Pacific Northwest
NATIONAL LABORATORY

PNNL-XXXXX

A Mathematical Approach to Analyzing ICP-MS and NMR Spectra

2023 NSIP Internship Final Report

October 2023

Emma S Sheppard

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-XXXXX

A Mathematical Approach to Analyzing ICP-MS and NMR Spectra

2023 NSIP Internship Final Report

October 2023

Emma S Sheppard

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Contents

1.0 Introduction

Pacific Northwest National Laboratory (PNNL) has been a forebearer of international scientific research since its establishment in 1965. PNNL is a United States Department of Energy (DOE) research site operated by the Battelle Memorial Institute, a scientific nonprofit for the benefit of national security. The lab's primary site in Richland, WA is partially located on the Hanford Nuclear Reservation, where all the plutonium for the world's first atomic bomb was produced under the Manhattan Project during World War II ("Hanford Site"). Once scientific research and development were separated from other activities at the Hanford site, operations at PNNL began. Within PNNL there are seven directorates: Business Services, Earth and Biological Sciences, Physical and Computational Sciences, Energy and Environment, Operational Systems and Technology, Computing and Information Technology, and National Security ("Lab Leadership").

The National Security Directorate (NSD) at PNNL contains four divisions: Artificial Intelligence (AI) and Data Analytics; Emerging Threats and Technologies; Nuclear, Chemistry, and Biosciences; and Physical Detection Systems and Deployment. Every summer, NSD accepts a limited number of interns into the National Security Internship Program (NSIP) for hiring into one of these four divisions. I was accepted into the 2023 cohort of NSIP into AI and Data Analytics as part of the Machine Learning (ML) and Mathematical Modeling Team under my mentor, Dr. Margaret (Maggie) Lund. The project on which Maggie needed my help involved utilizing mathematical analysis techniques on chemical spectroscopic and spectrometric data, which allowed us to synchronize two of the Department of Energy's key capabilities: Chemical and Materials Science, and Computational and Mathematical Sciences ("DOE Capabilities").

I am in my last semester at Gonzaga University, obtaining a BS in applied mathematics with a minor in chemistry on a pre-veterinary professional track. After a previous summer internship at a veterinary clinic that served clients with large, small, and exotic animals, I loved the job but didn't find many passionate veterinarians in the industry. The clients were difficult, the hours were long, and the vets suggested that I go into research due to my love of school. Thus, this summer I wanted to look for a more research-oriented internship experience. I found that my particular skillset didn't fit well with many internships – there was a need for students in chemical and biological laboratories, or for computational mathematics students with fluency in multiple coding languages and knowledge in different areas of applied math. While I had a good level of experience in both areas, I was not an expert in either. This made it difficult to be considered for solely mathematics-based or chemistry-based research internships, despite my work ethic and GPA. What I really wanted was to utilize my diverse skillset and learn more about the possible applications of an education in mathematics and chemistry, so I was straight-forward about this during my interview with Maggie. She then informed me that she already had an intern, Emily, and hadn't considered a second one until my recruiter had informed her of my interests. In essence, she only interviewed me because of my unique educational experience in mathematics and chemistry. This made me feel that the National Security Internship at PNNL was the perfect position for me to learn and flourish. With Maggie's skillset in data analytics, Emily's in machine learning, and mine in chemistry, we had the perfect team of mathematicians to conquer the challenges of our project.

The goal of the project is forensic analysis of food commodities to determine their geographic origin and authentication status. Geographic sourcing determines a food's region of origin using its chemical composition to identify biological signatures. Certain biological signatures are unique to geography, allowing analysts to determine where a food was originally harvested or

produced. Moreover, ensuring the authenticity of a food confirms that it hasn't been adulterated throughout the curation process, benefitting economic security. Adulteration can happen via dilution, additives, removals, or substitution of the natural materials within a food without the importer or manufacturer declaring that such changes have been made (Mumtaz). This may be done to mask the food's geographic origin, or to save on costs of production. Countries allowing imports of adulterated foods are considered victims of economically motivated adulteration, or food fraud (Center for Food Safety and Applied Nutrition). For example, the addition of sugar syrups, such as high fructose corn syrup, to pure maple syrup without the company declaring that there are additives constitutes food fraud. The American public is therefore under false pretenses that they're purchasing pure maple syrup. This leads to more profit for selling less of the pure product, and thus compromise of economic security within the US food industry. To prevent food fraud, a combination of chemical analysis techniques to analyze food products, called foodomics, is utilized (Valdes). This field of analytical chemistry is important to government organizations, consumers, researchers, and private food industry worldwide. An article published in American Chemical Society best defines this method of food analysis: "Foodomics' was defined to integrate the use of advanced omics technologies, such as transcriptomics, proteomics, [genomics,] and metabolomics, together with biostatistics, chemometrics, and bioinformatics, to allow the evaluation of complex biological systems, as well as the mechanisms of bioactive food compounds that may affect them," (Valdes). These various omics techniques study the RNA transcripts, proteins, genome, and metabolites of a food commodity, respectively.

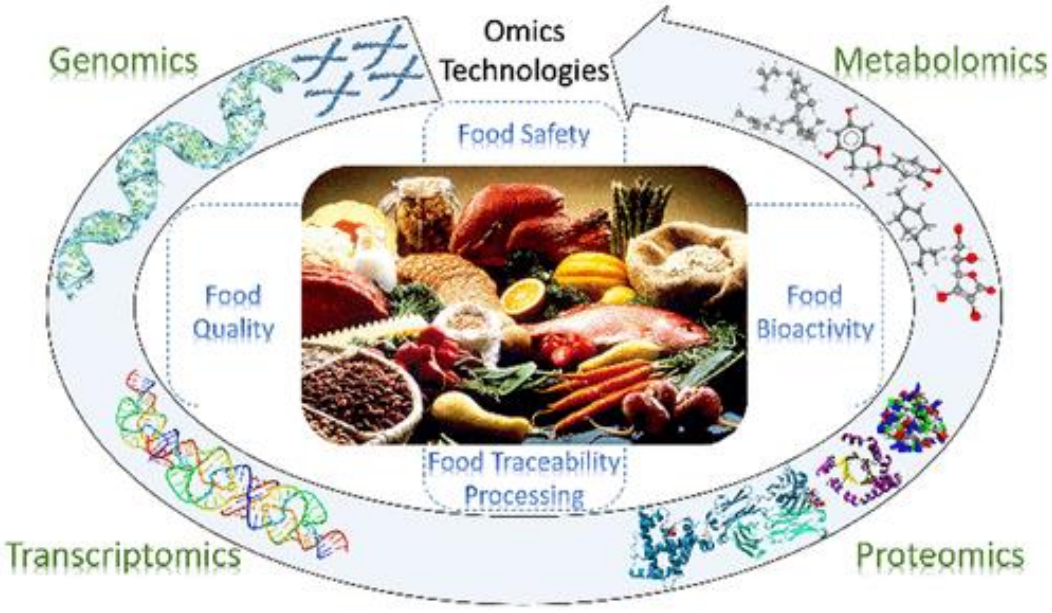

Figure 1. The purposes and names of various omics technologies comprising foodomics. From *Anal. Chem.* 2022, 94, 1, 366–381 by A. Valdés, et al., 2021. https://doi.org/10.1021/acs.analchem.1c04678

Utilizing the techniques and theories within foodomics, we were tasked to analyze spectral data of various food samples produced by inductively coupled plasma mass spectrometry (ICP-MS) and nuclear magnetic resonance (NMR) spectroscopy. When I came onto the project in May, Maggie had been working on a program in RStudio that analyzes ICP-MS data and determines geographic origin of a food commodity using machine learning algorithms. Thus, my tasks were

to refresh my knowledge on ICP-MS theory, understand the spectral dataset, and learn about the data science and machine learning techniques used to analyze the spectral data. Though I was able to fit this all into one sentence, learning really took up the majority of my first ten weeks on the project. I had only worked with ICP-MS spectra in classes and had never utilized my skills in mathematics with datasets or machine learning. Once the ICP-MS program was finalized and ready to submit to the sponsor, we moved along to exploratory research of the NMR dataset. I had more experience with NMR, but the goal to create a computer program that analyzes NMR spectra for geographic origin and adulteration analysis was daunting and brand new to me.

2.0 Inductively Coupled Plasma Mass Spectrometry

Spectrometry is a chemical analysis technique that produces a spectrum with quantitative measurements, as opposed to spectroscopy which requires analysis on the resulting spectra to come to a quantitative conclusion. Spectrometric instruments measure interactions between light and matter, as well as the “reactions and measurements of radiation intensity and wavelength,” (“Understanding Spectroscopy and Spectrometry”). Mass spectrometry produces an approximate mass measurement as the quantitative result, where inductively coupled plasma mass spectrometry (ICP-MS) uses plasma to atomize a chemical or biological sample, measures elemental composition at trace levels, and results in a mass-to-charge ratio of each element in the sample. There are two analytic methods utilizing inductively coupled plasma: inductively coupled plasma atomic emission spectroscopy (ICP-AES) and ICP-MS. ICP-AES has a higher detection limit than ICP-MS, requiring more of an element to be present in a sample for it to be detected (Wilchefski). Thus, it isn’t as useful for determining trace levels of elements in a sample, as it only measures in parts per million. On the other hand, ICP-MS has a much lower detection limit, a higher sensitivity, and can measure elemental composition within a wide range: 1000s of parts per million to 1 part per trillion (Wilchefski). Wilschefski and Baxter explain the important advantage of using ICP-MS: “From a laboratory perspective, perhaps the most significant advantage of ICP-MS is its multi-element capability, which allows multiple elements to be measured simultaneously in a single analysis.” Other element detection techniques, such as flame absorption and flame emission spectroscopies, can only measure single elements in each analysis. This makes ICP-MS a favorable technique for its efficiency and production of useful analytical information.

2.1 How the ICP-MS Instrument Works

The ICP-MS instrument uses an argon inductively coupled plasma source that atomizes a sample via ionization, where a mass spectrometer then distinguishes ions by their mass-to-charge (m/z) ratio (“Elemental Analysis Core”). Ionization occurs when ions are formed by the reduction or oxidation of an atom or molecule, thus breaking bonds with other atoms (“Ionization”). The chemical or biological sample of interest must be liquid to go through the inductively coupled plasma, which is an energy source supplied by electric currents (“Inductively Coupled Plasma”). Thus, any solid samples must be diluted or thermally digested, the breakdown of large molecules by heat (Ngo). After moving through the plasma, the detector counts the number of selected ions per second, allowing the instrument to determine concentration of each element. This detector measures in “counts per second,” which, intuitively, counts the number of ions hitting the detector every second. The conversion to units of concentration requires the use of an external calibration sample, where a sample with known elemental concentration values is measured and recorded in the instrument before any samples of interest. The resulting spectra has the mass or m/z ratio as the x-axis, indicating the element, and intensity, corresponding to concentration and counts-per-second, on the y-axis.

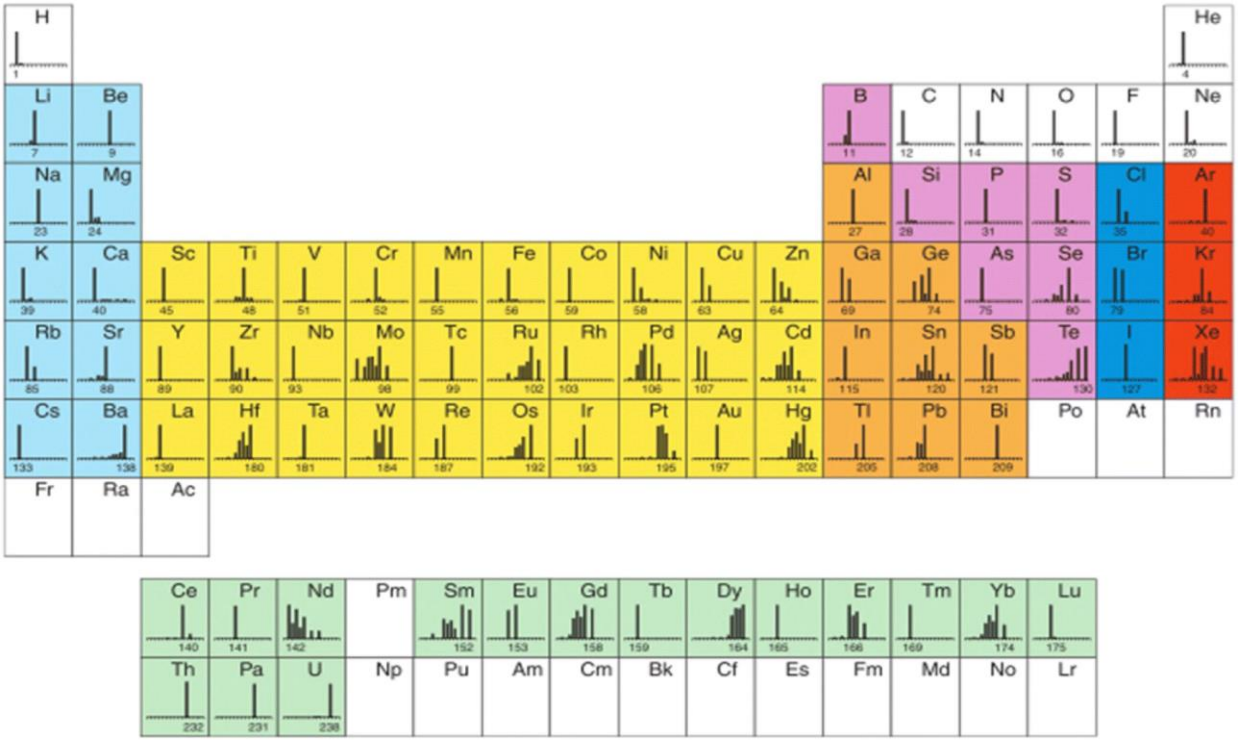

Figure 2. A guideline to ICP-MS spectra, where each element measured by the instrument has a standard number of peaks and m/z value. From *Guideline of inductively coupled plasma mass spectrometry* by M. F. Al-Hakkani, 2019. https://doi.org/10.1007/s42452-019-0825-5

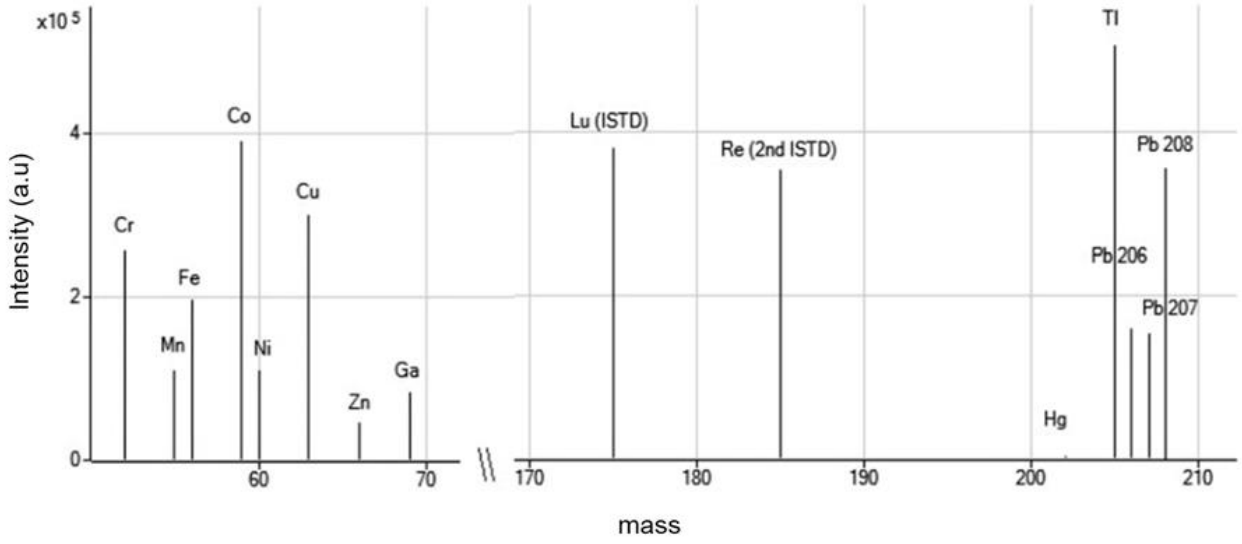

Figure 3. Example ICP-MS spectrum for detecting trace mercury (Hg) concentrations in samples of wood. From Inductively Coupled Plasma Mass Spectrometry Applications by A. Karttunen, 2022. https://wiki.aalto.fi/display/SSC/Inductively+Coupled+Plasma+Mass+Spectrometry

2.2 When to Use ICP-MS

The fundamental reason for a scientist to use ICP-MS as their analytical method is when the elemental composition of a biological liquid or dissolvable-solid is of interest. It can measure nearly every element in the periodic table, including many radioactively produced and non-natural elements. It is the only analytical technique with "such broad element coverage, low detection limits, and wide measurement range," ("An Introduction to the Fundamentals of ICP-MS"). While it does have significant strengths, there are drawbacks to ICP-MS. It is an extremely expensive method, considering the cost of the instrument, ranging from $50,000 to $500,000, as well as the cost of calibration samples (Wilbur). It also cannot measure every single element: hydrogen and helium are below its mass range; argon, nitrogen, and oxygen are present at too high a level due to the plasma and air; and fluorine and neon cannot be ionized in argon plasma," ("An Introduction to the Fundamentals of ICP-MS"). As can be seen in the periodic table of standards, there are also many heavy elements that can't be detected via ICP-MS. The most important drawback to an analyst is that ICP-MS cannot measure presence of chemical compounds and complex molecules or determine the bonding activity between elements. This analytical method should only be used when concerned with elemental composition, not with functional group or molecular composition. The table below best maps the benefits and drawbacks to ICP-MS in comparison to other elemental detection techniques.

Technique	Advantages	Disadvantages
ICP-MS	Multi-element technique Large analytical range Low detection limit High sample throughput Low sample volume Simple sample preparation High-resolution and tandem mass spectrometry (triple-quadrupole) instruments offer a very high level of interference control	Equipment cost Operating cost (argon) Multiple high purity gases required High level of staff expertise Interferences need to be controlled Laboratory set up costs (air-conditioning, HEPA filters, pipe work, dust reduction measures)
ICP-AES (also known as ICP-OES)	Multi-element technique Large analytical range High sample throughput Low sample volume Simple sample preparation	High detection limit Equipment cost High level of staff expertise Laboratory set up costs
Flame Atomic Emission	Equipment cost Low level of staff expertise Simple sample preparation Low laboratory set up cost	Single element technique Limited analytical range High detection limit Higher sample volume Flammable gases
Flame Atomic Absorption	Equipment cost Low level of staff expertise Simple sample preparation Low laboratory set up cost Reasonably high sample throughput Few interferences	Single element technique Limited analytical range High detection limit Higher sample volume Flammable gases
Graphite Furnace Atomic Absorption	Low detection limit Equipment cost Few interferences Low sample volume Low laboratory set up cost Simple sample preparation (in most cases)	Single element technique Limited analytical range Low sample throughput Some elements require acid digestion prior to analysis
Atomic Absorption (cold vapour/ hydride generation)	Low detection limit Equipment cost Few interferences Low laboratory set up cost	Suitable for limited elements Limited analytical range Low sample throughput High sample volume Complex acid digestions required for biological samples

Figure 4. Comparing the advantages and disadvantages of ICP-MS and alternative techniques for elemental detection in biological samples. From *Inductively Coupled Plasma Mass Spectrometry: Introduction to Analytical Aspects* by S. C. Wilschefski and M. R. Baxter, 2019. https://doi.org/10.33176%2FAACB-19-00024

2.3 ICP-MS in Analytical Chemistry and Research

2.3.1 Toxic, Therapeutic, Nutritional, and Metabolic Elements

ICP-MS is extremely useful in scientific research, especially in medical and national security applications. This technique can be used to detect levels of elements in the blood or urine that may be toxic to humans, so if a person's symptoms coincide with some sort of elemental toxicity, a doctor can confirm what element is present and treat accordingly (Wilchefski). In addition to determining concentration of harmful elements, it can measure concentration of therapeutic elements to ensure the effectiveness and responsiveness to a new medication or treatment. ICP-MS can also measure concentration of nutritional elements from food, and the body's naturally produced metabolic elements to confirm that metabolic processes are operating as expected. The figure below outlines relevant elements that can be detected by ICP-MS, as well as their classification of toxic, therapeutic, nutritional, or metabolic and the range of acceptable levels in the blood or urine.

Element	Clinical application	Approximate concentration range*
Aluminium	Toxic	0.1–10 µmol/L
Antimony	Toxic	1–30 nmol/L
Arsenic	Toxic	0.01–80 µmol/L
Barium	Toxic	7–700 nmol/L
Beryllium	Toxic	1–150 nmol/L
Bismuth	Toxic, Therapeutic	1–200 nmol/L
Bromide	Toxic, Therapeutic	0.1–40 mmol/L
Cadmium	Toxic	1–100 nmol/L
Chloride(sweat)	Metabolic	10–200 mmol/L
Chromium	Toxic	1–5000 nmol/L
Cobalt	Toxic	1–5000 nmol/L
Copper	Nutritional, Metabolic	1–50 µmol/L
Gold	Therapeutic	1–10 µmol/L
Iodine	Toxic, Nutritional	0.008–200 µmol/L
Lead	Toxic	0.01–10 µmol/L (~0.2–200 µg/dL)
Manganese	Nutritional	1–400 nmol/L
Mercury	Toxic	1–1000 nmol/L
Molybdenum	Toxic	1–20 nmol/L
Nickel	Toxic	1–200 nmol/L
Selenium	Toxic, Nutritional	0.1–10 µmol/L
Silver	Skin Pigmentation	1–500 nmol/L
Thallium	Toxic	1–50 nmol/L
Tin	Toxic	0.2–500 nmol/L
Vanadium	Toxic	1–1000 nmol/L
Zinc	Nutritional	1–40 µmol/L

*These ranges are meant as a guide only; toxic levels may occasionally exceed the upper limit of the quoted range.

Figure 5. Element concentration ranges that are often detected in biological samples of blood or urine. From *Inductively Coupled Plasma Mass Spectrometry: Introduction to Analytical Aspects* by S. C. Wilschefski and M. R. Baxter, 2019. https://doi.org/10.33176%2FAACB-19-00024

2.3.2 Geographical Sourcing of Food

In addition to applications in the medical field, our national security project utilizes ICP-MS spectral data to geographically source food commodities for the benefit of national economic security. Origin fraud is the most prominent motivation behind geographical sourcing: “Origin fraud, which occurs when plant food is misrepresented in its geographical origin, is a form of mislabeling that has a significant impact on the economy and is documented in many countries. Agricultural and food products are subject to strict control on their origin to ensure quality during import and export,” (Nguyen). The current standard for food traceability is the protected geographical indication (PGI) system, which relates a food product’s ingredients to the most significant location of production (“Geographical Indications and Quality Schemes Explained”). For example, a wine labeled as produced in the Provence region of France must be majority (85%) composed of grapes grown in Provence. While the PGI system sources origin using reported origin of ingredients, there is a push to approach food traceability from an analytical perspective using ICP-MS data (Nguyen). This way, a food’s origin can be confirmed by substantial chemical analysis and standards of elemental concentrations corresponding to a particular region. This would be particularly effective in plant-based foods: “Generally, trace elements represent the geographical tracer in a specific soil condition, and are absorbed via the roots and transferred to various parts of the plant. The distribution of trace elements reflects the elemental signature of the soil origin,” (Nguyen). Origin fraud is an international issue, so there must be an analytical approach to food traceability, as opposed to a standard agreement of claiming ingredient origins.

The push for ICP-MS in food traceability is well-researched and supported in literature, but with the limited data produced by the instrument, it may make sense to utilize additional analytical methods. Nuclear magnetic resonance spectroscopy provides many more data points for a single sample and includes much more information regarding chemical composition.

3.0 Nuclear Magnetic Resonance Spectroscopy

Spectroscopy is defined as the “study of the absorption and emission of light and other radiation by matter,” and involves the splitting of light in a similar way that a prism splits light into a rainbow (“Understanding Spectrometry and Spectroscopy”). The resulting spectrum is determined by measuring changes in the intensity or frequency of this radiative energy. In contrast to spectrometry, there are no analytical results or measurements without a hands-on or automated analysis of the spectrum. It is an inherently theoretical approach to studying the sample, as opposed to the practical measurement that spectrometry provides. NMR spectroscopy is used to determine the molecular identity and structure of a chemical sample and provides a chemist the ability to “characterize molecular structures, monitor the composition of mixtures, study molecular dynamics and interactions, and quantify known and unknown components” of a compound (“How NMR Works: Spectroscopy”).

3.1 How the NMR Instrument Works

The NMR instrument generates a strong magnetic field causing atoms in the nucleus of each desired element to spin in a manner that depends on the element’s environment. Electromagnetic radiation with a certain frequency is applied to generate this field, causing the nuclei to align their spins with it (“Physical Chemistry”). The instrument then detects the absorption signals within elemental nuclei containing an odd number of protons and/or neutrons because odd numbered nuclei “exhibit a built-in magnetic moment and angular momentum” that give the nuclei their spin (“How NMR Works: Spectroscopy”). Optional odd numbered nuclei include proton (1H), carbon (13C), nitrogen (15N), and phosphorus (31P) NMR, with 1H and 13C being the most common.

Proton (1H) NMR determines the different kinds of protons in the molecule by their absorption of the electromagnetic radiation, resulting in spectra that indicates protons in different environments in the compound. Bruker best describes the way an NMR instrument determines the unique environments of each proton in a compound: “Shift in the usual response frequency for a given isotope provide information about their immediate environment, including influences from nearby electrons and magnetic nuclei, making it possible to infer molecular identity, geometry, and more,” (“How NMR Works: Spectroscopy”). The shift in response frequency is read by the instrument as a signal produced by the nucleus returning to its “resting alignment,” which is unique to each nucleus.

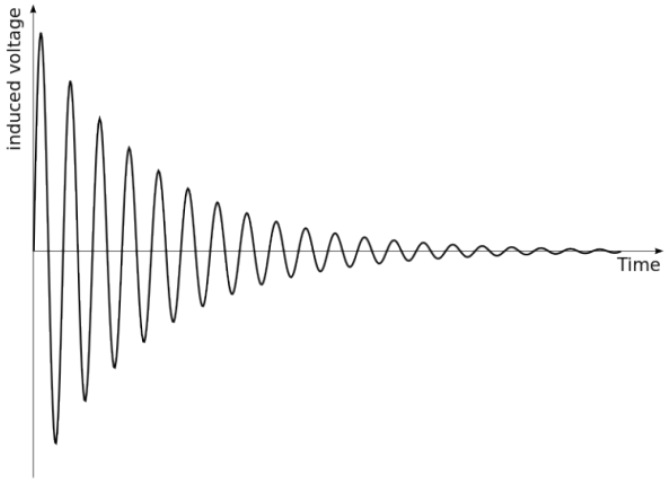

Figure 6. Raw signal from an NMR scan of ethanol. From *How NMR Works: Spectroscopy* by Bruker. www.bruker.com/de/resources/library/application-notes-mr/nmr-101.html.

The signal is then converted into an NMR spectrum that shows the frequencies that the nuclei responded at. The conversion occurs via a mathematical process called a Fourier transform. The resulting spectrum plots intensity, corresponding to concentration, on the y-axis, and shift, corresponding to functional group and each unique nuclei, on the x-axis.

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \qquad f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega$$

Figure 7. Equations for the Fourier Transform and the inverse Fourier Transform.

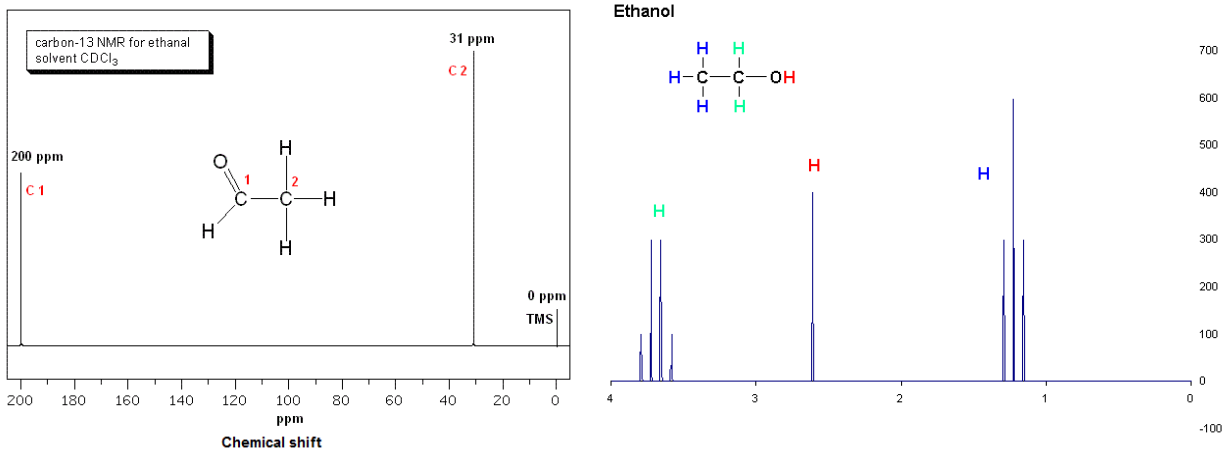

Figure 8. Carbon (13C) NMR spectrum for ethanol on the left, and proton (1H) NMR spectrum for the same ethanol sample on the right. Nuclei of interest (C1, C2, and hydrogen groups in blue, green, and red) are shown. From *How NMR Works: Spectroscopy* by Bruker. www.bruker.com/de/resources/library/application-notes-mr/nmr-101.html.

3.2 When to Use NMR

NMR is most useful when an unknown compound is of interest, especially in the case of a chemist monitoring a chemical reaction. Scanning a sample before, during, and after a reaction can help chemists determine reactants, intermediates, and products, especially if some compounds are known beforehand. The NMR instrument itself scans relatively quickly, though the analysis can take some time afterwards as the spectra contains a lot of data. This is a significant drawback of NMR – while extraordinary amounts of data contain a lot of information about a sample, it can be very complicated to analyze. In one NMR dataset, there can be upwards of 20,000 shift values ranging from 10 to 0 ppm after binning, which is a technique to shrink the size of a dataset without losing too much information. Often times, samples of interest are complicated, containing many functional groups and unique proton environments. This leads to peak overlap that can be difficult to distinguish from instrument noise, especially if the sample's structure is completely unknown. If less chemical information is needed to conduct an experiment, it makes more sense to use another technique with a smaller resulting dataset, such as ICP-MS.

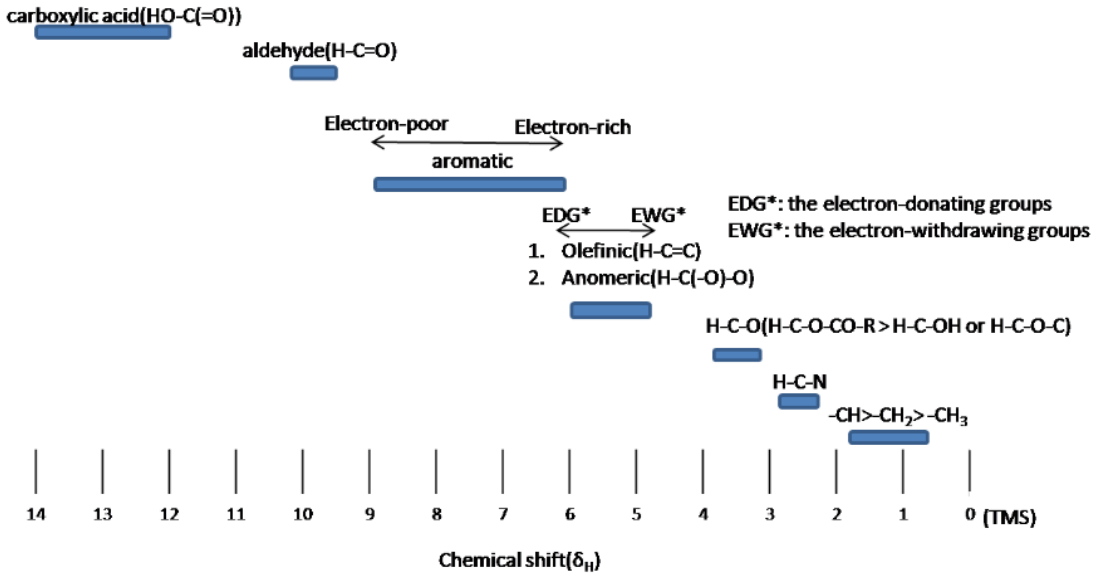

Figure 9. Proton (1H) NMR chemical shift functional group ranges for an organic compound. From *NMR – Interpretation* by You Jin Seo. https://chem.libretexts.org/@go/page/1812

3.3 NMR in Analytical Chemistry and Research

3.3.1 Nuclear Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a noninvasive medical imaging technique based on the physics of NMR that generates two-dimensional images instead of spectral frequencies (Radlib). These images are then stacked together via reconstruction to generate three-

dimensional images of objects. MRIs were originally called nuclear MRIs, but the “nuclear” portion was omitted due to the publicity around nuclear bombs during World War II when the MRI was gaining popularity. Though the “nuclear” in this sense is referencing the atomic spin within the nucleus, it was often met with hesitation as people assumed it indicated a dangerous nuclear reaction (“NMR v MRI”). The revolutionary thing about MRI machines is their ability to provide 3D diagnostic imaging of the body’s inner water-based tissues. Organs, connective tissue, and cartilage are all too low-density for an x-ray to pick up on, but an MRI can. This allows for use in “areas of cardiovascular, neurological, musculoskeletal, and oncological imaging,” (“Physical Chemistry”). In addition to the many abilities of MRI, it doesn’t use ionized radiation to image, like 3D computed tomography or x-ray. This makes the machine safe for patient exposure.

3.3.2 Plant Metabolomics

NMR is not only useful in the medical field, but in direct applications of analytical chemistry and research. Metabolomics identifies and quantifies the metabolites of an organism, which are “small molecules that participate in general metabolic reactions and that are required for maintenance, growth, and normal function of a cell,” (Schripsema). Plant metabolomics is a large area of research, as studying metabolites in plants can be easier than studying their proteins and lipids, via proteomics and lipidomics, because they have less of these macromolecules than animal products do. NMR is useful in metabolic profiling and metabolic fingerprinting, with definitions as follows:

> Metabolic profiling – “quantitative analysis of sets of metabolites in a selected biochemical pathway or a specific class of compounds. This includes target analysis, the analysis of a very limited number of metabolites, e.g. single analytes as precursors or products of biochemical reactions” (Schripsema).
>
> Metabolic fingerprinting – “unbiased, global screening approach to classify samples based on metabolite patterns or ‘fingerprints’ that change in response to disease, environmental or genetic perturbations with the ultimate goal to identify discriminating metabolites” (Schripsema).

NMR spectra can classify three quantities of metabolites in a plant’s metabolome: chemical nature (alkanes, carboxylic acids, amines, esters, peptides), solubility (water-soluble sugars to oil-soluble lipids), and concentration (high concentration to trace levels). The concentration can range from that of a sugar, which is about 50% composed of metabolites, down to picomolar.

Though the paper *Application of NMR in Plant Metabolomics: Techniques, Problems and Prospects* used NMR in their studies of plant metabolomics, NMR and ICP-MS are complementary and ideal in metabolomics studies. The figure below compares the different features of NMR and MS for use in metabolomics.

	Nuclear magnetic resonance (NMR)	Mass spectrometry (MS)
Sensitivity	Low	High
Reproducibility	Very high	Average
Number of detectable metabolites	30-100	300-1000+ (depending on whether GC-MS or LC-MS is used)
Targeted analysis	Not optimal for targeted analysis	Better for targeted analysis than NMR
Sample preparation	Minimal sample preparation required	More complex sample preparation required
Tissue extraction	Not required – tissues can be analysed directly	Requires tissue extraction
Sample analysis time	Fast – the entire sample can be analysed in one measurement	Longer than NMR – requires different chromatography techniques depending on the metabolites analysed
Instrument Cost	More expensive and occupies more space than MS	Cheaper and occupies less space than NMR
Sample Cost	Low cost per sample	High cost per sample

Figure 10. A comparison table between NMR spectroscopy and ICP-MS for use in metabolomics studies. From *Comparison of NMR and MS* by the European Molecular Biology Laboratory, 2020. https://doi.org/10.6019/TOL.MBS.2014.00001.1

3.3.3 Food Science and Foodomics

The last and most important example of NMR in research is within foodomics, where it is used to determine the structure of organic compounds in food. “NMR spectroscopy is used to determine structure of proteins, amino acid profile, carotenoids, organic acids, lipid fractions,” and water content, where it has recently been used to analyze “vegetable oils, fish oils, fish and meat, milk, cheese, wheat, fruit juices, coffee, green tea, foods such as wine and beer” (Parlak). As an example, food scientists can determine the age and quality of meat via its intramuscular fat and water contents. For some meats, such as a ribeye steak, a certain amount of fat is necessary to maintain the steak’s quality and taste to the consumer’s expectations. Thus, it is important to the consumer that a food scientist has rigorously checked quality prior to purchasing and grilling their ribeye. NMR is also used to monitor the effects of cooking; once a steak is cooked, an endothermic reaction has occurred, denaturing proteins, and changing its taste, texture, and chemical composition. New or different peaks correspond to new or changed functional groups within the meat’s chemistry after the input of high-energy heat. It is important to understand what people are consuming, and whether a piece of meat is safe to eat before and/or after cooking based on its chemical composition.

Following the use of ICP-MS data to geographically source and authenticate food, our project aspires to utilize the mass amounts of data in 1H NMR spectra to better accomplish these goals. With more information in NMR spectra comes the need for more complicated mathematical techniques to analyze and extract useful chemical information. Then, we can combat food fraud, adulteration, and masking of geographical source with higher accuracy.

4.0 Using Data Science Techniques on Spectral Data

Seeing as my internship was centered around applied mathematics, it was important for me to apply my skills in mathematics to spectral data via techniques in data science. I had never worked on datasets prior to this, so approaching chemical datasets felt like a natural application of my skillset. From my perspective, there are three essential processes to working with datasets: preprocessing, analysis, and interpreting results. Preprocessing, also known as pretreatment, entails the cleaning and transforming of data to improve its quality and make it more suitable for analysis ("Data Preprocessing in Data Mining"). This can be done by removing noise, reducing dimensions, and binning data to make it smaller and more manageable. The pertinent part of preprocessing is making data easier to analyze without omitting important information. In spectral data, this would mean binning such that there are less points or noise without removing any chemically relevant peaks. Subsequently, data analysis is the heart of data science, where the broad goal is to find patterns and useful information within a dataset to draw conclusions from real experimental data. In NMR, this means analyzing each peak and corresponding shift value to determine what functional groups are in a sample. Finally, visualizing and interpreting results is arguably the most important part. Why are we analyzing this data in the first place? Once patterns are found, how do they affect our goal? If our data is uninformative, how should we change our experimental process to improve the outcome? The goal here is to sum up the moral of the story. We want to make results interpretable for everyone without too much technical information. In NMR, the point may be to determine a product from a novel reaction, while in ICP-MS, the goal might be to detect toxic elements in a blood sample to determine a treatment plan. The final interpretation of the results is why data science exists.

4.1 Implementation of Geographic Sourcing with ICP-MS Data

The program our team developed to implement geographic sourcing of food commodities using ICP-MS data is called Predictive Algorithms for Commodity Traceability (PACT). PACT starts with a preprocessing method called Pareto Scaling, which is frequently used in metabolomics analysis to normalize the data (van den Berg). Pareto Scaling scales the data by dividing each variable, the element detected by ICP-MS, by the square root of the standard deviation. This reduces the relative importance of any large data points, and keeps the data structure majority intact, staying close to the original measurement. The one downfall of Pareto Scaling is its sensitivity to changes in the data. Scaling data is an important technique to balance the influence of each variable, and makes it easier for a ML model to analyze the data.

Then, another preprocessing method called Principal Component Analysis (PCA) is applied. PCA is used to reduce the dimensionality of large datasets to "increase interpretability" while "minimizing information loss" (Jolliffe). It finds new variables, called principal components, that are linear combinations of variables in the original dataset by inputting each variable into an eigenvalue/eigenvector problem, then solving using matrix algebra. This preserves important statistical information while making a dataset more manageable. PCA is an extremely adaptable technique that defines variables depending upon each individual dataset. For example, ICP-MS has been used to determine atmospheric trace element concentrations in Norwegian moss samples due to atmospheric deposition (Berg). Atmospheric deposition is the process "whereby precipitation (rain, snow, fog), particles, aerosols, and gases move from the atmosphere to the earth's surface," ("Atmospheric Deposition"). PCA was used on this Norwegian moss ICP-MS dataset consisting of 33 elements and 495 samples, resulting in the following principal components listed in Table 1.

Principal Component	Elements in Linear Combination(s)
Long-range atmospheric transported elements	Bi, Pb, Sb, Mo, Cd, V, As, Zn, Tl, Hg, Ga
Windblown mineral particles	Y, La, Al, Li, U, Th, Ga, Fe, V, Cr
Local emission sources	Ni, Cu, Co, and As Zn, Cd and Hg Fe, Cr, and Al
Transport from the marine environment	Mg, B, Na, Sr, Ca
Contribution from higher plants	Cs, Rb, Ba, Mn

Table 1. Principal components and their corresponding linear combinations of elements from an ICP-MS dataset of 33 elements and 495 samples. From *Atmospheric Trace Element Deposition: Principal Component Analysis of ICP-MS Data from Moss Samples* by Berg, T., et al., 2000. www.sciencedirect.com/science/article/pii/026974919591049Q

PCA thus reduced the dataset from 33 variables (elements) to 7 without losing predictive power for each sample, making it an extremely useful preprocessing technique on ICP-MS data.

Finally, after scaling and performing PCA on the food commodity dataset, the data is analyzed using a machine learning algorithm called Support Vector Machines (SVM). SVM is useful for categorical prediction, where it trains on ICP-MS data with known geographic source and finds unifying features between each sample belonging to a known region. Once the model is trained, it is tested with samples from unknown regions and compares to two possible regions, resulting in a pairwise similarity metric. The model can determine whether a food sample is more similar to samples from region A or region B, where the resulting statistical values add up to 1. The reason it's a similarity metric and not a probability is that there's a possibility that the sample is from neither region A nor B. There are plans to move forward with the program to implement a worldwide model that can determine whether a food sample is most likely from region A, B, C, D, E, etc. This model would compare the unknown sample to all possible regions in the model's training data, as opposed to just two regions at a time.

Overall, the program was a success, and the analysis of ICP-MS data has been effectively completed. The next step is to move onto the more complicated dataset, which contains a lot more information regarding geographic origin and adulteration but is much harder to analyze.

4.2 Innovation of a Theoretical Dual-Angle Approach to NMR Analysis

Applying similar ideas to NMR is more challenging because the spectra are extremely complicated rather than a distinct set of elements being measured. Researchers on the project (data scientists, chemists, and machine learnists) have tried many different approaches to working with the data. When Maggie presented me and Emily with a subset of the NMR dataset to begin exploratory analysis near the end of the summer, we felt overwhelmed. When I plotted the data, I knew I was looking at NMR spectra, but it was much denser than anything I'd worked with during organic chemistry courses. I began by doing research on significant geographic

markers that can be found in NMR, along with possible adulterant peaks. I was able to highlight regions where these peaks may occur, but couldn't analyze the spectra by hand due to its complexity. This sparked a brainstorming session among the three of us, where we tried to figure out a way to tackle the dataset using machine learning and chemistry.

Often times, data scientists will input data into a machine learning model to see what it spits out without understanding the features on which the model uses to distinguish between datasets. The best example of this is well-known in the world of computer science – a machine learning model was built to distinguish between images of huskies and wolves (Besse). A lot of the time, the model was successful, but occasionally there'd be a misclassification. It was then discovered, via machine learning explainability tools, that the model was not looking for features on the canines but for snow in the background. A lot of the training images turned out to have wolves in snow, and the model never saw huskies in the snow as they were often in grass or indoors. Thus, when an image of a husky in the snow was input, the ML model classified it as a wolf (Besse). This blind input of data into ML can be a helpful method if an analyst knows what output they're looking for, but can be difficult to decipher if not. With food NMR spectra, peaks and regions that a model picks out are meaningless without a chemistry-backed explanation. Thus, the idea for a dual-angle approach to NMR analysis was born.

The idea is simple: data scientists can analyze NMR spectra with machine learning models, but chemists must be involved in the preprocessing and interpretation of results. In addition to chemists attempting to reveal what the models are looking at in the data, we also want to implement explainability tools. An explainability tool is an algorithm that is applied to a machine learning model that reveals what features on which it is focusing. This allows for a deeper understanding of the output so we can adjust training data or change the focus features.

With the NMR spectra, we would start by consulting a chemist on preprocessing methods. It is important to reduce the size of the data and analyze at smaller shift widths without omitting peaks. Often times, preprocessing methods will evenly space out sections of data for a model to analyze, i.e. analyze each 0.5 ppm section individually. A chemist will understand that this doesn't make sense – a peak could be centered at 0.5 ppm, thus omitting important chemical data by splitting it. Then, after preprocessing such that a couple peaks are analyzed at a time by a ML model, we want to implement explainability tools to tell us what features the model picks out as being associated with geographic origin, for example. Then, when a chemist looks at the model's output, they can determine whether the features were actually indicative of region of origin. This allows for chemistry-backed conclusions and, ideally, a faster development of a method that works to determine geographic origin.

5.0 Conclusion

I was able to contribute a significant amount to this project, especially since Maggie offered every opportunity for me to participate in official reports, presentations, and code for the sponsor. I am a co-author on the PACT code as well as its user guide. The code was the most significant portion of my summer due to the technical skills I gained, and writing technical reports became more of a streamlined process for me. Additionally, going through the Information Release and Invention Disclosure processes with the program and its user guide were tedious and allowed me to see the more bureaucratic side of national research. Along with the project deliverables for the sponsor, I held a seminar on NMR spectroscopy for my team, as I was most well-versed in chemistry. This NMR seminar allowed me, Maggie, and Emily to be on the same page to move forward with analysis from a chemistry-informed position. The NMR seminar was a huge turning point and sparked our idea for the dual-angle approach to analysis of the NMR dataset. This idea was novel because there was an opening for collaboration between chemists in the lab and data scientists working with the numbers, and we took advantage of that opportunity. I am proud to have contributed a novel idea to a project in my first summer of government research, especially as an intern. I also presented on this project to fellow interns, researchers from all my projects, project managers, and administrative employees at the lab. “The Intersection of Math and Chemical Sciences: *Geographical sourcing and authentication of food commodities*” was the first research presentation I’d ever given, and I am extremely proud of how it turned out. Learning how to tell a cohesive story and engage the crowd was an extremely valuable learning experience for my future in national security. I learned how important it is to not only perform research and please the sponsors, but also be able to share it with the public.

Now that I’ve been shown ways that a diverse skillset can really fit into the world, I’ve decided to continue this path and get my PhD in applied mathematics. I hope to find an area of research that integrates my chemistry skills as well, but I know that even if my dissertation is heavy in math, there are always applications that involve working with chemical data and materials science. My current path is to maintain work as an undergraduate intern through next July, then transition to a PhD intern once I’ve been accepted into a program. Then, I will return to PNNL every summer throughout graduate school to continue gaining experience and working with peers and mentors of whom I’ve become so fond. Finally, I’d like to be hired full-time as a PhD mathematician at a national lab or alternative government research facility.

I know I can extend this fundamental skillset in analytical chemistry to multiple fields of interest: nuclear chemistry, materials science, mathematical modeling, and cheminformatics are just a few. An additional project I worked on was focused on quantitative radiography, analyzing x-ray radiographs, to characterize materials. We were given a radiograph of an unknown object consisting of five concentric spheres, only having knowledge of the diameter beforehand. Maggie taught me the mathematical skills to analyze this image with the goal of determining the material in each layer. My chemistry knowledge was invaluable on this challenge because I understood the likelihood of each layer being one material, or element, over another. For example, through radiographic analysis we were able to determine approximate densities of each layer in the mystery object. Though this is a small use of my chemistry knowledge, it is proving valuable in areas outside of analytical chemistry.

This internship has been a turning point for me in my education and life in general. Before this summer, I was against the idea of getting a PhD in math as I thought I didn’t like research. My only research experience has been in fundamental analysis problems and, although they’re very

important, my heart truly lies in solving real world problems. Working on mathematics problems for the benefit of national security makes me feel like I'm doing something important for the people around me. I've discovered that I can use my interdisciplinary skills in mathematics and chemistry at a national lab, and I want to hold onto this opportunity with white knuckles. In this project focused on geographical sourcing and food authentication, I've learned invaluable skills in analytical chemistry that will benefit me far beyond my education at Gonzaga.

6.0 References

"Atmospheric Deposition." Maryland Department of Natural Resources, dnr.maryland.gov/streams/Pages/atmosphericdeposition.aspx. Accessed 30 Aug. 2023.

Berg, T., et al. "Atmospheric Trace Element Deposition: Principal Component Analysis of ICP-MS Data from Moss Samples." Environmental Pollution, Elsevier, 20 Apr. 2000, www.sciencedirect.com/science/article/pii/026974919591049Q.

Besse, Philippe, et al. *Can Everyday AI Be Ethical? Machine Learning Algorithm Fairness (English Version)*. Unpublished, 2018, https://doi.org10.13140/RG.2.2.22973.31207.

Center for Food Safety and Applied Nutrition. "Economically Motivated Adulteration (Food Fraud)." U.S. Food and Drug Administration, FDA, www.fda.gov/food/compliance-enforcement-food/economically-motivated-adulteration-food-fraud#:~:text=Economically%20motivated%20adulteration%20(EMA)%20occurs,better%20or%20of%20greater%20value. Accessed 30 Aug. 2023.

"Data Preprocessing in Data Mining." GeeksforGeeks, 6 May 2023, www.geeksforgeeks.org/data-preprocessing-in-data-mining/.

"Doe Capabilities." DOE Capabilities, www.pnnl.gov/doe-capabilities. Accessed 30 Aug. 2023.

"Elemental Analysis Core." Oregon Health & Science University, www.ohsu.edu/elemental-analysis-core/icp-ms-technique. Accessed 30 Aug. 2023.

"Geographical Indications and Quality Schemes Explained." Agriculture and Rural Development, agriculture.ec.europa.eu/farming/geographical-indications-and-quality-schemes/geographical-indications-and-quality-schemes-explained_en. Accessed 30 Aug. 2023.

"Hanford Site." Wikipedia, Wikimedia Foundation, 23 Aug. 2023, en.wikipedia.org/wiki/Hanford_Site.

"How NMR Works: Spectroscopy." Bruker, www.bruker.com/de/resources/library/application-notes-mr/nmr-101.html. Accessed 30 Aug. 2023.

"Inductively Coupled Plasma." Inductively Coupled Plasma - an Overview | ScienceDirect Topics, www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/inductively-coupled-plasma. Accessed 30 Aug. 2023.

"An Introduction to the Fundamentals of Inductively Coupled Plasma – Mass Spectrometry (ICP-MS)." Chemical Analysis, Life Sciences, and Diagnostics, www.agilent.com/en/product/atomic-spectroscopy/inductively-coupled-plasma-mass-spectrometry-icp-ms/what-is-icp-ms-icp-ms-faqs#:~:text=ICP%2DMS%20is%20usually%20used%20for%20relatively%20low%20matrix%20samples,ensure%20they%20give%20reliable%20data. Accessed 30 Aug. 2023.

"Ionization." Ionization - Energy Education, energyeducation.ca/encyclopedia/Ionization#:~:text=Ionization%20is%20the%20process%20by,the%20formation%20of%20an%20ion. Accessed 30 Aug. 2023.

Jolliffe, Ian T, and Jorge Cadima. "Principal Component Analysis: A Review and Recent Developments." The Royal Society Publishing, 13 Apr. 2016, royalsocietypublishing.org/doi/10.1098/rsta.2015.0202.

"Lab Leadership." Lab Leadership, www.pnnl.gov/lab-leadership. Accessed 30 Aug. 2023.

Momtaz, M, et al. "Mechanisms and Health Aspects of Food Adulteration: A Comprehensive Review." Foods (Basel, Switzerland), U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/36613416/. Accessed 30 Aug. 2023.

Ngo, Phuong Linh, et al. "Mechanisms, Status, and Challenges of Thermal Hydrolysis and Advanced Thermal Hydrolysis Processes in Sewage Sludge Treatment." Chemosphere, Pergamon, 16 May 2021, www.sciencedirect.com/science/article/pii/S0045653521013618.

Nguyen, Quang Trung, et al. "Towards a Standardized Approach for the Geographical Traceability of Plant Foods Using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) and Principal Component Analysis (PCA)." MDPI, Multidisciplinary Digital Publishing Institute, 29 Apr. 2023, www.mdpi.com/2304-8158/12/9/1848.

"NMR V MRI." Questions and Answers in MRI, mriquestions.com/mr-vs-mri-vs-nmr.html. Accessed 30 Aug. 2023.

Parlak, Yeliz, and Nuray Güzeler. "Nuclear Magnetic Resonance Spectroscopy Applications in Foods." Current Research in Nutrition and Food Science Journal, 25 Oct. 2016, www.foodandnutritionjournal.org/vol04nospl-issue-conf-october-2016/nuclear-magnetic-resonance-spectroscopy-applications-in-foods/.

"Physical Chemistry." Chemistry LibreTexts, 26 Aug. 2023, chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Physical_Chemistry_(LibreTexts).

Radlib. "Featured History: Magnetic Resonance Imaging." UW Radiology, 25 Sept. 2021, rad.washington.edu/blog/featured-history-magnetic-resonance-imaging/.

Schripsema, Jan. Application of NMR in Plant Metabolomics: Techniques, Problems and Prospects, 6 Oct. 2009, analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/pca.1185.

"Understanding Spectrometry and Spectroscopy." ATA Scientific, 17 Jan. 2020, www.atascientific.com.au/spectrometry/#:~:text=Spectrometry%20is%20the%20measurement%20of,spectroscopic%20analysis%20of%20sample%20materials.

Valdes, Alberto, et al. “Foodomics: Analytical Opportunities and Challenges.” American Chemical Society Publications, pubs.acs.org/doi/10.1021/acs.analchem.1c04678. Accessed 30 Aug. 2023.

van den Berg, Robert A, et al. “Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data.” BMC Genomics, U.S. National Library of Medicine, 8 June 2006, www.ncbi.nlm.nih.gov/pmc/articles/PMC1534033/.

Wilbur, Steve. “A Comparison of the Relative Cost and Productivity of Traditional Metal Analysis Techniques Versus ICP-MS in High Throughput Commercial Laboratories.” Agilent Technologies, 17 Jan. 2005, www.agilent.com/cs/library/applications/5989-1585EN.pdf.

Wilschefski , SC, and MR Baxter. “Inductively Coupled Plasma Mass Spectrometry: Introduction to Analytical Aspects.” The Clinical Biochemist. Reviews, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/31530963/. Accessed 30 Aug. 2023.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov