CRISP-DM: Predicting Molecular Properties

Amin Fesharaki

University of San Diego

Master of Science, Applied Data Science

Course: Foundations of Data Science

Date: October 2021

Business Understanding

Background

Researchers across the globe conduct nuclear magnetic resonance experiments (NMR) to understand the dynamics and structures of molecules. Furthermore, using NMR to gain insight into molecular properties depends on accurately predicting scalar couplings. Currently, it is possible to accurately calculate scalar coupling constants given only 3D molecular structure as input by using state-of-the-art methods from quantum machines. However, these quantum mechanics calculations are costly and time-consuming, and thus, have limited applicability in day-to-day workflows.

The rise of machine learning has created an opportunity to implore potential strategies using data to make scientific predictions. In the past, the chemistry machine learning community has focused on predicting molecular or atomic properties; therefore, no prior research has been conducted to predict atom-pair properties like scalar coupling constants. Thus, the Chemistry and Mathematics in Phase Space (CHAMPS) aims to utilize modern-day technology to incorporate machine learning and predictive analytics to revolutionize scientists' understanding of chemical transformations. By developing a fast and reliable method to predict coupling interactions, chemists can gain structural insights quicker and cheaper, which will ultimately allow scientists to understand further how the 3D chemical structure of a molecule affects its properties and behavior. Such tools will ultimately impact researchers to make significant progress in a range of important problems, such as designing molecules to carry out specific cellular tasks or creating a better drug molecule to fight diseases. In addition, developing machine learning models will profoundly impact industries that depend on chemical change analysis.

The Chemistry and Mathematics in Phase Space have requested our expertise to develop this machine-learning algorithm to assist chemists with their chemical research without NMR experimentation. Moreover, CHAMPS is a multidisciplinary organization involving Bristol University, Cardiff University, Imperial College, and Leeds University. This organization is dedicated to bringing together mathematicians, chemists, and data scientists to provide machine learning models that would significantly improve researchers' current understanding of chemical transformation and ultimately impact all industries that rely on understanding chemical change. In addition, CHAMPS aims to "utilize and develop methods in nonlinear dynamics and machine learning to cope with 6n-dimensional representations of molecular dynamics datasets" (CHAMPS, 2018). Creating analytical models will enable researchers to capture the critical dynamics of complex systems and provide insights into the fundamental microscopic behavior of dynamical systems.

Furthermore, CHAMPS is sponsored by the Engineering and Physical Sciences Research Council (EPSRC). The EPSRC is the "main funding body for engineering, and physical sciences research in the UK [with the goal to] build the knowledge and skills base needed to address the scientific and technological challenges facing the nation" (EPSRC, 2021). By being part of the UK's Research and Innovation, the council is funded by a grant-in-aid from the UK government and is dedicated to supporting research that has an impact across all sectors. Therefore, the EPSRC has given CHAMPS a Programme Grant to address the "urgent need to provide a framework for understanding and exploiting the explosion in dynamical information coming out of modern experiments and simulations in chemistry and biology" (UK Research and Innovation, 2017). The project outlined in the CRISP-DM will have a long-term impact on a wide variety of EPSRC research themes, including environmental science, healthcare, pharmaceutical industries,

energy, and chemical change. However, the immediate effect of this project will allow CHAMPS' researchers to replace NMR experimentation with a cheaper and faster alternative using a machine learning approach.

Business objectives and success criteria

This project aims to develop an algorithm to predict the magnetic interaction between two atoms in a molecule instead of relying on an NMR machine to analyze the scalar coupling constant. In other words, if successful, this project will replace the need to use NMR machines to calculate the scalar coupling constants of molecular molecules. Moreover, the prediction of the parameters, especially in 3-dimensional molecular structures, are "increasingly moving towards the quantitative comparison between computed values for proposed chemical structures and experiment," so the use of accurate and quick NMR prediction methods is crucial to compare the two (Gerrad et al., 2020). Thus, comparing the experimental values with the computed values will determine the success of this project. The project outcome will be considered a success if the calculated values from the algorithm accurately resemble experimental values with a low standard error.

Moreover, NMR machines and systems are quite costly. Lower field and benchtop NMR prices start at about $25,000 and increase to about $150,000. However, as the strength of the magnetic field of a compound increase, so does the price of an NMR machine. Typically, 300 MHz NMR prices start around $150,000 and increase up to $5 million for more powerful instruments like 900 MHz. In other words, increasing the magnetic field strength of an NMR machine to analyze specific molecular structures will increase the machine's price. The most powerful NMR machine, which can handle up to 1.2 GHz, is priced at $17.8 million

(Bettenhausen, 2020). Therefore, utilizing a machine learning algorithm instead of NMR machines can greatly cut down on equipment costs. By implementing this algorithm, we can expect to save CHAMPS roughly $250,000 based on their current use of NMR machines.

Furthermore, NMR experiments are also time-consuming. The run time of an NMR experiment depends on the complexity of the molecular structure. For example, a simple 1D spectrum of small molecule data can typically be collected in 15 minutes and run times for insensitive 3D and 4D chemical compounds can take about a week for a single NMR experiment. Moreover, data collection to calculate a three-dimensional structure can vary from a week to a few months (Mullen, 2021). Thus, incorporating a predictive model instead of relying on NMR experiments can greatly increase our chemist's productivity by reducing the time it takes to analyze a scalar coupling of molecular structures. Regarding the impact within CHAMPS's research department, the algorithm would reduce NMR experimentation rundown by 90%.

In conclusion, creating a reliable algorithm to predict scalar coupling interactions quickly will enable our chemists to effectively understand a chemical compound's molecular properties. As a result, not only can we cut down the costs of acquiring NMR machines, but we will also be able to analyze our chemicals significantly faster, which will substantially improve productivity, increasing our profit margins. In addition, another option for profitability would be to sell the algorithm to environmental science, materials science, pharmaceutical industries, and energy sectors. By doing so, chemists across the globe can utilize this algorithm and further advance our understanding of chemical structures and molecular properties.

Inventory of resources

Data:

Kaggle Dataset: The dataset contains data for 2,358,657 atoms collected through CHAMPS (Bristol University, Cardiff University, Imperial College, and Leeds University). The dataset was separated into 8 different sets of data (CSV files). In addition, the training and test splits are by molecule, so that no molecule in the training data is found in the test data.

CSV Files:

1. Training - A dataset containing molecular names, atoms, and scalar coupling constants to train the predictive model
2. Testing - The test set contains the same info as the training set, but without the target variable to test the predictive model
3. Structures - Contains the molecular structure for every molecule in x,y,z coordinates
4. Dipole Moments - Contains the molecular electric dipole moment data. These are the three dimensional vectors that indicate charge distribution in the molecule.
5. Magnetic Shielding Tensors - Contains the magnetic shielding tensors data for all atoms in molecules
6. Mulliken Charges - Contains data for the mulliken charges for all atoms in the molecules
7. Potential Energy - Contains data for the potential energy of the molecules

8. Scalar Coupling Contributions - Contains data for the scalar coupling constants from the training dataset

Hardware:

Various models need to be used on different machines depending on the quantity and complexity of each of the datasets provided from CHAMPS. In addition, each PC is running a Linux OS operating system.

Base Hardware:

1. Linux OS, Windows 10, or macOS PC with at least 16 gb of ram

Suggested Hardware for Optimal Data Mining (Linux OS)

1. PC(s) with 4 GPUs, each a NVIDIA GeForce RTX 2080 Ti
2. PC(s) with 1 GPU NVIDIA Tesla V100 with 32 GB memory
3. PC(s) with 1 GPU NVIDIA Tesla V100 with 16 GB Memory

Software:

1. Python 3.5+ - To load .csv file for data analysis and modeling
2. Pytorch - Optimized tensor library primarily used for Deep Learning Applications
3. NVIDIA APEX - A repository that holds NVIDIA-maintained utilities to streamline mixed precision and distributed training in Pytorch

Personnel:

1. Project Sponsor

- CHAMPS - 6 year EPSRC - sponsored Programme Grant involving Bristol University, Leeds University, Cardiff University, and Imperial College

2. Principal Investigators
    - Craig Butts - Professor of Structural and Mechanistic at University of Bristol
    - Will Gerad - Phd student under Craig Butts
3. CHAMPS Researcher
    - Lars Bratholm - Post-Doctoral Researcher

Requirements, Assumptions, Constraints

The following section describes the requirements, assumptions, and constraints for this project. All parties (e.g., management and sponsorship) must fully comprehend this section before starting the project. Requirements of the project will include a schedule of completion, comprehensibility, quality of results, and security. Moreover, the assumptions will be made by the project about the data. However, team members can later verify the assumptions on data during the data mining phase. Lastly, the constraints are also included in the project; constraints can be considered inefficiency or a roadblock to the project.

Requirements

The specific target for this project is for our chemists to cut downtime and company costs when conducting NMR experiments to examine molecular properties. In the future, we can discuss potential deployment strategies to allow researchers around the globe who conduct NMR experiments to have access to this model. This project can potentially impact all industries that

rely on understanding chemicals: environmental science, pharmaceutical industries, and energy sectors.

Furthermore, CHAMPS is an interdisciplinary organization comprising two central departments: chemistry and mathematics. However, only the chemistry researchers and staff members will need to assist our programming team with the fundamentals of targeted chemistry knowledge. The following table highlights the suggested schedule for optimal project performance and success. Outsourcing/hiring additional coders for the programming team should be completed before following the outlined schedule. Assuming there are no significant setbacks, this project is expected to take around 13 weeks to complete.

Time Frame	**Phase**	**Task**
Weeks 1 - 2	Data Understanding	Collect initial data, describe data, and explore data
Weeks 2 - 3	Data Preparation	Select, construct, integrate, and format data
Weeks 4 - 8	Modeling	Select modeling technique, generate test design, build model, and assess model
Weeks 9 - 11	Evaluation	Evaluate results, review process, and if needed, determine next steps to optimize model
Weeks 12 - 13	Deployment	Plan monitoring and maintenance, produce final report, and review project

Additionally, the platform on which the model is deployed should be user-friendly, where researchers and chemists should be able to use the model without prior technical coding knowledge. At least one coder should be tasked with maintaining the platform and roll out any updates as needed. In addition, the marketing team will be responsible for deploying the platform/software to the CHAMPS organizations of chemical researchers and potential future recipients in different industries outside of CHAMPS.

Also, during the algorithm development, only members of the programming team will have access to the code to ensure no code tampering. Security will be achieved by two-factor authentication, including employee ID/password and a confirmation code sent via text message or email. In addition, sharing any information (i.e., code, emails, etc.) is prohibited during the duration of this project. However, sharing information is allowed after completing the project. This project is part of CHAMPS' primary mission to "enable the formulation of comprehensible analytic models which capture the key dynamics of complex systems and provide fundamental microscopic insight into the behavior of dynamical systems" (CHAMPS, 2018). Therefore, allowing access and knowledge to others is detrimental to advancing our (anyone and everyone in the chemistry field) understanding of molecular properties and potentially optimizing the model.

Assumptions

As stated previously, this project's goal is to develop an algorithm that can predict the scalar coupling constants, the magnetic interaction between two atoms in a molecule, using various chemistry data constants for chemical researchers.

Assumptions in direct regards to our goal:

- “scalar coupling constant” - The scalar coupling constant is previously determined and proven for any molecule within the dataset provided to ensure the model's accuracy when comparing the predicted and actual values.
- “various chemistry data constants” - Data for structures, dipole moments, magnetic shielding charges, Mulliken charges, potential energy, and scalar coupling constants are assumed to be accurate within all the datasets.
- “chemical researchers” - These people are assumed to be knowledgeable in their respective fields that can understand the data used for this model.

Data quality assumptions:

- All samples, with their respective data constants, are assumed to be accurate in the data set provided.
- Data is assumed to follow a normal distribution to achieve the best possible model with the highest accuracy.
- The data for the project is also assumed to be available through CHAMPS and other sources and databases containing scalar coupling constants and atom indexes of chemical molecules that are experimentally proven and reviewed.

External factor assumptions:

- Financials - The sponsorship funding, along with our current resources, are enough to provide all financial funding needed to complete this project (e.g., hardware, software, and personnel).
- Technology - Modern-day computer systems and current analytical methods are assumed to be sufficient to develop a machine learning model to predict the target variable.

Assumptions regarding whether it is necessary to understand or explain the model:

- CHAMPS' senior management is assumed to have a high chemical understanding of each variable presented in the dataset.
- People using the algorithm are assumed to have a fundamental understanding of the topics pertaining to nuclear magnetic resonance (NMR) experiments.

Constraints

General constraints:

- Not all molecules known to man have had their scalar coupling constants determined. Therefore, predicting unknown molecules' scalar coupling constant cannot be genuinely verified without experimental data.
- No apparent legal issues for completing this project. However, a disclaimer should be added explaining how the model is only a prediction and should not be considered 100% accurate without further experimentation for support.
- The budget is not unlimited; management should consider the best cost-efficient hardware/software and a team that is best suited for the task at hand.
- The availability of staff and team members are not certain due to absences (e.g., emergencies and sickness); therefore, the timetable might be increased.

Technical accessibility of data constraints:

- Current operating systems/software available may not accurately predict the target variable with a negligible standard of error.

- When collecting data or transferring the data .csv files to the software mentioned in the “inventory of resource” section, uncoordinated data management can lead to inaccurate model development.

Relevant knowledge constraints:

- Not understanding the relevant knowledge of the dataset can serve as a roadblock when developing the model.
- Current knowledge of the scientific topics related to this project is available through scholastically peer-reviewed journals. However, our understanding is continuously expanding through constant research of chemical change.

RESOLVEDD Strategy

The RESOLVEDD strategy is a step-by-step decision-making method based on "professional ethics and approaches ethics from the point of view of personal ethical problems in work contexts" (Vakkuri and Kemell, 2019). The strategy consists of nine steps illustrating the rational ethical decision-making process. The RESOLVEDD procedure shown below highlights each step and how it can be related to the context of this project.

1. *Review*: The Chemistry and Phase Space is a research organization sponsored by the Engineering and Physical Sciences Research Council (EPRSC). The purpose of this project is to reduce the need to conduct NMR experiments by incorporating a machine-learning algorithm to predict the magnetic interaction between two atoms in a molecule (i.e., scalar coupling constant). Furthermore, CHAMPS is a non-profit organization whose intentions are to save time and money for their chemists and further progress chemical research by publishing this project for scientists to use around the world.

However, as previously mentioned, CHAMPS' is funded by the EPRSC. Both these organizations are expected to follow the Global Chemist Code of Ethics. The Global Chemist code of Ethics states that "research in chemical sciences should benefit humankind and improve quality of life, while protecting the environment and preserving it for future generations. [In addition], researchers should conduct their work with the highest integrity and transparency, avoid conflicts of interest, and practice collegiality in the best way "(American Chemical Society, 1965). Thus, the team members of this project are also expected to conduct research that upholds these same standards to progress chemical research. Therefore, the ethical concerns within the RESOLVEDD strategy are not directly aimed at the project itself but at CHAMPS' motivations for creating this endeavor.

2. *Estimate*: Ethical concerns can arise if CHAMPS does not follow their mission statement to "provide new models that revolutionize our understanding of chemical transformation" (CHAMPS, 2018). One concern is that the motivation for this project is for CHAMPS to boost their credibility or prestige status within the chemical research field by publicizing the model and labeling it as accurate information, even though the model is incorrect or insufficient enough to replace NMR experimentation. Another way for the model to be inaccurate is if CHAMPS uses incorrect data or a biased sample of molecular data to achieve the goal of this project, knowing that it will fail when using data outside of the dataset. By publishing false information, the model will not benefit humankind and potentially harm the recipients of this information. For example, if a chemist assumes a wrong molecular property, the experiment can fail, wasting time and resources, or result in safety issues. Moreover, another potential ethical concern is if CHAMPS goes against

their EPSRC sponsorship and decides to withhold their research and instead sell the model for profit in secrecy while restricting access to others.

3. *Solutions*: One solution to checking inaccuracies is to cross-reference the data with other chemical databases using the same molecules as CHAMPS provided. Another possible solution is reviewing any documentation of how and who collected the data used in this project. Additionally, the EPSRC can check the completed model to ensure the model's accuracy for chemical use. Moreover, with the ethical concern of limiting the model's access for profit, we can notify the EPSRC that CHAMPS is involved with this project and have them send a representative to CHAMPS to monitor its progression.
4. *Outcomes*: The consequences of disproving CHAMPS' supposedly correct data are a loss of the organization's reputation and potentially having their sponsorship from EPSRC removed. A result of sending an EPSRC representative to CHAMPS could be a loss of potential revenue.
5. *Likely*: The impact of checking CHAMPS data is ensuring CHAMPS is a reputable organization. Having an EPSRC representative to check on the project's progression would be further advancements towards chemical research and allowing chemists to gain structural insights faster and cheaper than conducting NMR experimentation.
6. *Values*: The values of scientific research include a "duty to society, beneficence, conflict of interest, integrity, non-exploitation, professional competence, and professional discipline" (Weinbaum et al., 2019). Integrity goes with the ethical concern that CHAMPS is skewing the data to promote their reputation. Beneficence, conflict of interest, and non-exploitation relate to CHAMPS selling their model for their financial greed instead of progressing in chemical research. Lastly, a lack of professional

competence and discipline is caused by CHAMPS using inaccurate or false data to get their model working sufficiently while knowing that the model is not suitable enough to replace NMR experimentation.

7. *Evaluate*: Proceeding with the solutions mentioned previously, CHAMPS will be held accountable for providing correct data and allowing model access for chemists across the world to use. This will ensure the model's accuracy and its overall contribution to chemical research. Additionally, CHAMPS will uphold their mission statement and values of scientific research.
8. *Decide*: The best solution is to notify the Engineering and Physical Sciences Research Council of CHAMPS's involvement with this project. By doing so, the council will make sure that the data is correct not to harm their reputation and to promote their mission statement to "maximize the contribution of each of [their] component parts" (Engineering and Physical Sciences Research Council, 1994). The solution concerning cross-referencing the data with another chemical database and reviewing how the samples were collected for more than 2 million chemical compounds will take too much time and resources.
9. *Defend*: The most significant weakness to the representative solution is the simple fact that CHAMPS can lie to the council or hide their involvement with the project. However, as one of the primary sponsorships for the organization, it would be irrational for CHAMPS to be nothing but transparent to EPSRC. The funding CHAMPS receives from their sponsors, as well as their reputation and scientific goal, should far outweigh the financial gains from this one project.

Risks and Contingencies

Risk	**Contingency Plan**
Hardware Malfunction (i.e., computer systems)	At least one backup machine with the same hardware should be available in case of a hardware malfunction. In addition, existing code and data should be backed up or saved onto an external hard drive or cloud storage periodically to prevent the loss of code/data.
Software Issues (i.e., inability to provide accurate models)	If one of the softwares fails to provide a desired outcome, use a different software capable of machine learning models. Other softwares sufficient for deep learning include, but not limited to, TensorFlow, Microsoft Cognitive Tooklit, Keras, ConvNetJS, Torch, and MXNet.
System Outages (i.e., power loss, internet connectivity)	A backup generator should be available on site for power loss issues. Machines should be hardwired into the router to prevent internet disruptions. However, mobile wireless hotspots should also be in place for internet outages.
Staffing Issues (i.e., staff member is unavailable)	Each team should consist of two or more people to prevent the disruption of workflow when a member becomes unavailable due to an emergency. Team members must also record key thought processes, results, and annotate code. By doing so, if a member of

	the team becomes unavailable, another member can continue his or her work.
Insufficient Hardware Performance (i.e., slow or unusable machine to effectively analyze data)	If the hardware used cannot effectively analyze the data or build an accurate model, then the hardware should be upgraded by purchasing or renting additional hardware capable of using the software at optimal performance. Hardware upgrades can include, but not limited to, upgrading ram, cpu cooling, processors, and graphics processing unit.
Memory issues (i.e, CUDA out of memory)	This error is due to running out of memory on the GPU. One solution is to use the *--batch_chunk* option to cut down large batches into a few smaller (equal) shares. Another option is to use few *--n_layer*, or smaller *--batch_size*. Additionally, *if* multi-GPU machines are being utilized, use the *--multi_gpu* flag, which controls the minibatch size passed to GPU.

Terminology

6n-dimensional phase space: “Six-dimensional space is any space that has six dimensions, six degrees of freedom, and that needs six pieces of data, or coordinates, to specify a location in this space” (Davis, 2020)

Algorithm: Machine learning algorithm that can predict the target variable using a set of parameters.

Big Data Analytics: Analyzing big data using “a collection of different tools types, including those based on predictive analytics, data mining, statistics, artificial intelligence, natural language processing, and so on” (Russom, 2011).

Chemical Researchers: Researchers who conduct NMR experiments or chemists who use principles of chemical change.

Chemical Transformations: “A chemical transformation can occur when molecules collide with sufficient energy” (National Research Council, 2003).

Dipole Moments: “Dipole moments occur when there is a separation of charge between atoms and electrons” (Blaber, 2020).

Machine Learning: “Machine learning is a branch of computer science that broadly aims to enable computers to ‘learn’ without being directly programmed” (Bi et al., 2019).

Magnetic Interactions: An interaction between two magnetic dipoles. This is a necessary component for a Nuclear Magnetic Resonance experiment.

Magnetic Shielding Tensor: “The shielding or more properly the magnetic shielding is the tensor that describes the relative change in the local magnetic field at the nucleus position relative to the external magnetic field” (Facelli, 2011).

Model: Model and results should be presented to senior management/sponsor in a clear but descriptive manner. The process of how the model was made and the significance of the results should be described in detail.

Nuclear Magnetic Resonance: “Nuclear magnetic resonance (NMR) spectroscopy is a robust method, which can rapidly analyze mixtures at the molecular level without requiring separation and/or purification steps” (Hatzakis, 2018).

Potential Energy: “Potential Energy is the energy due to position, composition, or arrangement. Also, it is the energy associated with forces of attraction and repulsion between objects” (Harbick et al., 2020).

Quantum Machines: Quantum computation using "variational algorithms that employ classical optimization cou- pled with quantum hardware to evaluate the quality of each candidate solution" (Gokhale et al., 2019).

Scalar Coupling Constants: An atomic property that is "critically dependent on the 3D structure of the molecule for which they are being measured" (Bratholm et al., 2021). In addition, the scalar coupling coefficient is a measure of the magnetic interaction between two atoms in a molecule.

Scalar Coupling Interactions: Scalar coupling "is an interaction between nuclei containing spin" (Kaseman, 2020)

Data mining goals and success criteria

In order to remove the need for NMR systems to analyze molecular properties, an algorithm will be developed to predict the magnetic interaction between two atoms in a molecule. Specifically, the algorithm needs to predict the scalar coupling constant between atom pairs, given two atom types, (i.e., O and C), the coupling type (i.e., 1JHN), and any features made from the molecular structure. Furthermore, to ensure the algorithm can predict this interaction, the model needs to sufficient\\ly predict the explicitly listed pairs in the train and test files rather than all the molecules. For example, the scalar coupling constant will predict some molecules that contain Oxygen (O); however, the algorithm should initially predict the scalar coupling constant for molecules that contain Oxygen (O) in the train and test files instead of

predicting the scalar coupling constant for any pair that includes O. By comparing the train and test data, the accuracy of the predictive model can be determined.

The success criteria for this project will be evaluated on the Log of the Mean Absolute Error, calculated for each scalar coupling type, and then averaged across types. Equation 1 indicates the accuracy score value for the scalar coupling constant.

$$Score = \frac{1}{T}\sum_{t=1}^{T} \log\left(\frac{1}{n_t}\sum_{i=1}^{n_t}|y_i - \hat{y}_i|\right) \tag{1}$$

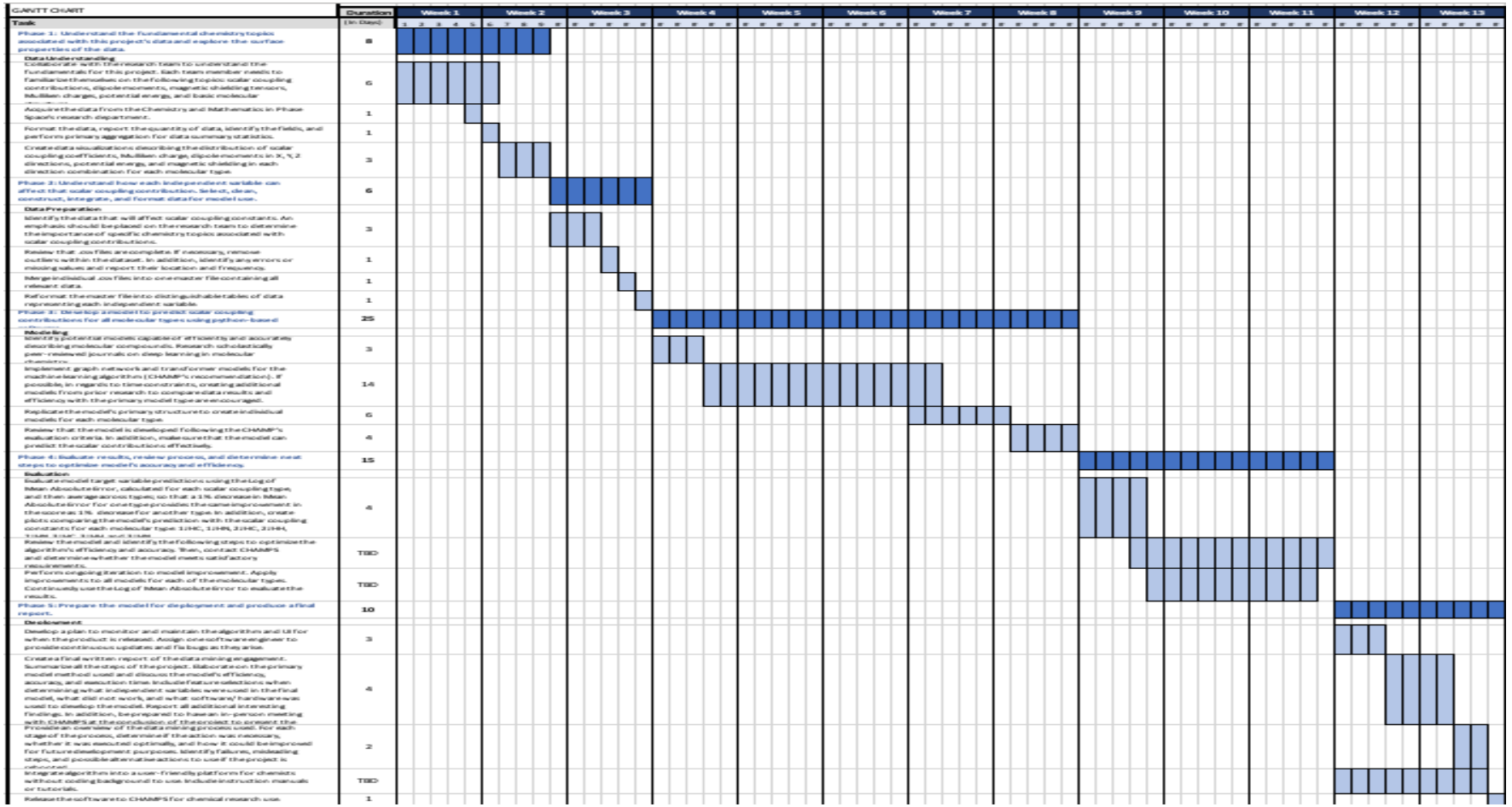
Where:

- **T** is the number of scalar coupling types
- $\boldsymbol{n_t}$ is the number of observations of type t
- $\boldsymbol{y_i}$ is the actual scalar coupling constant for the observation
- $\boldsymbol{\hat{y}_i}$ is the predicted scalar coupling constant for the observation

For reference, the minimum *best* possible score for perfect predictions is approximately -20.7232 for this metric, as shown in Equation 1 (Kaggle, 2019). Moreover, for this project, the target variable to predict is the scalar coupling constant. Therefore, the model's accuracy will follow the Log of the Mean Absolute Error, calculated for each coupling type, and then average across types. Afterward, tests will be performed on random atom indices of the atom-pair for any molecule, given the experimental scalar coupling constant, to compare the accuracy of individual cases rather than the average. This will confirm the accuracy of the model and highlight potential concerns when outliers, or extremes, data is present. By the completion date, the model should ensure consistent and repeatable target variable predictions sufficiently accurate for NMR analysis.

Project plan / Order of tasks

A Gantt chart is provided to outline the project plan and task estimates for each phase. Due to the extensive length of this outline, the tasks will be included underneath the chart to read each task easily.



GANTT CHART	Duration
Task	(In Days)
Phase 1: Understand the fundamental chemistry topics associated with this project's data and explore the surface properties of the data.	**8**
Data Understanding	
Collaborate with the research team to understand the fundamentals for this project. Each team member needs to familiarize themselves on the following topics: scalar coupling contributions, dipole moments, magnetic shielding tensors, Mulliken charges, potential energy, and basic molecular structure.	6
Acquire the data from the Chemistry and Mathematics in Phase Space's research department.	1
Format the data, report the quantity of data, identify the fields, and perform primary aggregation for data summary statistics.	1
Create data visualizations describing the distribution of scalar coupling coefficients, Mulliken charge, dipole moments in X, Y, Z directions, potential energy, and magnetic shielding in each direction combination for each molecular type.	3
Phase 2: Understand how each independent variable can affect that scalar coupling contribution. Select, clean, construct, integrate, and format data for model use.	**6**
Data Preparation	
Identify the data that will affect scalar coupling constants. An emphasis should be placed on the research team to determine the importance of specific chemistry topics associated with scalar coupling contributions.	3
Review that .csv files are complete. If necessary, remove outliers within the dataset. In addition, identify any errors or missing values and report their location and frequency.	1
Merge individual .csv files into one master file containing all relevant data.	1
Reformat the master file into distinguishable tables of data representing each independent variable.	1
Phase 3: Develop a model to predict scalar coupling contributions for all molecular types using python-based software.	**25**
Modeling	
Identify potential models capable of efficiently and accurately describing molecular compounds. Research scholastically peer-reviewed journals on deep learning in molecular chemistry.	3
Implement graph network and transformer models for the machine learning algorithm (CHAMP's recommendation). If possible, in regards to time constraints, creating additional models from prior research to compare data results and efficiency with the primary model type are encouraged.	14

Replicate the model's primary structure to create individual models for each molecular type.	6
Review that the model is developed following the CHAMP's evaluation criteria. In addition, make sure that the model can predict the scalar contributions effectively.	4
Phase 4: Evaluate results, review process, and determine next steps to optimize model's accuracy and efficiency.	**15**
Evaluation	
Evaluate model target variable predictions using the Log of Mean Absolute Error, calculated for each scalar coupling type, and then average across types; so that a 1% decrease in Mean Absolute Error for one type provides the same improvement in the score as 1% decrease for another type. In addition, create plots comparing the model's prediction with the scalar coupling constants for each molecular type: 1JHC, 1JHN, 2JHC, 2JHH, 2JHN, 3JHC, 3JHH, and 3JHN.	4
Review the model and identify the following steps to optimize the algorithm's efficiency and accuracy. Then, contact CHAMPS and determine whether the model meets satisfactory requirements.	TBD
Perform ongoing iteration to model improvement. Apply improvements to all models for each of the molecular types. Continuesly use the Log of Mean Absolute Error to evaluate the results.	TBD
Phase 5: Prepare the model for deployment and produce a final report.	**10**
Deployment	
Develop a plan to monitor and maintain the algorithm and UI for when the product is released. Assign one software engineer to provide continuous updates and fix bugs as they arise.	3
Create a final written report of the data mining engagement. Summarize all the steps of the project. Elaborate on the primary model method used and discuss the model's efficiency, accuracy, and execution time. Include feature selections when determining what independent variables were used in the final model, what did not work, and what software/ hardware was used to develop the model. Report all additional interesting findings. In addition, be prepared to have an in-person meeting with CHAMPS at the conclusion of the project to present the project's key results.	4
Provide an overview of the data mining process used. For each stage of the process, determine if the action was necessary, whether it was executed optimally, and how it could be improved for future development purposes. Identify failures, misleading steps, and possible alternative actions to use if the project is rebooted.	2
Integrate algorithm into a user-friendly platform for chemists without coding background to use. Include instruction manuals or tutorials.	TBD
Release the software to CHAMPS for chemical research use.	1

Data Understanding

Initial data collection report

The following list of datasets were acquired from Kaggle and can be downloaded using the Kaggle API:

API: kaggle competitions download -c champs-scalar-coupling

Files:

- train.csv
- test.csv
- structures.csv & structures.zip

Additional Data:

Note: Additional Data is provided for the molecules in Train only

- dipole_moments.csv
- magnetic_shielding_tensors.csv
- mulliken_charges.csv
- potential_energy.csv
- scalar_coupling_contributions.csv

The datasets can also be accessed through the master directory located at C:/Users/datascience/Documents/Kaggle Datasets/champs-scalar-coupling. Furthermore, the data was collected through laboratory experiments using various chemistry methods, measuring instruments, and software. The experimental data were then organized into datasets provided by CHAMPS.

Note: No problems were encountered investigating the csv file using Jupyter Notebook on a macOS with an Apple M1 Processor and 16gb of RAM. However, accurate predictive modeling cannot be achieved exclusively through this system, and therefore, additional hardware should be used for optimal performance

Data description report

The entire dataset requires 1.22 GB to download with a total of 2,505,192 samples of atoms. Each file, with the exemption of the train and test, in the datasets each contain 130,789 unique values. Of the total 130,789 unique values, 65% (85012) are stored in the training file and the remaining 35% (45,777) are stored in the testing file. The testing and training splits are by molecule, so that no molecule in the test data is found in the training data. Additionally, the scalar_coupling_constributions file has the same number of unique values as the training set (85012). In total, there are 47 columns for the 130,789 observations.

The following description report was examined in python through Jupyter Notebook. In addition, each table is describing the column's properties under its respective csv file. It is important to note that the table serves as a surface-level summary and organization of columns within the dataset. The first row of the table indicates names of the data with its datatype (i.e., # is integer).

Note: The tables shown below are not the full actual representation of the file itself. In addition, the table legend for datatypes (integers, string, decimals) is located at the end of this section.

File 1(a): structures.zip

A folder containing the molecular structure of all compounds used in the dataset. The "first line is the number of atoms in the molecule, followed by a blank line, and then a line for every atom, where the first column contains the atomic element (C for carbon, H for hydrogen etc.) and the remaining columns contain the X, Y, and Z, cartesian coordinates" (Kaggle, 2019).

File 1(b): structures.csv - Same info as the structures zip but in one csv file instead of individual xyz files.

Table 1: Representation of "structures.csv" – 6 columns

A **molecule_name**	# **atom_index**	***A*** **atom**	# **x**	# **y**	# **z**
Names of the molecule	Index of the atom in the molecule	Atomic element	Cartesian coordinates	Cartesian coordinates	Cartesian coordinates

File 2: dipole_moments.csv - Contains the molecular electric dipole moments. The dipole moments are three dimensional vectors that indicate the charge distribution in the molecule.

Table 2: Representation of "dipole_moments.csv" – 4 columns

A **molecule_name**	# **X**	# **Y**	# **Z**
Names of the molecule	X component respectively of the dipole moment	Y component respectively of the dipole moment	Z components respectively of the dipole moments

File 3: Magnetic_shielding_tensors.csv- Contains the magnetic shielding tensors data for all atoms in molecules.

Table 3: Representation of "magnetic_shielding_tensor.csv" – 11 columns (The table is transposed to fit all the information in 1 table (i.e, the columns and rows are flipped)

A **molecule_name**	Names of the molecule
# atom_index	Index of the atom in the molecule
# XX	XX elements of the tensor/matrix
# YX	YX elements of the tensor/matrix
# ZX	ZX elements of the tensor/matrix
# XY	XY elements of the tensor/matrix
# YY	YY elements of the tensor/matrix
# ZY	ZY elements of the tensor/matrix
# XZ	XZ elements of the tensor/matrix
# YZ	YZ elements of the tensor/matrix
# ZZ	ZZ elements of the tensor/matrix

File 4: muliken_charges.csv -Contains data for the mulliken charges for all atoms in the molecules.

Table 4: Representation of "muliken_charges.csv" – 3 columns

A **molecule_name**	# **atom_index**	# **muliken_charge**
Names of the molecule	Index of the atom in the molecule	muliken charge of the atom

File 5: potential_energy.csv - Contains data for the potential energy of the molecules.

Table 5: Representation of "potential_energy.csv" – 2 columns

A **molecule_name**	# **potential_energy**
Names of the molecule	Potential energy of the molecule

File 6: scalar_coupling_contributions.csv - Contains data for the scalar coupling constants from the training dataset. The scalar couplings constants are a sum of four terms: fc, sd, pso, dso.

Table 6: Representation of "scalar_coupling_contributtions.csv" – 8 columns (The table is transposed in order to fit all the information in 1 table (i.e, the columns and rows are flipped)

A **molecule_name**	Names of the molecule
# atom_index_0	Atom indices of the atom pair
# atom_index_1	Atom indices of the atom pair
A **type**	Type of coupling
# fc	Fermi Contact contribution
# sd	Spin-dipolar contribution
# pso	Paramagnetic spin-orbit
# dso	Diamagnetic spin-orbit

File 7: train.csv - A dataset containing molecular names, atoms, and scalar coupling constants to train the predictive model

Table 7: Representation of "train.csv" – 6 columns

O **id**	***A*** **molecule_name**	# **atom_index_0**	# **atom_index_1**	***A*** **type**	# **scalar_coupling_cont.**
Sequential ID number of each sample observation	Names of the molecule	Atom indices of the atom-pair creating the coupling (atom_index 0 & 1)	Atom indices of the atom-pair creating the coupling (atom_index 0 & 1)	Type of coupling	Scalar coupling constant that we want to be able to predict

File 8: test.csv - The test set contains the same info as the training set, but without the target variable to test the predictive model

Table 8: Representation of "test.csv" – 5 columns

O **id**	***A*** **molecule_name**	**#** **atom_index_0**	**#** **atom_index_1**	***A*** **type**
Sequential ID number of each sample observation	Names of the molecule	Atom indices of the atom-pair creating the coupling (atom_index 0 & 1)	Atom indices of the atom-pair creating the coupling (atom_index 0 & 1)	Type of coupling

Table Symbols Legend:

#: Decimal - 22 columns of decimal datatypes

A: String – 12 columns of string datatypes

#: Integer – 10 columns of integer datatypes

O: Other – 2 columns of ID numbers

Data exploration report

The purpose is to visualize the distributions of each variable to predict the target variable (scalar coupling constant). The data exploration report was done in python using Jupyter notebook. The exploration report includes created summary statistics and graphical representation for essential characteristics for variables in each file to find the distribution of each variable and other patterns and behaviors. Determining distributions is a critical component in developing the appropriate model. Furthermore, the hypothesis for the data exploration report is that each dataset, under its respective variable, follows a normal distribution. Having a normal distribution will allow us to implement an appropriate model to predict the target variable.

Moreover, key variables of each file type are shown below. The numerical and graphical analysis is initially required to how to create an appropriate machine learning algorithm. In addition, the types of J-couplings that exist in the datasets are 1JHC, 2JHC, 3JHC, 1JHN, 2JHN, 3JHN, 2JHH.

Test & Train

Table 1: Test & Train Summary Statistics *(Test = A, Train= B)*

	Variable	**Mean**	**Standard Deviation**	**Minimum**	**Q1 (25%)**	**Median (50%)**	**Q3 (75%)**	**Maximum**
A	Atom index 0	13.34	3.27	1.00	11.00	13.00	16.00	28.00
A	Atom index 1	5.88	5.00	0.00	2.00	5.00	8.00	28.00
B	Atom index 0	13.36	3.27	1.00	11.00	13.00	16.00	28.00
B	Atom index 1	5.88	4.99	0.00	2.00	5.00	8.00	28.00

B	Scalar coupling Constant	15.92	34.93	-44.76	-0.25	2.28	7.39	207.71

The atom indexes of 0 and 1 have nearly identical summary statistics (mean, standard deviations, min and maximum). Therefore, we can accurately use both datasets (testing and training) to develop a machine learning algorithm to predict the scalar coupling constant. The testing set does not include a scalar coupling constant since that is what we are trying to predict. After predicting the scalar coupling constant from the testing set, we can then compare it with the training set's scalar coupling constant to determine if our model is sufficient in predicting the target variable.

Structures

Table 2: Structure Summary Statistics

Variable	**Mean**	**Standard Deviation**	**Minimum**	**Q1 (25%)**	**Median (50%)**	**Q3 (75%)**	**Maximum**
x	-0.10	1.66	-9.23	-0.09	0.05	1.12	9.38
y	-0.34	1.98	-9.93	-1.83	-0.40	1.37	10.18
z	-0.06	1.45	-9.14	0.84	0.01	0.94	7.89

From Table 2, we can see that the mean is less than the median for x and z. Therefore, based on the numerical results of the summary statistics, we can predict that the distribution is slightly skewed to the left. On the other hand, y has a mean greater than the median, which

indicates that the distribution can slightly be skewed to the right. Therefore, to better understand the distribution, the data should be graphed to see the shape of the distribution of each variable.

As shown in the figures below, the molecular structure's x (blue) coordinate appears to be a unimodal distribution. However, the y (red) and z (green) coordinates appear to be multimodal distributions. From the figure, the y coordinate shows a distribution with three distinct peaks, and the z coordinate shows a distribution with five peaks.

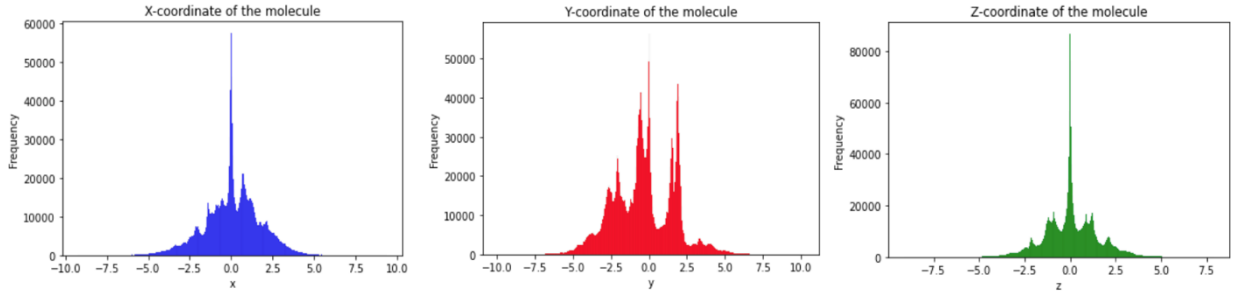


Figure 1: Histogram to visualize the distribution of dipole moments in X, Y and Z direction

Figure 2 illustrates the x coordinate (blue & purple), y coordinate (red & dark pink), and z coordinate (maroon & green) for each coupling type. Thus, the structures for each coupling type's distribution appear to be normally distributed.

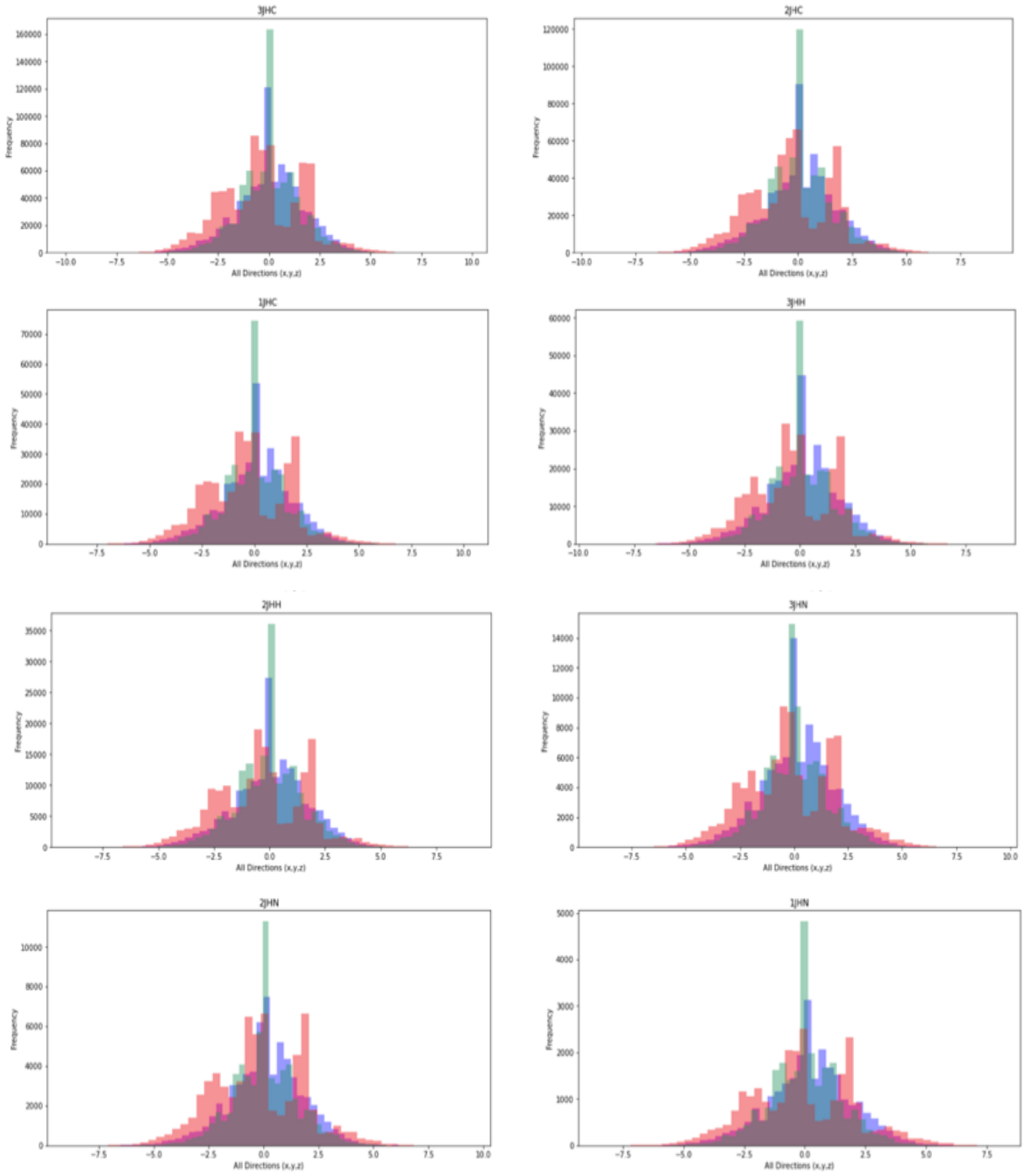


Figure 2: Three different histograms layered into one graph to show the distribution of structures in all directions (X,Y,Z) for each molecule type (J coupling)

Scalar Coupling Contributions

Table 3: Scalar Coupling Contributions Summary Statistics

Variable	Mean	Standard Deviation	Minimum	Q1 (25%)	Median (50%)	Q3 (75%)	Maximum
Fermi Contact	15.69	34.41	0.4171	-0.213	2.33	7.47	207.42
Spin-dipolar	0.08	0.14	-4.420	-0.1437	0.05	0.13	7.67
Paramagnetic spin-orbit	0.38	0.74	-3.288	-0.03	0.16	0.44	8.20
Diamagnetic spin-orbit	-0.23	0.92	-6.86	-0.37	-0.01	0.14	1.70

Table 3 shows the summary statistics for each component for scalar coupling contributions. However, what we are primarily concerned about is the distribution of the scalar coupling constant. From Table 1 (test/train table), we can see that the mean (15.92) is roughly three times bigger than the median (5.00). Therefore, one can conclude that the distribution is heavily skewed to the right. However, graphing the scalar coupling constant shows that the distribution is bimodal, which dramatically impacts the skewness of the distribution.

From Figure 3, we can see that the distribution of the scalar coupling coefficient is skewed to the left, but the distribution is not perfectly unimodal. There is a significantly smaller peak, close to 100, making it bimodal. The mode and mean are approximately 0.

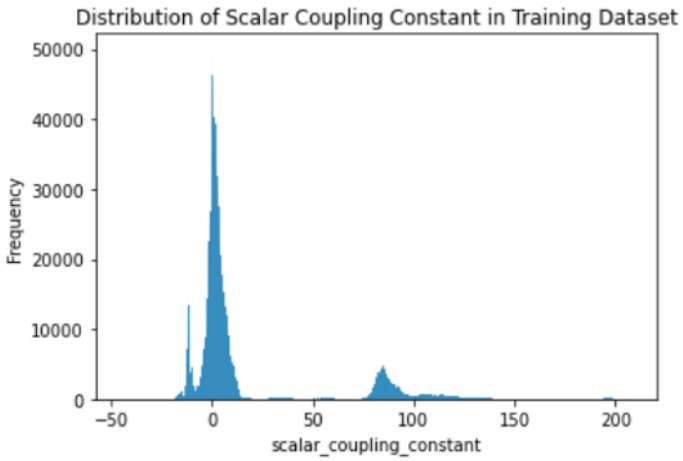


Figure 3: Histogram of the distribution of scalar coupling coefficient in train dataset

In the figures below, there are the distributions of the target (scalar coupling constant) for different molecular categories. Three distributions are unimodal (2JHC, 3JHC, 2JHH); however, 2JHC is slightly skewed to the right. Three distributions are heavily skewed to the left (1JHC, 3JHH, 2JHN), and two distributions are bimodal (3JHN, 1JHN).

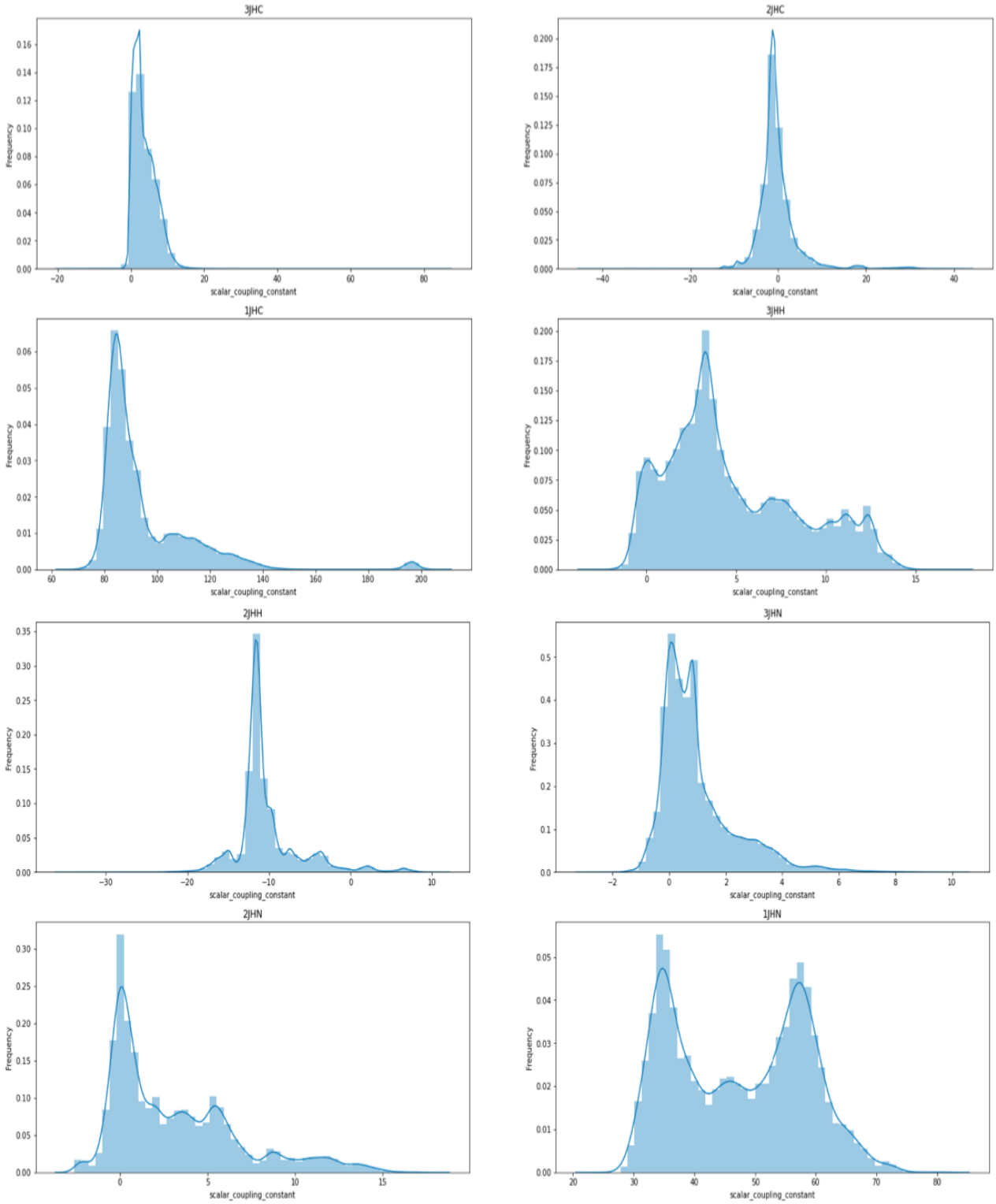


Figure 4:Smooth histograms were made to show the distribution for each molecule type's scalar coupling constant (J coupling)

Dipole Moment

Table 4: Dipole Moment Summary Statistics

Variable	Mean	Standard Deviation	Minimum	Q1 (25%)	Median (50%)	Q3 (75%)	Maximum
X	-0.01	2.29	-22.96	-1.35	0.00	1.32	29.59
Y	0.09	1.74	-9.25	-0.94	0.08	1.16	13.01
Z	0.25	1.03	-6.03	-0.25	0.12	0.90	7.70

The dipole moment summary statistics shows that the mean is approximately the same as the median for all the variables (X, Y, Z). In other words, we can expect a normal distribution for each of these variables. To gain further insight, we should continually graph the data points to illustrate the distribution better, as shown in Figure 5.

The distributions of dipole moment along the X and Y axes are approximately normal with a mean of 0. However, the X-axis distribution has a more significant standard deviation and range. In contrast, the dipole moment along the Z-axis has a slightly skewed distribution (skewed to the right), with a secondary peak around 1 in addition to the primary peak (mode) above 0.

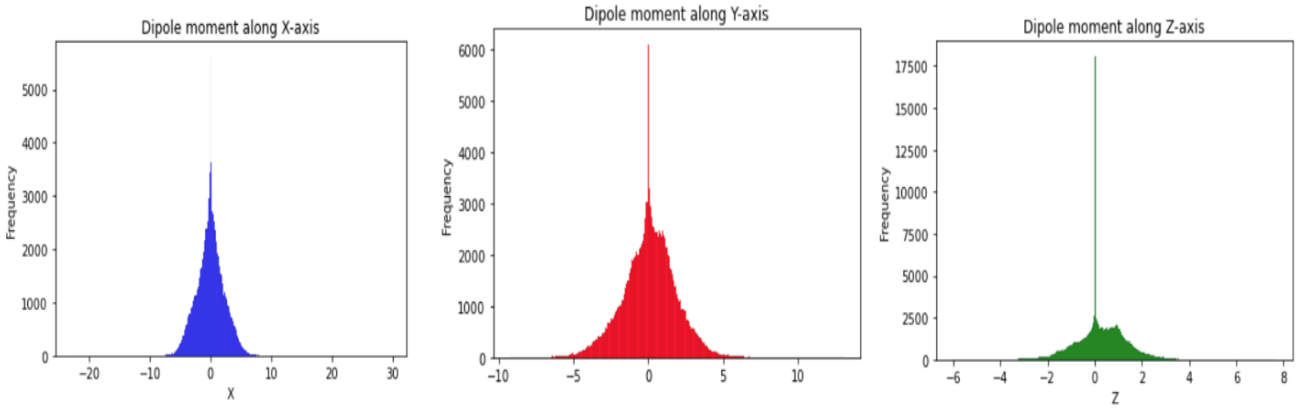


Figure 5: Three histograms to show the distribution of dipole moments in X, Y and Z direction

In Figure 6, the blue (X), red (Y), and green (Z) distributions represent the dipole moment distributions along the X, Y, and Z axes, respectively. They all are normal distributions with a mean of 0, but the standard deviation increases from Z (green) to Y (red) to X (blue).

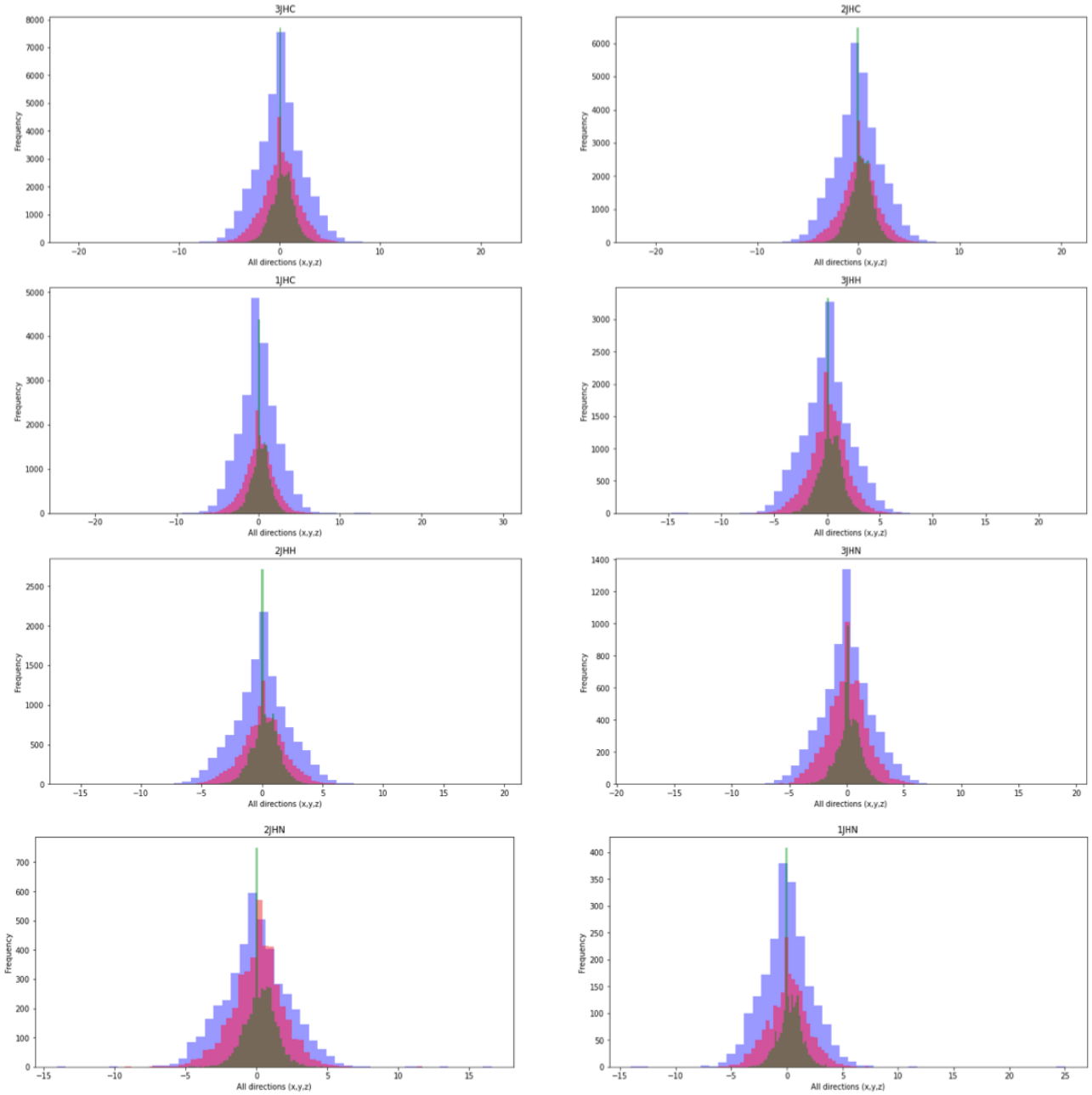


Figure 6: Three histograms layered into 1 graph to show the distribution of each dipole moment in all directions (X,Y,Z charges) for each molecule type (J coupling)

Mullikan Charges

Table 5: Muliken Charges Summary Statistics

Variable	Mean	Standard Deviation	Minimum	Q1 (25%)	Median (50%)	Q3 (75%)	Maximum
Mullikan Charge	-3.26e-10	0.27	-0.73	-0.192	-0.099	0.13	0.72

According to Table 5, the mean is significantly smaller than the median. This will indicate a distribution that is heavily skewed to the left. We can confirm this notion by graphing the Mullikan Charge data points.

The distribution of Mullikan charges peaks at around 0.175. However, further examination of the Figure 7 shows that the tails distribution is very (small peaks and valleys) appearance on the tails. Taking this into consideration, the Mullikan charge was concluded to be a left-skewed unimodal distribution.

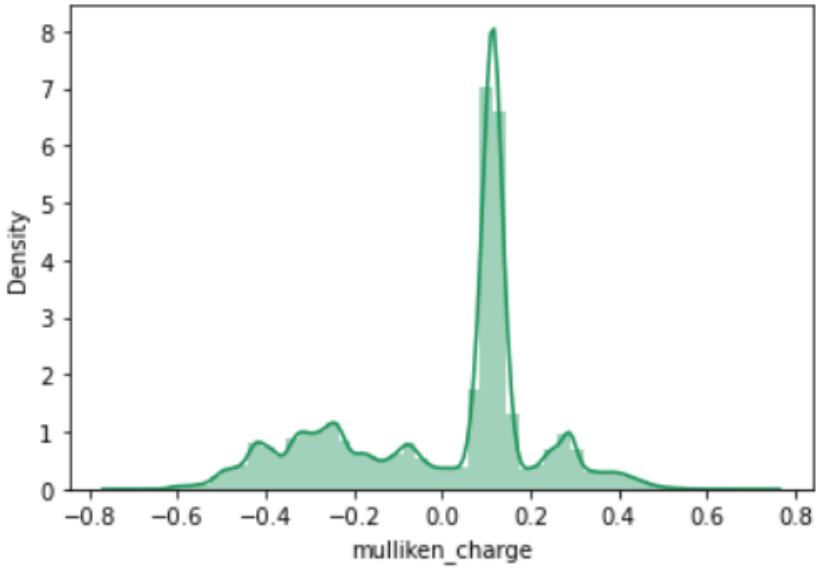


Figure 7: Smooth histogram to better visualize the distribution of Mullikan charge

As one can see from the distributions below, shown in Figure 8, atom index 0 and atom index 1 have a quad modal (4 peaks) distribution. Additionally, atom indexes 2, 3, and 4 exhibits a similar right-skewed multimodal distribution. The means of the distributions tend to decrease as the atom index increases. In other words, the Mulliken charges are higher for lower atom indices.

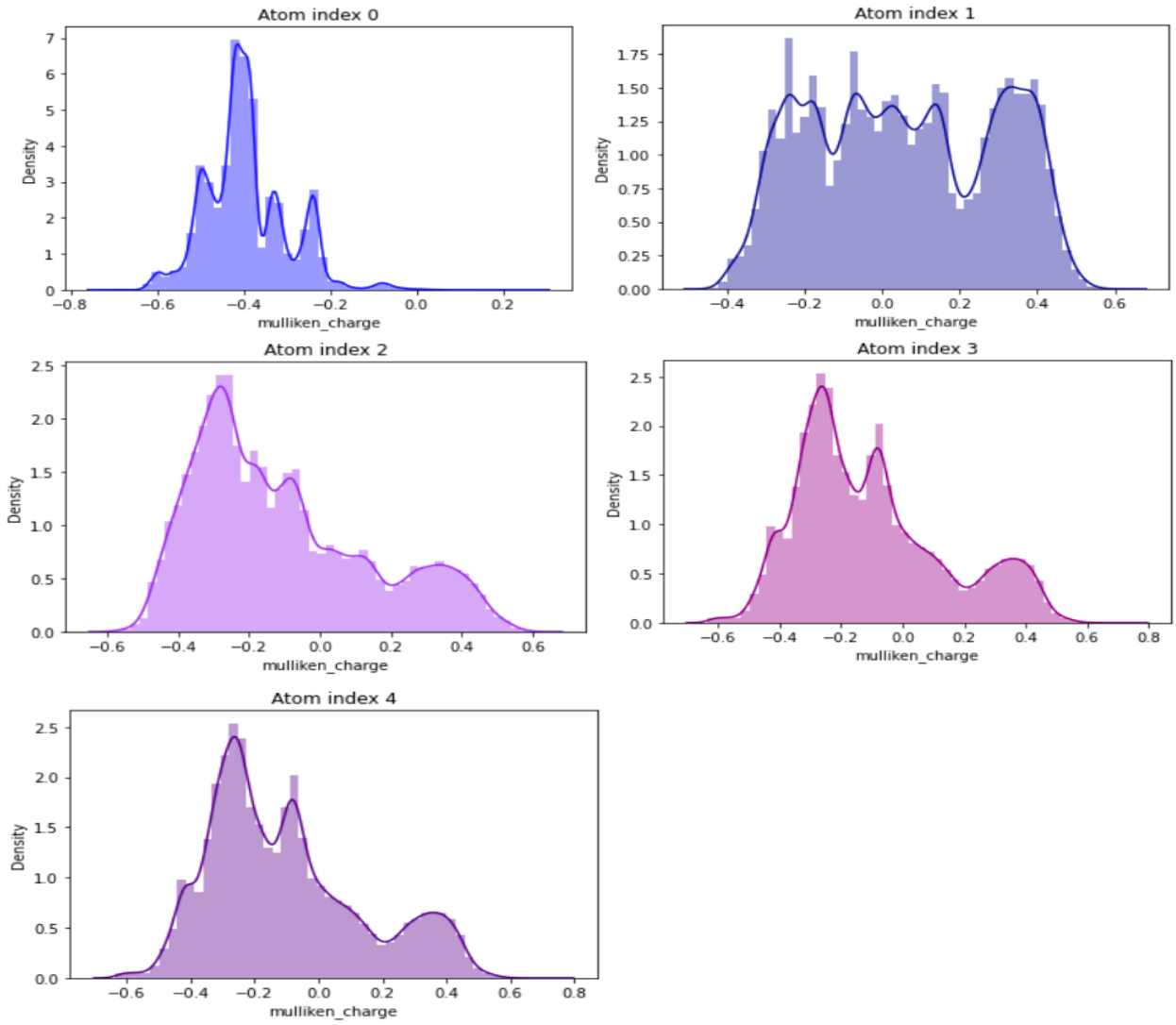


Figure 8: Visualizing distribution of Mullikan charge for each atom index

Potential Energy

Table 6: Potential Energy Summary Statistics

Variable	**Mean**	**Standard Deviation**	**Minimum**	**Q1 (25%)**	**Median (50%)**	**Q3 (75%)**	**Maximum**
Potential Energy	-410.95	39.84	-714.63	-438.00	-416.92	-387.22	-40.52

The summary statistics table for potential energy indicates a mean that is approximately equal to the median. Therefore, we can predict a normal distribution of data. Figure 9, confirms this notion by illustrating a normal distribution of data points when graphed. The distribution of potential energy of the molecules is approximately normal with a mean of around -40 shown in Figure 9.

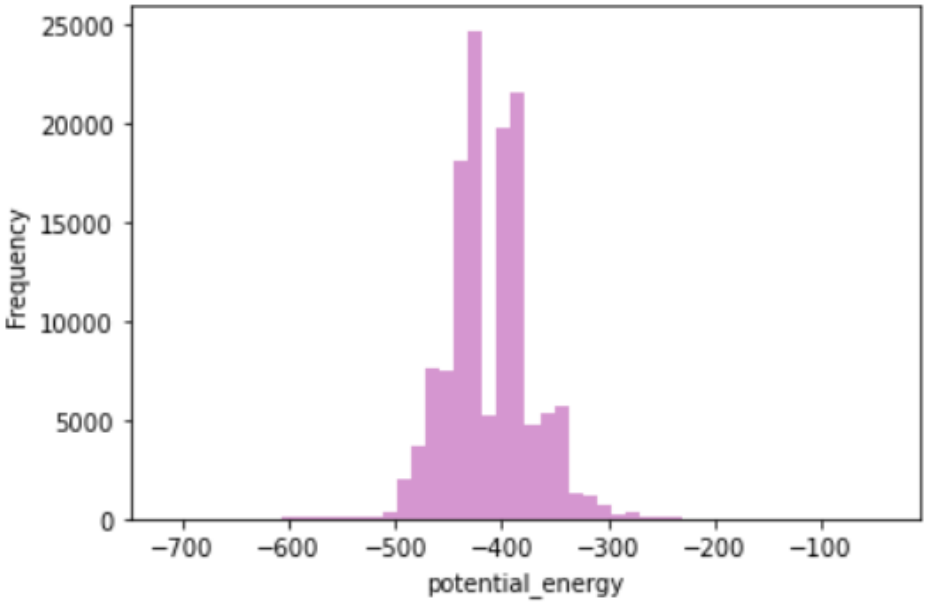


Figure 9: Histogram of the potential energy for all the molecules

Figure 10 illustrates the distributions of potential energy for each molecule type. One can see that the distributions are very different for each molecule type. However, the general distribution shape of each kind of molecule type is roughly normal.

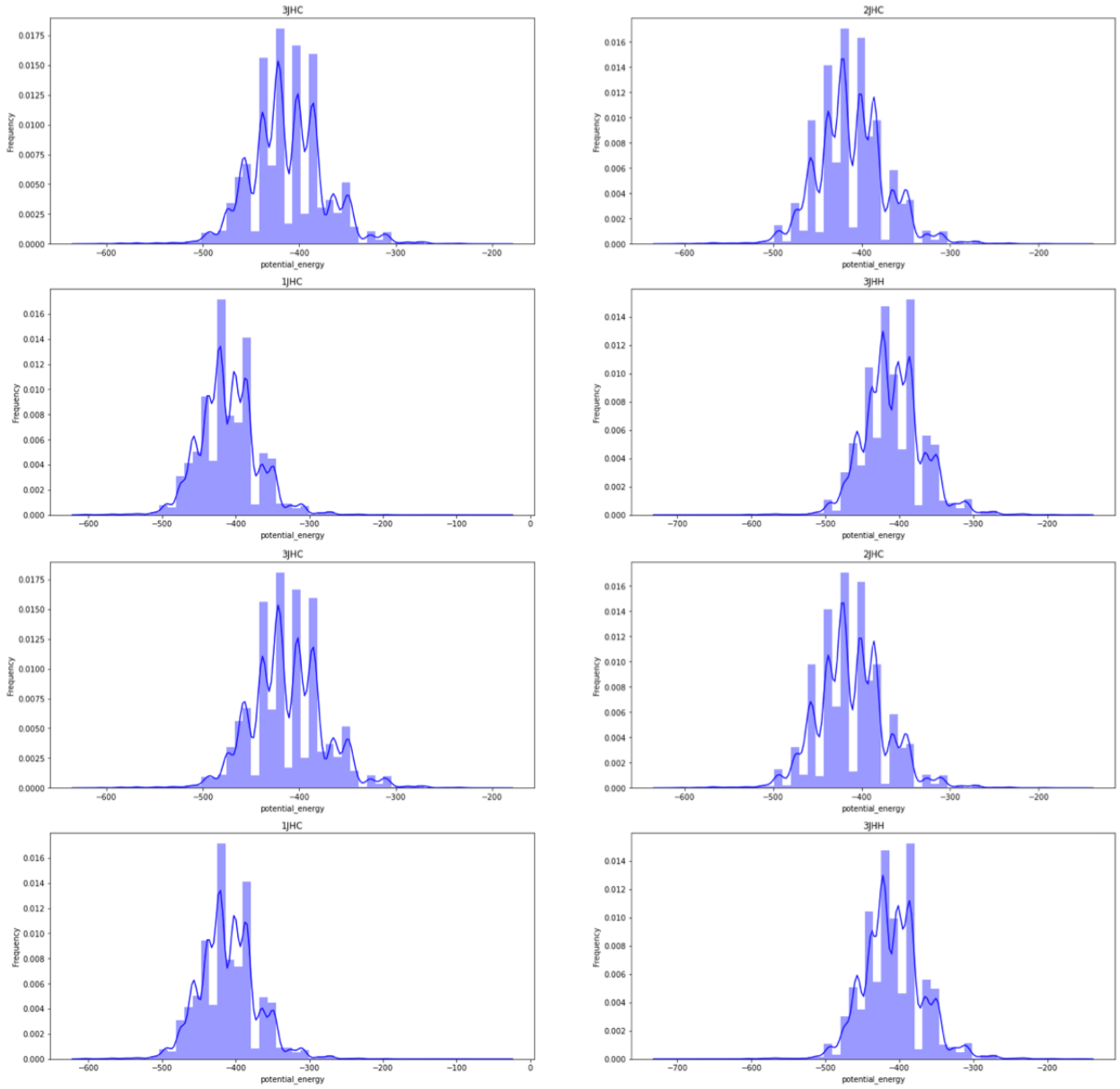


Figure 10: Six histograms visualizing the distribution of potential energy for each molecule type (J Coupling)

Magnetic Shielding Tensors

Table 7: Magnetic Shielding Tensors Summary Statistics

Variable	Mean	Standard Deviation	Minimum	Q1 (25%)	Median (50%)	Q3 (75%)	Maximum
XX	68.70	114.77	-3452.00	28.04	31.90	147.94	425.40
YX	-0.12	36.80	-758.29	-3.339	0.00	3.37	1080.02
ZX	0.08	34.05	-738.96	-2.544	0.00	2.61	662.04
XY	-0.013	36.63	-0.09	-3.204	0.00	3.23	1273.98
YY	65.34	106.35	-2146.60	27.18	31.73	136.02	425.46
ZY	-0.016	33.68	-652.32	-2.79	0.00	2.84	673.91
XZ	-0.08	34.77	-743.25	-2.484	0.00	2.55	863.36
YZ	0.04	34.15	-654.22	-2.80	0.00	2.87	738.41
ZZ	82.75	85.50	-948.60	27.35	33.79	141.68	556.88

According to Table 7, every variable has approximately the same mean and median values value, except for variables YY and ZZ, which indicates a normal distribution. We can also predict that the YY and ZZ variables are skewed to the right since the mean is greater than the median. However, further analyzing the table shows outliers when looking at the minimum and maximum values. Therefore, outliers need to be removed in this data set to illustrate the distribution of each variable within the dataset.

The outliers were removed from the magnetic shielding tensor data to accurately visualize the distributions of each magnetic shielding variation, shown in Figure 11. The distributions of magnetic shielding in each direction seem to be roughly linear with a mean of 0. Furthermore, some of graphs shows the distribution with steep, sharp slopes from the peak to the

tails (i.e., YX), while others have smooth, bulgy tails (i.e., ZY). All of variables have unique shapes in their distribution, however most of them can be considered under one of these two categories: steep sharp slopes or smooth bulgy tails.

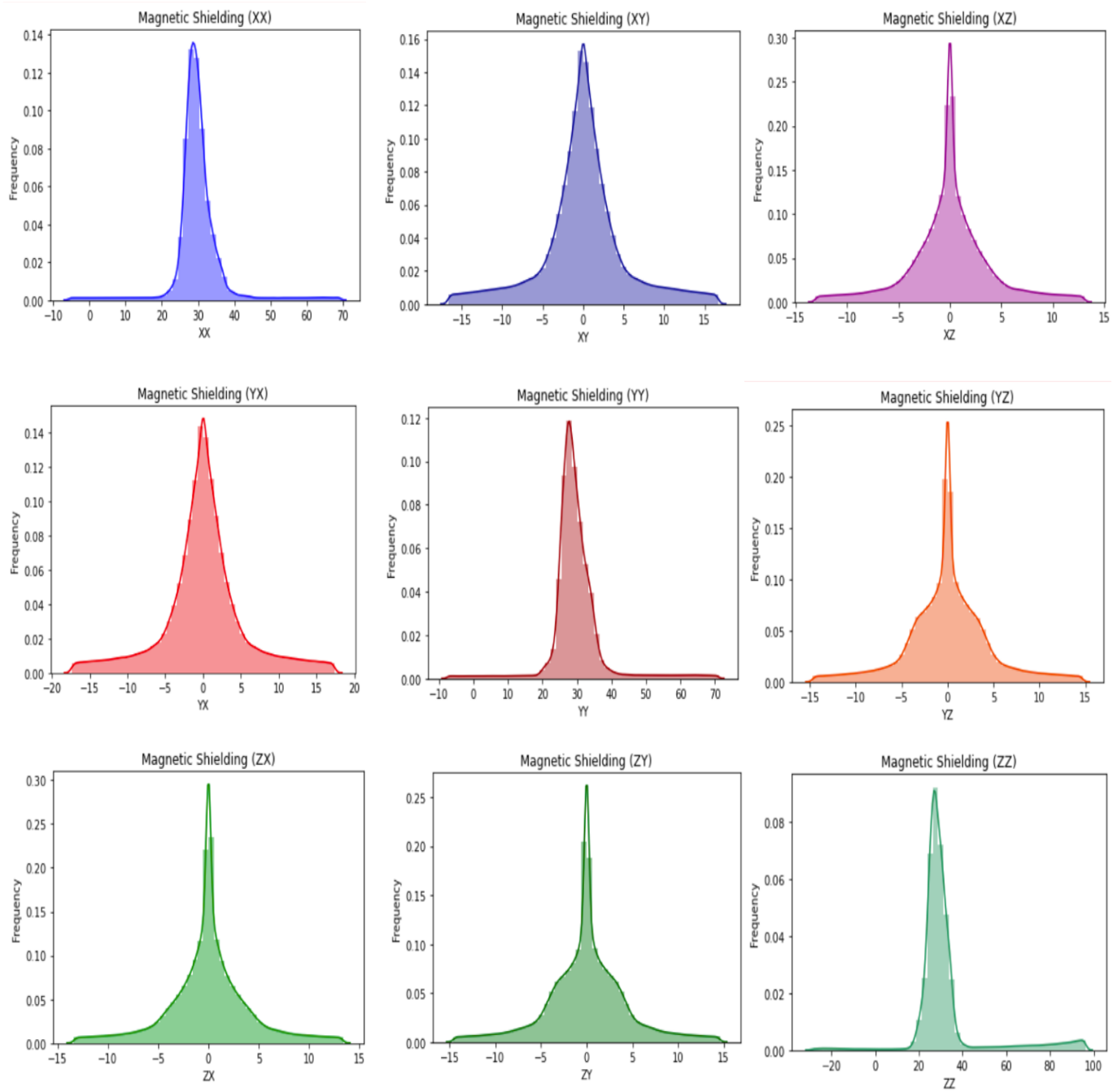


Figure 11: Visualizing the distribution of the magnetic shield tensors in each direction combination without outliers using smooth histogram

After examining the dataset, we can conclude that our hypothesis was correct, where each variable followed a normal distribution. The only exception to this was the data for magnetic shielding tensors, where we had to take out the outlier for the data to follow a normal distribution.

Data quality report

The goal for the initial exploratory report was to visualize the distributions of the critical variables from each file by itself and when grouped by their respective coupling types. Visualizations were created for all scenarios necessary to begin designing a machine learning model. All data were reviewed beforehand by CHAMPS for data accuracy (i.e., no missing data, valid data points). Therefore, outliers were deemed necessary to keep in the initial data collection report because each data point is significant in the overall goal of this project. If we removed outliers, our future model might not accurately predict molecular compounds with more extreme data points. However, the outliers were only removed for the magnetic shielding tensors to visualize their distributions accurately.

Additionally, a lack of scientific understanding of each variable may hinder the creation of an accurate model. In other words, one may struggle to figure out how each variable impacts one another, which could potentially make the algorithm difficult to code. In addition, the data sets contain a very large quantity of datapoints allowing the dataset to be resistant to outliers. Therefore, there were no issues in determining the distribution shape for each variable except for the magnetic shielding tensor data. However, the presence of outliers was present in the data, therefore two models can be made, one including outliers and one excluding outliers, to determine which method yields the best results. Moreover, an advanced understanding of

chemical knowledge is crucial when designing the model. Therefore, it is my recommendation to hire a chemical expert to assist with the coding team. An optimal candidate for this project would be a coder with machine learning experience with high chemical knowledge/understanding. Another limitation to this project is the multitude of data separated into different files. Combining the data into one file may save time and prove beneficial for this project's success; however, merging the data is not necessary. Furthermore, no other errors or concerns are apparent in the initial data exploration process.

References

About Us. About us - EPSRC website. (n.d.). Retrieved October 18, 2021, from https://epsrc.ukri.org/about/.

Bettenhausen, C. A. (n.d.). *Bruker installs world's first 1.2 GHz NMR*. Cen.acs.org. Retrieved October 18, 2021, from https://cen.acs.org/business/instrumentation/Bruker-installs-12-GHz-NMR/98/i19.

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A Primer for the epidemiologist. *American Journal of Epidemiology*. https://doi.org/10.1093/aje/kwz189

Blaber , M. (2020, August 21). *Dipole moments*. Chemistry LibreTexts. Retrieved October 18, 2021, from https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Physical_Properties_of_Matter/Atomic_and_Molecular_Properties/Dipole_Moments.

Bratholm, L. A., Gerrard, W., Anderson, B., Bai, S., Choi, S., Dang, L., Hanchar, P., Howard, A., Kim, S., Kolter, Z., Kondor, R., Kornbluth, M., Lee, Y., Lee, Y., Mailoa, J. P., Nguyen, T. T., Popovic, M., Rakocevic, G., Reade, W., … Glowacki, D. R. (2021). A community-powered search of machine learning strategy space to find NMR property prediction models. *PLOS ONE*, *16*(7). https://doi.org/10.1371/journal.pone.0253612

Champs. CHAMPS. (n.d.). Retrieved October 18, 2021, from https://champsproject.com/.

Davis, B. (2020, October 17). *Home*. MVOrganizing. Retrieved October 18, 2021, from https://www.mvorganizing.org/are-there-6-dimensions/.

Facelli, J. C. (2011, May). *Chemical shift tensors: Theory and application to molecular structural problems*. Progress in nuclear magnetic resonance spectroscopy. Retrieved October 18, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3058154/%5C/.

Gokhale, P., Ding, Y., Propson, T., Winkler, C., Leung, N., Shi, Y., Schuster, D. I., Hoffmann, H., & Chong, F. T. (2019). Partial compilation of variational algorithms for noisy intermediate-scale quantum machines. *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*. https://doi.org/10.1145/3352460.3358313

Graph networks as a universal machine learning framework ... (n.d.). Retrieved October 18, 2021, from https://materialsvirtuallab.org/pubs/10.1021_acs.chemmater.9b01294.pdf.

Hatzakis, E. (2018). Nuclear magnetic resonance (NMR) spectroscopy in Food Science: A comprehensive review. *Comprehensive Reviews in Food Science and Food Safety*, *18*(1), 189–220. https://doi.org/10.1111/1541-4337.12408

Kaseman, D. (2020, August 21). *J-coupling (scalar)*. Chemistry LibreTexts. Retrieved October 18, 2021, from https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Spectroscopy/Magn

etic_Resonance_Spectroscopies/Nuclear_Magnetic_Resonance/NMR_-_Theory/NMR_Interactions/J-Coupling_(Scalar).

Mullen, S. (2017, October 31). *Frequently asked questions: NMR Structural Biology Facility & Biophysical Core Facility*. NMR Structural Biology Facility Biophysical Core Facility. Retrieved October 18, 2021, from https://health.uconn.edu/structural-biology/frequently-asked-questions/.

National Research Council. (n.d.). *"Beyond the molecular frontier: Challenges for chemistry and Chemical Engineering" at nap.edu*. National Academies Press: OpenBook. Retrieved October 18, 2021, from https://www.nap.edu/read/10633/chapter/6.

Russom, P. (2011). BIG DATA ANALYTICS. *TDWI BEST PRACTICES REPORT*.

Ukri. (n.d.). *Gateway to research (GTR) - explore publicly funded research*. GtR. Retrieved October 18, 2021, from https://gtr.ukri.org/.

Vakkuri, V., & Kemell, K.-K. (2019). Implementing AI ethics in practice: An empirical evaluation of the resolvedd strategy. *Lecture Notes in Business Information Processing*, 260–275. https://doi.org/10.1007/978-3-030-33742-1_21

Weinbaum, C., Landree, E., Blumenthal, M., Piquado, T., & Gutierrez, C. (2019). Ethics in scientific research: An examination of ethical principles and emerging topics. https://doi.org/10.7249/rr2912

Welcome to Amerian Chemical Society. American Chemical Society. (n.d.). Retrieved October 18, 2021, from https://www.acs.org/content/acs/en.html.