

The College of Wooster

Open Works

Senior Independent Study Theses

2021

The Application of Machine Learning in Analyzing Organic Compounds from NMR Spectral Data

Nicole Maia Powell

The College of Wooster, npowell21@wooster.edu

Follow this and additional works at: <https://openworks.wooster.edu/independentstudy>



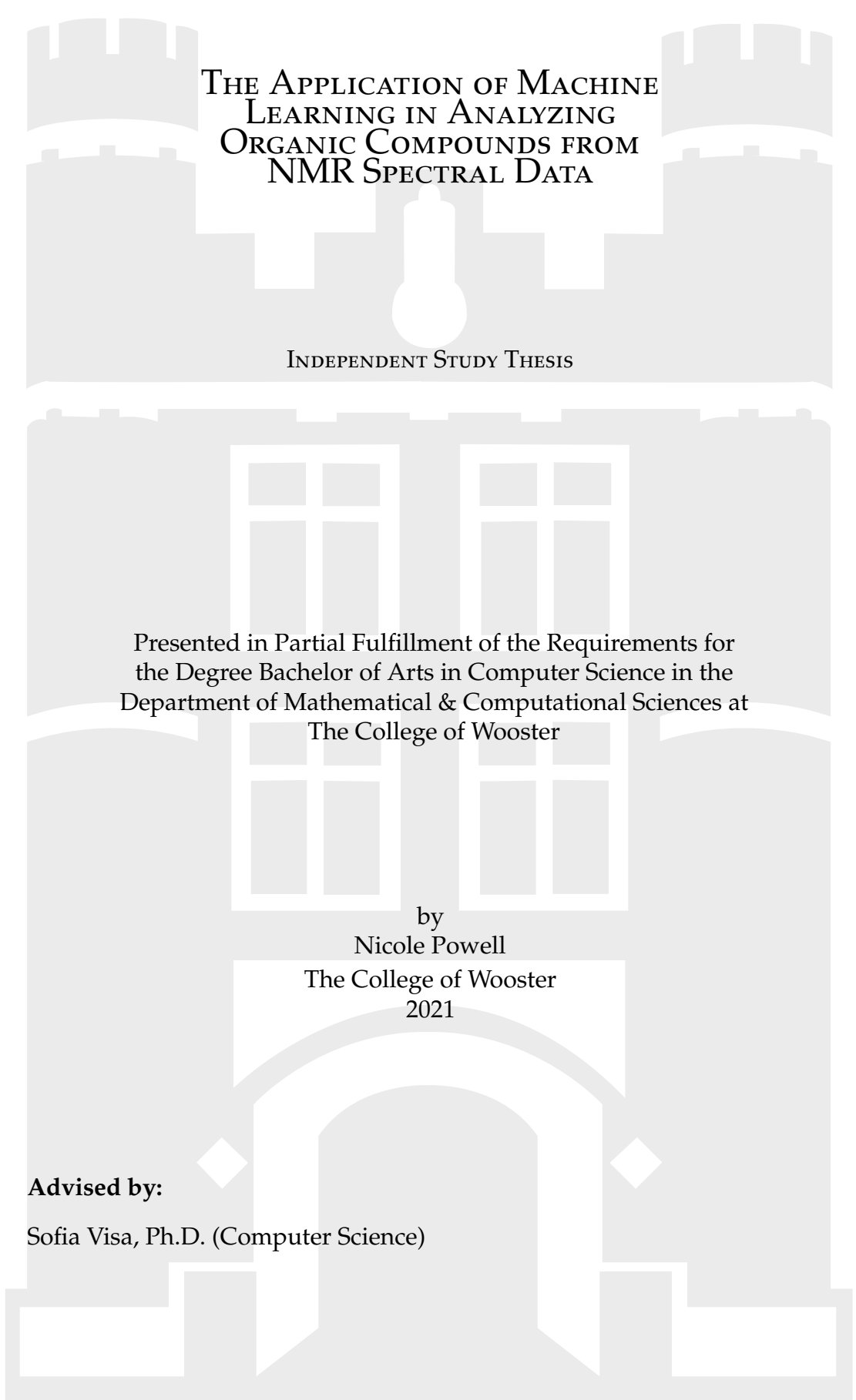
Part of the [Analytical Chemistry Commons](#), [Artificial Intelligence and Robotics Commons](#), and the [Organic Chemistry Commons](#)

Recommended Citation

Powell, Nicole Maia, "The Application of Machine Learning in Analyzing Organic Compounds from NMR Spectral Data" (2021). *Senior Independent Study Theses*. Paper 9471.

This Senior Independent Study Thesis Exemplar is brought to you by Open Works, a service of The College of Wooster Libraries. It has been accepted for inclusion in Senior Independent Study Theses by an authorized administrator of Open Works. For more information, please contact openworks@wooster.edu.

© Copyright 2021 Nicole Maia Powell



THE APPLICATION OF MACHINE LEARNING IN ANALYZING ORGANIC COMPOUNDS FROM NMR SPECTRAL DATA

INDEPENDENT STUDY THESIS

Presented in Partial Fulfillment of the Requirements for
the Degree Bachelor of Arts in Computer Science in the
Department of Mathematical & Computational Sciences at
The College of Wooster

by
Nicole Powell

The College of Wooster
2021

Advised by:

Sofia Visa, Ph.D. (Computer Science)



THE COLLEGE OF

WOOSTER

© 2021 by Nicole Powell

ABSTRACT

Nuclear magnetic resonance (NMR) is used in organic chemistry to identify unknown organic compounds. The data obtained from an NMR spectrometer are typically shown in the form of a spectrum, which is then analyzed by an analytical chemist. The action of analyzing a spectrum, especially one of a large and complex molecule, is a long and tedious process. In this project, Python is used to implement hierarchical clustering on NMR data obtained from an NMR spectrometer at the College of Wooster to explore its application in NMR analysis. MATLAB is used to build a decision tree from the same data, whose accuracy is compared to that of the hierarchical clustering. The decision tree is also examined to gain information about how to better automate the analysis process. These data clustering and classification processes are used to identify major functional groups within the compound from the spectral data, once feature extraction has been performed. Once these functional groups are identified, the compounds are clustered via hierarchical clustering, or classified with a decision tree. This processes provides insight into how to identify unknown organic molecules in a faster and more accurate manner, a much needed improvement in organic chemistry experimental research. It was found that decision trees are a much more accurate machine learning method to classify the organic compounds, when doing so based on present functional groups.

ACKNOWLEDGMENTS

I am deeply thankful to Dr. Visa for being my advisor and mentor, for offering constant support whenever I needed it. I would like to also acknowledge Dr. Sommer for being my temporary advisor and setting me up for success at the very beginning of my independent study. But I could not have gotten to this point in my computer science career without Dr. Byrnes. Dr. Byrnes pushed me as a student and advisee from the moment I declared my major to the summer she sadly passed away. She will always be a large part of my success when it comes to computer science and programming. I also owe thanks to the chemistry department for their support of the interdisciplinary aspects of my research. I'd like to thank Dr. Bonvallet for providing me with not only the interest to pursue automating NMR analysis, but also for the knowledge necessary to get started. I also would like to thank Prof. Arnholt and Dr. Sobeck for the providing me with the spectral data, which were the basis of this thesis, and for answering my NMR- and chemistry-related questions.

I want to express my gratitude for all of the friends I made at the College of Wooster over the last four years. They were there when I needed motivation to work, when I needed a break, and when I needed support during difficult times. Thank you to the biology majors for being with me through organic chemistry and lending me their thoughts on the chemistry components of my research; thank you to the economics major for keeping me working hard during early computer science classes and having math conversations with me; thank you to the history major for keeping me company during all of last winter; and thank you to the English major

for sticking with me through absolutely everything and taking care of me through my lowest points.

Finally, I want to acknowledge everything my family has done for me to get me to this point. My siblings have been by my side my entire life, whether I wanted them to be or not, and have always pushed me to do my best at what I love. I cannot express the extent to which my parents have helped me grow. They have always been incredibly supportive in all ways, not only these past four years at Wooster, but my entire life. I owe so much of my success to them. And I know I can count on them to continue encouraging me as I continue past this thesis, this college, and this chapter of my life.

CONTENTS

Abstract	v
Acknowledgments	vii
Contents	ix
List of Figures	xi
List of Tables	xv
CHAPTER	PAGE
1 Introduction	1
2 Nuclear Magnetic Resonance	5
2.1 Properties of Organic Molecules	6
2.2 Analysis of ¹ H NMR Spectra in Organic Chemistry	7
2.2.1 The ¹ H NMR Spectrum	8
2.2.2 The Analysis Process	8
2.3 Analysis of ¹³ C NMR Spectra in Organic Chemistry	11
2.3.1 The ¹³ C NMR Spectrum	11
2.3.2 The Analysis Process	12
2.4 Obstacles in NMR Analysis	14
3 Hierarchical Clustering, Decision Trees, and Machine Learning Tools	17
3.1 Hierarchical Clustering	17
3.1.1 Heat Maps	21
3.2 Decision Trees	23
3.3 Tools and Libraries	25
3.3.1 SpinWorksJ and Jcamp	25
3.3.2 NumPy, SciPy, and Matplotlib	26
3.3.3 Pandas and Seaborn	26
3.3.4 Decision Trees in MATLAB	26
4 Preparing Spectral Data for Hierarchical Clustering and Decision Trees	29
4.1 Formatting Data	29
4.1.1 Discretizing Data	31
4.2 Performing Hierarchical Clustering and Building a Decision Tree . .	33

5	Analysis of Results	35
5.1	Results of Hierarchical Clustering of the Organic Compounds	35
5.2	Non Computational Clustering of the Organic Compounds	42
5.3	Discussion of Findings	53
5.3.1	Proton NMR Hierarchical Clustering	53
5.3.2	Carbon NMR Hierarchical Clustering	54
5.3.3	Analysis of Combined Results	55
5.3.4	Discussion of Error in Hierarchical Clustering of the Spectral Data	56
5.3.5	Examination of Linkage Methods	57
5.4	Results and Discussion of Decision Tree Formation	58
6	Conclusion	65
6.1	Limitations in this Project	65
6.2	Future Work	66
	References	69

LIST OF FIGURES

Figure		Page
2.1	A quartet: a peak with a multiplicity of four	9
2.2	The identification of ethylbenzene using ^1H NMR analysis [19]	10
2.3	The molecular structure of cholesterol [20]	10
2.4	The NMR spectrum of cholesterol [1]	11
2.5	The identification of ethylbenzene using ^{13}C NMR analysis [19]	13
2.6	The NMR analysis process [23]	14
3.1	Example of clustered data [26]	18
3.2	Example of a dendrogram [26]	19
3.3	Dendrogram split at the largest vertical distance with no merging of classes [17]	19
3.4	Example of a heat map with student exam scores [14]	22
3.5	Example of a decision tree that classifies animal types	24
4.1	^1H NMR spectra of 2-pentanol and benzil before normalization (a and b respectively) and after normalization (c and d respectively)	31
4.2	Locations of hydrogens in specific functional groups on a spectrum [2]	32
4.3	Locations of carbons in specific functional groups on a spectrum [2, 9, 11, 28]	32
5.1	Dendrogram visualizing the clustering of 24 organic compounds from ^1H NMR spectral data. Single linkage was used to form this dendrogram.	36
5.2	Dendrogram visualizing the clustering of 24 organic compounds from ^{13}C NMR spectral data. Ward linkage was used to form this dendrogram.	36
5.3	Dendrogram visualizing the clustering of 24 organic compounds from combined ^1H NMR and ^{13}C NMR spectral data. Ward linkage was used to form this dendrogram.	37
5.4	Heat map visualizing the clustering of 24 organic compounds with the functional groups they possess. Created using the ^1H NMR spectral data	38

5.5	Heat map visualizing the clustering of 24 organic compounds with the functional groups they possess. Created using the ^{13}C NMR spectral data	38
5.6	Heat map visualizing the clustering of 24 organic compounds with the functional groups they possess. Created using the combined ^1H NMR and ^{13}C NMR spectral data	39
5.7	Dendrogram showing the six clusters of organic compounds from ^1H NMR spectral data	40
5.8	Dendrogram showing the six clusters of organic compounds from ^{13}C NMR spectral data	41
5.9	Dendrogram showing the six clusters of organic compounds from combined ^1H NMR and ^{13}C NMR spectral data	42
5.10	Spectra of organic compounds clustered by the human chemist . . .	44
5.11	Structure of 4-methylbenzyl alcohol [5]	44
5.12	Structure of 4-nitrobenzaldehyde [5]	45
5.13	Structure of 2-pentanone	45
5.14	^1H NMR dendrogram formed using average linkage	46
5.15	^1H NMR dendrogram formed using ward linkage	47
5.16	^1H NMR dendrogram formed using single linkage	47
5.17	^1H NMR dendrogram formed using complete linkage	48
5.18	^{13}C NMR dendrogram formed using average linkage	48
5.19	^{13}C NMR dendrogram formed using ward linkage	49
5.20	^{13}C NMR dendrogram formed using single linkage	49
5.21	^{13}C NMR dendrogram formed using complete linkage	50
5.22	^{13}C NMR dendrogram formed using weighted linkage	50
5.23	^{13}C NMR dendrogram formed using median linkage	51
5.24	Combined data dendrogram formed using average linkage	51
5.25	Combined data dendrogram formed using ward linkage	52
5.26	Combined data dendrogram formed using single linkage	52
5.27	Clusters from ^1H NMR data formed using hierarchical clustering with average, ward, single, and complete linkage methods. Colors indicate when the group the compound is in matches the groups formed by the chemist.	53
5.28	Clusters from ^{13}C NMR data formed using hierarchical clustering with average, ward, single, complete, weighted, and median linkage methods. Colors indicate when the group the compound is in matches the groups formed by the chemist.	54
5.29	Clusters from combined ^1H NMR and ^{13}C NMR data formed using hierarchical clustering with average, ward, and single linkage methods. Colors indicate when the group the compound is in matches the groups formed by the chemist.	56

5.30	Decision tree formed from combined spectral data. Attributes x3, x2, and x19 represent the presence of an aromatic ring in $^1\text{HNMR}$, the presence of an aldehyde group in $^1\text{HNMR}$, and the presence of an R_3CH alkyl group in $^{13}\text{CNMR}$ respectively. The classes at the leaf nodes correspond to those formed by the human chemist.	58
5.31	Decision tree formed from classes 1 and 3 of the combined spectral data. Attribute x5 represents the presence of a neighboring halogen, O, or NO_2 in $^1\text{HNMR}$. The classes at the leaf nodes correspond to those formed by the human chemist.	61
5.32	Decision tree formed from combined spectral data, with <i>MinParentSize</i> = 5.	62

LIST OF TABLES

Table		Page
3.1	The rules generated by the decision tree in figure 3.5	23
5.1	The six clusters formed when the organic compounds were clustered by a chemist viewing their structures	43
5.2	The rules generated by the decision tree in figure 5.30	59
5.3	The rules generated by the decision tree in figure 5.31	61
5.4	The rules generated by the decision tree in figure 5.32	62

CHAPTER 1

INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is a method of identifying unknown organic compounds. It is one of the most common methods, and often crucial in the identification process. The use of NMR in compound identification is often coupled with other forms of spectroscopy, or multiple methods of NMR spectroscopy are used in tandem to reduce uncertainty of compound identity [28].

While the process through which NMR spectroscopic data are analyzed is relatively efficient and reliable, which causes NMR's popularity, it has plenty of room for improvement and potential to be even more powerful than it currently is. The spectra produced by a spectrometer after the insertion of a compound is typically viewed by an analytical or organic chemist specializing in NMR and analyzed by hand. This process is tedious and long for compounds of even moderate complexity, and takes years of schooling to learn how to do effectively. Even when performed by experts, much of the analysis process contains room for error in identification. A program that fully automates – or even semi automates – the analysis process would save much time and many resources. This project explores the application of machine learning with NMR spectra analysis to attain that automation.

This thesis is organized in such a way that ensures the results may be understood after an introduction into both organic chemistry and NMR analysis, as well as the machine learning tools used in this project. Chapter 2 is comprised of four

main sections. Section 2.1 provides a brief introduction to and description of organic molecules and their properties with respect to their interaction with nuclear magnetic resonance. Sections 2.2 and 2.3 describe the analysis process of ^1H NMR and ^{13}C NMR respectively. The information that can be extracted from NMR spectra is explained in order to provide an understanding of the approach taken in this project. Finally, section 2.4 summarizes the issues that accompany NMR spectra and their analysis. Chapter 3 presents further background to provide an understanding of the tools used in this project. Section 3.1 explains hierarchical clustering, a method of unsupervised learning that is used in this project with NMR spectral data. In section 3.2, a form of supervised learning is explained: decision trees. This method is also used with NMR spectral data. Finally, the tools with which these machine learning methods are implemented are described in section 3.3. Chapter 4 acts as an introduction to the project. The first section, section 4.1, outlines the data preparation procedure and the way in which the data were formatted. Section 4.2 is a short introduction into the hierarchical clustering and decision tree formation. Chapter 5 presents the results of all the analyses performed. Section 5.1 includes a description of the outcome of hierarchical clustering with ^1H NMR data, ^{13}C NMR data, and combined ^1H NMR and ^{13}C NMR data, as well as the figures of dendrograms and heat maps that summarize those results. Section 5.2 describes how the organic compounds from the dataset were clustered by a chemist, whose clusters are used in section 5.3 for the analysis of the hierarchical clustering results from section 5.1. In section 5.3, the results of hierarchical clustering are compared to the clusters formed by the chemists and the comparison is discussed. The error that accompanies hierarchical clustering in this project specifically as well as in a broader sense is also touched upon. Finally, a closer examination of the linkage methods used in hierarchical clustering is performed. The final section of chapter 5 is section 5.4, in which the formation of the decision tree is described. In this section,

the results of the decision tree are displayed and then discussed with respect to their comparison with the clusters formed by the chemist, and their usefulness in further organic chemical compound analysis. This thesis is concluded with chapter 6, which summarizes all that has been discussed in previous chapters, touches on problems encountered and errors discovered, and closes with a description and explanation of future work that could be done following what has been learned in this project.

CHAPTER 2

NUCLEAR MAGNETIC RESONANCE

When an organic compound is in an oscillating magnetic field, the frequency of that magnetic field may cause the nuclei within the compound to emit electromagnetic signals. This is referred to as nuclear magnetic resonance, or NMR. Different nuclei emit electromagnetic signals at different frequencies of oscillation depending on their structure and surrounding electron cloud. Subjecting an organic compound to a magnetic field oscillating at a range of frequencies and recording the frequencies at which electromagnetic signals are emitted is known as NMR spectroscopy. NMR spectroscopy is used to identify unknown organic compounds, since the patterns of electromagnetic signals emitted are distinct between molecules [28].

There are many forms of NMR, each examining a different atom within a compound, the most common forms being ^1H NMR and ^{13}C NMR. ^1H NMR provides information about the hydrogen-1 atoms (atomic weight = 1 amu) within the molecule, whereas ^{13}C NMR provides information about the carbon-13 atoms (atomic weight = 13 amu) within the molecule. These are most common because organic molecules are largely made up of carbon and hydrogen atoms. ^1H NMR provides a greater amount of information about the structure of the molecule, though ^1H NMR and ^{13}C NMR are often used in tandem to obtain the largest amount of information possible [28]. Because ^1H NMR provides the largest amount of structural information (detailed in section 2.2.1), as well as the fact that ^1H NMR was

historically the first form of NMR performed, the use of the term "nuclear magnetic resonance" or "NMR" is assumed to refer to ^1H NMR unless otherwise specified [28].

While the most common use of NMR is organic compound identification, which is the process explored in this project, it has other important applications as well. Magnetic resonance imaging (MRI) is used daily to obtain images and scans of living tissue such as the brain and other organs, tendons, bones, etc. for medical purposes. The MRI scanner is a large-scale NMR spectrometer. MRI is the least invasive process for imaging the interior of the human body, due to the way in which NMR spectroscopy affects molecules without breaking any bonds or destroying them in any way [28, 10].

2.1 PROPERTIES OF ORGANIC MOLECULES

While not all compounds that contain carbon are organic, all organic compounds contain carbon. Compounds that contain exclusively carbon and hydrogen atoms are referred to as hydrocarbons, but it is also common for organic compounds to contain nitrogen, oxygen, and/or halogens such as chlorine or bromine. The majority of organic compounds are made up of mostly carbon and hydrogen [28].

Atoms that have an odd atomic number or an odd mass number have a nonzero nuclear spin (I). This occurs when the number of protons or the number of neutrons an atom has is odd; I is only zero when the number of both the protons and neutrons are even. If this is the case, the atom has no magnetic moment, and can therefore not be observed with nuclear magnetic resonance spectroscopy [28, 23]. Some of the most commonly present atoms in organic compounds (hydrogen, carbon, and nitrogen) have spin active isotopes, allowing them to respond to NMR. This allows for organic compounds to always be detected through NMR spectroscopy, which contributes to the practicality of the method in identifying unknown organic

compounds [23]. The most abundant isotope of hydrogen is ^1H , where ^1H makes up 99.99% of all hydrogen atoms. This means that almost all hydrogen atoms are spin active, and therefore organic molecules are very sensitive to NMR. Carbon, conversely, has a highly abundant isotope with no magnetic moment, and the spin active isotope is more rare. With the spin active ^{13}C only making up 1.07% of all carbon atoms, carbon is insensitive to NMR when compared to hydrogen atoms. Because of this, an organic sample must be approximately 100 times more concentrated when performing ^{13}C NMR spectroscopy than when performing ^1H NMR spectroscopy [28, 25].

Organic compounds are often characterized by reactive sections called functional groups. Functional groups are distinct groups of atoms and bonds that are more reactive than hydrocarbons, which consist of only single bonds (these parts of the compound are called alkanes or alkyl groups). Functional groups may be a double bond between two carbons, an oxygen bonded to a carbon, a nitrogen bonded to a carbon, a halogen bonded to a carbon, or other similar bonds. The term *functional group* gets its name because functional groups are the parts of the compound where reactions occur [28].

2.2 ANALYSIS OF ^1H NMR SPECTRA IN ORGANIC CHEMISTRY

Spectral analysis is the most common method of organic compound identification, and ^1H NMR is the most commonly used form of NMR when identifying those compounds [28]. In an ^1H NMR spectrum, the x-axis is in terms of parts per million (ppm), a representation of the frequency at which an electromagnetic signal may be emitted. The y-axis represents the relative intensity of the emitted signals, and is therefore unitless.

2.2.1 THE ^1H NMR SPECTRUM

An ^1H NMR spectrum gives insight into the structure of a molecule in a variety of ways. A peak on a spectrum has three important properties:

- **Chemical shift:** the position of the peak on the x-axis, denoting the frequency at which electromagnetic energy was emitted.
- **Intensity:** the area under the peak, or integration, denoting how strong of a signal was emitted at that peak's frequency.
- **Splitting pattern or multiplicity:** the way in which the peak is split, denoting how many hydrogen atoms "neighbor" the atom(s) corresponding to the peak in question. Hydrogen atoms are "neighbors" if their respective carbon atoms to which they are bonded are also bonded to each other.

As an example of how these properties present themselves in a spectrum, figure 2.1 shows a quartet, or a peak split into four, that occurs at a ppm of 4.28, and has an integration (intensity) of 1.

The multiplicity tells us that the hydrogen atom represented by that peak has three neighboring hydrogen atoms. The chemical shift tells us that the atom is relatively shielded by electrons, creating a frequency that is below a hydrogen in a highly electronegative area, but above a hydrogen in a non-electronegative area. Specifically, this hydrogen is likely bonded to the same carbon to which a halide (F, Cl, Br, or I) or oxygen is also bonded. The integration tells us that this peak is describing one hydrogen atom, as opposed to a group of more than one.

2.2.2 THE ANALYSIS PROCESS

Analysis of ^1H NMR spectra can be a very complex process, especially when performed for a particularly large or complex molecule. In its most simple form,

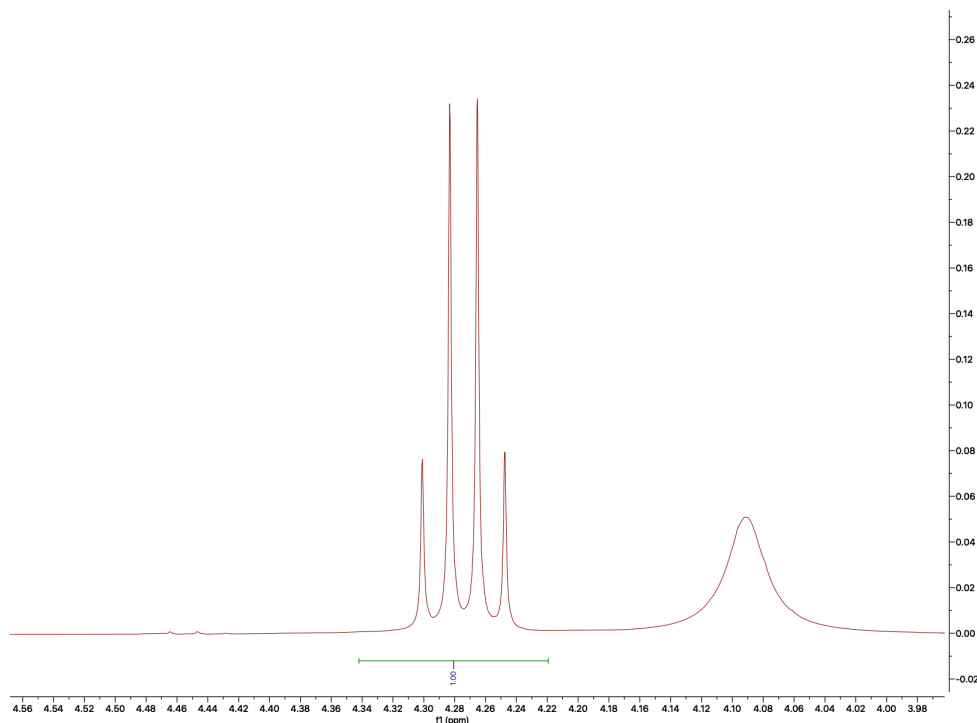


Figure 2.1: A quartet: a peak with a multiplicity of four

each peak corresponds to one group of hydrogen atoms in the compound, and an identification can easily be made. This is the case with ethylbenzene in figure 2.2, where the three groups of hydrogen atoms in ethylbenzene are easily captured by ^1H NMR.

While this seems to be a simple task, that is not always the case. A cholesterol molecule (depicted in figure 2.3) for example, would be much more difficult to identify from the ^1H NMR spectrum alone. The cholesterol molecule is much larger than the ethylbenzene molecule, and contains almost five times more hydrogen atoms. Furthermore, the cholesterol molecule has no symmetry, causing almost every hydrogen atom to react differently to the NMR and therefore create a separate peak. Some hydrogens, while they are different enough to create separate peaks, are still very similar, which causes their peaks within the spectrum to overlap. This overlapping of peaks creates difficulties in analysis because integration is altered,

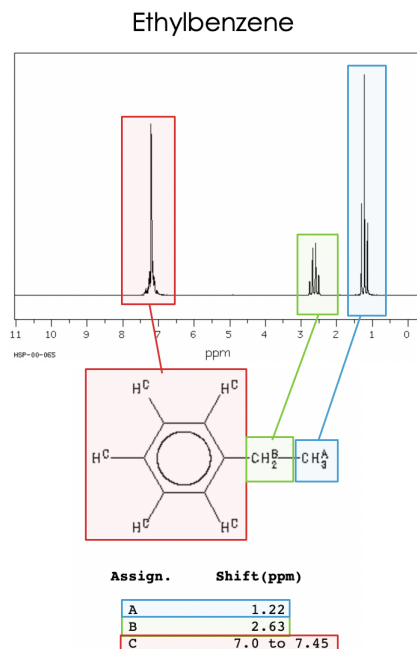


Figure 2.2: The identification of ethylbenzene using ^1H NMR analysis [19]

and two peaks close to one another can be mistaken for a doublet (a single peak split into two). The ^1H NMR spectrum of cholesterol is shown in figure 2.4. Peaks at higher ppm are distinct and clear, but peaks at lower ppm are clustered and difficult to discern. This is common with large molecules and creates difficulties in analysis that lead to errors in compound identification.

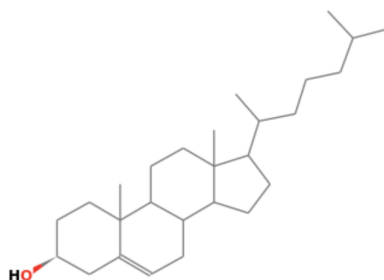


Figure 2.3: The molecular structure of cholesterol [20]

Analyzing the identity of an organic compound from NMR often begins with

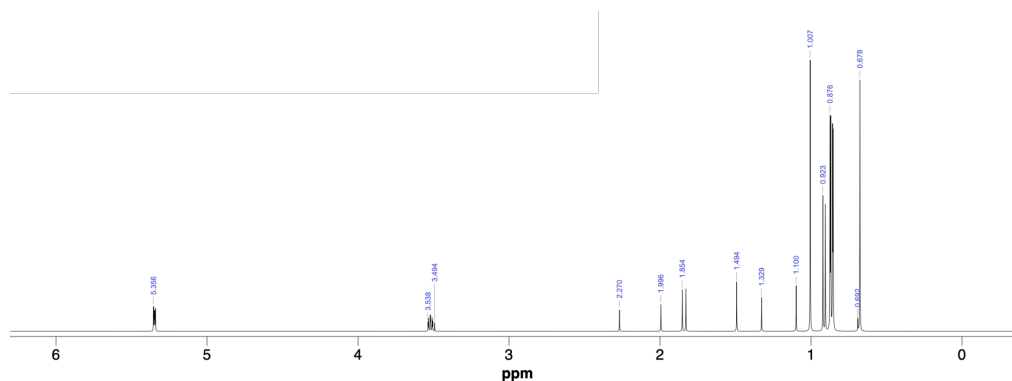


Figure 2.4: The NMR spectrum of cholesterol [1]

^1H NMR spectrum analysis, but is supplemented with ^{13}C NMR analysis. ^{13}C NMR is the second most common form of NMR [28]. The process of analyzing these spectra is described below.

2.3 ANALYSIS OF ^{13}C NMR SPECTRA IN ORGANIC CHEMISTRY

^{13}C NMR analysis is rarely used alone in identifying organic compounds. It is often used in conjunction with ^1H NMR analysis in order to gain confidence in the identified functional groups present in the organic compound. In a ^{13}C NMR spectrum, the x-axis is also in terms of parts per million (ppm), and the y-axis again represents the unitless relative intensity of the emitted signals.

2.3.1 THE ^{13}C NMR SPECTRUM

The most common isotope of carbon is carbon-12, which is not spin-active and therefore does not react to NMR. The carbon isotope that is being represented in ^{13}C NMR is carbon-13, which makes up about 1.07% of all carbon atoms. Because only 1.07% of the carbon atoms in the unknown organic compound are spin active, whereas 99.99% of the hydrogen atoms in the compound are spin active, a much higher concentration of the compound is needed to generate a readable spectrum

[28, 25]. Additionally, this low concentration of spin active carbons causes a lot of noise in the spectrum. To remedy this, hundreds of spectra are taken and averaged to form a spectrum that is usable in analysis [28]. The spectrum itself provides less information than the ^1H NMR spectrum, but still allows for a better understanding of the present functional groups. Each peak in a ^{13}C NMR spectrum describes a carbon atom or a group of carbon atoms, similarly to how the ^1H NMR does so for hydrogen atoms. However, unlike ^1H NMR spectra, ^{13}C NMR spectra lack splitting patterns. Therefore, ^{13}C NMR gives no direct insight into neighboring hydrogen or carbon atoms. Intensity, or integration, of ^{13}C NMR peaks correlate to the number of carbon atoms being described, but not as precisely as ^1H NMR peak intensity shows the number of hydrogen atoms being described. Chemical shift is the most important property of ^{13}C NMR spectra. Like ^1H NMR spectra, different areas along the x-axis correspond to different functional groups. While there is quite a bit of overlap of these areas, causing some confusion as to which functional group the peak is describing, there are many distinct functional groups that can be described by a ^{13}C NMR spectrum. Additionally, when cross-referencing these functional groups with those extracted from the ^1H NMR spectrum for the same compound, the overlap becomes inconsequential [28].

The values of chemical shift specifically are approximately 15-20 times larger than those in the ^1H NMR spectra. This is because the carbon atoms in a molecule are less shielded by electrons than their hydrogen counterparts [28, 9, 11].

2.3.2 THE ANALYSIS PROCESS

Because ^{13}C NMR is very rarely used alone, the analysis process often includes cross-referencing with the ^1H NMR spectrum of the same organic molecule. To better understand the technique of using both forms of NMR to identify a compound,

an example is presented. The ^{13}C NMR spectrum of ethylbenzene, whose ^1H NMR spectrum is shown in figure 2.2, can be seen in figure 2.5.

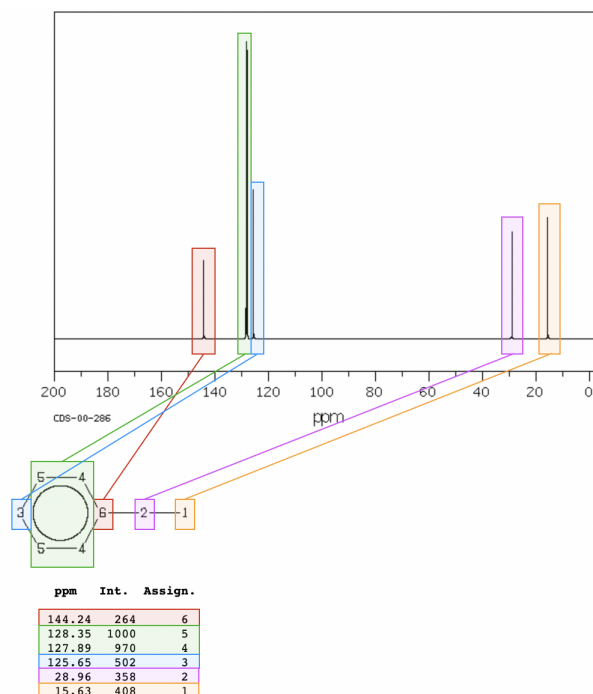


Figure 2.5: The identification of ethylbenzene using ^{13}C NMR analysis [19]

Both spectra show the presence of functional groups in different ways. Therefore, if one has ambiguous peaks, the other can make clear what is present. In the example of ethylbenzene, the peaks in the ^{13}C NMR spectrum showing the presence of an aromatic ring (shown in red, green, and blue) may be mistaken for an alkene, whose peaks occur between 100 and 150 ppm in ^{13}C NMR [9]. Referencing the ^1H NMR spectrum, however, it can be concluded that there is an aromatic ring, and no alkene, whose peaks occur between 4.5 and 6.8 in ^1H NMR, where no peaks are present [2]. While ethylbenzene is a simple example and would not need both forms of NMR to identify, the cross-referencing process described is necessary for larger and more complex molecules.

2.4 OBSTACLES IN NMR ANALYSIS

NMR analysis is not a straightforward procedure. It has an experimental approach, in which structures are often built or drawn based on the spectral information, and either rejected or accepted upon comparison to the spectrum. A simple flow diagram of what this process looks like is seen in figure 2.6.

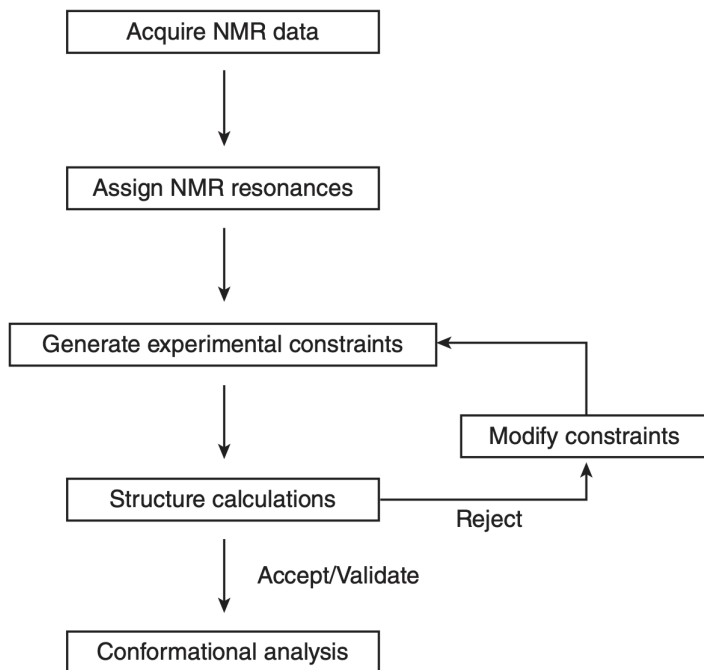


Figure 2.6: The NMR analysis process [23]

This method of analysis is time consuming and difficult, and becomes even more so the more complex a molecule is.

What makes NMR spectrum analysis more complicated is missing or indiscernible information. In ^1H NMR spectra, a hydrogen atom that is bonded to an oxygen or nitrogen atom is sometimes visible and sometimes is not. The information gained from the area in which that peak may appear is therefore not completely reliable. In ^{13}C NMR, baseline noise causes small peaks to occasionally get lost in the baseline, or noise to be mistaken for a peak that is part of the unknown compound.

Additionally, the absence of splitting pattern in ^{13}C NMR creates a lack of clear structural information [28]. Impurities in the sample impedes the analysis even further, as it is often not known which peaks refer to the compound in question, and which to impurities in the solution.

The NMR instrument itself is large and quite expensive, with prices of \$500,000 and above [24], which creates problems for researchers with little funding or capital.

CHAPTER 3

HIERARCHICAL CLUSTERING, DECISION TREES, AND MACHINE LEARNING TOOLS

This chapter introduces hierarchical clustering and decision tree classification techniques, and explores the libraries that were used for data preprocessing and analysis.

3.1 HIERARCHICAL CLUSTERING

Machine learning can be broken into three categories: supervised learning, unsupervised learning, and reinforcement learning [7]. Supervised learning allows data to be sorted into labeled classes, by providing an already sorted set of similar data, called the training set. Unsupervised learning is used to find connections and patterns in data, without knowing what those connections or patterns will be beforehand. The goals of supervised and unsupervised learning are largely the same – grouping data into classes or clusters – the difference being whether it is known prior what those classes or clusters will be. Reinforcement learning uses a reward and punishment system to train a software agent, and is often used in game development to simulate intelligent behavior. In this project, a common form of unsupervised learning called clustering will be used. When clustering is applied,

data are grouped into clusters such that members of a cluster are similar to each other in some way, and dissimilar to data points in other clusters [7]. An example of this can be seen in figure 3.1, where data points are clustered by position in the x-y plane.

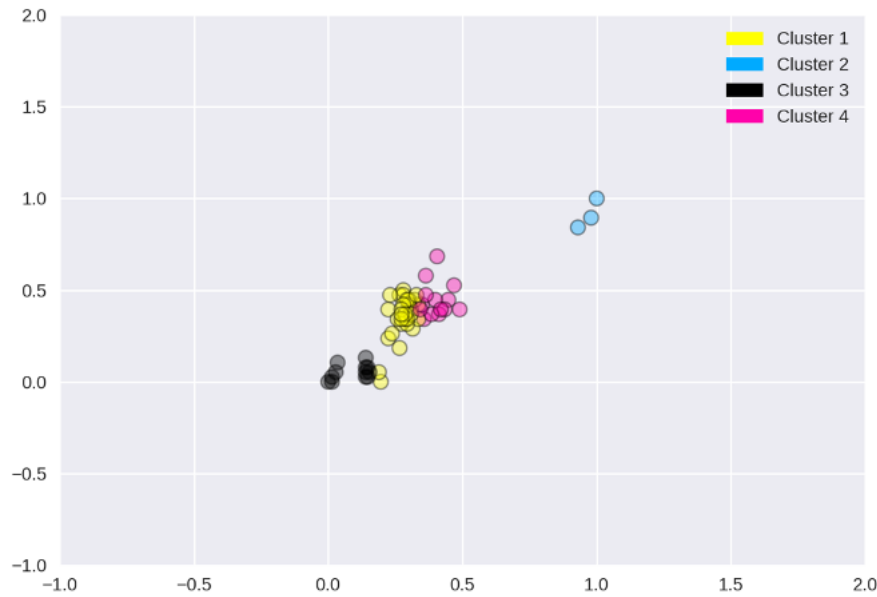


Figure 3.1: Example of clustered data [26]

One of the many forms of clustering is hierarchical clustering. Agglomerative hierarchical clustering builds clusters slowly, allowing each combination of data points/sets to be recorded into a hierarchy. Divisive hierarchical clustering does the opposite of agglomerative clustering, breaking clusters apart one at a time to form a hierarchy. The hierarchy that is formed by this method of clustering, which is called a dendrogram, allows for the splitting of data to obtain the desired number of clusters. Splitting a dendrogram close to the base creates a larger number of clusters that are smaller and more specific. Splitting a dendrogram close to the top creates fewer clusters that contain more data and are broader in the type of data they include. A visual representation of this can be seen in figure 3.2 [26, 17]. Often, the most effective location to split a dendrogram is through the largest vertical area

in which no classes are clustered. An example of this can be seen in figure 3.3. In this example, there would be four classes, since four vertical lines are crossed by the split going through the horizontal band denoted by AB [17]. These classes would include data elements 9, 23, 17, 6, 11, 3, 15, and 7 in the first cluster, elements 14, 19, 16, 24, and 22 in the second cluster, and so on.

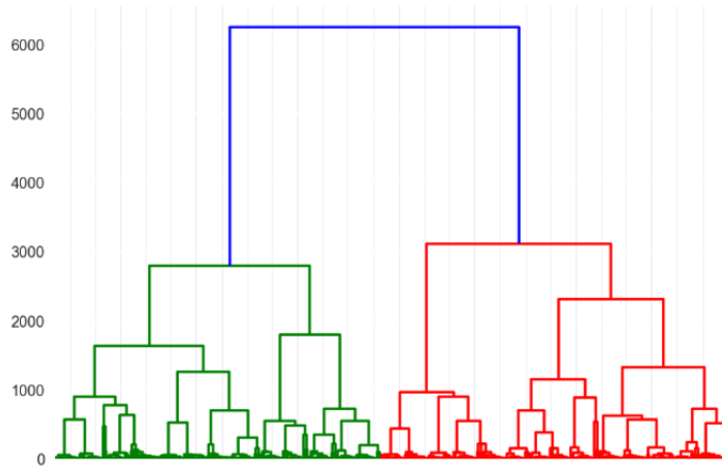


Figure 3.2: Example of a dendrogram [26]

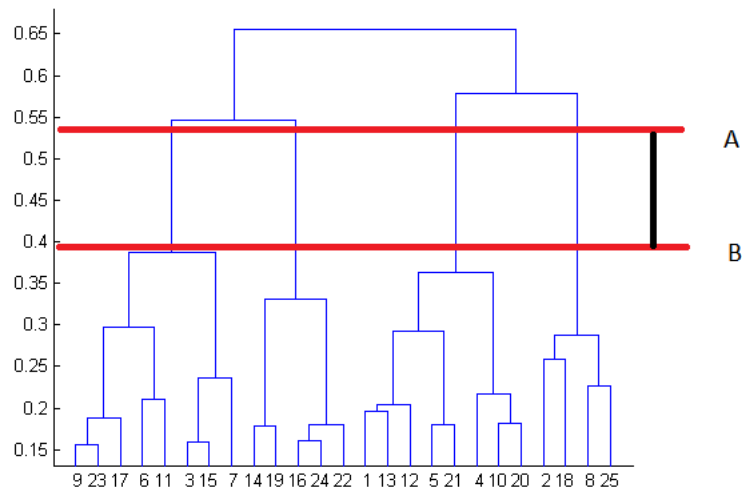


Figure 3.3: Dendrogram split at the largest vertical distance with no merging of classes [17]

The way in which the hierarchy is formed is that, at each step, the two most similar clusters are combined into one cluster. This process depends on the type of data and the similarity measure. The criteria followed to determine how the clusters relate to each other are called linking criteria. Seven ways of determining the distance between clusters are the following:

- **Average linkage** defines the distance between two clusters to be the average distance between each of the data points in one cluster with each of the data points in the other cluster.
- **Ward linkage** defines the distance between two clusters to be the sum of the squares of the distances between all data points within each cluster.
- **Single linkage** defines the distance between two clusters to be the distance between the two data points closest together, where one data point is in one cluster and the other data point is in the other.
- **Complete linkage** defines the distance between two clusters to be the distance between the two data points farthest apart, where one data point is in one cluster and the other data point is in the other.
- **Weighted linkage** defines the distance between two clusters to be the average distance between one cluster to each of the two subclusters of the other.
- **Centroid linkage** defines the distance between two clusters to be the distance between the centroids computed from all elements in each cluster.
- **Median linkage** defines the distance between two clusters to be the distance between the centroids computed from the average of the centroids of the two subclusters in each cluster.

The similarity measure used to calculate the distance between the data points themselves is also important. Common methods of doing so are using Euclidian and Manhattan distances. Euclidian distance is simply the shortest distance between two points, $x = (x_1, x_2)$ and $y = (y_1, y_2)$:

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (3.1)$$

Manhattan distance is the distance between two points traveling along axes at right angles, and is calculated with the following equation:

$$|x_1 - y_1| + |x_2 - y_2| \quad (3.2)$$

for points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ [17, 8].

Every method of machine learning has advantages and disadvantages. Hierarchical clustering is inefficient in time as well as memory relative to other forms of clustering, with a time complexity of $O(n^3)$ and a space complexity of $O(n^2)$. This causes issues when hierarchical clustering is required for a large dataset, and there is a limited amount of time and/or storage [21]. Hierarchical clustering is, however, a very useful way of clustering data when it is not known how many clusters there will be, or what will identify those clusters, as it is a form of unsupervised learning. The dendrogram formed through hierarchical clustering is also a distinct advantage, as it allows for any number of clusters to be extracted, depending on where the dendrogram is chosen to be split.

3.1.1 HEAT MAPS

A heat map, also called a double dendrogram, is a method of data visualization in matrix form. The rows of the data matrix are clustered to form a dendrogram, which

is then used to determine the ordering of the rows within the heat map, causing similar rows to be near each other. Similarly, the columns of the data matrix are clustered to determine their ordering in the heat map as well. Finally, a color scale is used to denote values of each data point, allowing the viewer to visualize and understand the clustering of, and differences between, points within the dataset [14]. An example of a heat map is shown below, in figure 3.4.

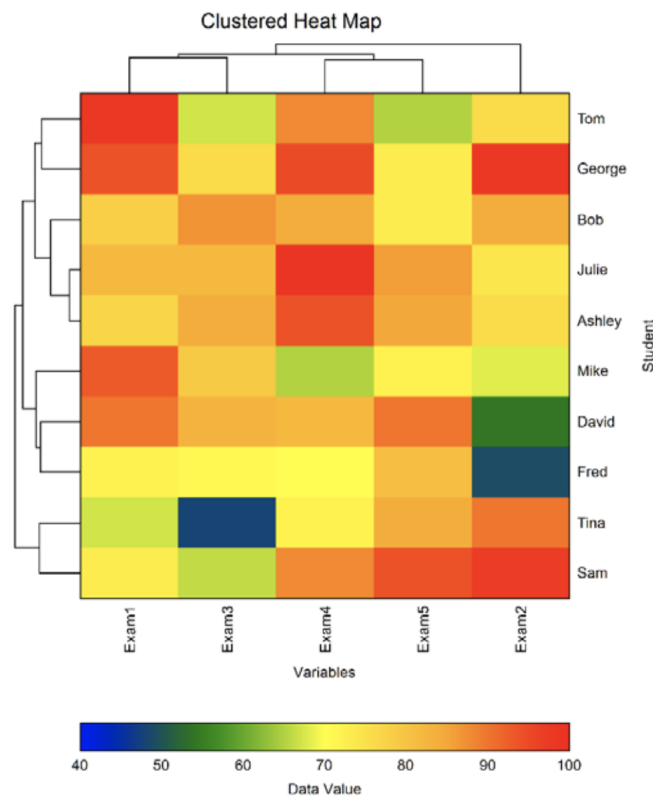


Figure 3.4: Example of a heat map with student exam scores [14]

Any of the linkage methods described previously in section 3.1 may be used when clustering the rows and columns of the heat map. The linkage methods with which rows and columns are clustered may be different from each other, and these methods are often chosen based on the 'goodness-of-fit' of the method to the data. The 'goodness-of-fit' can be determined by calculating the Cophenetic correlation coefficient or the delta of each of the methods and choosing the method that results

in the highest or lowest value respectively [14]. Due to the small size of the dataset used in this project, neither of these methods are used to determine the best linkage method. Instead, all linkage methods are used and compared to the way in which a chemist would cluster the compounds, and the most accurate linkage method is determined by which set of clusters has the most overlap with the set of clusters formed by the chemist.

3.2 DECISION TREES

Decision trees, unlike hierarchical clustering, are a form of supervised learning. Class labels are known, and data points are grouped and given those labels by the decision tree. From a set of training data, a tree is formed, and classification error is minimized. A set of test data is then passed through the tree to ensure that there is little error when classifying data outside of the training examples. A decision tree can be described as a set of rules, that are in an "if-else" format. An example of this is in figure 3.5, with a table of rules in table 3.1.

	Rule
R1	If no vertebrae \Rightarrow Insect
R2	If vertebrae and fur \Rightarrow Mammal
R3	If vertebrae, no fur, and no wings \Rightarrow Amphibian
R4	If vertebrae, no fur, and wings \Rightarrow Bird

Table 3.1: The rules generated by the decision tree in figure 3.5

Each rule in the table follows from the root of the tree to a leaf, where a classification is made. Each node within the tree corresponds to an attribute of

Animal	Wings	Vertebrae	Fur	Eggs	Class
Dog	0	1	1	0	Mammal
Parrot	1	1	0	1	Bird
Ladybug	1	0	0	1	Insect
Pig	0	1	1	0	Mammal
Penguin	1	1	0	1	Bird
Frog	0	1	0	1	Amphibian
Ant	0	0	0	1	Insect

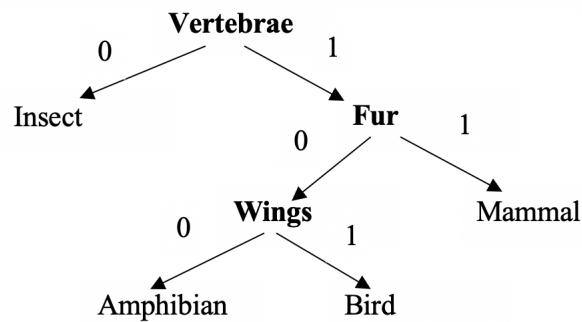


Figure 3.5: Example of a decision tree that classifies animal types

the data. In the example in figure 3.5, the attributes are *Wings*, *Vertebrae*, *Fur*, and *Eggs*. The attributes closer to the root node split the data in such a way that each group has the least entropy possible. Entropy refers to the impurity of the group of data points, so the greater the entropy, the less one class of data points dominates the other classes. This means that the attributes closer to the root node of the decision tree are more useful in partitioning the dataset into the desired classes. This example does not make use of the attribute *Eggs*. This omitting of an attribute occurs when its information can be found in other attributes in an equally or more efficient way. While this example has only two options for each attribute, that is not always the case for other datasets: there can be any number of options in the dataset for attributes, even an infinite amount, as is the case with continuous data.

Datasets used with decision trees are typically much longer than that in the given example in figure 3.5. With a larger dataset, more accurate information can be extracted from the decision tree.

3.3 TOOLS AND LIBRARIES

In this project, multiple Python libraries were used in conjunction to obtain dendrograms and heat maps of the NMR spectral data. After the raw data were converted to a usable format through the use of the SpinWorksJ software and the open-source program jcamp, NumPy was used to store and manipulate the data, and matplotlib was used to perform agglomerative clustering on that data, as well as to display the dendrograms formed from that clustering. Pandas was used to form a dataframe, which was then formed into a heat map with the use of seaborn. MATLAB was also used to form a decision tree from the data formatted by SpinWorksJ, jcamp, and NumPy.

3.3.1 SPINWORKSJ AND JCAMP

The raw data output from the NMR spectrometer is in a format that is difficult to read. SpinWorksJ was used to convert the data from a jdf format to a jcamp format. SpinWorksJ allows the user to view the NMR spectrum and manipulate it for ease of analysis. The data could be saved in a variety of formats; this project made use of the option to save the file in the jcamp format [6]. The jcamp format data files were still difficult to read, and needed to be restructured in a way that allowed the data to be more easily obtainable. The jcamp program allows the user to insert a jcamp file, and get a Python dictionary of the data in return. The Python dictionary can then be iterated and read with more ease than the jcamp file. The data were extracted from the Python dictionary and saved as a set of NumPy arrays to allow for manipulation and calculation [12].

3.3.2 NUMPY, SCIPY, AND MATPLOTLIB

The Python library NumPy provides the ability to create arrays, which are similar to lists but easier to manipulate for mathematical purposes. Many other open-source libraries are built upon NumPy, including the libraries used in this project (i.e. `scipy`, `matplotlib`, `pandas`, and `seaborn`) [13]. The Python library `matplotlib` contains the capability to form dendrograms from data that are formatted by the SciPy Python library. The *linkage* method of `scipy.cluster.hierarchy` takes in a set of data and a linkage method (e.g. `ward`), and outputs a linkage matrix encoded with the hierarchical clustering [27]. This linkage matrix can then be passed to the `matplotlib` library method called *dendrogram*, which forms a dendrogram from the given data, with optional formatting specifications [15].

3.3.3 PANDAS AND SEABORN

The Python library `seaborn` was created for simple, clear, and well-designed data visualizations. It is based on `matplotlib`, allowing intuitive interaction between the two. `Pandas` allows the formation of dataframes, a data structure for large datasets that the `seaborn` method *clustermap* reads in order to form a heat map of the given data. The `Pandas` dataframe structure was used, holding data transferred from the NumPy array of spectral data, in the *clustermap* method to form the heat maps displayed in this thesis. [30, 29]

3.3.4 DECISION TREES IN MATLAB

MATLAB was used to form a decision tree from the spectral data formatted by `SpinWorksJ`, `jcamp`, and NumPy. A numerical matrix was created from the data in the NumPy array, and that matrix, as well as an additional matrix holding the target classifications, were passed into the MATLAB function *fitctree*. This method

produced a decision tree from the given data, with information about how many data points from each class were given each class label [3].

CHAPTER 4

PREPARING SPECTRAL DATA FOR HIERARCHICAL CLUSTERING AND DECISION TREES

This chapter describes and explains data preprocessing steps and introduces the methods used for hierarchical clustering and decision trees. The spectral data examined in this chapter was obtained using a Jeol NMR spectrometer with a reference frequency of 400 MHz.

4.1 FORMATTING DATA

In order to extract usable and relevant data from the raw data obtained from the spectrometer, a series of calculations were completed. The raw ^1H NMR data contained data points at frequencies above 5000 MHz, which would always have an intensity of zero. This is because no hydrogen atoms can have a vibrational frequency above 5000 MHz in a 400 MHz spectrometer [28]. These data points were ignored, as were any extraneous data points that happened to appear below 0 MHz. The raw ^{13}C NMR data contained data points at frequencies above 22,000 MHz, which would always have an intensity of zero because no carbon atoms can have a vibrational frequency above 22,000 MHz in a 400 MHz spectrometer [28]. These data points were also ignored, as were any extraneous data points that happened to appear below 0 MHz.

The frequency was then converted into units of parts per million (ppm), by dividing the ^1H NMR frequencies by the reference frequency of the spectrometer (400 MHz), and the ^{13}C NMR frequencies by one fourth of the reference frequency (100 MHz). Any ^1H NMR data points with an intensity below zero were rounded up to zero, and any ^{13}C NMR data points below 5% of the maximum intensity were rounded to zero, in order to mitigate the large amount of noise typically present in ^{13}C NMR spectra.

Finally, the data was normalized for easy comparison by dividing each intensity value by the largest in that dataset. This normalization causes the largest intensity to be equal to 1, and any other peaks to have intensities that are a fraction of the largest. Differences in the y-axis of the spectra occur when different concentrations of the sample are measured. Variation in concentration does not affect the identity of the compound itself and only causes difficulties in comparing compounds to each other, due to the differences in y-axis scale. Normalization ensures that all datasets have an equal weight, so that meaningful relationships can be drawn between similar compounds without the inconsistency in how each sample was measured in the lab. An example of the effect data normalization has on the spectrum is shown in figure 4.1.

Figures 4.1 a and b show how dissimilar the y-axis scale can be for samples of different concentration, where 2-pentanol has a maximum intensity of approximately 50,000 (a), and benzil has a maximum intensity of approximately 30,000 (b). Once the data are normalized, the maximum intensities for 2-pentanol and benzil (c and d respectively) are equal; they are both 1. The spectrum itself, independent of the y-axis, is exactly the same between the original graph and the normalized graph. This shows that normalizing the data does not affect the dataset itself, only the ability to compare different datasets that may have been obtained using varied concentrations of sample.

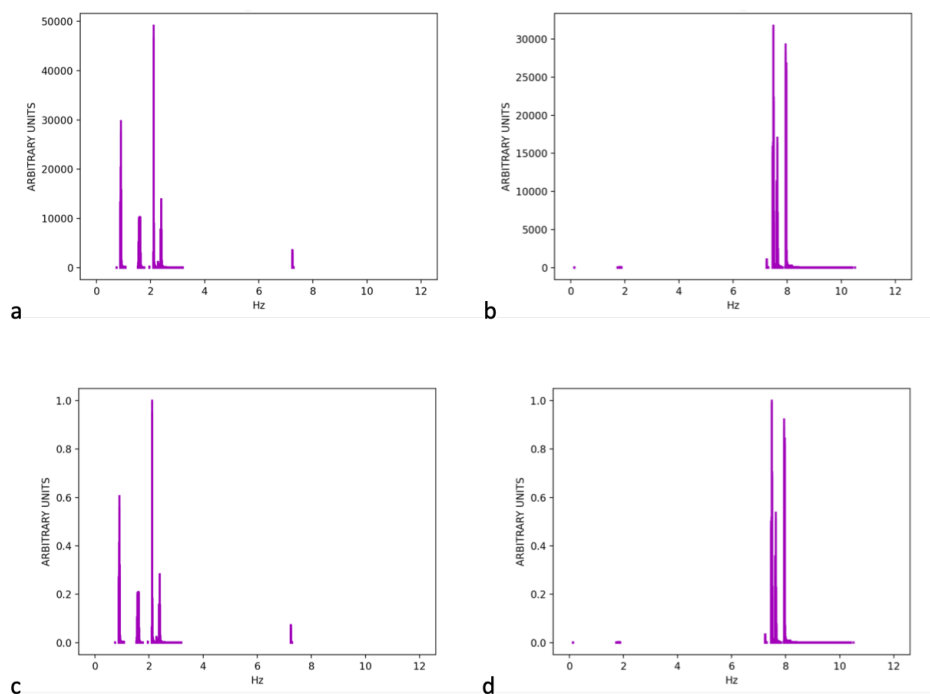


Figure 4.1: ^1H NMR spectra of 2-pentanol and benzil before normalization (a and b respectively) and after normalization (c and d respectively)

4.1.1 DISCRETIZING DATA

The datasets contain thousands of data points for each compound. Comparing each singular data point to one another would take an immense amount of computational power and time for a result whose greater accuracy would not be worth that sacrifice. Instead, the data were discretized into seven or fourteen chunks, allowing for faster comparisons and a lower dimensional clustering and heat map. In order to discretize the data in a meaningful way, current techniques of spectral analysis were examined. Due to the way specific atoms or bond types distribute electrons within a molecule, signals of hydrogen and carbon atoms in certain functional groups will appear in certain frequency ranges, or locations on the x-axis of the spectrum. A visual representation of this can be seen in figure 4.2 for ^1H NMR, and figure 4.3 for ^{13}C NMR.

The spreading of functional groups over the spectrum was used as a guideline

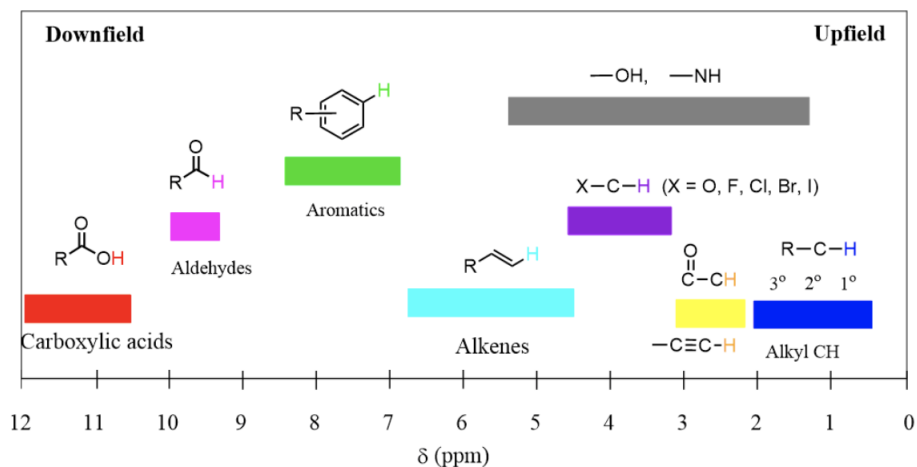


Figure 4.2: Locations of hydrogens in specific functional groups on a spectrum [2]

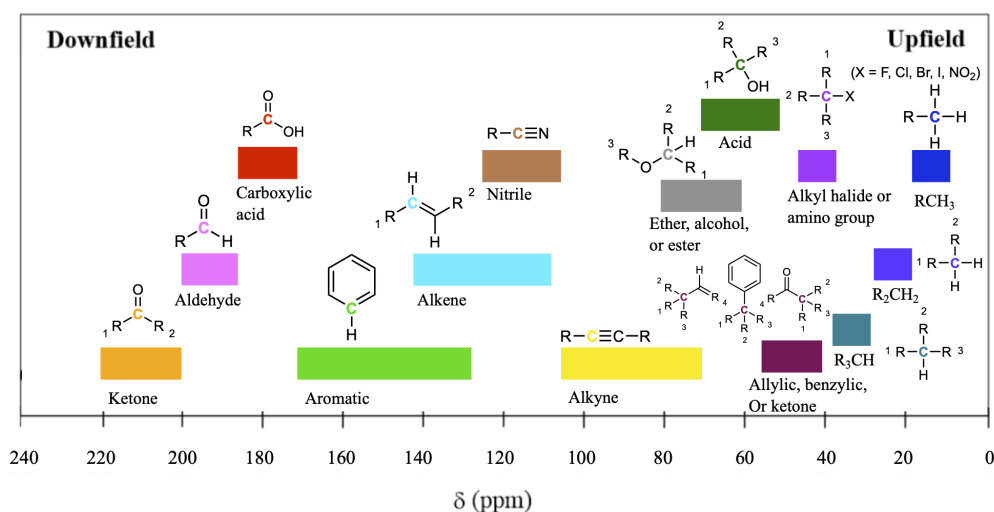


Figure 4.3: Locations of carbons in specific functional groups on a spectrum [2, 9, 11, 28]

for discretizing the data: the data were split into groups that correspond to each functional group, and a sum of signal intensities was taken over each range. This provided a complete dataset with seven data points from ^1H NMR and fourteen from ^{13}C NMR, as opposed to the thousands existing originally.

While this discretization was largely beneficial and necessary, there are some disadvantages that accompany it. Firstly, with the combining of hundreds of data

points into one, some data is lost. This is unavoidable in discretization, as it is not possible to contain the information of hundreds of data points in just one datum. Additionally, the way in which functional groups appear on a spectrum is not completely discrete. There are sections in which ranges overlap each other, allowing some signals to have the possibility of referring to two or more types of hydrogens. For ^1H NMR, there is also the alcohol/amine (-OH and -NH) range in which hydrogen atoms bonded to oxygen or nitrogen atoms may appear, which was not included as a discretized group in this project. The reason for omission is due to the fact that the signals are often varyingly small or don't appear at all. The omission can become an issue if the signal does appear, as it will then fall into one of the other regions.

4.2 PERFORMING HIERARCHICAL CLUSTERING AND BUILDING A DECISION TREE

The compounds within the dataset were first clustered by a chemist, based on functional groups present in the molecules. It is important to note that there is not a single correct way of doing this. While some sets of clusters are better than others, different chemists may cluster the compounds in slightly different ways. For the dataset in this project, the chemist formed the clusters in an appropriate way, with collaboration with other chemists. The clusters formed in this way were used as a basis for measuring error in the hierarchical clustering, and as target classifications for the decision tree.

As outlined in section 3.1, there are many linkage types that can be used in the clustering of data. In order to obtain the most accurate set of clusters, and to explore the effects of each linkage type on the clusters formed, hierarchical clustering was repeated with each linkage type and the resulting clusters were compared. A heat

map was formed using the most accurate linkage type, and error was calculated. The results of this clustering are shown in chapter 5, and an analysis of the linkage types is detailed in section 5.3.

Using the MATLAB command *fitctree*, a decision tree was formed from the combined spectral data. The clusters formed by the human chemist were used as class labels to provide target classification and to determine the error in the tree. The accuracy of the decision tree was compared to that of the hierarchical clustering. Results of the tree are shown in section 5.4, and implications of what was learned are discussed subsequently.

CHAPTER 5

ANALYSIS OF RESULTS

The results of the hierarchical clustering performed and decision trees built in this project, along with the discussion of them, are described below.

5.1 RESULTS OF HIERARCHICAL CLUSTERING OF THE ORGANIC COMPOUNDS

Hierarchical clustering was performed on ^1H NMR and ^{13}C NMR spectral data from 24 distinct organic compounds. The ^1H NMR and ^{13}C NMR data for each compound were then combined and hierarchical clustering was performed for a third time. Dendrograms were formed for the resulting clusters, and can be seen in figures 5.1, 5.2, and 5.3. Single linkage was used to form the dendrogram for the ^1H NMR data, and ward linkage was used to form the dendrograms for the ^{13}C NMR data and the combined data. The linkage methods used in these dendrograms differ due to their accuracy in clustering the data. The calculation of the accuracies and a deeper discussion of this discrepancy in them can be found in section 5.3.

Heat maps were constructed for each NMR spectral dataset as well, which are shown in figures 5.4, 5.5, and 5.6. Again, single linkage was used to form the heat map for the ^1H NMR data, and ward linkage was used to form the heat maps for the

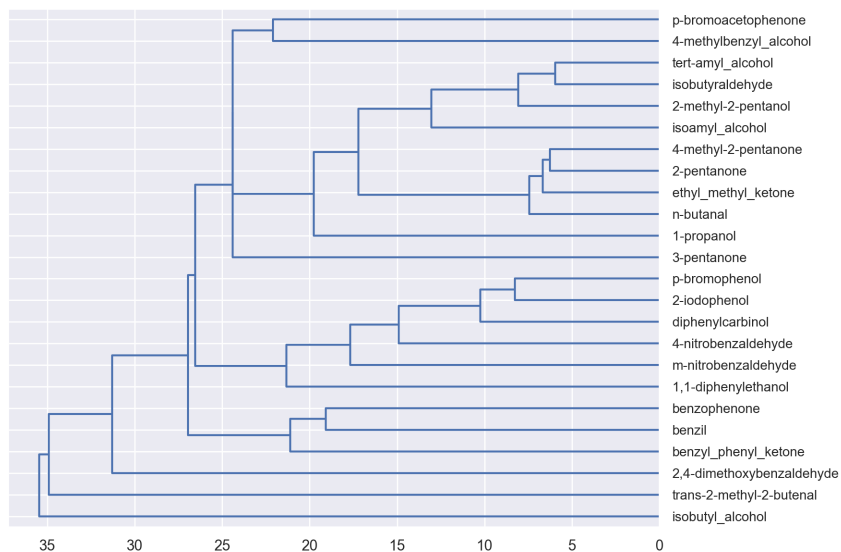


Figure 5.1: Dendrogram visualizing the clustering of 24 organic compounds from ^1H NMR spectral data. Single linkage was used to form this dendrogram.

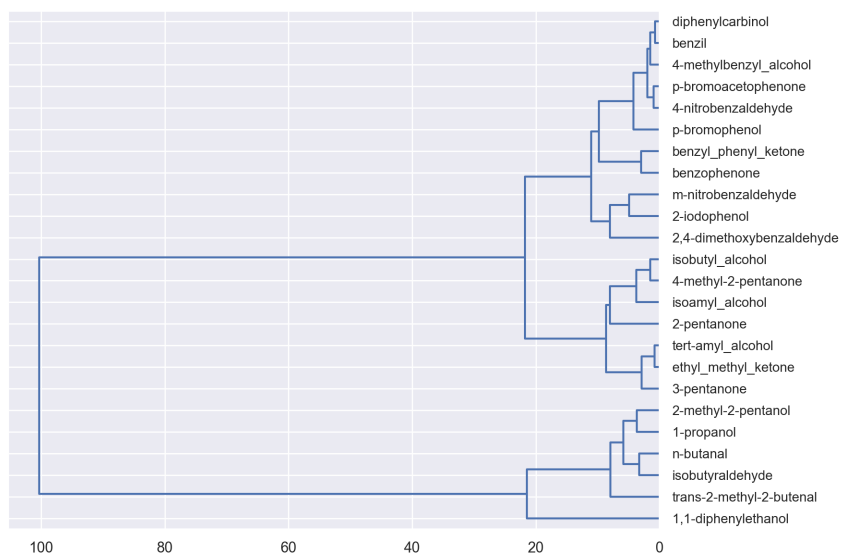


Figure 5.2: Dendrogram visualizing the clustering of 24 organic compounds from ^{13}C NMR spectral data. Ward linkage was used to form this dendrogram.

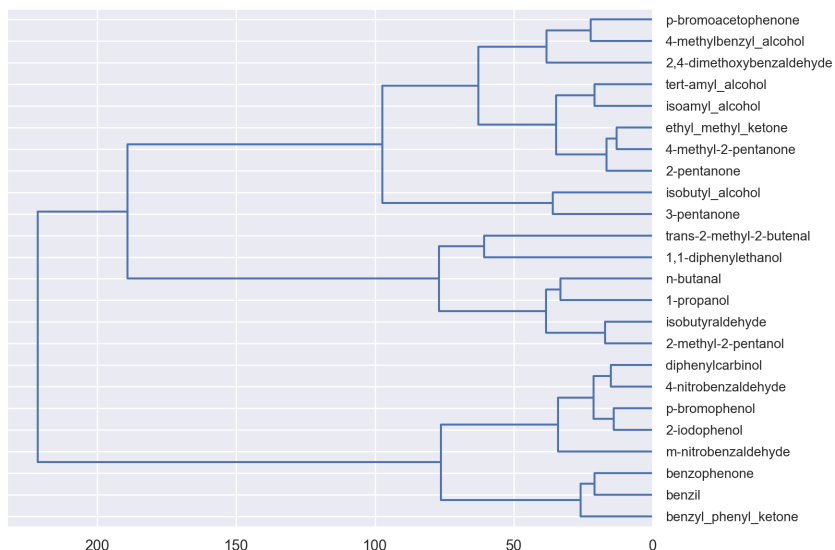


Figure 5.3: Dendrogram visualizing the clustering of 24 organic compounds from combined ^1H NMR and ^{13}C NMR spectral data. Ward linkage was used to form this dendrogram.

^{13}C NMR data and the combined data. The dendrograms for each dataset can be seen on the left side of their respective heat maps.

From each of the three dendrograms, which are also shown in their respective heat maps, clusters of similar molecules can be created. If six clusters were formed from the ^1H NMR spectral data by using a cutoff of 25 in the dendrogram, the clusters would be:

- 4-methyl-2-pentanone, 2-pentanone, ethyl methyl ketone, n-butanal, trans-2-methyl-2-butenal
- p-bromoacetophenone, 4-methylbenzyl alcohol, 2,4-dimethoxybenzaldehyde
- tert-amyl alcohol, isobutyraldehyde, 2-methyl-2-pentanol, isoamyl alcohol
- 3-pentanone, 1-propanol, isobutyl alcohol

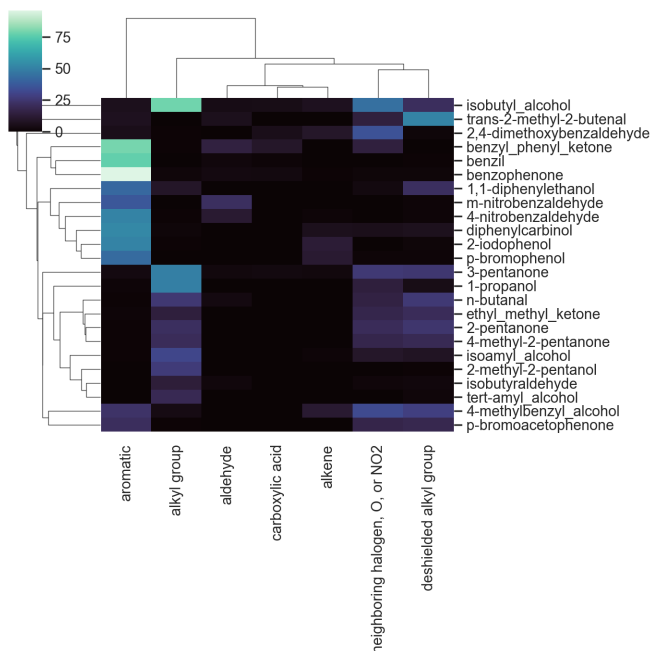


Figure 5.4: Heat map visualizing the clustering of 24 organic compounds with the functional groups they possess. Created using the ^1H NMR spectral data

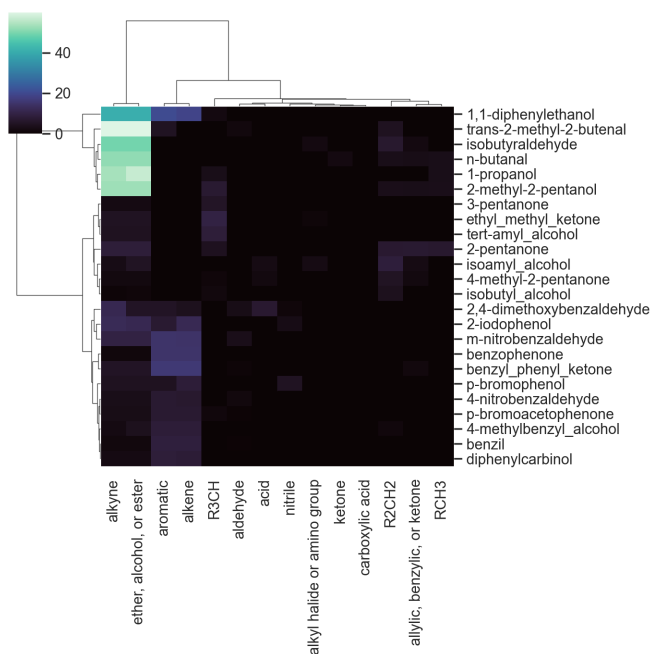


Figure 5.5: Heat map visualizing the clustering of 24 organic compounds with the functional groups they possess. Created using the ^{13}C NMR spectral data

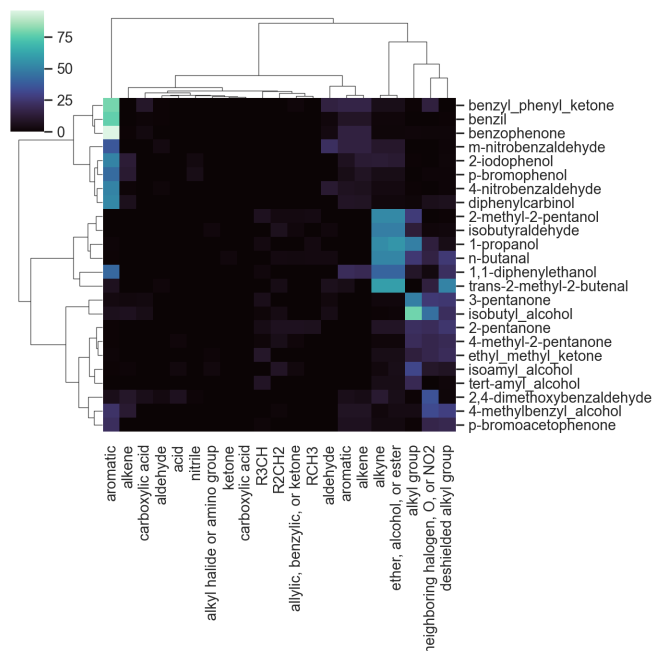


Figure 5.6: Heat map visualizing the clustering of 24 organic compounds with the functional groups they possess. Created using the combined ^1H NMR and ^{13}C NMR spectral data

- p-bromophenol, 2-iodophenol, diphenylcarbinol, 1,1-diphenylethanol, m-nitrobenzaldehyde, 4-nitrobenzaldehyde
- benzophenone, benzil, benzyl phenyl ketone

as seen in figure 5.7.

If six clusters were formed from the ^{13}C NMR spectral data by using a cutoff of 9 in the dendrogram, the clusters would be:

- diphenylcarbinol, benzil, 4-methylbenzyl alcohol, p-bromoacetophenone, 4-nitrobenzaldehyde, p-bromophenol
- benzyl phenyl ketone, benzophenone
- m-nitrobenzaldehyde, 2-iodophenol, 2,4-dimethoxybenzaldehyde
- isobutyl alcohol, 4-methyl-2-pentanone, isoamyl alcohol, 2-pentanone, tert-amyl alcohol, ethyl methyl ketone, 3-pentanone

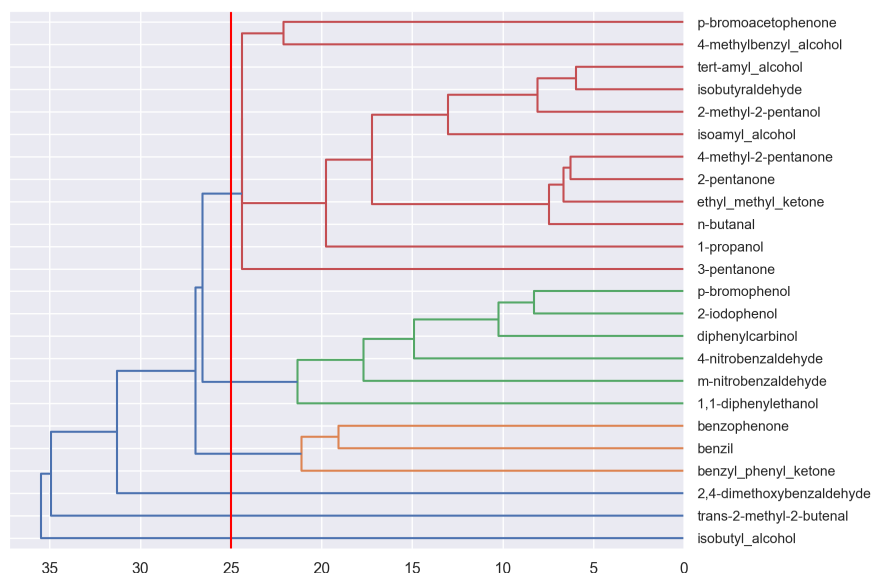


Figure 5.7: Dendrogram showing the six clusters of organic compounds from ^1H NMR spectral data

- 2-methyl-2-pentanol, 1-propanol, n-butanal, isobutyraldehyde, trans-2-methyl-2-butenal
- 1,1-diphenylethanol

as seen in figure 5.8.

If six clusters were formed from the combined ^1H NMR and ^{13}C NMR spectral data by using a cutoff of 70 in the dendrogram, the clusters would be:

- p-bromoacetophenone, 4-methylbenzyl alcohol, 2,4-dimethoxybenzaldehyde, tert-amyl alcohol, isoamyl alcohol, ethyl methyl ketone, 4-methyl-2-pentanone, 2-pentanone
- isobutyl alcohol, 3-pentanone
- trans-2-methyl-2-butenal, 1,1-diphenylethanol
- n-butanal, 1-propanol, isobutyraldehyde, 2-methyl-2-pentanol

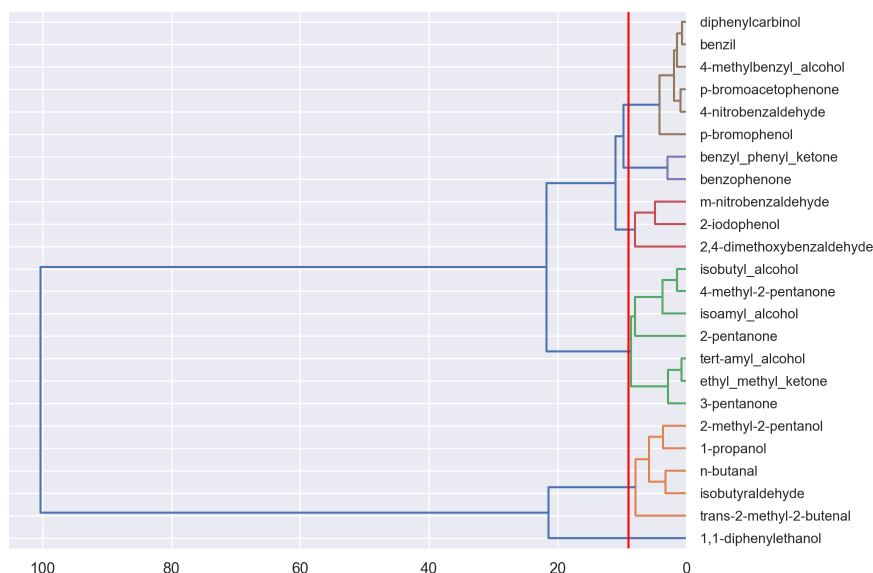


Figure 5.8: Dendrogram showing the six clusters of organic compounds from ^{13}C NMR spectral data

- diphenylcarbinol, 4-nitrobenzaldehyde, p-bromophenol, 2-iodophenol, m-nitrobenzaldehyde
- benzophenone, benzil, benzyl phenyl ketone

as seen in figure 5.9.

While similarity between molecules may be an abstract and unmeasurable quality of organic compounds, it is still possible to analyze the correctness of this clustering. This analysis may be done by comparing the results obtained through the applied method of clustering to the way a chemist may cluster these 24 molecules by hand, as is done in the next sections.

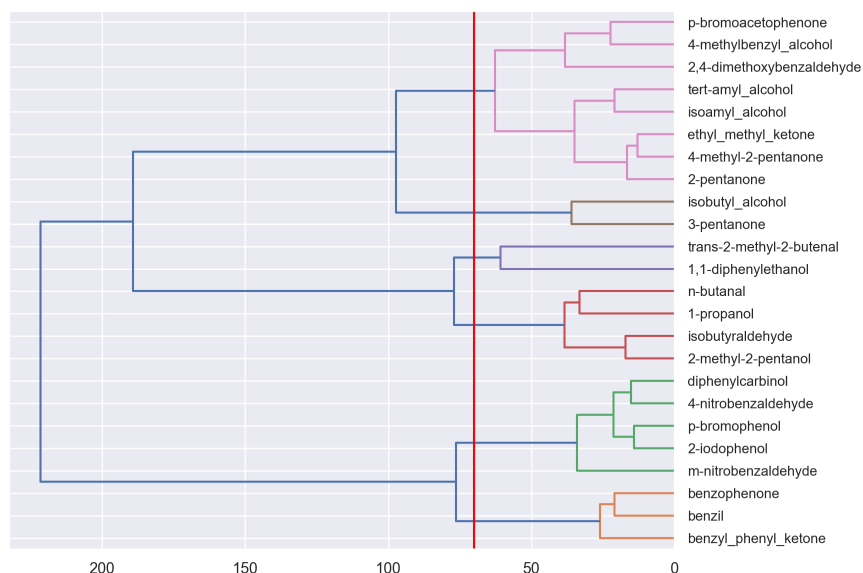


Figure 5.9: Dendrogram showing the six clusters of organic compounds from combined ^1H NMR and ^{13}C NMR spectral data

5.2 NON COMPUTATIONAL CLUSTERING OF THE ORGANIC COMPOUNDS

When clustering the 24 compounds non computationally, the chemical structure of the compounds were analyzed and compared. This resulted in the groups depicted in table 5.1.

This clustering of compounds was done based on the structure of the compounds, namely the functional groups contained within them. The spectra, therefore, should look similar to one another within the clusters. A visual representation of the clusters can be seen in figure 5.10, where the spectra for each compound is shown in its corresponding cluster.

Visually, the clusters seem to be accurate, even with no knowledge of the actual

A	1-propanol, 2-methyl-2-pentanol, isoamyl alcohol, isobutyl alcohol, tert-amyl alcohol
B	2-iodophenol, diphenylcarbinol, p-bromophenol, 4-methylbenzyl alcohol, 1,1-diphenylethanol
C	2-pentanone, 3-pentanone, 4-methyl-2-pentanone, p-bromoacetophenone, ethyl methyl ketone
D	2,4-dimethoxybenzaldehyde, isobutyraldehyde, n-butanal, trans-2-methyl-2-butenal
E	4-nitrobenzaldehyde, m-nitrobenzaldehyde
F	benzyl phenyl ketone, benzil, benzophenone

Table 5.1: The six clusters formed when the organic compounds were clustered by a chemist viewing their structures

structure of the compounds that the spectra represent. This shows that compounds have the possibility of being clustered by structure based on their spectra alone.

Certain groups look very similar, namely B and E, but are structurally different enough to be placed in separate groups. Groups A and D also share quite a few similarities, but are again structurally distinct. Even so, the structure differences between groups B and E, and those between groups A and D, are less significant than the structural differences between groups A and F for example, which have many more clear differences in their spectra (figure 5.10). An example of this can be seen in figures 5.11, 5.12, and 5.13.

The compound shown in figure 5.11 is 4-methylbenzyl alcohol and is in group B, and that in figure 5.12 is 4-nitrobenzaldehyde and is in group E. They are structurally very similar, but the difference of the alcohol group (-OH) of 4-methylbenzyl alcohol and the aldehyde (O double bonded to C-H) of 4-nitrobenzaldehyde causes them to be placed in different groups by chemists. Similarities in compound structure like



Figure 5.10: Spectra of organic compounds clustered by the human chemist

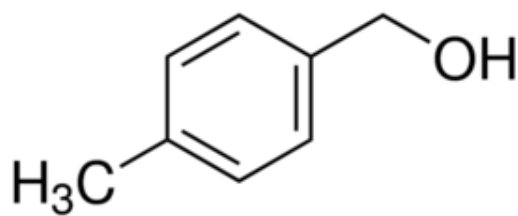


Figure 5.11: Structure of 4-methylbenzyl alcohol [5]

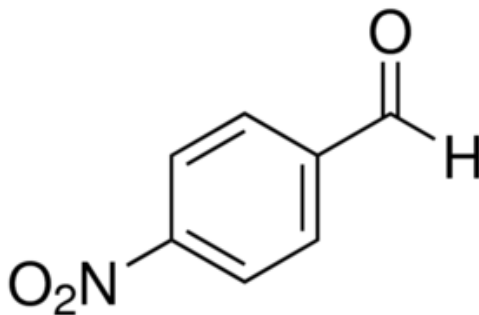


Figure 5.12: Structure of 4-nitrobenzaldehyde [5]

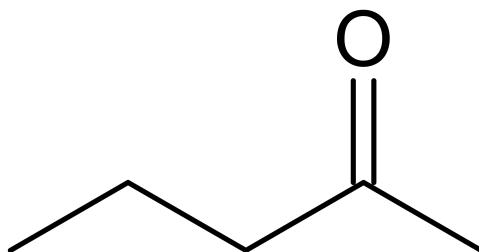


Figure 5.13: Structure of 2-pentanone

this cause groups B and E to be closer clusters than B and C for example, where 2-pentanone, the compound in figure 5.13 is located. The structure of 2-pentanone and other compounds in the same group have many more differences to 4-methylbenzyl alcohol and other compounds in group B.

Hierarchical clustering was used to cluster the compounds, whose clusters were compared to those of the chemist. The hierarchical clustering was performed six times for each NMR method as well as the combined data from both NMR methods, once using each linkage type (average, ward, single, complete, weighted, centroid, and median linkage). Figures 5.14, 5.15, 5.16, and 5.17 show the dendrograms for average, ward, single, and complete linkage hierarchical clustering of ¹HNMR data respectively. The clusters formed from weighted, centroid, and median linkage hierarchical clustering were equivalent to those using average linkage. Figures

5.18, 5.19, 5.20, 5.21, 5.22, and 5.23 show the dendrograms for average, ward, single, complete, weighted, and median linkage hierarchical clustering of ^{13}C NMR data respectively. The clusters formed from centroid linkage hierarchical clustering were equivalent to those using single linkage. Figures 5.24, 5.25, and 5.26 show the dendrograms for average, ward, and single linkage hierarchical clustering respectively of combined ^1H NMR and ^{13}C NMR data. The clusters formed from complete linkage hierarchical clustering were equivalent to those using ward linkage, and the clusters formed from weighted, centroid, and median linkage hierarchical clustering were equivalent to those using average linkage.

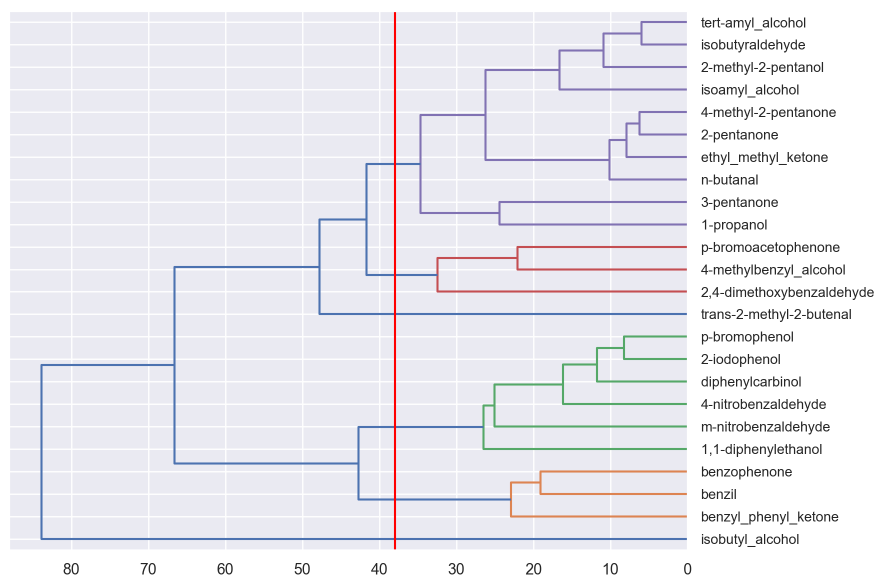


Figure 5.14: ^1H NMR dendrogram formed using average linkage

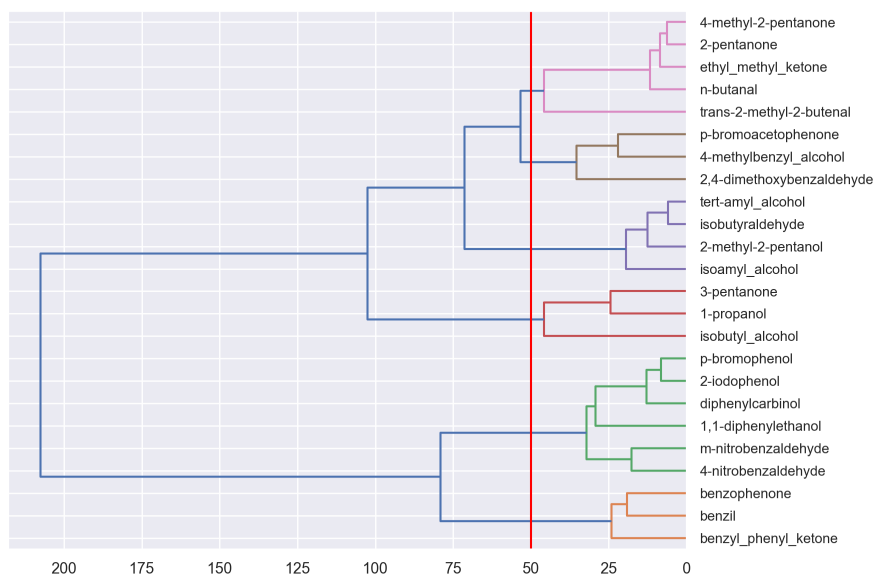


Figure 5.15: ^1H NMR dendrogram formed using ward linkage

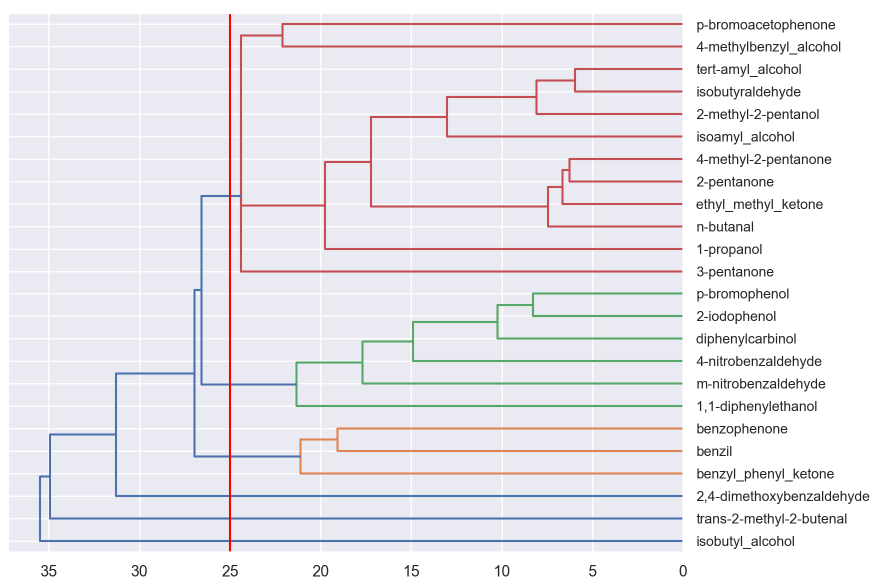


Figure 5.16: ^1H NMR dendrogram formed using single linkage

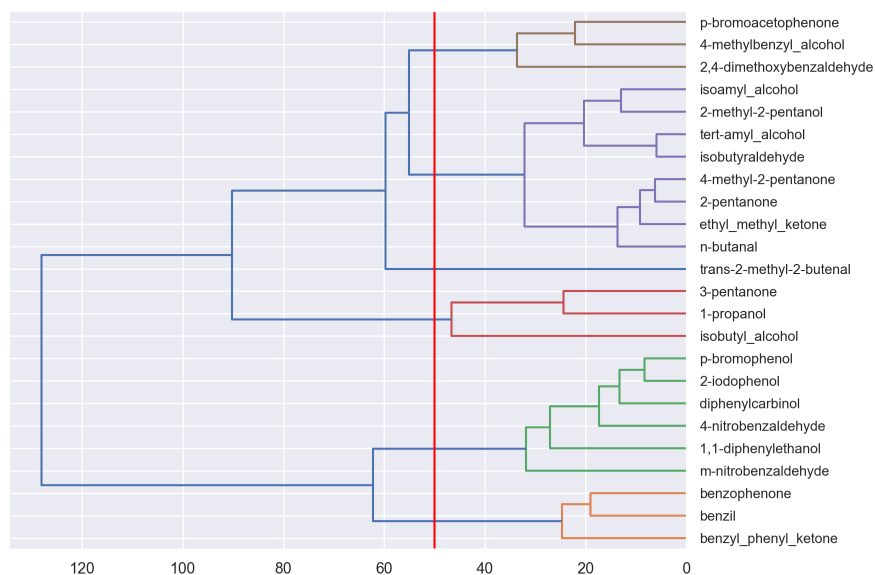


Figure 5.17: ^1H NMR dendrogram formed using complete linkage

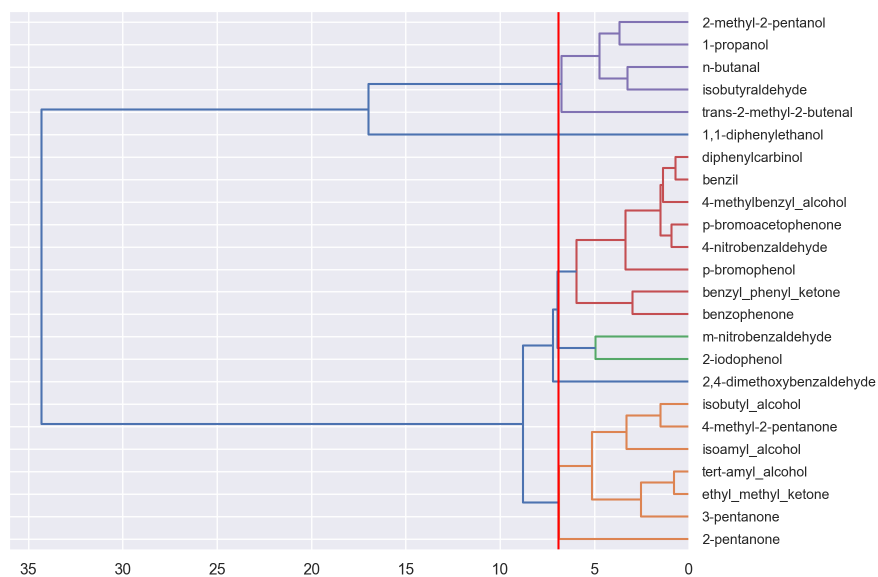


Figure 5.18: ^{13}C NMR dendrogram formed using average linkage

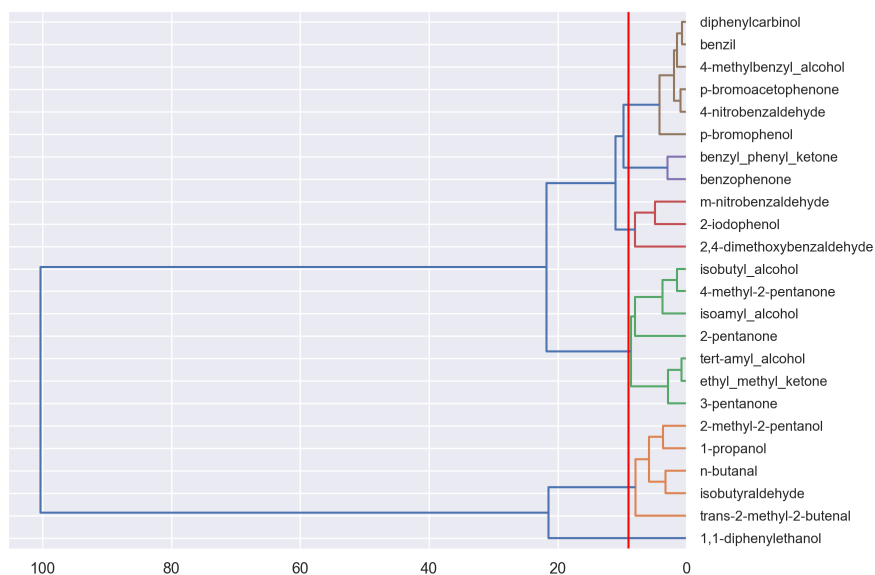


Figure 5.19: ^{13}C NMR dendrogram formed using ward linkage

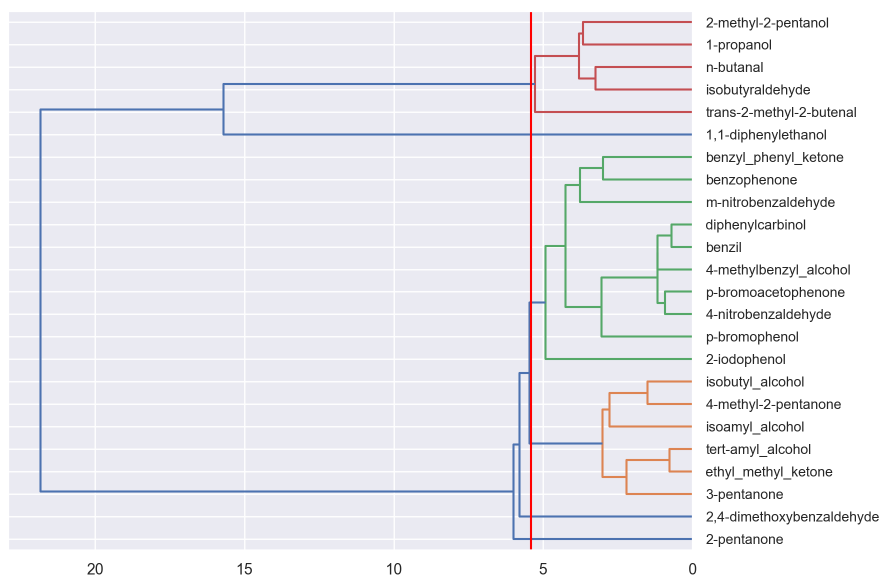


Figure 5.20: ^{13}C NMR dendrogram formed using single linkage

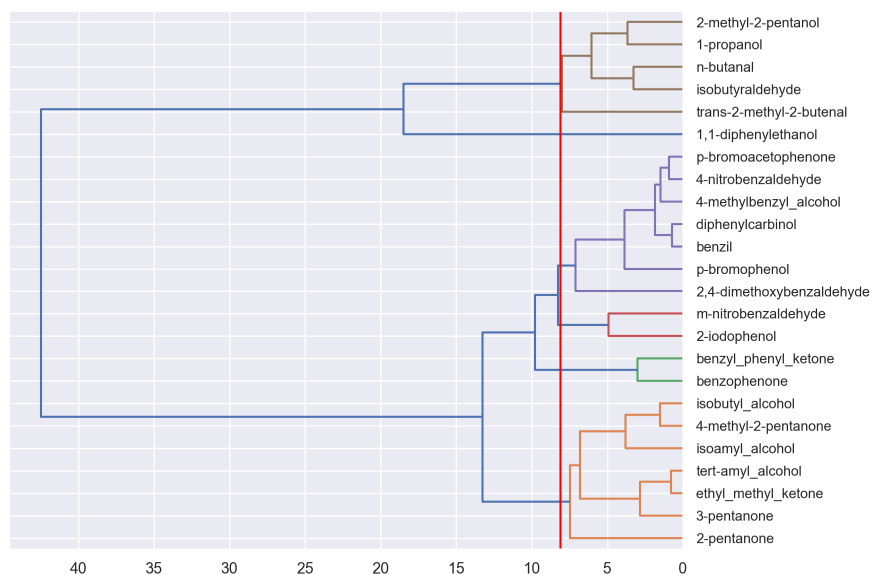


Figure 5.21: ^{13}C NMR dendrogram formed using complete linkage

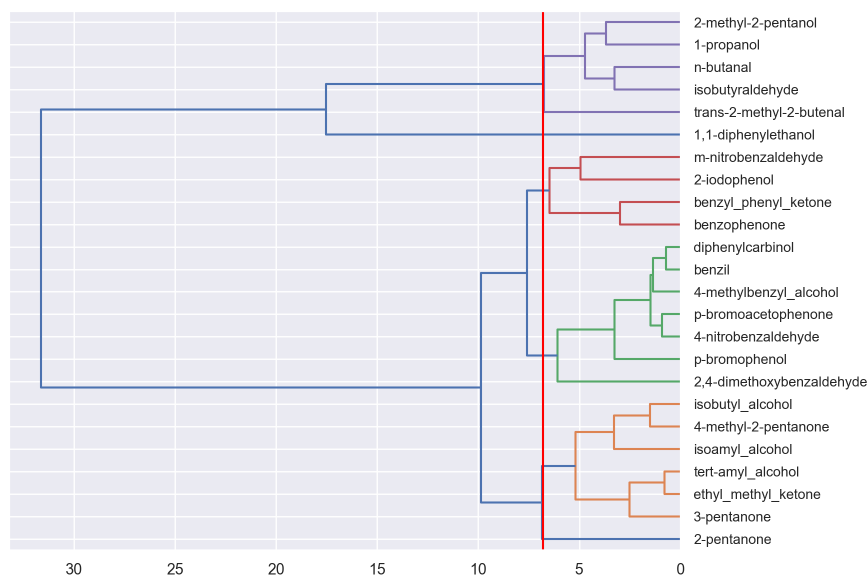


Figure 5.22: ^{13}C NMR dendrogram formed using weighted linkage

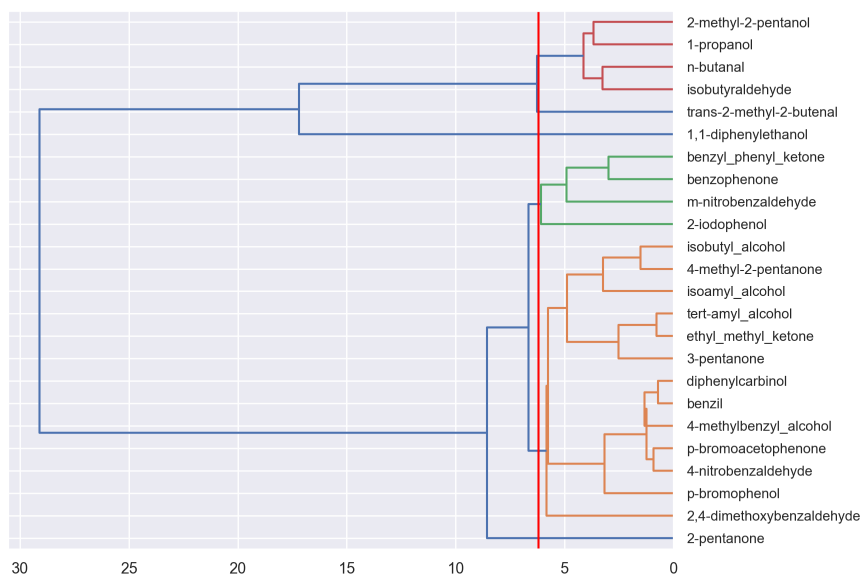


Figure 5.23: ^{13}C NMR dendrogram formed using median linkage

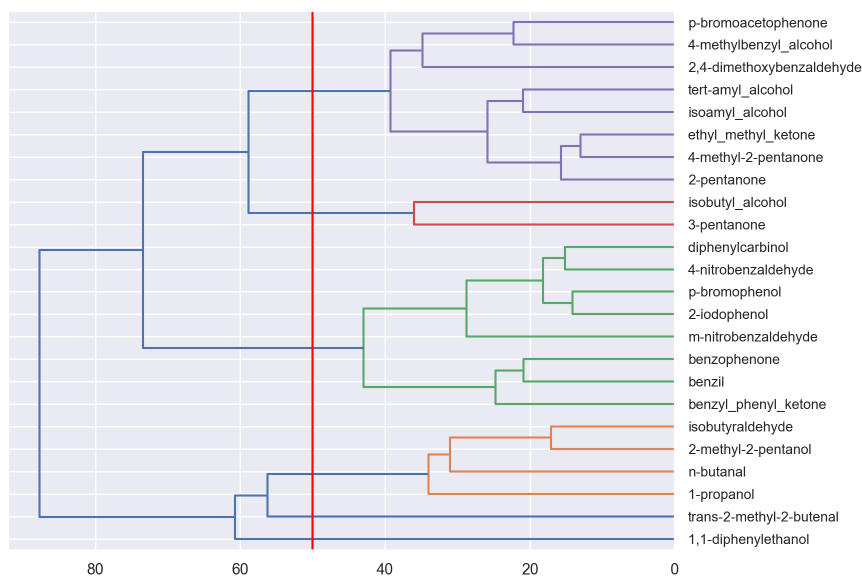


Figure 5.24: Combined data dendrogram formed using average linkage

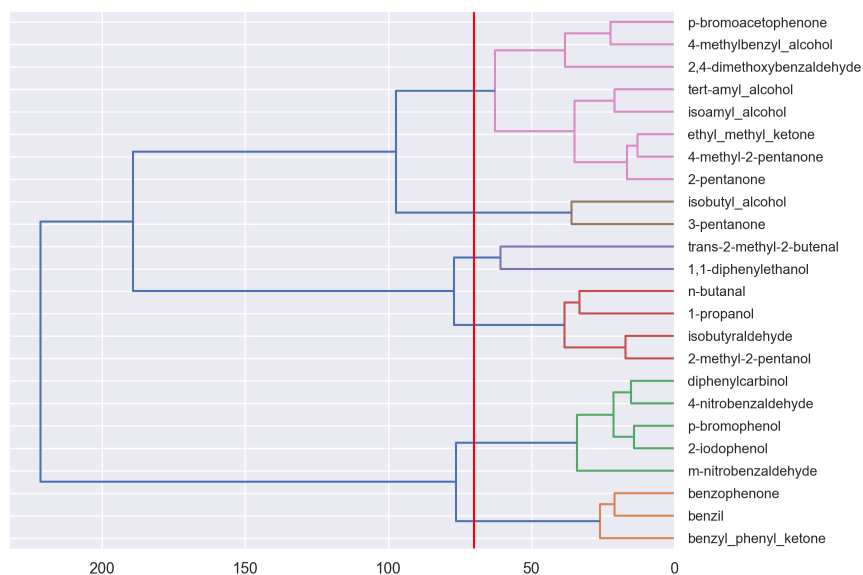


Figure 5.25: Combined data dendrogram formed using ward linkage

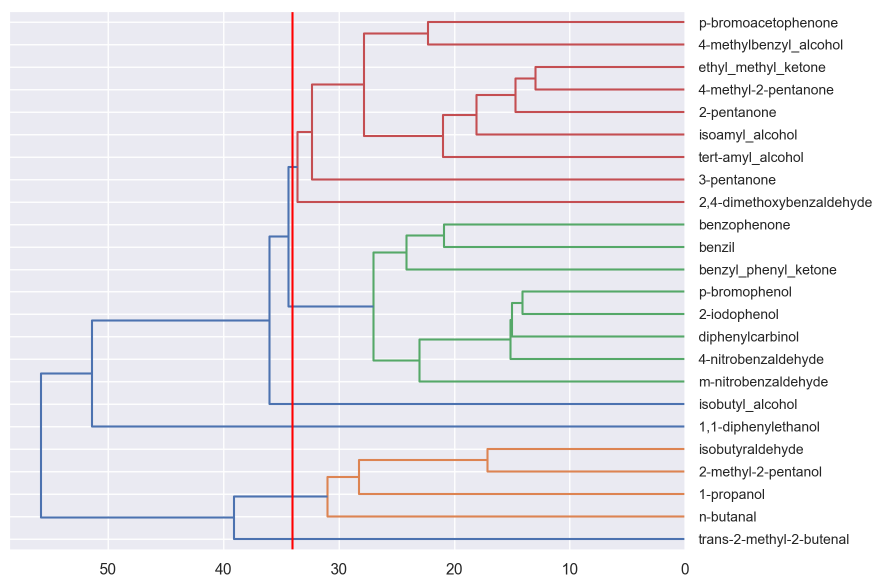


Figure 5.26: Combined data dendrogram formed using single linkage

5.3 DISCUSSION OF FINDINGS

5.3.1 PROTON NMR HIERARCHICAL CLUSTERING

Figure 5.27 shows the clusters of average, ward, single, and complete linkage hierarchical clustering from ^1H NMR spectral data compared to the groups formed by the chemist. The clusters formed from weighted, centroid, and median linkage hierarchical clustering were equivalent to those formed using average linkage.

Human expert	Average	Ward	Single	Complete
1-propanol 2-methyl-2-pentanol isoamyl alcohol isobutyl alcohol tert-amyl alcohol	isobutyl alcohol	2-methyl-2-pentanol isoamyl alcohol tert-amyl alcohol isobutyraldehyde	isobutyl alcohol	1-propanol isobutyl alcohol 3-pentanone
2-iodophenol diphenylcarbinol p-bromophenol 4-methylbenzyl alcohol 1,1-diphenylethanol	2-iodophenol diphenylcarbinol p-bromophenol 1,1-diphenylethanol 4-nitrobenzaldehyde m-nitrobenzaldehyde	2-iodophenol diphenylcarbinol p-bromophenol 1,1-diphenylethanol 4-nitrobenzaldehyde m-nitrobenzaldehyde	2-iodophenol diphenylcarbinol p-bromophenol 1,1-diphenylethanol 4-nitrobenzaldehyde m-nitrobenzaldehyde	2-iodophenol diphenylcarbinol p-bromophenol 1,1-diphenylethanol 4-nitrobenzaldehyde m-nitrobenzaldehyde
2-pentanone 3-pentanone 4-methyl-2-pentanone p-bromoacetophenone ethyl methyl ketone	4-methyl-2-pentanone 2-pentanone ethyl methyl ketone 3-pentanone tert-amyl alcohol 2-methyl-2-pentanol isoamyl alcohol 1-propanol isobutyraldehyde n-butanal	3-pentanone 1-propanol isobutyl alcohol	p-bromoacetophenone 4-methyl-2-pentanone 2-pentanone ethyl methyl ketone 3-pentanone tert-amyl alcohol 2-methyl-2-pentanol isoamyl alcohol 1-propanol n-butanal isobutyraldehyde 4-methylbenzyl alcohol	4-methyl-2-pentanone 2-pentanone ethyl methyl ketone 2-methyl-2-pentanol isoamyl alcohol tert-amyl alcohol isobutyraldehyde n-butanal
2,4-dimethoxybenzaldehyde isobutyraldehyde n-butanal trans-2-methyl-2-butenal	trans-2-methyl-2-butenal	n-butanal trans-2-methyl-2-butenal 2-pentanone 4-methyl-2-pentanone ethyl methyl ketone	trans-2-methyl-2-butenal	trans-2-methyl-2-butenal
4-nitrobenzaldehyde m-nitrobenzaldehyde	p-bromoacetophenone 4-methylbenzyl alcohol 2,4-dimethoxybenzaldehyde	p-bromoacetophenone 4-methylbenzyl alcohol 2,4-dimethoxybenzaldehyde	2,4-dimethoxybenzaldehyde	p-bromoacetophenone 4-methylbenzyl alcohol 2,4-dimethoxybenzaldehyde
benzyl phenyl ketone benzil benzophenone	benzyl phenyl ketone benzil benzophenone	benzyl phenyl ketone benzil benzophenone	benzyl phenyl ketone benzil benzophenone	benzyl phenyl ketone benzil benzophenone

Figure 5.27: Clusters from ^1H NMR data formed using hierarchical clustering with average, ward, single, and complete linkage methods. Colors indicate when the group the compound is in matches the groups formed by the chemist.

Single linkage produced the most accurate clusters when compared to those formed by the chemist, though they were not much more accurate than all other linkage methods. The compounds were placed into the same clusters as the chemist 58.3% of the time. While this is not very accurate, the compounds that were incorrectly clustered together tend to belong to similar groups created by the

chemist. If four groups were formed by the chemist in place of six, and the similar groups (A and D, and B and E) were merged, the accuracy of hierarchical clustering with single linkage would be 79.2%.

5.3.2 CARBON NMR HIERARCHICAL CLUSTERING

Figure 5.28 shows the clusters of average, ward, single, complete, weighted, and median linkage hierarchical clustering from ^{13}C NMR spectral data compared to the groups formed by the chemist. The clusters formed from centroid linkage hierarchical clustering were equivalent to those formed using average linkage.

Human expert	Average	Ward	Single	Complete	Weighted	Median
1-propanol 2-methyl-2-pentanol isoamyl alcohol isobutyl alcohol tert-amyl alcohol	1-propanol 2-methyl-2-pentanol isobutyraldehyde n-butanal trans-2-methyl-2-butenal	1,1-diphenylethanol	isobutyl alcohol isoamyl alcohol tert-amyl alcohol 4-methyl-2-pentanone ethyl methyl ketone 3-pentanone	1-propanol 2-methyl-2-pentanol isobutyraldehyde n-butanal trans-2-methyl-2-butenal	isoamyl alcohol isobutyl alcohol tert-amyl alcohol 4-methyl-2-pentanone ethyl methyl ketone 3-pentanone	1-propanol 2-methyl-2-pentanol isobutyraldehyde n-butanal
2-iodophenol diphenylcarbinol p-bromophenol 4-methylbenzyl alcohol 1,1-diphenylethanol	1,1-diphenylethanol	diphenylcarbinol p-bromophenol 4-methylbenzyl alcohol p-bromoacetophenone 4-nitrobenzaldehyde benzil	1,1-diphenylethanol	diphenylcarbinol p-bromophenol 4-methylbenzyl alcohol benzil p-bromoacetophenone 4-nitrobenzaldehyde 2,4-dimethoxybenzaldehyde	diphenylcarbinol p-bromophenol 4-methylbenzyl alcohol benzil p-bromoacetophenone 4-nitrobenzaldehyde 2,4-dimethoxybenzaldehyde	1,1-diphenylethanol
2-pentanone 3-pentanone 4-methyl-2-pentanone p-bromoacetophenone ethyl methyl ketone	4-methyl-2-pentanone ethyl methyl ketone 2-pentanone 3-pentanone isobutyl alcohol isoamyl alcohol tert-amyl alcohol	4-methyl-2-pentanone ethyl methyl ketone 2-pentanone 3-pentanone isobutyl alcohol isoamyl alcohol tert-amyl alcohol	2-pentanone	4-methyl-2-pentanone ethyl methyl ketone 2-pentanone 3-pentanone isobutyl alcohol isoamyl alcohol tert-amyl alcohol	3-pentanone	4-methyl-2-pentanone ethyl methyl ketone 3-pentanone p-bromoacetophenone isobutyl alcohol isoamyl alcohol tert-amyl alcohol diphenylcarbinol benzil 4-methylbenzyl alcohol 4-nitrobenzaldehyde p-bromophenol 2,4-dimethoxybenzaldehyde
2,4-dimethoxybenzaldehyde isobutyraldehyde n-butanal trans-2-methyl-2-butenal	2,4-dimethoxybenzaldehyde	isobutyraldehyde n-butanal trans-2-methyl-2-butenal 1-propanol 2-methyl-2-pentanol	isobutyraldehyde n-butanal trans-2-methyl-2-butenal 1-propanol 2-methyl-2-pentanol	1,1-diphenylethanol	isobutyraldehyde n-butanal trans-2-methyl-2-butenal 1-propanol 2-methyl-2-pentanol	trans-2-methyl-2-butenal
4-nitrobenzaldehyde m-nitrobenzaldehyde	m-nitrobenzaldehyde 2-iodophenol	m-nitrobenzaldehyde 2-iodophenol 2,4-dimethoxybenzaldehyde	2,4-dimethoxybenzaldehyde	m-nitrobenzaldehyde 2-iodophenol	1,1-diphenylethanol	2-pentanone
benzyl phenyl ketone benzil benzophenone	benzil benzyl phenyl ketone benzophenone diphenylcarbinol 4-methylbenzyl alcohol p-bromoacetophenone 4-nitrobenzaldehyde p-bromophenol	benzyl phenyl ketone benzophenone	benzyl phenyl ketone benzophenone benzil m-nitrobenzaldehyde 4-methylbenzyl alcohol p-bromoacetophenone 4-nitrobenzaldehyde 2-iodophenol p-bromophenol	benzyl phenyl ketone benzophenone	benzyl phenyl ketone benzophenone m-nitrobenzaldehyde 2-iodophenol	benzyl phenyl ketone benzophenone m-nitrobenzaldehyde 2-iodophenol

Figure 5.28: Clusters from ^{13}C NMR data formed using hierarchical clustering with average, ward, single, complete, weighted, and median linkage methods. Colors indicate when the group the compound is in matches the groups formed by the chemist.

Ward linkage produced the most accurate clusters when compared to those formed by the chemist, though it was not much more accurate than average, complete, and weighted linkage. The compounds were placed into the same clusters as the chemist 54.2% of the time. Again, this is not very accurate, but the compounds

that were incorrectly clustered together tend to belong to similar groups created by the chemist. If four groups were formed by the chemist in place of six, and the similar groups (A and D, and B and E) were merged, the accuracy of hierarchical clustering with ward linkage would be 70.8%.

5.3.3 ANALYSIS OF COMBINED RESULTS

Figure 5.29 shows the clusters of average, ward, and single linkage hierarchical clustering from the combined ^1H NMR and ^{13}C NMR spectral data compared to the groups formed by the chemist. The clusters formed from complete linkage hierarchical clustering were equivalent to those formed using ward linkage, and the clusters formed from weighted, centroid, and median linkage hierarchical clustering were equivalent to those formed using average linkage.

Again, ward linkage produced the most accurate clusters when compared to those formed by the chemist, though it was not much more accurate than single linkage. The compounds were placed into the same clusters as the chemist 54.2% of the time. While this is not very accurate, once again the compounds that were incorrectly clustered together tend to belong to similar groups by the chemist. If four groups were formed by the chemist in place of six, and the similar groups (A and D, and B and E) were merged, the accuracy of hierarchical clustering with ward linkage would be 75.0%.

Interestingly, the combining of spectral data produced clusters with an accuracy between that of the ^1H NMR and ^{13}C NMR data clustering, rather than a higher accuracy than both. Possibilities as to why this is the case are discussed in the next section.

Human expert	Average	Ward	Single
1-propanol 2-methyl-2-pentanol isoamyl alcohol isobutyl alcohol tert-amyl alcohol	isobutyl alcohol 3-pentanone	isobutyl alcohol 3-pentanone	isobutyl alcohol
2-iodophenol diphenylcarbinol p-bromophenol 4-methylbenzyl alcohol 1,1-diphenylethanol	1,1-diphenylethanol	1,1-diphenylethanol trans-2-methyl-2-butenal	1,1-diphenylethanol
2-pentanone 3-pentanone 4-methyl-2-pentanone p-bromoacetophenone ethyl methyl ketone	p-bromoacetophenone ethyl methyl ketone 2-pentanone 4-methyl-2-pentanone 4-methylbenzyl alcohol 2,4-dimethoxybenzaldehyde isoamyl alcohol tert-amyl alcohol	2-pentanone 4-methyl-2-pentanone ethyl methyl ketone p-bromoacetophenone 4-methylbenzyl alcohol 2,4-dimethoxybenzaldehyde isoamyl alcohol tert-amyl alcohol	2-pentanone 3-pentanone 4-methyl-2-pentanone ethyl methyl ketone p-bromoacetophenone 4-methylbenzyl alcohol 2,4-dimethoxybenzaldehyde isoamyl alcohol tert-amyl alcohol
2,4-dimethoxybenzaldehyde isobutyraldehyde n-butanal trans-2-methyl-2-butenal	isobutyraldehyde n-butanal 1-propanol 2-methyl-2-pentanol	n-butanal isobutyraldehyde 1-propanol 2-methyl-2-pentanol	n-butanal isobutyraldehyde 1-propanol 2-methyl-2-pentanol
4-nitrobenzaldehyde m-nitrobenzaldehyde	trans-2-methyl-2-butenal	4-nitrobenzaldehyde m-nitrobenzaldehyde 2-iodophenol diphenylcarbinol p-bromophenol	trans-2-methyl-2-butenal
benzyl phenyl ketone benzil benzophenone	benzyl phenyl ketone benzil benzophenone diphenylcarbinol p-bromophenol 2-iodophenol 4-nitrobenzaldehyde m-nitrobenzaldehyde	benzyl phenyl ketone benzil benzophenone	benzyl phenyl ketone benzil benzophenone diphenylcarbinol p-bromophenol 2-iodophenol 4-nitrobenzaldehyde m-nitrobenzaldehyde

Figure 5.29: Clusters from combined ^1H NMR and ^{13}C NMR data formed using hierarchical clustering with average, ward, and single linkage methods. Colors indicate when the group the compound is in matches the groups formed by the chemist.

5.3.4 DISCUSSION OF ERROR IN HIERARCHICAL CLUSTERING OF THE SPECTRAL DATA

Inaccuracies in the clustering are due to a mixture of both physical and computational issues. When NMR spectroscopy is performed on organic compounds, a solution of the compound must be created. Impurity of this solution can affect the spectrum produced, as the hydrogen atoms in the impurities would be read just as any hydrogen atoms in the compound of interest are, and the same is true for carbon atoms in any present impurities. Changes in spectra due to these impurities, as well as any systematic errors that may have been present in the spectrometer for some of the compounds, could affect the clustering of the compounds from the spectral data.

Computationally, the largest obstacle is the feature extraction and data discretization. The condensation of the data is unavoidable, but causes a loss of precision. A larger set of features could increase the accuracy of the clusters. Additionally, a larger dataset would provide a clearer view of which clusters are more distinct. These changes could be implemented into a future project and the effects of those changes could be explored. A further discussion of the obstacles that accompany hierarchical clustering can be seen in section 6.1, and a further discussion of the work that could potentially follow this thesis can be read in section 6.2.

5.3.5 EXAMINATION OF LINKAGE METHODS

The ^1H NMR data had the most accurate clustering when single linkage was used, and ^{13}C NMR data had the most accurate clustering when ward linkage was used. The differences in accurate clustering between linkage methods is minute, though an examination of single and ward linkage may still be useful. Single linkage defines distance between two clusters u and v as:

$$d(u, v) = \min(\text{dist}(u[i], v[j]))$$

for all points i in u and j in v . Ward linkage defines distance between two clusters u and v as:

$$d(u, v) = \sqrt{\frac{|v| + |s|}{T}d(v, s)^2 + \frac{|v| + |t|}{T}d(v, t)^2 - \frac{|v|}{T}d(s, t)^2}$$

where s and t are subclusters of u , $T = |v| + |u| + |t|$, and $|x|$ denotes cardinality [4].

These equations show that single linkage takes only one datapoint from each cluster into account, while the computation for ward linkage includes all subclusters, which include all data points within them. One of these linkage methods may be more accurate in some situations, and seen as 'better', but, as seen in this project, one is not always 'better' than the other. While the ^1H NMR and ^{13}C NMR spectral

data are distinct within the focus of organic chemistry, they are mathematically the same: a numerical matrix of sums of intensities of peaks within sections of a spectrum. Yet different linkage methods performed differently when clustering the compounds. It may be predicted that this disparity in accurate linkage methods is incidental, but this cannot be certain without a larger set of compounds with which to test this hypothesis.

5.4 RESULTS AND DISCUSSION OF DECISION TREE FORMATION

The discretized data was formatted into a numerical matrix, in which each row was a compound and each column was an attribute (the apparent presence of a functional group in the spectral data). The target column matrix was made up of integers representing the classes formed by the human chemist. Using the MATLAB command *fitctree*, a decision tree was formed from the data. This tree can be seen in figure 5.30, and its rules are listed in table 5.2.

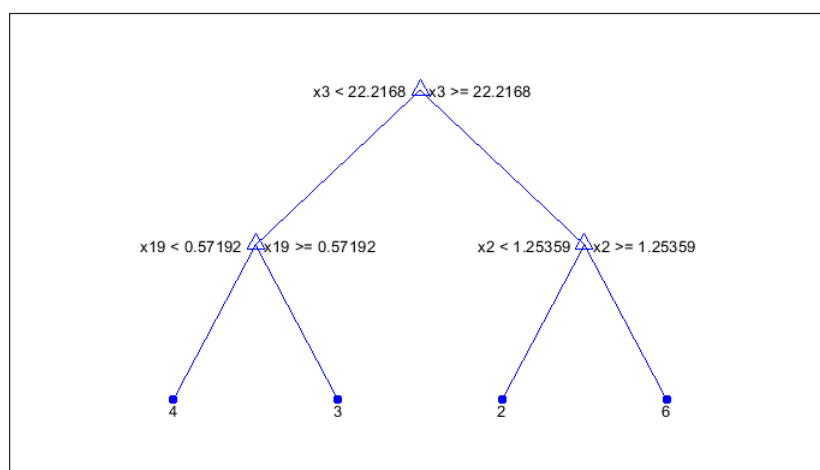


Figure 5.30: Decision tree formed from combined spectral data. Attributes x3, x2, and x19 represent the presence of an aromatic ring in ^1H NMR, the presence of an aldehyde group in ^1H NMR, and the presence of an R_3CH alkyl group in ^{13}C NMR respectively. The classes at the leaf nodes correspond to those formed by the human chemist.

	Rule	Data Points in Classification
R1	$x_3 < 22.21$ and $x_{19} < 0.57 \Rightarrow C_4$	4 data points in C4, 1 data point in C1
R2	$x_3 < 22.21$ and $x_{19} \geq 0.57 \Rightarrow C_3$	5 data points in C3, 4 data points in C1
R3	$x_3 \geq 22.21$ and $x_2 < 1.25 \Rightarrow C_2$	5 data points in C2
R4	$x_3 \geq 22.21$ and $x_2 \geq 1.25 \Rightarrow C_6$	3 data points in C6, 2 data points in C5

Table 5.2: The rules generated by the decision tree in figure 5.30

There are two notable aspects of this tree: 18 of the 21 attributes not being considered as part of the classification, and classes 1 and 5 not being represented. Attributes x_3 , x_2 , and x_{19} are the only three attributes that are considered when classifying the 24 data points, or compounds, in this decision tree. This shows that, in this case, these three attributes are the most important to view when identifying these organic compounds. Attribute x_3 represents the presence of an aromatic ring in ^1H NMR. The aromatic ring is one of the most recognizable functional groups in both the physical structure of a molecule and the ^1H NMR spectrum of the molecule. In the heat map in figure 5.6, the aromatic column is also visibly distinct from the other functional group columns. Attributes x_2 and x_{19} represent the presence of an aldehyde group in ^1H NMR, and the presence of an R_3CH alkyl group in ^{13}C NMR respectively. These are also recognizable functional groups, though not as recognizable as the aromatic ring in ^1H NMR. Rule R3, as seen in the table 5.2, perfectly recognizes class 2. This shows that attributes x_3 and x_2 are most important when recognizing specifically the types of compounds present in class 2 (alcohols with aromatic rings).

If the dataset used to form the decision tree was larger, the attributes of importance may change. With a dataset as small as 24 compounds, it is possible that there is

one or more functional group(s) in much higher presence proportionally than there would be in a dataset that is orders of magnitude larger, and therefore would more closely resemble the distribution of functional groups over the set of all organic compounds. With this small dataset, it is possible to get an idea of what to expect when viewing more compounds, even if the most important attributes themselves are not the same.

Another consequence of the small size of the dataset used in the formation of this decision tree is the missing classification leaf nodes. To avoid overfitting, MATLAB prevents nodes from splitting further if they have fewer than 10 data points. This causes classes of small size to be lost. The smallest classes present in the dataset are classes 5 and 6, which contain two and three data points respectively. The leaf node classified as class 6 contains all five of the data points that belong to classes 5 and 6, but because class 5 only has two data points and class 6 has three, the tree assumes that all should belong to class 6. Class 1 is also missing from the decision tree, although not due only to its size. Data points from class 1 were split on attribute x19, causing one of them to be classified with class 4, and the other four to be classified with class 3. Class 3 having five data points caused all nine of the data points (five from C3 and four from C1) at that leaf node to be classified as class 3.

Altogether, the misclassifications produced an error of 29.2%, giving the decision tree an accuracy of 70.8%. Comparing this to the accuracy of the hierarchical clustering described in the previous sections (54.2%), the decision tree has significantly higher accuracy.

The misclassifications that cause this error can be further examined in another decision tree. As mentioned previously, classes 1 and 3 are classified almost completely the same. A decision tree can be formed from isolated data from only those two classes to identify the attribute that is needed to separate the two. When this is done, the decision tree in figure 5.31 is formed and the rules in table 5.3 can

be written. This shows that attribute x5 (the presence of a neighboring halogen, O, or NO₂ in ¹HNMR) is the most important attribute in separating these two classes.

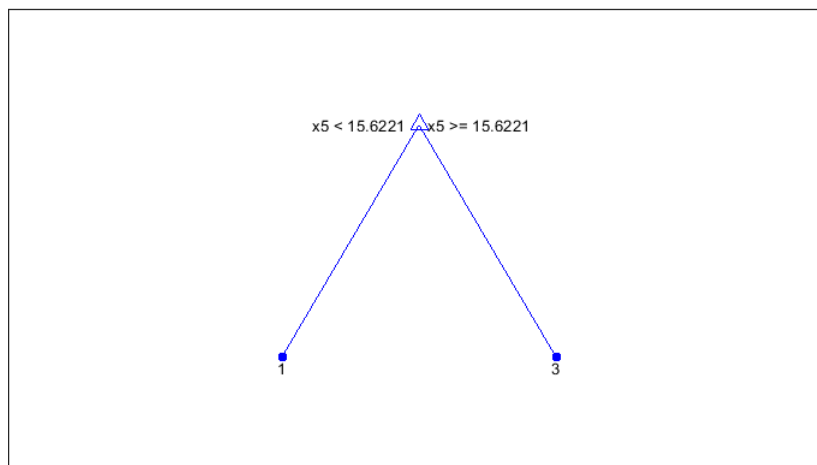


Figure 5.31: Decision tree formed from classes 1 and 3 of the combined spectral data. Attribute x5 represents the presence of a neighboring halogen, O, or NO₂ in ¹HNMR. The classes at the leaf nodes correspond to those formed by the human chemist.

	Rule	Data Points in Classification
R1	$x5 < 15.62 \Rightarrow C1$	4 data points in C1
R2	$x5 \geq 15.62 \Rightarrow C3$	5 data points in C3, 1 data point in C1

Table 5.3: The rules generated by the decision tree in figure 5.31

Ideally, this examination of misclassification would be repeated with classes 5 and 6, but because there are a total of only five data points on those two classes combined, the *fitctree* command would not form a tree.

While the default for *fitctree* is to not allow splitting of nodes with fewer than ten data points, the parameter *MinParentSize* can be changed in order to allow the nodes to split as small as necessary. Changing *MinParentSize* to 5 to allow for the

splitting of classes 5 and 6 produces the decision tree depicted in figure 5.32, with rules described in table 5.4.

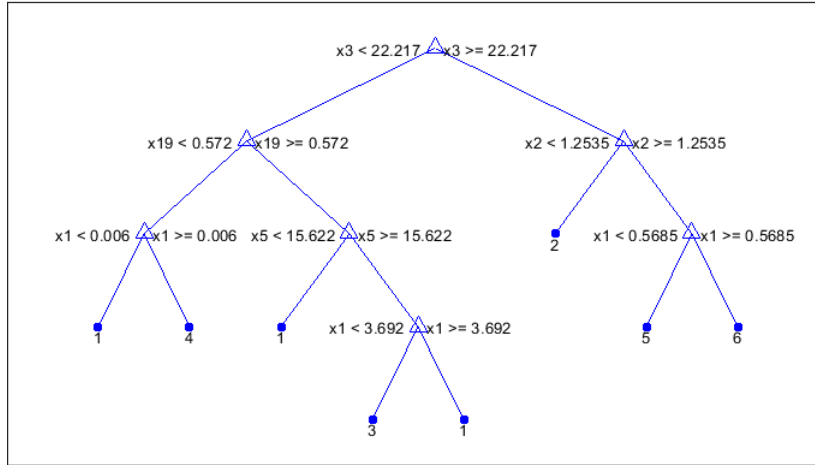


Figure 5.32: Decision tree formed from combined spectral data, with *MinParentSize* = 5.

	Rule	Data Points in Classification
R1	$x_3 < 22.22$ and $x_{19} < 0.57$ and $x_1 < 0.006 \Rightarrow C1$	1 data point in C1
R2	$x_3 < 22.22$ and $x_{19} < 0.57$ and $x_1 \geq 0.006 \Rightarrow C4$	4 data points in C4
R3	$x_3 < 22.22$ and $x_{19} \geq 0.57$ and $x_5 < 15.622 \Rightarrow C1$	3 data points in C1
R4	$x_3 < 22.22$ and $x_{19} \geq 0.57$ and $x_5 \geq 15.622$ and $x_1 < 3.692 \Rightarrow C3$	5 data points in C3
R5	$x_3 < 22.22$ and $x_{19} \geq 0.57$ and $x_5 \geq 15.622$ and $x_1 \geq 3.692 \Rightarrow C1$	1 data point in C1
R6	$x_3 \geq 22.22$ and $x_2 < 1.25 \Rightarrow C2$	5 data points in C2
R7	$x_3 \geq 22.22$ and $x_2 \geq 1.25$ and $x_1 < 0.569 \Rightarrow C5$	2 data points in C5
R8	$x_3 \geq 22.22$ and $x_2 \geq 1.25$ and $x_1 \geq 0.569 \Rightarrow C6$	3 data points in C6

Table 5.4: The rules generated by the decision tree in figure 5.32

The decision tree generated with a *MinParentSize* of 5 has an error of 0%, or an accuracy of 100%. While this specific decision tree overfits this small dataset of

only 24 examples, it provides clear and useful rules for each of the eight classes. Furthermore, it gives important information about the five most important attributes out of the 21 total attributes in defining these classes. For instance, attribute x1 (the presence of carboxylic acid in $^1\text{HNMR}$) is important for discerning between classes 5 and 6. It is also important when identifying compounds in class 1. Attribute x5 (the presence of a neighboring halogen, O, or NO_2 in $^1\text{HNMR}$) separates class 1 data points from class 4 and class 3.

It is important to note that no validation or testing set was used in this project. This is, again, due to the limited available data. Error in the decision trees was calculated from misclassifications in the training data, which included all available data in order to produce the most accurate tree possible. With a larger dataset, validation and testing sets could more easily be generated, and a clearer understanding of the accuracy of the decision tree could be established.

The constructing of decision trees in this way can provide chemists with the best places to focus on when identifying organic compounds. Attributes close to the root of the tree divide the set of compounds clearly and simply, and examining attributes farther down the tree separate compounds less distinct from each other. Allowing a decision tree to classify a compound completely independently is a possibility, though decision trees with larger error would be put to better use providing attributes at which to look more closely to analyze and identify compounds manually.

CHAPTER 6

CONCLUSION

The machine learning methods of hierarchical clustering and decision trees were applied to a dataset containing ^1H NMR and ^{13}C NMR spectral data. The results were analyzed to determine the accuracy and usefulness of each of these methods, and it was found that decision trees not only form more accurate clusters, but also provide valuable information about which attributes are most relevant in the identification process. The purpose of this project was to begin exploring the application of machine learning methods with NMR data analysis, but much more work is to be done. There were significant limitations encountered throughout this project, which are discussed below in section 6.1. The work that is necessary to reduce and overcome these limitations, as well as the work that logically follows what was learned in this project, is discussed in section 6.2.

6.1 LIMITATIONS IN THIS PROJECT

As explained in section 4.1.1, discretizing data with feature extraction is necessary with a dataset with as many attributes as there are in a (virtually) continuous signal. However, data is lost through discretization. Some forms of data may not need feature extraction, and others may have clear and distinct features such that discretization causes little to no loss in data, but full NMR signals are unfortunately too complex to have lossless feature extraction. Additionally, if the number of

attributes is relatively large, while the number of data points (or compounds in this case) is relatively small, both hierarchical clustering and decision tree forming are difficult to make accurate. It is analogous to plotting very few points in a very high dimensional space. This is the case in this project. NMR data for only 24 compounds were available, and 21 attributes were present for the data with combined ^1H NMR and ^{13}C NMR spectral data.

An additional problem with such a small dataset is lack of representation. There are over nine million organic compounds currently known by scientists [18]. Therefore, the dataset used in this project contains less than 0.0003% of all organic compounds. The more representation present in the dataset used in machine learning, the more accurate the result of that learning method will be.

6.2 FUTURE WORK

This work could easily be expanded upon simply with the use of more organic compounds. A larger dataset would increase the accuracy of both hierarchical clustering and decision trees, and introduce the types of organic compounds that were missing from this project due to the size of the dataset available. If a dataset of NMR spectra with multiple spectra for each compound were available, a decision tree or other form of supervised learning could be implemented with each class label as a different compound. This would take enormous amounts of data, and therefore storage, time, and processing power as well, but would result in a virtually automatic organic compound identifier, assuming little error occurs. While this is ideal, it is not very attainable. However, further research can be done to determine the best way to achieve semi- or close to full automation with less data. Machine learning algorithms other than hierarchical clustering and decision trees would be valuable to implement for analyzing NMR data. Unsupervised learning is especially

beneficial when completing tasks similar to hierarchical clustering, where similar compounds are identified and clustered. Supervised learning would require clear labels which, as discussed in section 4.2 can be difficult to define. As mentioned above, these labels could simply be the names of the compounds themselves, but again, large amounts of data would be required to obtain an accurate classifier of the compounds.

It is important to note that this work can never be truly complete. New organic compounds are discovered constantly, with an annual rate of approximately 4.4% as of 2015 [16]. As these new discoveries are made, databases of organic compounds must be updated, and machine learning identification algorithms must then be re-tested and possibly re-trained.

With the NMR analysis process semi- or fully automated, as this project began to pursue, the field of practical organic chemistry would be drastically altered. Semesters worth of education would not be required to focus on the tedious process of analyzing an NMR spectrum by hand, hours of work identifying unknown compounds from their spectra could be spent continuing the experiment, and intelligent chemists would not need to spend hour after hour completing the busy work of analyzing NMR spectra and identifying compounds. While this project only scratched the surface of what is possible to accomplish in the area of organic compound spectral analysis using machine learning, it has prompted a discussion of what can be improved and how those improvements may be accomplished.

REFERENCES

1. 1hnmr spectrum, September 2014. URL https://hmdb.ca/spectra/nmr_one_d/2491. xi, 11
2. Nmr spectroscopy - chemical shift and integration, 2018. URL <https://i.pinimg.com/originals/82/20/91/82209194284a145fe6f88bdb962ab2a.png>. xi, 13, 32
3. Matlab optimization toolbox, 2019. The MathWorks, Natick, MA, USA. 27
4. Scipy.cluster.hierarchy.linkage, November 2020. URL <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html>. 20, 57
5. Millipore sigma, 2020. URL <https://www.sigmaaldrich.com/united-states.html>. xii, 44, 45
6. *SpinWorks Documentation, Java Version*, February 2020. 25
7. Tarek Amr. *Introduction to Machine Learning*. Packt Publishing, 2020. 17, 18
8. Paul E. Black. Manhattan distance, February 2019. URL <https://www.nist.gov/dads/HTML/manhattanDistance.html>. 21
9. Jim Clark. Interpreting c-13 nmr spectra. *LibreTexts*, August 2020. xi, 12, 13, 32
10. Brian M. Dale, Mark A. Brown, and Richard C. Semelka. *MRI: Basic Principles and Applications*. John Wiley & Sons, Incorporated, 5 edition, 2015. 6
11. Kevin P. Gable. 13c nmr chemical shifts, January 2014. URL <https://sites.science.oregonstate.edu/~gablek/CH335/Chapter10/CarbonChemicalShift.htm>. xi, 12, 32
12. Nathan Hagen. jcamp. PyPI, July 2017. URL <https://pypi.org/project/jcamp/>. 25

13. Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2. 26
14. Jerry L. Hintze. *User's Guide IV*. NCSS Statistical System, 329 North 1000 East Kaysville, Utah 84037, 2007. xi, 22, 23
15. J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55. 26
16. Eugenio J. Llanos, Wilmer Leal, Duc H. Luu, Jürgen Jost, Peter F. Stadler, and Guillermo Restrepo. Exploration of the chemical space and its three historical regimes. *Proceedings of the National Academy of Sciences*, 116(26): 12660–12665, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1816039116. URL <https://www.pnas.org/content/116/26/12660>. 67
17. Cory Maklin. Hierarchical agglomerative clustering algorithm example in python. *Towards Data Science*, December 2018. xi, 18, 19, 20, 21
18. Richard O.C. Norman, Melvyn C. Usselman, Carl R. Noller, and Steven S. Zumdahl. Chemical compound. *Encyclopedia Britannica*, June 2020. URL <https://www.britannica.com/science/chemical-compound>. 66
19. National Institute of Advanced Industrial Science and Technology. Spectral database for organic compounds sdb, 2018. URL https://sdb.db.aist.go.jp/sdb/cgi-bin/direct_frame_top.cgi. xi, 10, 13
20. National Institute of Standards and Technology. Cholesterol. <https://webbook.nist.gov/cgi/inchi?ID=C57885>, 2018. xi, 10
21. Chaitanya Reddy Patlolla. Understanding the concept of hierarchical clustering technique. *Towards Data Science*, December 2018. 21
22. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
23. Vasudevan Ramesh, editor. *Biomolecular and Bioanalytical Techniques: Theory, Methodology and Applications*. John Wiley & Sons Ltd, 2019. xi, 6, 7, 14

24. Marc S. Reisch. Nmr instrument price hikes spook users, June 2015. URL <https://cen.acs.org/articles/93/i26/NMR-Instrument-Price-Hikes-Spook.html>. 15
25. K. J. R. Rosman and P. D. P. Taylor. Isotopic compositions of the elements. *Pure and Applied Chemistry*, 70(1):217–235, 1998. 7, 12
26. Jake Teo. Data science in python. *Read the Docs*, 2017. xi, 18, 19
27. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2. 26
28. L. G. Wade. *Organic Chemistry*. Pearson Education, Inc., 9 edition, 2017. xi, 1, 5, 6, 7, 11, 12, 15, 29, 32
29. Michael Waskom. Seaborn. pydata, 2020. URL <https://seaborn.pydata.org>. 26
30. Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a. 26

