oil based on volatile and non-volatile fractions with various analytical techniques: a metabolomic approach. Food chemistry, 143, 325-335.
50. Mitropoulou, A., Hatzidimitriou, E., & Paraskevopoulou, A. (2011). Aroma release of a model wine solution as influenced by the presence of non-volatile

components. Effect of commercial tannin extracts, polysaccharides and artificial saliva. Food research international, 44(5), 1561-1570

51. Murdoch, D. J., and E. D. Chow. "A graphical display of large correlation matrices." The American Statistician 50.2 (1996): 178-180.
52. Murphy, C., & Cain, W. S. (1980). Taste and olfaction: independence vsinteraction. Physiology and Behavior, 24, 601–605.
53. Murphy, C., Cain, W. S., & Bartoshuk, L. M. (1977). Mutual action oftaste and olfaction. Sensory Processes, 1, 204–211.
54. Ochi, H., Naito, H., Iwatsuki, K., Bamba, T., & Fukusaki, E. (2012). Metabolomics-based component profiling of hard and semi-hard natural cheeses with gas chromatography/time-of-flight-mass spectrometry, and its application to sensory predictive modeling. Journal of bioscience and bioengineering, 113(6), 751-758.
55. Olson, J. C. (1972). What is an esthetic response?. In Proceedings of the Third Annual Conference of the Association for Consumer Research (pp. 167-179).
56. Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." The Journal of Machine Learning Research 12 (2011): 2825-2830.
57. Perez-Cacho, P. R., & Rouseff, R. L. (2008). Fresh squeezed orange juice odor: a review. Critical reviews in food science and nutrition, 48(7), 681-695.
58. Persson, Tyko, Erik, Sydow, and Caj, Åkesson. "The Aroma of canned beef: models for correlation of instrumental and sensory data." *Journal of Food Science* 38, no. 4 (1973): 682-689.

59. Piggott, J. .R, S. .JPiggott,Simpson, andWilliams. 1998. Sensory analysis. International journal of food science & technology 33 (1): 7-12.

60. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment.Communications of the ACM, 45(4), 211-218.

61. Rabi, I. I., S. Millman, P. Kusch, and J. R. Zacharias. "The Molecular Beam Resonance Method for Measuring Nuclear Magnetic Moments. The Magnetic Moments of Li 6 3, Li 7 3 and F 19 9." *Physical review* 55, no. 6 (1939): 526.

62. Rodríguez-Bencomo, J. J., Muñoz-González, C., Andújar-Ortiz, I., Martín-Álvarez, P. J., Moreno-Arribas, M. V., & Pozo-Bayón, M. Á. (2011). Assessment of the effect of the non-volatile wine matrix on the volatility of typical wine aroma compounds by headspace solid phase microextraction/gas chromatography analysis. Journal of the Science of Food and Agriculture, 91(13), 2484-2494.

63. Rouseff, R., Gmitter, F., & Grosser, J. (1994). Citrus breeding and flavour. In Understanding Natural Flavors (pp. 113-127). Springer US.

64. Reineccius GA. 2008. Flavoromics - the next frontier? Abstracts of Papers, 235th ACS National Meeting, New Orleans, LA, United States. American Chemical Society. p. AGFD-061.

65. Salles, C., Hollowood, T. A., Linforth, R. S. T., Taylor, A. J., Le Quéré, J. L., & Étiévant, P. X. (2003). Relating real time flavour release to sensory perception of soft cheeses. In Flavour Research at the Dawn of the Twenty-first Century-Proceedings of the 10th Weurman Flavour Research Symposium, Beaune, France, 25-28 June, 2002. (pp. 170-175). Editions Tec & Doc.

66. Schifferstein, H. N., & Verlegh, P. W. (1996). The role of congruency and pleasantness in odor-induced taste enhancement. Acta Psychologica, 94, 87–105.

67. Shin, Y. K., Martin, B., Golden, E., Dotson, C. D., Maudsley, S., Kim, W., ... & Munger, S. D. (2008). Modulation of taste sensitivity by GLP□1 signaling. Journal of neurochemistry, 106(1), 455-463.

68. Schutz, H. G., & Wahl, O. L. (1981). Consumer perception of the relative importance of appearance, flavor and texture to food acceptance. In *Criteria of Food Acceptance Symposium Proceedings* (pp. 97-116).

69. Shaw, Philip E., James H. Tatum, and Charles W. Wilson III. "Improved flavor of navel orange and grapefruit juices by removal of bitter components with. beta.-cyclodextrin polymer." *Journal of Agricultural and Food Chemistry* 32.4 (1984): 832-836.

70. Shmueli, Galit. "To explain or to predict?." *Statistical science* (2010): 289-310.

71. Simpson, Nigel JK. *Solid-phase extraction: principles, techniques, and applications*. CRC press, 2000.

72. Smith, Colin A., Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification." *Analytical chemistry* 78, no. 3 (2006): 779-787.

73. Snyder, Lloyd R., Joseph J. Kirkland, and John W. Dolan. *Introduction to modern liquid chromatography*. John Wiley & Sons, 2011.2

74. Srivastava, Santosh, Maya R. Gupta, and Béla A. Frigyik. "Bayesian Quadratic

Discriminant Analysis." Journal of Machine Learning Research 8.6 (2007): 1277-1305.

75. Sutton, Oliver. "Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction." University lectures, University of Leicester (2012).
76. Tautenhahn, R., Patti, G. J., Kalisiak, E., Miyamoto, T., Schmidt, M., Lo, F. Y., McBee, J., Baliga, N. S., Siuzdak, G. (2010). metaXCMS: second-order analysis of untargeted metabolomics data. Analytical chemistry, 83(3), 696-700.
77. Tikunov, Y., Lommen, A., de Vos, C. R., Verhoeven, H. A., Bino, R. J., Hall, R. D., & Bovy, A. G. (2005). A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiology, 139(3), 1125-1137.
78. Tewari, Jagdish C., Vivechana Dixit, Byoung-Kwan Cho, and Kamal A. Malik. "Determination of origin and sugars of citrus fruits using genetic algorithm, correspondence analysis and partial least square combined with fiber optic NIR spectroscopy." *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 71, no. 3 (2008): 1119-1127.
79. Totlani, Vandana M., and Devin G. Peterson. "Epicatechin carbonyl-trapping reactions in aqueous maillard systems: identification and structural elucidation."*Journal of agricultural and food chemistry* 54, no. 19 (2006): 7311-7318.
80. Togari, N., A. Kobayashi, and T. Aishima. "Relating sensory properties of tea

aroma to gas chromatographic data by chemometric calibration methods." Food Research International 28, no. 5 (1995): 485-493.

81. Trygg, J., Holmes, E., & Lundstedt, T. (2007). Chemometrics in metabonomics. *Journal of proteome research*, *6*(2), 469-479.
82. Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). Journal of chemometrics, 16(3), 119-128.
83. USDA Foreign Agricultural Service. 2015. Citrus: World Markets and Trade- http://apps.fas.usda.gov/psdonline/circulars/citrus.pdf
84. Van der Heijden, A., Brussel, L. B. P., Kosmeijer, J. G., & Peer, H. G. (1983). Effects of salts on perceived sweetness. *Zeitschrift für Lebensmittel-Untersuchung und Forschung*, *176*(5), 371-375.
85. Weckwerth, W. (2007). Metabolomics: methods and protocols (Vol. 358). Springer Science & Business Media.
86. Weinberger, Kilian Q., John Blitzer, and Lawrence K. Saul. "Distance metric learning for large margin nearest neighbor classification." In *Advances in neural information processing systems*, pp. 1473-1480. 2005.
87. Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.
88. Wold, S., Eriksson, L., Trygg, J., & Kettaneh, N. (2004). The PLS method–partial least squares projections to latent structures–and its applications in industrial RDP (research, development, and production). Unea University.
89. Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component

analysis."*Chemometrics and intelligent laboratory systems* 2, no. 1 (1987): 37-52.

90. Wong, Pak Chung, and R. Daniel Bergeron. "30 Years of Multidimensional Multivariate Visualization." Scientific Visualization. 1994.

91. Zhang, Harry. "The optimality of naive Bayes." *AA* 1, no. 2 (2004): 3.

Appendix I.

Raw Sensory Data

Solvent Assisted Flavor Evaporation Data

Sample	Panelist	Sweet	Sour	Bitter	Astringency	Orange Character	Orange Peel	Cooked	Green	Floral	Green Bean	Rep
Blank	1	5	0	3	1	7	4	2	4	5	1.5	1
Blank	1	6	2	2	0	6	5	3	4	4	1	2
Blank	2	4	5	2	0	6	3	1	4	3	0	1
Blank	2	3.5	5	4	1	6.5	3	2	3.5	2	2	2
Blank	3	6	1	2	0	5	1	3	1.5	2	0	1
Blank	3	6	0	3	0	5	2	1	1.5	2	0	2
Blank	4	4	1	3	0	3	4	3	2	1	1	1
Blank	4	4	1	4	0	4	3	3	0	1	2	2
Blank	5	5	4	4	1	5	2	0	1	2	0	1
Blank	5	5	4	4	1	5	1	0	1	2	0	2
Blank	6	5	3	3	0	6	4	1	2	1	0	1
Blank	6	7	2	3	1	5.5	2	2	1	1	2	2
Blank	7	5	1	2	0	6	2	3	2	1.5	1	1
Blank	7	4	1	3.5	0	3.5	1	3	0	1.5	1.5	2
Compound 383	1	6	0	4	1	5	4	6	2	3	3	1
Compound 383	1	6	2	4	1	4	3	6.5	1.5	2	3.5	2
Compound 383	2	3	5	2	0	6	3	3	3	2	3	1
Compound 383	2	4.5	5	4	0	5.5	3	5	3	2.5	2	2
Compound 383	3	7	0	3	0	2	1	2	0	1	1	1
Compound 383	3	8	0	3	0	3	1	1	0	1	1	2
Compound 383	4	4	1	3	0	2	4	4	0	1	3.5	1
Compound 383	4	4	1	3	0	3	3	4	1	1	2.5	2
Compound 383	5	4	5	4	2	1	1	0	1	1.75	0.5	1
Compound 383	5	4	5	3	2	3	1	1	2	2	0.5	2
Compound 383	6	6	4	3	0	5	1	3	1	0	3	1
Compound 383	6	7	3	2	0	5	4	1	3	0	2	2

Compound 383	7	3.5	2	0.5	0	5.5	1	2	1	1	0	1
Compound 383	7	5	1	1	0	2	1	1	1	1	0	2
Compound 413	1	7	2	2	2	5	4	7	2	3	5	1
Compound 413	1	6	0	4	0	3	1	7	1	1	5.5	2
Compound 413	2	4	6	3.5	1	6	3	4	2	3	2	1
Compound 413	2	3	6	4.5	0	5	2	5	2	1	3	2
Compound 413	3	7	0	4	0	2	0	4	0	1	0	1
Compound 413	3	8	0	2	0	1	0	3	0	1	0	2
Compound 413	4	2	1	4	0	4	3	4	1	1	2	1
Compound 413	4	4	1	3	0	2	4	5	1	0	3	2
Compound 413	5	6	3	2	0	4	2	2	3	2	1	1
Compound 413	5	3.5	3	2	1	3	2	1	1	1	1	2
Compound 413	6	7	3	2	1	4	4	1	2	0	1	1
Compound 413	6	6	3	2	0	4.5	1	4	0	0	1.5	2
Compound 413	7	5	1	1	0	5	1	4	1	1	1	1
Compound 413	7	5	1	2	0	5	2	4	1	1	0	2
Compound191	1	8	2	1	2	3	2	7	2	1	5	1
Compound191	1	5.5	2	2	2	7	6	1	3	5	1	2
Compound191	2	3	4	4.5	0	5.5	2	4.5	2	1.5	3	1
Compound191	2	3.5	3.5	3	0	6	2	5	1	1	3	2
Compound191	3	5	0	3	0	2	0	1	1	1	0	1
Compound191	3	6	0	1	0	3	2	0	0	1	0	2
Compound191	4	2	1	6	0	1	1	1	0	1	2	1
Compound191	4	5	0	3	0	5	3	3	1	1	2	2
Compound191	5	4	3	5	1	2	1	2	3	3	2	1
Compound191	5	6	4	3	0	1	2	2	2	2	0	2

Compound191	6	5	2	3	0	5	2	2	1	0	1.5	1
Compound191	6	6	3	2	0	5	2	4	0	0	1.5	2
Compound191	7	3	1	1	0	6	1	2	2	1	1	1
Compound191	7	5	1	2	1	5	2	3	2	2	1	2
Compound693	1	7	4	4	2	6	5	2	4	5	1	1
Compound693	1	5	0	6	0	5	7	4	2	2	2	2
Compound693	2	4	5	4	0	6	2	5	2	2	2.5	1
Compound693	2	4	6	2	0	6	3	2.5	3	3	1.5	2
Compound693	3	6	0	2	0	3	0	2	0	1	0	1
Compound693	3	9	0	3	0	3	1	1	1	1	0	2
Compound693	4	4	2	2	0	2	3	4	2	2	2	1
Compound693	4	5	0	2	0	3	3	3	1	0	1	2
Compound693	5	5	4	5	2	2	1	1	2	2	1	1
Compound693	5	4	2	3	1	4	3	2	1	2	2	2
Compound693	6	7	3	3	0	5	2	4	1	0	2	1
Compound693	6	5	3	3	0	5	1	4	1	1	3	2
Compound693	7	4	2	4	0	3	1	4	1	0	2	1
Compound693	7	5	1	1	0	3	1	2	0	1	1	2

Commercial Volatile Aroma Flavor

Sample	Panelist	Sweet	Sour	Bitter	Astringency	Orange Character	Orange Peel	Cooked	Green	Floral	Green Bean	Rep
Compound 383	1	6	3	1	1	5	3	1	3	4	2	1
Compound 383	1	6	3	1	3	6	3	0	3	4	0	2
Compound 383	2	4	2	1	3	4	1	2	2	0	1	1
Compound 383	2	6	2	2	3	4	2	1	3	3	1	2
Compound 383	3	5	3	1	1	4	3	2	1	0	1	1

Compound 383	3	6	3	1	1	6	2	1	2	2	0	2
Compound 383	4	6	2	2	1	4	3	1	1	2	1	1
Compound 383	4	6	2	2	1	4	2	1	2	2	1	2
Compound 383	5	6	0	0	0	5	1	1	0	1	1	1
Compound 383	5	5	1	2	0	4	3	2	2	3	1	2
Compound 383	6	5	2	1	1	3	3	2	0	1	1	1
Compound 383	6	5	2	0	1	5	4	0	3	2	0	2
Compound 383	7	6	2	0	0	3	4	1	0	1	0	1
Compound 383	7	5	3	0	0	5	3	1	1	0	0	2
Compound 383	8	4	2	2	1	4	3	1	2	1	2	1
Compound 383	8	5	2	2	1	4	3	0	2	2	1	2
Compound 383	9	4	3	1	2	4	2	1	2	0	1	1
Compound 383	9	5	2	1	2	4	2	2	3	2	0	2
Compound 413E1	1	5	1	2	1	5	3	0	3	4	1	1
Compound 413E1	1	6	3	2	1	5	3	1	2	4	1	2
Compound 413E1	2	5	2	1	4	4	1	2	1	1	1	1
Compound 413E1	2	5	2	2	4	5	2	1	3	2	0	2
Compound 413E1	3	6	2	1	1	5	3	2	1	1	1	1
Compound 413E1	3	6	2	1	1	6	3	2	2	1	1	2
Compound 413E1	4	6	2	2	1	4	3	2	2	2	1	1
Compound 413E1	4	5	2	1	1	4	2	1	1	2	1	2
Compound 413E1	5	6	1	0	0	4	1	2	0	1	0	1
Compound 413E1	5	5	1	0	0	4	2	1	2	2	0	2
Compound 413E1	6	5	2	1	1	5	4	1	2	2	1	1
Compound 413E1	6	6	1	0	1	4	3	2	1	2	2	2
Compound 413E1	7	5	3	1	0	3	3	4	0	0	1	1
Compound 413E1	7	5	3	0	0	4	3	2	1	1	0	2
Compound 413E1	8	5	2	2	1	4	2	1	2	2	1	1
Compound 413E1	8	4	2	1	1	4	2	2	2	1	2	2

Compound 413E1	9	4	1	0	2	5	3	0	2	1	0	1
Compound 413E1	9	5	2	1	1	4	1	2	1	1	0	2
Compound 413E2	1	6	2	3	2	4	3	2	1	3	1	1
Compound 413E2	1	6	1	1	1	4	2	1	1	3	0	2
Compound 413E2	2	5	1	0	3	4	2	1	1	1	1	1
Compound 413E2	2	5	1	0	2	5	2	0	2	1	0	2
Compound 413E2	3	6	2	1	1	4	2	1	1	1	0	1
Compound 413E2	3	6	2	1	1	4	3	2	1	1	0	2
Compound 413E2	4	5	2	2	1	4	2	1	1	2	1	1
Compound 413E2	4	5	3	2	1	4	2	1	1	3	1	2
Compound 413E2	5	6	2	1	0	4	2	2	0	2	0	1
Compound 413E2	5	5	1	0	0	4	3	1	2	1	0	2
Compound 413E2	6	5	2	1	1	5	3	1	2	1	0	1
Compound 413E2	6	5	2	0	1	5	4	1	2	1	0	2
Compound 413E2	7	5	3	0	0	3	2	3	1	2	0	1
Compound 413E2	7	5	2	0	0	3	2	4	0	1	1	2
Compound 413E2	8	5	3	0	1	4	3	0	1	2	0	1
Compound 413E2	8	5	2	1	1	4	3	0	1	2	1	2
Compound 413E2	9	5	0	1	0	5	2	0	2	0	0	1
Compound 413E2	9	6	1	1	1	4	0	1	2	3	0	2
Compound 457	1	6	3	2	2	5	4	2	2	3	2	1
Compound 457	1	5	3	2	1	5	2	1	2	3	0	2
Compound 457	2	5	2	2	3	5	2	0	3	2	0	1
Compound 457	2	6	2	1	3	4	1	2	1	1	2	2
Compound 457	3	6	3	1	1	4	2	2	1	1	1	1
Compound 457	3	6	2	1	1	6	3	3	1	1	1	2
Compound 457	4	6	3	2	1	4	3	1	2	2	1	1
Compound 457	4	5	3	2	1	4	3	1	1	2	1	2
Compound 457	5	6	1	0	0	5	1	2	0	2	0	1

Compound 457	5	5	2	0	0	3	1	2	0	2	0	2
Compound 457	6	5	2	1	0	3	3	1	0	1	0	1
Compound 457	6	6	2	0	0	4	3	0	1	1	0	2
Compound 457	7	6	1	1	0	4	2	4	1	0	2	1
Compound 457	7	4	2	0	0	3	3	4	0	0	2	2
Compound 457	8	5	2	1	1	3	3	2	1	1	2	1
Compound 457	8	5	2	1	1	4	2	2	1	1	1	2
Compound 457	9	4	2	1	1	4	2	2	0	2	0	1
Compound 457	9	5	2	1	2	5	2	0	2	2	0	2
Compound 661	1	5	3	3	4	4	3	2	3	3	1	1
Compound 661	1	4	2	1	1	5	4	0	3	4	0	2
Compound 661	2	4	2	2	3	4	1	3	2	0	1	1
Compound 661	2	6	2	1	4	4	1	2	2	1	1	2
Compound 661	3	6	2	1	1	5	2	1	1	1	0	1
Compound 661	3	6	2	1	1	6	3	1	1	2	0	2
Compound 661	4	5	2	2	1	4	2	1	2	1	1	1
Compound 661	4	5	2	2	1	4	3	2	1	1	1	2
Compound 661	5	5	2	1	0	4	1	2	1	1	2	1
Compound 661	5	5	3	1	1	5	3	2	3	2	2	2
Compound 661	6	5	2	1	1	5	4	1	1	2	1	1
Compound 661	6	4	3	0	1	5	4	0	1	1	0	2
Compound 661	7	4	3	0	0	4	1	1	1	0	1	1
Compound 661	7	4	3	0	0	4	2	2	1	0	1	2
Compound 661	8	5	2	1	1	4	3	1	2	2	1	1
Compound 661	8	5	2	1	1	4	3	1	1	1	1	2
Compound 661	9	5	1	0	0	4	1	2	0	0	1	1
Compound 661	9	5	2	1	1	4	1	2	1	1	1	2
Compound 693	1	5	4	4	1	5	3	3	3	4	1	1
Compound 693	1	5	2	3	1	5	4	0	3	3	1	2

Compound 693	2	6	2	0	2	4	1	1	1	1	1	1
Compound 693	2	4	1	1	3	5	2	1	2	1	0	2
Compound 693	3	5	3	1	1	6	2	2	1	1	1	1
Compound 693	3	5	3	1	1	4	3	2	1	1	0	2
Compound 693	4	5	2	2	1	4	2	1	1	2	1	1
Compound 693	4	5	2	2	1	4	2	2	1	2	1	2
Compound 693	5	5	1	0	0	4	2	1	0	2	0	1
Compound 693	5	5	1	0	1	3	2	2	3	3	1	2
Compound 693	6	6	2	1	1	4	4	2	1	2	1	1
Compound 693	6	5	2	1	0	4	3	1	2	2	0	2
Compound 693	7	4	2	0	0	3	2	4	1	0	3	1
Compound 693	7	4	3	0	0	5	3	1	1	0	0	2
Compound 693	8	4	2	1	1	4	3	2	2	1	1	1
Compound 693	8	5	2	2	1	4	3	1	2	1	1	2
Compound 693	9	4	3	1	2	4	2	1	1	0	2	1
Compound 693	9	5	3	1	2	4	2	2	3	1	1	2
Sample Blank	1	5	2	1	1	6	3	1	3	4	0	1
Sample Blank	1	4	2	1	1	5	3	0	3	4	1	2
Sample Blank	2	6	2	1	4	5	3	1	3	3	0	1
Sample Blank	2	4	1	1	3	4	1	2	2	1	2	2
Sample Blank	3	5	2	1	1	4	3	3	1	1	2	1
Sample Blank	3	6	2	1	1	4	2	3	1	1	2	2
Sample Blank	4	5	3	2	1	4	2	1	2	2	1	1
Sample Blank	4	5	2	2	1	4	2	1	2	2	1	2
Sample Blank	5	5	1	0	0	3	1	2	1	2	0	1
Sample Blank	5	5	2	2	0	3	1	2	0	3	0	2
Sample Blank	6	5	2	0	1	4	2	3	1	2	2	1
Sample Blank	6	4	2	1	0	4	3	3	1	2	2	2
Sample Blank	7	6	2	0	0	6	3	3	1	0	1	1

Sample Blank	7	4	2	0	0	4	3	2	1	1	2	2
Sample Blank	8	5	2	1	1	4	3	2	1	1	1	1
Sample Blank	8	4	2	1	1	4	2	2	1	1	1	2
Sample Blank	9	4	2	1	1	4	2	2	0	2	0	1
Sample Blank	9	4	2	1	2	5	2	1	2	2	0	2

SAFE extract:

Fit:aov(formula = Orange.Character ~ Sample + Panelist, data = sensory)

	Estimate	Std. Error	t value	Pr(>|t|)
Compound 383 - Blank== 0	-1.5357	0.4029	-3.812	0.00125
Compound 413E1 - Blank== 0	-1.4286	0.4029	-3.546	0.00292
Compound191 - Blank == 0	-1.2143	0.4029	-3.014	0.01358
Compound693 - Blank == 0	-1.25	0.4029	-3.103	0.01065

Fit: aov(formula = Cooked ~ Sample + Panelist, data = sensory)

	Estimate	Std. Error	t value	Pr(>|t|)
Compound 383 - Blank== 0	0.8929	0.4923	1.814	0.221
Compound 413E1 - Blank== 0	2	0.4923	4.062	<0.001
Compound191 - Blank == 0	0.75	0.4923	1.523	0.363
Compound693 - Blank == 0	0.9643	0.4923	1.959	0.167

Fit: aov(formula = Green.Bean ~ Sample + Panelist, data = sensory)

	Estimate	Std. Error	t value	Pr(>|t|)
Compound 383 - Blank == 0	0.9643	0.3537	2.726	0.0292
Compound 413E1 - Blank == 0	1	0.3537	2.827	0.0223
Compound191 - Blank == 0	0.7857	0.3537	2.222	0.0968
Compound693 - Blank == 0	0.6429	0.3537	1.818	0.2191

Fit: aov(formula = Floral ~ Sample + Panelist, data = sensory)

	Estimate	Std. Error	t value	Pr(>|t|)
Compound 383 - Blank == 0	-0.6964	0.2815	-2.474	0.05449
Compound 413E1 - Blank == 0	-0.9286	0.2815	-3.298	0.00603
Compound191 - Blank == 0	-0.6071	0.2815	-2.157	0.11147
Compound693 - Blank == 0	-0.5	0.2815	-1.776	0.23665

Model Orange Beverage:

Fit: aov(formula = Sweet ~ Sample + Panelist, data = sensory)

	Estimate	Std. Error	t value	Pr(>|t|)
383 Recombination - Sample Blank	0.5	0.1975	2.532	0.0604 .
413E1 Recombination - Sample Blank	0.44444	0.1975	2.25	0.1171
413E2 Recombination - Sample Blank	0.55556	0.1975	2.813	0.0291 *
457 Recombination - Sample Blank	0.55556	0.1975	2.813	0.0291 *
661 Recombination - Sample Blank	0.11111	0.1975	0.563	0.9848
693 Recombination - Sample Blank	0.05556	0.1975	0.281	0.9996

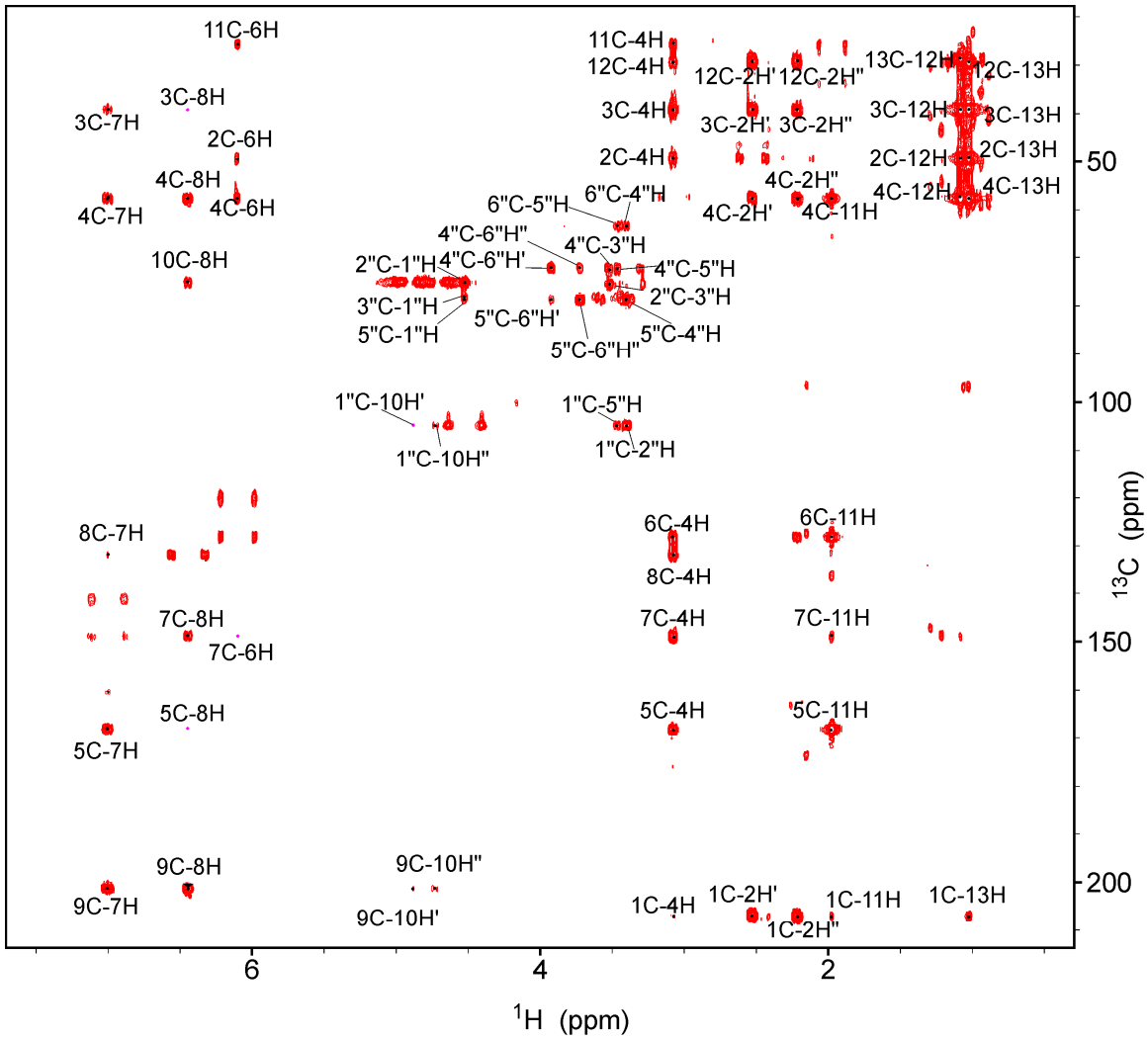
Fit: aov(formula = Cooked ~ Sample + Panelist, data = sensory)				
	Estimate	Std. Error	t value	Pr(>|t|)
383 Recombination - Sample Blank	-0.7778	0.2773	-2.805	0.0297 *
413E1 Recombination - Sample Blank	-0.3333	0.2773	-1.202	0.6838
413E2 Recombination - Sample Blank	-0.6667	0.2773	-2.404	0.0823 .
457 Recombination - Sample Blank	-0.1667	0.2773	-0.601	0.979
661 Recombination - Sample Blank	-0.4444	0.2773	-1.603	0.4019
693 Recombination - Sample Blank	-0.2778	0.2773	-1.002	0.82

Fit: aov(formula = Floral ~ Sample + Panelist, data = sensory)				
	Estimate	Std. Error	t value	Pr(>|t|)
383 Recombination - Sample Blank	-0.2222	0.2181	-1.019	0.809
413E1 Recombination - Sample Blank	-0.2222	0.2181	-1.019	0.809
413E2 Recombination - Sample Blank	-0.2222	0.2181	-1.019	0.809
457 Recombination - Sample Blank	-0.3889	0.2181	-1.783	0.298
661 Recombination - Sample Blank	-0.6111	0.2181	-2.802	0.030 *
693 Recombination - Sample Blank	-0.3889	0.2181	-1.783	0.298

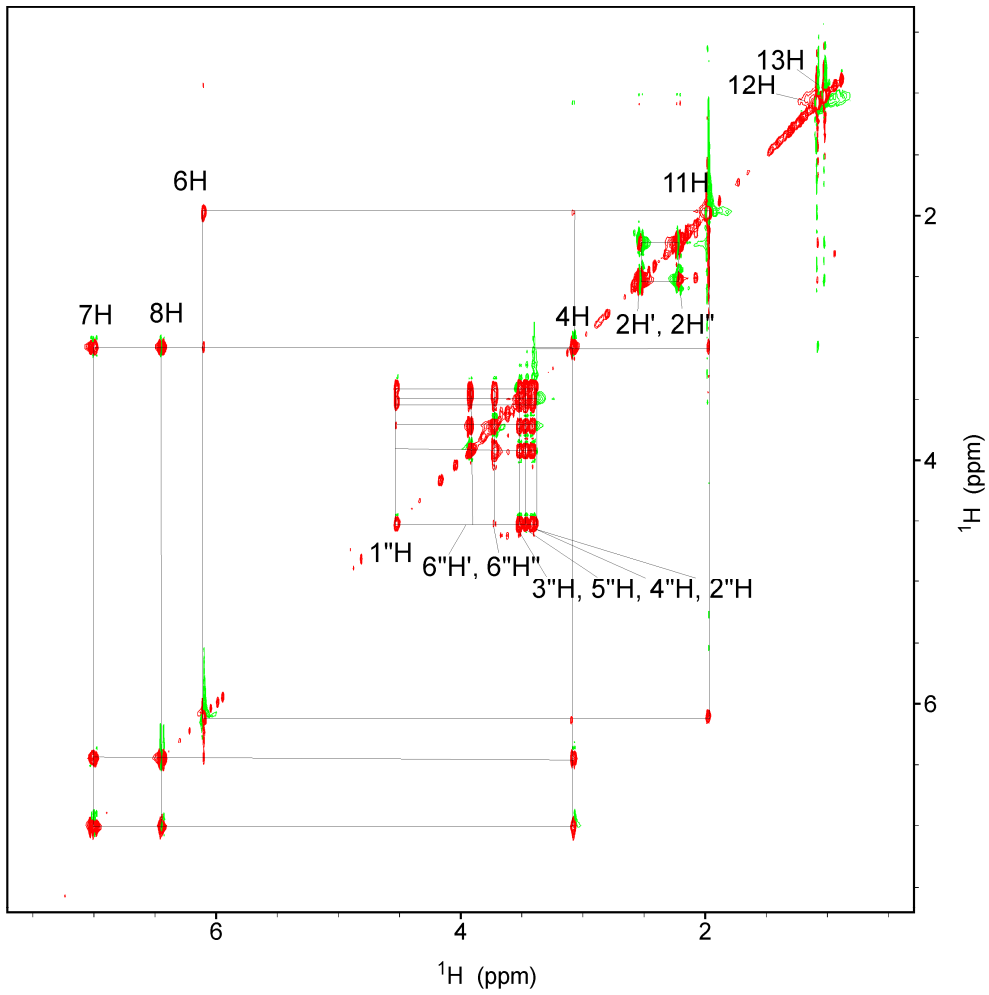
Fit: aov(formula = Green_Bean ~ Sample + Panelist, data = sensory)				
	Estimate	Std. Error	t value	Pr(>|t|)
383 Recombination - Sample Blank	-0.2222	0.2284	-0.973	0.8374
413E1 Recombination - Sample Blank	-0.2222	0.2284	-0.973	0.8375
413E2 Recombination - Sample Blank	-0.6667	0.2284	-2.918	0.0217 *
457 Recombination - Sample Blank	-0.1667	0.2284	-0.73	0.9484
661 Recombination - Sample Blank	-0.1111	0.2284	-0.486	0.9928
693 Recombination - Sample Blank	-0.1111	0.2284	-0.486	0.9928

Appendix II. NMR Data

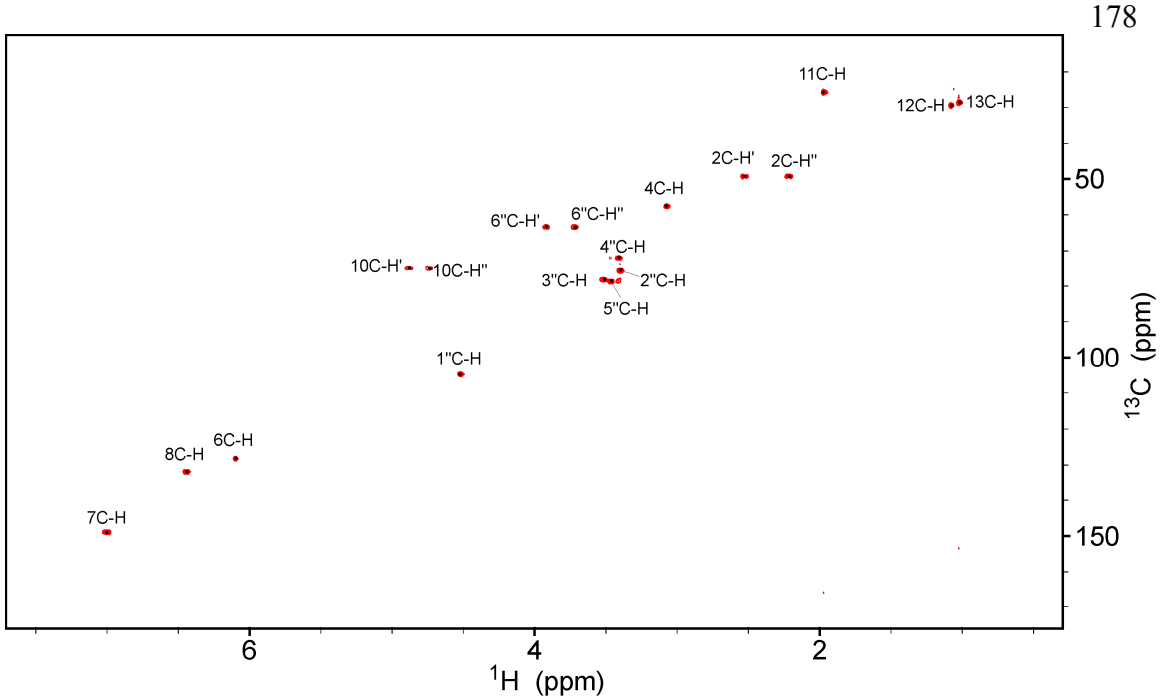
Compound 383:



HMBC of compound 383



TOCSY of Compound 383



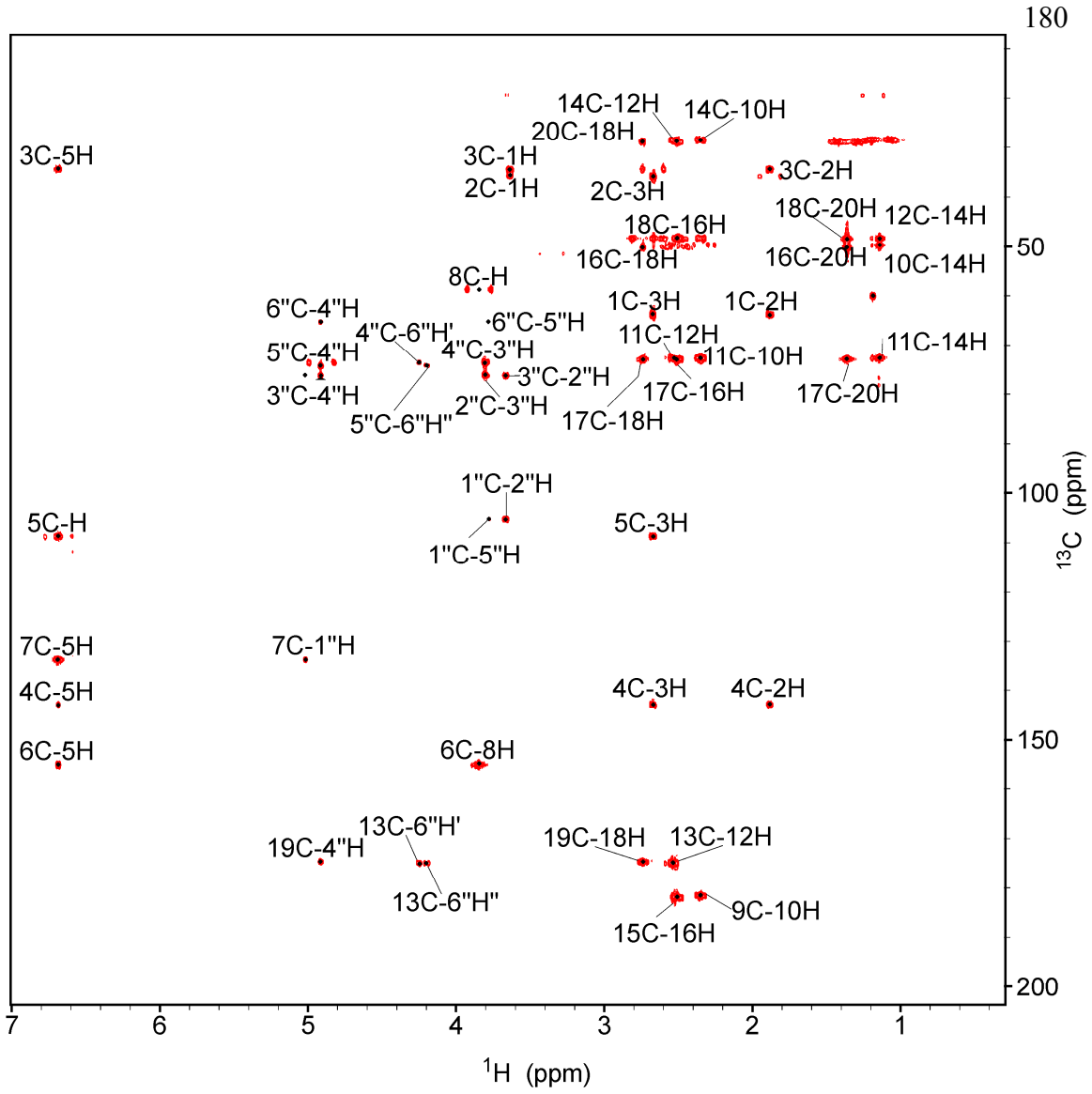
HSQC of Compound 383

Correlation Table:

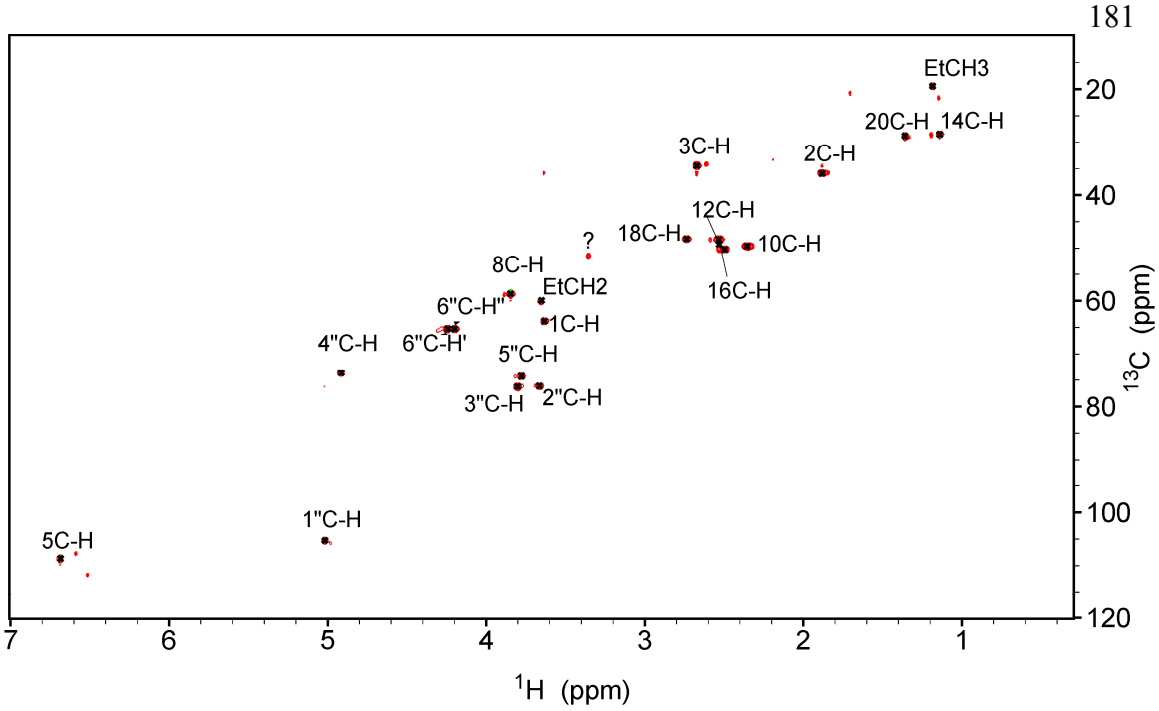
Position	^{13}C (ppm)	^{1}H (ppm)	multiplicity (J in Hz)	integral/number of ^{1}H	^{1}H-^{1}H correlation in TOCSY	^{1}H-^{13}C correlation in HMBC
1	207.0					
2	49.3	2.21, 2.53	d (16.8), d(16.8)	1, 1	2H', 2H"	1C, 3C, 4C, 12C, 13C
3	39.1					
4	57.6	3.08	d (9.3)	1	7H, 8H	1C ,2C, 3C, 5C, 6C, 7C, 8C, 11C, 12C
5	168.2					
6	128.1	6.10	s	1	11H, 4H	2C, 4C, 7C, 11C
7	148.9	7.00	dd (15.8, 9.3)	1	4H, 8H	3C, 4C, 5C, 8C, 9C
8	131.8	6.45	d (15.8)	1	4H, 7H	3C, 4C, 5C, 7C, 9C, 10C
9	201.3					
10	75.0	4.73, 4.88	d (8.0), d(8.0)	1, 1	10H', 10H"	1"C, 9C
11	25.6	1.98	s	3	4H, 6H	1C, 4C, 5C, 6C, 7C,
12	29.3	1.08	s	3		2C, 3C, 4C, 13C
13	28.5	1.02	s	3		2C, 3C, 4C, 12C

1”	104.8	4.53	d (6.9)	1	1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”	2"C, 3"C, 5"C
2”	75.4	3.40	m	1	1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”	1"C, 3"C
3”	78.2	3.52	t (8.5)	1	1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”	2"C, 4"C
4”	72.2	3.41	t (8.5)	1	1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”	3"C, 5"C, 6"C
5”	78.7	3.47	t (8.8)	1	1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”	4"C, 6"C
6”	63.3	3.72, 3.92	m, d (12.7)	1, 1	1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”	4"C, 5"C

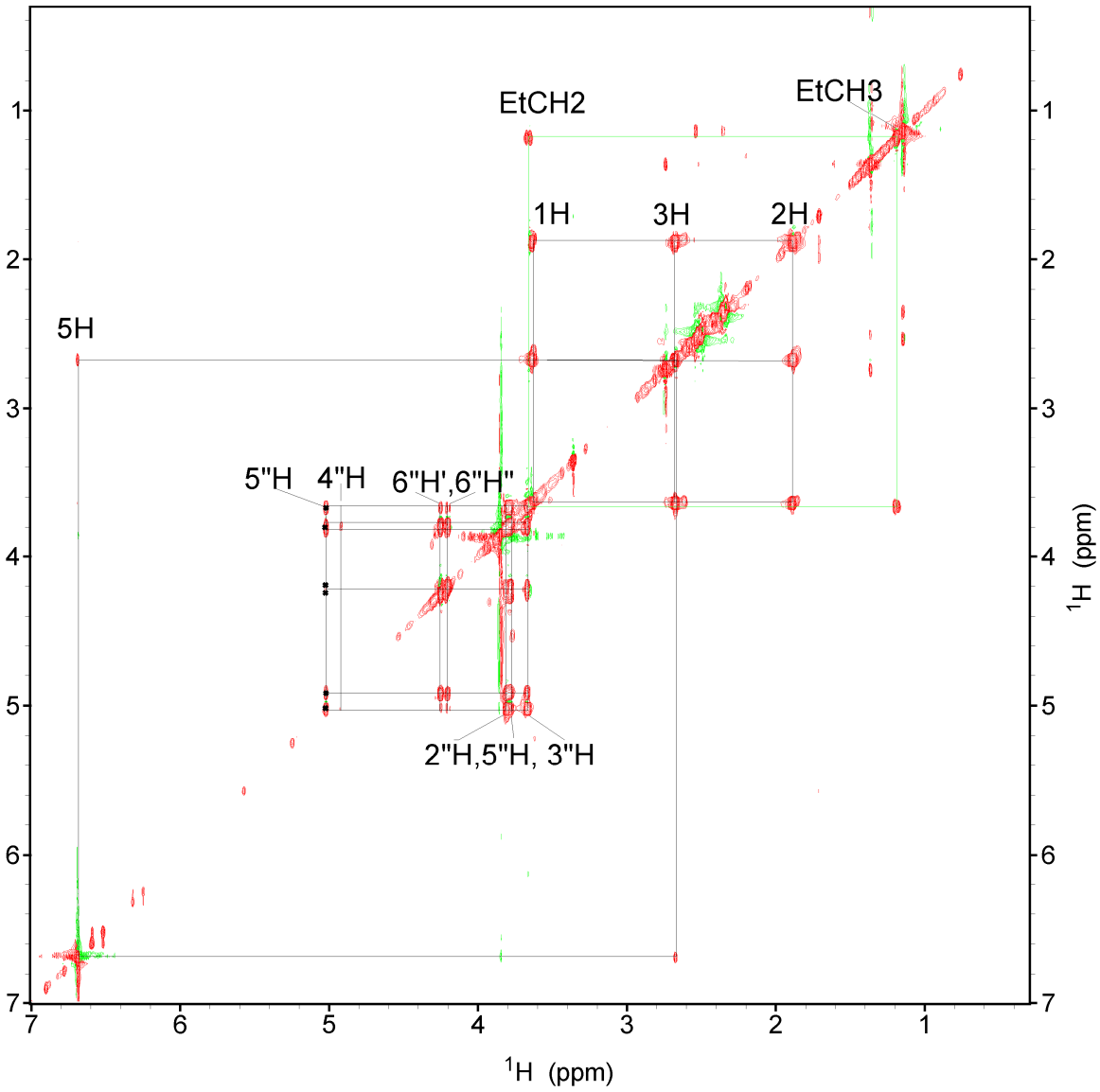
NMR data for Compound 661:



HMBC of compound 661



HSQC of compound 661

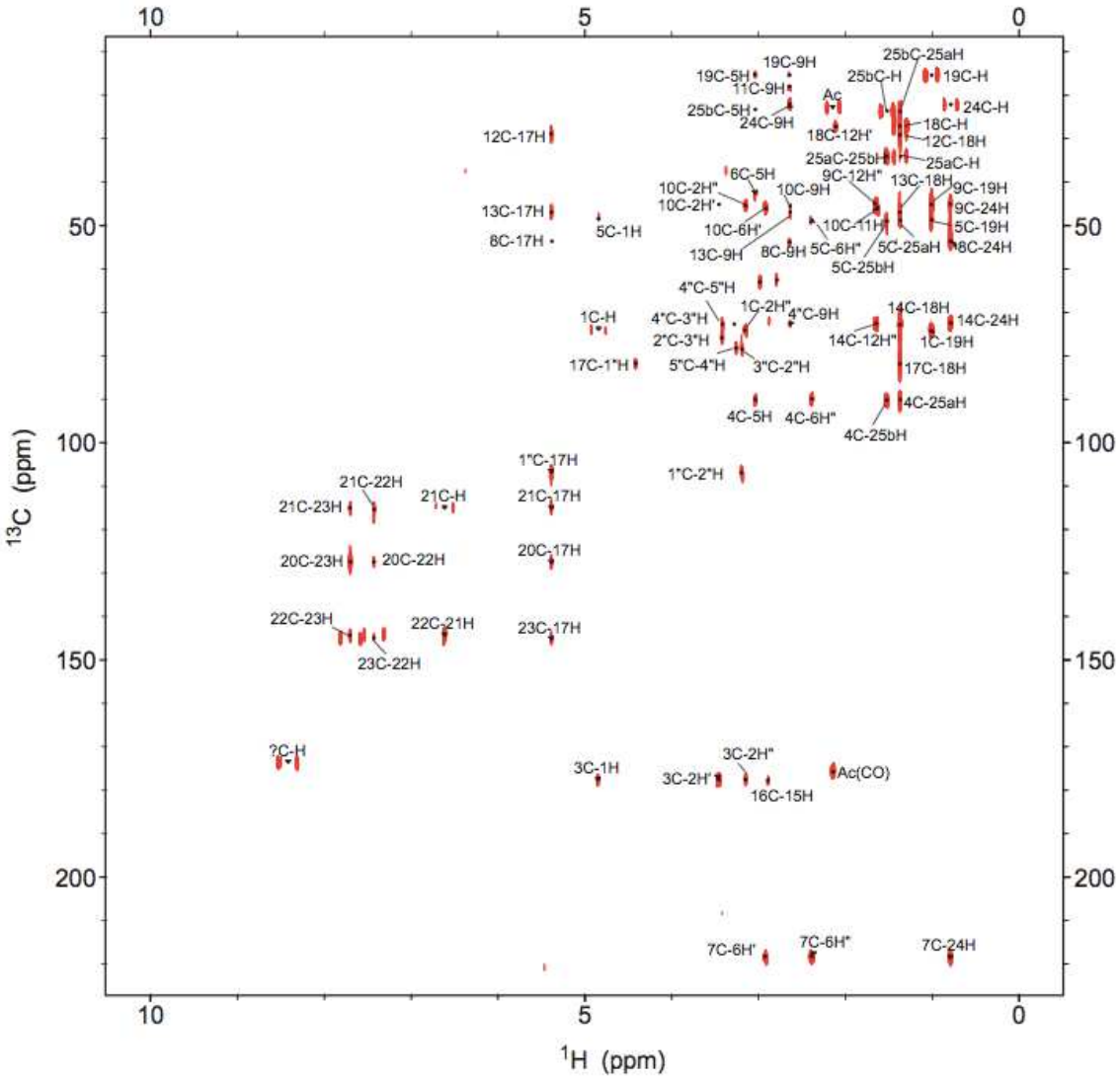


TOCSY of compound 661

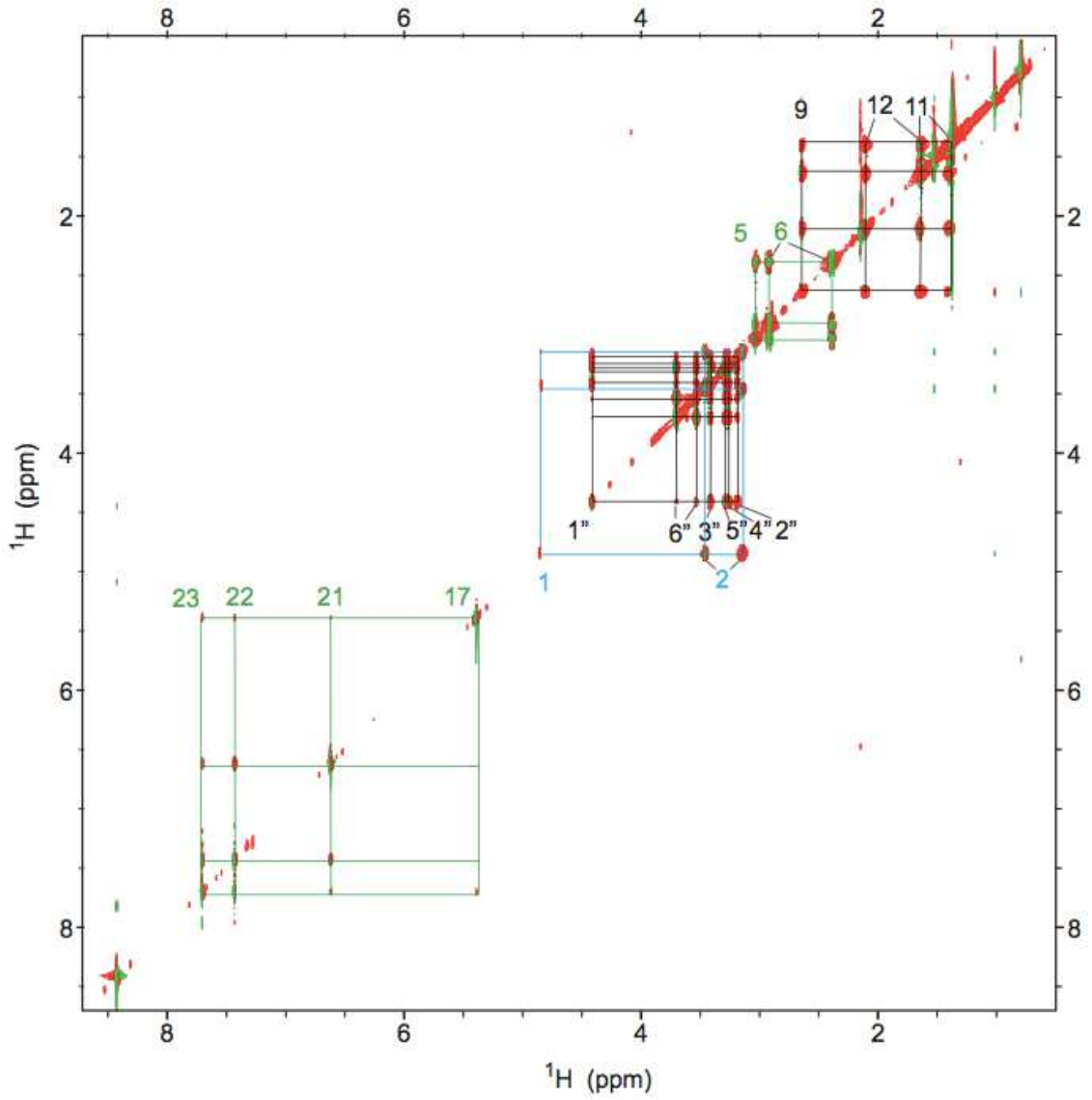
Position	^{13}C (ppm)	^{1}H (ppm)	multiplicity (J in Hz)	integral/number of ^{1}H	^{1}H-^{1}H correlation in TOCSY	^{1}H-^{13}C correlation in HMBC
1	63.9	3.63	t (6.6)	2	1H, 2H, 3H	2C, 3C
2	35.8	1.88	p (6.6, 7.8)	2	1H, 2H, 3H	1C, 3C, 4C
3	34.4	2.67	t (7.8)	2	1H, 2H, 3H	1C, 2C, 4C, 5C
4	143.0					
5	108.7	6.68	s	2	5H, 3H	3C, 4C, 5C, 6C, 7C

6	**155.0**					
7	**133.7**					
8	**58.7**	**3.84**	**s**	**6 (two CH3)**		**6C**
9	**181.5**					
10	**49.6**	**2.35**	**q (12.6, 15.0)**	**2**		**9C, 11C, 14C**
11	**72.7**					
12	**48.3**	**2.53**	**q(14.3, 5.1)**	**2**		**11C, 13C, 14C**
13	**174.9**					
14	**28.6**	**1.14**	**s**	**3**		**10C, 11C, 12C**
15	**181.8**					
16	**50.2**	**2.51**	**q (16.7, 14.6)**	**1**		**15C, 17C, 18C**
17	**72.9**					
18	**48.3**	**2.74**	**s**	**2**		**16C, 17C, 19C, 20C**
19	**174.6**					
20	**28.9**	**1.36**	**s**	**3**		**16C, 17C, 18C**
1”	**105.3**	**5.02**	**d (8.1)**	**1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**7C**
2”	**76.1**	**3.67**	**m**	**1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**1"C, 3"C**
3”	**76.2**	**3.8**	**t (9.5)**	**1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**2"C, 4"C**
4”	**73.6**	**4.92**	**t (9.8)**	**1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**3"C, 5"C, 6"C, 19C**
5”	**74.2**	**3.78**	**m**	**1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**1"C, 6"C**
6”	**65.3**	**4.20, 4.24**	**dd (12.3, 2.0), dd (12.3, 6.6)**	**1, 1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**4"C, 5"C, 13C**

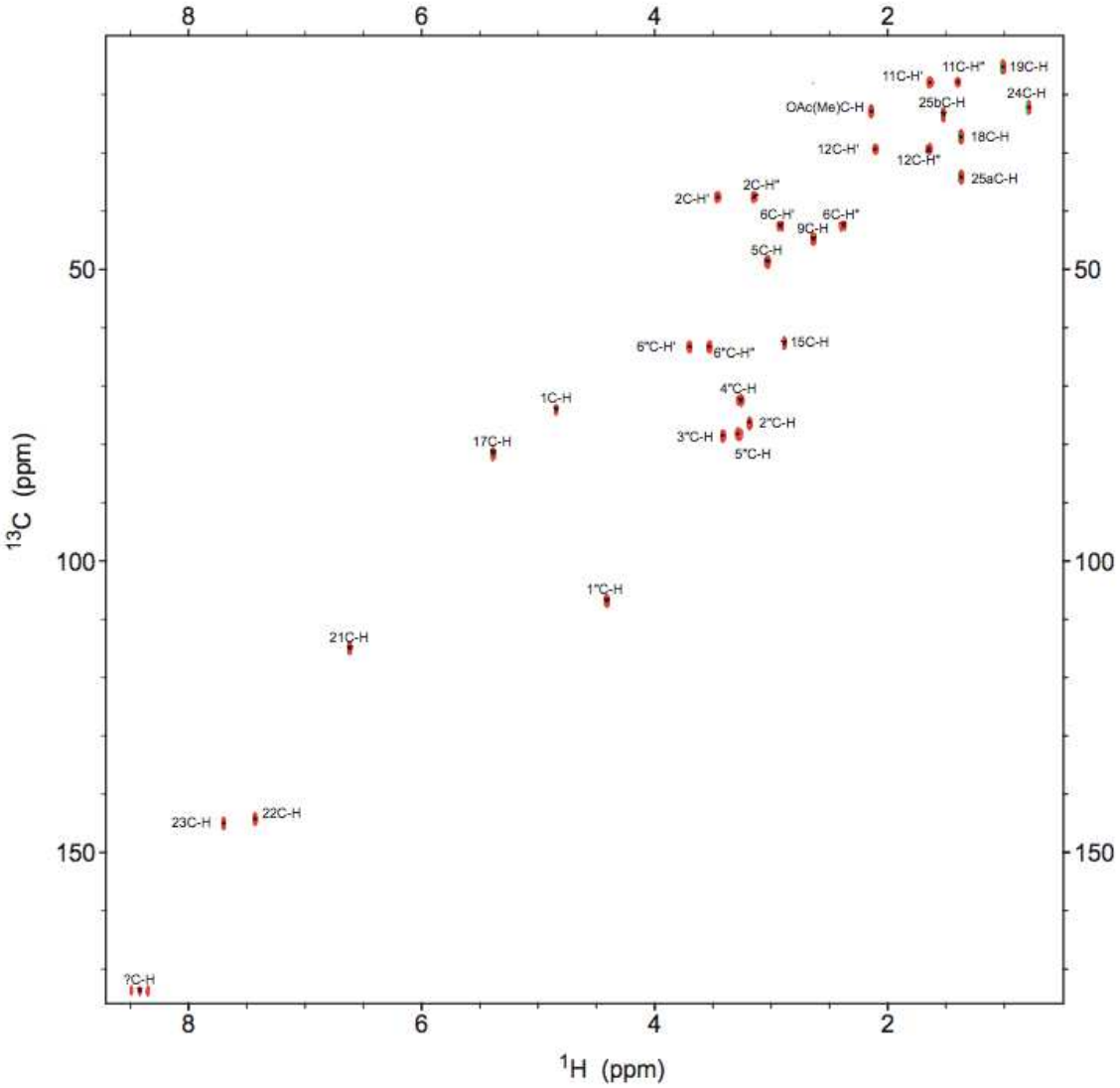
Compound 693 NMR



HMBC of compound 693



693 TOCSY



Compound 693 HSQC

Position	^{13}C (ppm)	^{1}H (ppm)	multiplicity (J in Hz)	integral/number of ^{1}H	^{1}H-^{1}H correlation in TOCSY	^{1}H-^{13}C correlation in HMBC
1	74.0	4.84	d (7.7)	1.1/1	1H, 2H', 2H"	3C, 5C
2	37.5	3.46, 3.15	d (15.9), dd (15.9, 7.7)	1.0/1, 1.0/1	1H, 2H', 2H"	1C, 3C, 10C
3	177.4					
4	90.1					
5	48.6	3.04	m	1.0/1	5H, 6H', 6H"	4C, 6C, 19C, 25bC

6	**42.5**	**2.92, 2.38**	**dd (20.3, 7.4), dd (20.3, 12)**	**0.7/1, 1.0/1**	**5H, 6H', 6H”**	**4C, 5C, 7C, 10C**
7	**218.4**					
8	**53.8**					
9	**44.7**	**2.64**	**dd (12.0, 7.7)**	**1.0/1**	**9H, 11H', 11H”, 12H', 12H”**	**8C, 11C, 13C, 19C, 24C**
10	**45.5**					
11	**17.8**	**1.64, 1.40**	**m, m**	**2.1/2, 1.2/1**	**9H, 11H', 11H”, 12H', 12H”**	**9C**
12	**29.3**	**2.11, 1.65**	**m, m**	**1.1/1, 2.1/2**	**9H, 11H', 11H”, 12H', 12H”**	**9C, 14C, 18C**
13	**46.7**					
14	**72.4**					
15	**62.6**	**2.89**	**s**	**1.1/1**		**16C**
16	**177.5**					
17	**81.6**	**5.39**	**s**	**0.9/1**		**8C, 12C, 13C, 21C, 23C, 1”C**
18	**27.2**	**1.37**	**s**	**6.0/6**		**12C, 13C, 14C, 17C**
19	**15.2**	**1.01**	**s**	**3.0/3**		**1C, 5C, 9C**
20	**127.4**					
21	**114.9**	**6.62**	**s**	**1.0/1**	**21H, 22H, 23H**	**22C**
22	**144.2**	**7.43**	**s**	**1.0/1**	**21H, 22H, 23H**	**20C, 21C, 23C**
23	**145.0**	**7.70**	**s**	**1.0/1**	**21H, 22H, 23H**	**20C, 21C, 22C**
24	**22.1**	**0.79**	**s**	**3.0/3**		**5C, 7C, 9C, 14C**
25a	**34.1**	**1.37**	**s**	**6.0/6**		**4C, 5C, 25bC**
25b	**23.4**	**1.52**	**s**	**3.1/3**		**4C, 5C, 25aC**
OAc(Me)	**22.8**	**2.14**	**s**	**3.0/3**		**OAc(CO)**
OAc(CO)	**175.7**					
1”	**106.8**	**4.41**	**d (8.0)**	**1.0/1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**17C**
2”	**76.3**	**3.19**	**d (8.5)**	**1.1/1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**1”C, 3”C**
3”	**78.5**	**3.41**	**t (9.3)**	**1.1/1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**2”C, 4”C**
4”	**72.4**	**3.26**	**t (9.3)**	**1.1/1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**5”C**

5”	**78.2**	**3.28**	**m**	**1.1/1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	**4”C**
6”	**63.2**	**3.70 , 3.54**	**dd (12.3, 2.0), dd (12.3, 5.6)**	**1.1/1, 1.1/1**	**1”H, 2”H, 3”H, 4”H, 5”H, 6”H', 6”H”**	