# Prediction of $^1$H and $^{13}$C NMR Chemical Shifts of Small Molecules Using Machine Learning

by

Zinat Sayeeda

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

# Abstract

For more than 70 years, chemists have used Nuclear Magnetic Resonance (NMR) spectroscopy to characterize the atomic structure and dynamics of molecules. Key to performing the NMR analysis of almost any molecule is a process called “chemical shift assignment”. This involves matching specific peaks or chemical shifts in the NMR spectrum with specific atoms in the molecule. Using a variety of NMR techniques, chemists have performed chemical shift assignments for hundreds of thousands of organic compounds over the past few decades. However, the chemical shift assignment process can be time-consuming and difficult. It can also be fraught with errors. Because of these challenges, NMR spectroscopists have long been interested in predicting NMR chemical shifts. Having accurate methods to predict $^{1}H$ (hydrogen) and $^{13}C$ (carbon) NMR chemical shifts of organic molecules would greatly improve the speed and accuracy with which chemical shift assignments could be made. Over the past two decades a variety of methods, ranging from *Ab initio* approaches to database search methods to machine learning (ML) techniques have been applied to improve chemical shift prediction. The most promising of these are the ML methods. However, most ML methods do not achieve the level of accuracy required for consistent chemical shift assignments of small molecules, nor do they properly handle diasterotopic protons, solvent effects, pH effects, or alternate chemical shift referencing schemes. In this thesis, I will describe my efforts to develop an ML-based NMR chemical shift predictor that can accurately predict $^{1}H$ and $^{13}C$ NMR chemical shifts while at the same time accommodating diasterotopic protons, solvent effects, pH effects, and alternate chemical shift referencing schemes. In developing this predictor, called NMRPred, I assembled and curated a large dataset of carefully assigned and carefully referenced experimental $^{1}$H and $^{13}$C NMR assignments from 953 molecules. I also tested a variety of feature extraction and ML methods to develop two separate predictors, one for $^{1}$H and another for $^{13}$C chemical shifts. The best performing $^{1}$H predictor, which used a Random Forest Regressor, obtained a Mean Absolute Error (MAE) of 0.11 ppm with a standard deviation of 0.18 ppm on a validation set of 272 $^{1}$H assignments and MAE of 0.36 ppm with a standard deviation of 0.56 ppm on a second validation set of 442 $^{1}$H assignments. The best performing $^{13}$C predictor, which used a Gradient Boost Regressor, obtained an MAE of 2.94 ppm with a standard deviation of 4.2 ppm on a validation set of 1087 $^{13}$C assignments and MAE of 6.65 ppm with a standard deviation of 8.65 ppm on a validation set of 653 $^{13}$C assignments. On the first validation set the $^{1}$H shift predictor outperformed other chemical shift predictors in terms of its accuracy (MAE), and its ability to handle diasterotopic protons, solvent effects, pH effects, and alternate chemical shift referencing schemes. Unfortunately, the $^{13}$C shift predictor did not match the performance of the most recent and widely used $^{13}$C shift predictors.  In this thesis I discuss some of the reasons why this may have happened and I present evidence that suggests that by using a larger and more varied dataset that it would be possible to improve the performance of both the $^{1}$H and $^{13}$C shift predictors.

# Preface

This thesis is an original work by Zinat Sayeeda. Dr. David Wishart, my supervisor, provided direction for the research described in this thesis and helped with editing the thesis document. Dr. Brian Lee, Dr. An Chi Guo and Dr. Manoj Rout, all of whom are research associates in Dr. Wishart's laboratory, helped with the NMR data collection and curation, as detailed in chapters 2 and 3. Dr. Russ Greiner provided advice regarding the ML methods and feature selection process mentioned in Chapters 2 and 3. A portion of the material described in Chapter 2 was incorporated into a paper called "NP-MRD: the Natural Products Magnetic Resonance Database", which was published in Nucleic Acids Research in January 2022. Chapters 2 and 3 will be modified at a future date so that they could appear in a published paper on ML and chemical shift prediction.

# Acknowledgements

I would like to convey my profound gratitude to my supervisor, Dr. David Wishart, for his tremendous help and patient supervision throughout my research program. I would not have been able to resume my education or my plan to change my career path without his encouragement. He has provided me with wise counsel, insightful observations, and many life lessons for which I am most grateful and will remember for the rest of my life. His generosity along with his family's kindness toward me during some of my most trying circumstances, will always be remembered. Dr. Wishart provided tremendous motivation for me, gave constructive criticism on my thesis, had faith in my work, and genuinely cared about my research. I truly hope that my efforts were worthy of his kindness and caring.

In addition, I want to express my gratitude to my nephews Tanvir and Touquir Sajed for guiding me into this new field of study, something that was entirely unrelated to my previous area of expertise. I also wish to thank the members of Dr. Wishart's lab, including Drs. Brian L. Lee, Dipanjan Bhattacharyya, An Chi Guo, Manoj Rout, Mark Berjanskii, Vasuk Gautam and Marcia LeVatte along with Mr. Xuan Cao and Ms. Galina Durant, for their help and assistance.

I also want to express my sincere thanks to my family for their love and confidence in me. Last, but not least, I would like to express my gratitude to my friends Daniel Caminhas, Sheila Schoepp, Colton MacLean and Veni Suresh for being there for me throughout my thesis journey, sharing in both my successes and failures and for offering support during my most challenging times.

# Contents

CHAPTER 1: INTRODUCTION ........................................................................................ 1

PRINCIPLES OF THE NMR PHENOMENON .................................................................... 7

STRUCTURE ELUCIDATION BY NMR ............................................................................ 12

CHEMICAL SHIFT PREDICTION ................................................................................... 17

*Ab initio Techniques* ............................................................................................ *18*

*Rule-based Techniques* ......................................................................................... *19*

*Database Approaches* ........................................................................................... *21*

*Machine Learning Methods* .................................................................................... *26*

THESIS HYPOTHESIS AND AIMS ................................................................................ 40

THESIS OUTLINE ..................................................................................................... 41

CHAPTER 2: APPLICATION OF MACHINE LEARNING TO THE PREDICTION OF $^1$H CHEMICAL SHIFTS OF SMALL ORGANIC MOLECULES ................................................... 42

INTRODUCTION ........................................................................................................ 42

PROBLEM DEFINITION .............................................................................................. 43

TERMINOLOGY AND BACKGROUND ............................................................................ 43

METHODOLOGICAL OUTLINE ..................................................................................... 45

PERFORMANCE EVALUATION METRIC ......................................................................... 46

THE INITIAL $^1$H NMR CHEMICAL SHIFT DATASET ...................................................... 47

*The Training Dataset* ........................................................................................... *48*

*The Holdout Dataset* ............................................................................................ *49*

*Atom and Chemical Shift Labeling* ......................................................................... *50*

*Correcting Diastereotopic Proton Assignments* ........................................................ *63*

Feature Identification ........ 66

Methodology ........ 75

*Support Vector Regressor* ........ *76*

*Random Forest Regressor (RFR)* ........ *77*

*XGBoost Regressor* ........ *78*

*CatBoost Regressor* ........ *79*

Coding Details ........ 80

Results and Discussion ........ 84

*Performance Comparison Against Popular Methods* ........ *91*

$^1$H NMR Chemical Shift Predictions in $CDCl_3$ and DMSO ........ 104

Conclusion ........ 106

CHAPTER 3: APPLICATION OF MACHINE LEARNING TO THE PREDICTION OF $^{13}$C CHEMICAL SHIFTS OF SMALL ORGANIC MOLECULES ........ 110

Introduction ........ 110

Problem Definition ........ 112

Methodological Outline ........ 113

Performance Evaluation Metric ........ 114

The Initial $^{13}$C NMR Chemical Shift Dataset ........ 114

*The Training Dataset* ........ *115*

*The Holdout Dataset* ........ *117*

*Atom and Chemical Shift Labeling* ........ *117*

Feature Identification ........ 125

Methodology ........ 131

Coding Details ........ 132

RESULTS AND DISCUSSION ........................................ 134

*Performance Comparison Against Popular Methods* ........................................ *142*

CONCLUSION ........................................ 150

CHAPTER 4: SUMMARY AND FUTURE DIRECTIONS ........................................ 154

*Contributions* ........................................ *157*

*Future Work* ........................................ *158*

REFERENCES ........................................ 161

# List of Tables

TABLE 1.1: $^{1}$H CHEMICAL SHIFT REFERENCES AND RESPECTIVE CHEMICAL SHIFT DIFFERENCES RELATIVE TO DSS. REFERENCE: WISHART ET AL. (30)....................................................24

TABLE 1.2: $^{13}$C CHEMICAL SHIFT REFERENCES AND RESPECTIVE CHEMICAL SHIFT DIFFERENCES RELATIVE TO DSS. REFERENCE: WISHART ET AL (30).....................................................25

TABLE 1.3: A TABLE ILLUSTRATING THE PERFORMANCE OF VARIOUS ML BASED PREDICTORS FOR $^{1}$H AND $^{13}$C CHEMICAL SHIFT PREDICTION. ..........................................................................40

TABLE 2.1: THE LIST OF THE CHEMICAL FUNCTIONAL GROUPS AND CHEMICAL SUBSTRUCTURES (WRITTEN IN SMART STRINGS) THAT WERE USED TO ANNOTATE THE FEATURE SPACE...........73

TABLE 2.2: THE SEQUENCE OF ALL FEATURES, USED TO TRAIN THE DIFFERENT ML MODELS. ........74

TABLE 2. 3: PAIRED T-TEST RESULT AMONG RFR, XGBR, CBR AND SVR MODELS......................85

TABLE 2.4: THE CHEMICAL CLASS DISTRIBUTION IN THE NMRPRED TRAINING DATASET, THE 1ST HOLDOUT DATASET (FROM HMDB) AND THE 2ND HOLDOUT DATASET (FROM NP-MRD) AS INDICATED BY MEMBERSHIP IN CHEMICAL SUBCLASSES. ......................................................103

TABLE 3.1: THE LIST OF THE CHEMICAL FUNCTIONAL GROUPS AND PARTICULAR CHEMICAL SUBSTRUCTURES CONSIDERED IN THE DIFFERENT MOLECULAR NEIGHBORHOOD WRITTEN OUT AS SMART STRINGS. 130

TABLE 3. 2: PAIRED T-TEST RESULT AMONG GBR, SVR, XGBR AND RFR MODELS 138

# List of Figures

Figure 1.1: An NMR spectrometer (left side) and its computer terminal (right side)......2

Figure 1.2: A schematic diagram illustrating the main components an NMR spectrometer, including the superconducting magnet (on the left) and the RF probe (on the right). ......3

Figure 1.3: A schematic diagram of an NMR spectrometer illustrating the magnet assembly (right), the probe (inside the magnet), the amplifier and console/transceiver. The measured FID is shown on the lower left along with the resulting Fourier transformed NMR spectrum. ......4

Figure 1.4: $^1$H NMR spectrum of ethanol. The methyl (-CH3) and methylene (-CH2-) groups couple with each other which results in a triplet multiplet for the -CH3 signal and a quartet for the  -CH2-. The -OH group does not couple with any other protons and as a result the signal for the -OH is singlet. ......8

Figure 1.5: An image illustrating the spin of the protons in ethanol (left side). The right side shows the Free Induction Decay (FID) plotted against the time axis, arising from the spinning protons after they have been perturbed by an external RF signal......10

Figure 1.6: An example of a Fourier transformed NMR spectrum arising from the hydrogens (protons) in ethanol. This spectrum is characterized by sharp peaks plotted against their resonance frequencies (chemical shifts). ......11

Figure 1.7: An example of $^1$H NMR spectrum for a compound with the chemical formula: $C_9H_{10}O_2$. 13

Figure 1.8: Approximate range of proton chemical shift values for some common functional groups. 14

Figure 1.9: Individual possible substructure structures derived from the NMR peaks in [Figure 1.7]. 15

Figure 1.10: The derived structure (3-Phenyl-propionic acid) from the $^1$H NMR spectrum in [Figure 1.7]. 15

Figure 1.11: The HOSE CODE for atom number 1 up to three spheres. The dashed lines represent the path generated for different spheres from atom number 1. The 1st, 2nd and 3rd spheres from atom number 1 are represented by red, green and purple dashed lines, respectively. 23

Figure 1.12: A basic flow chart outlining the Machine Learning process. 31

Figure 2.1: A table of $^1$H chemical shift assignments for hydrogen atoms connected to carbon atoms in 2-Isopropylmalic acid (HMDB00402). The numbers in the Atom column refer to carbon atom positions drawn in Figure 2.2. 52

FIGURE 2.2: THE ATOM-NUMBERED STRUCTURAL IMAGE OF 2-ISOPROPYLMALIC ACID (HMDB00402) WITH HEAVY ATOMS NUMBERED IN THE FIGURE. THESE NUMBERS ARE USED TO MAP THE MEASURED $^1$H CHEMICAL SHIFTS ASSIGNMENTS IN FIGURE 2.1............................52

FIGURE 2.3: THE STRUCTURE OF 2-ISOPROPYLMALIC ACID (HMDB00402) AS DOWNLOADED FROM THE HMDB SERVER WITH ATOM NUMBERING GENERATED VIA MARVIN SKETCH...................53

FIGURE 2.4: THE FINAL CHEMICAL SHIFT ASSIGNMENT FILE FOR 2-ISOPROPYLMALIC ACID (HMDB00402) USING THE PROCESS OF MANUAL ATOM. AS SHOWN HERE, ONE MUST BEGIN BY MAPPING THE ATOMS OF FIGURE 2.2 AND FIGURE 2.3, THEN ONE MUST REPLACE THE ATOM COLUMN IN FIGURE 2.1 WITH THE ATOMIC POSITIONS FROM FIGURE 2.3. ATOMIC FEATURES FOR THE ML MODEL MUST THEREFORE CALCULATED FROM THE STRUCTURE IN FIGURE 2.3 SINCE THE STRUCTURE FROM FIGURE 2.2 EXISTS ONLY AS AN IMAGE FILE, NOT AN SDF FILE. 55

FIGURE 2.5: ERYTHRITOL (LEFT SIDE) AND D-THREITOL (RIGHT SIDE). THEY BOTH HAD SAME MOLECULAR SDF FILES.........................................................................................56

FIGURE 2.6: MOLECULAR STRUCTURE OF GLYCERALDEHYDE (HMDB01051) IN THE STORED IMAGE FILE. ..........................................................................................................................57

FIGURE 2.7: MOLECULAR STRUCTURE OF GLYCERALDEHYDE (HMDB01051) DRAWN USING MARVIN SKETCH (LEFT SIDE). THE TABLE OF CHEMICAL SHIFT ASSIGNMENTS CORRESPONDS TO THE ATOM NUMBERING IN FIGURE 2.6 (RIGHT SIDE). ............................................................57

Figure 2.8: The spectrum of Glyceraldehyde (HMDB01051) as displayed in HMDB's Jpectra viewer. The auto assignment function could not confirm the correct value of that hydrogen atom 2.......... 58

Figure 2.9: The atom numbering difference between the BMRB (left image) and the atom numbering system the HMDB (middle image) for Glyceraldehyde. Atom number mapping between the HMDB SDF file and the BMRB SDF file for Glyceraldehyde (right image). .......... 59

Figure 2.10: Carbonyl function of aldehydes and ketones creating a reversible equilibrium with a hydrate with the present of acid or base. The aldehyde converts to a diol and vice versa.......... 60

Figure 2.11: Structure of the SDF containing HC(OH)2 (left side) and the structure of the SDF containing HC=O (right side). .......... 60

Figure 2.12: MNOVA could not assign the chemical shift value for hydrogen atom 10 from the provided spectrum.......... 61

Figure 2.13: Chemical structure of Pipecolic Acid (HMDB00070) and atom numbering scheme in the HMDB image file.......... 62

Figure 2.14: Atom numbering system for Pipecolic Acid in the molecular SDF file (left side) and the image file for the chemical shift assignments table (right side) corresponding to the Figure 2.13.......... 62

FIGURE 2.15: 1,1-DIMETHYLETHYLEN (LEFT SIDE). TWO OLEFINIC PROTONS A**H** = **H**B HAVE IDENTICAL CHEMICAL SHIFTS IN THE MIRROR PLANE OF SYMMETRY (RIGHT SIDE) ..................63

FIGURE 2.16: AFTER REPLACING THE METHYL GROUP'S HYDROGENS WITH A DIFFERENT HEAVY ATOM E, THE MOLECULE IS NO LONGER SYMMETRICAL. ........................................................65

FIGURE 2.17: IF $_{A}$**H** HAS CHEMICAL SHIFT VALUE **X** AND **H**B HAS **Y**, THOSE VALUES CAN SWITCH BETWEEN EACH OTHER. ..........................................................................................65

FIGURE 2.18: TWO HYDROGEN ATOMS (OR THE TWO METHYL: $CH_3$ GROUPS) IN DIMETHYLETHYLENE ARE ON THE SAME SIDE OF THE DOUBLE BOND INDICATING A CIS-DIMETHYLETHYLENE (LEFT SIDE). WHEN THE METHYL GROUPS ARE ON THE OPPOSITE SITE OF THE DOUBLE BOND THIS INDICATES A TRANS-DIMETHYLETHYLENE (RIGHT SIDE). ..................66

FIGURE 2.19: ATOMIC FEATURES THAT CAN BE CALCULATED FROM A MOLECULAR STRUCTURE USING THE CDK PACKAGE. ALL VALUES ARE NUMERIC. ........................................................70

FIGURE 2.20: IF THE ATOM OF INTEREST IS #10, THE NEAREST THREE ATOMS ARE #3, #4, AND #6 (LEFT SIDE). IF THE ATOM OF INTEREST IS #10 (RIGHT SIDE), THEN LOOKING AT ATOMS UP TO FOUR BONDS AWAY WE CAN SEE THAT AN OH GROUP APPEARED 2 TIMES, A C=O GROUP APPEARED 1 TIME, WHILE COOH, AND NH2 GROUPS NEVER APPEARED AND SO ON. ..............74

FIGURE 2.21: AN OPTIMISTICALLY SKEWED ESTIMATE OF THE MODEL PERFORMANCE RESULTS FROM COMBINING HYPER-PARAMETER ADJUSTMENT AND MODEL EVALUATION IN THE SAME CV LOOP.........................................................................................................................82

FIGURE 2.22: AN EXPLANATION OF NESTED CV. THE OUTER LOOP IS USED TO ESTIMATE THE MODEL ERROR WHILE THE INNER LOOP IS USED TO SEARCH FOR HYPER-PARAMETERS. ......................83

FIGURE 2.23: THE SIZE OF THE DATA SAMPLES IN THE INNER AND OUTER LOOPS OF THE NESTED CV AS USED IN OUR EXPERIMENT. ........................................................................................84

FIGURE 2.24: COMPARISON OF THE 5-FOLD CROSS VALIDATION PERFORMANCE FOR ALL 4 ALGORITHMS. RANDOM FOREST DOES THE BEST WITH THE SMALLEST MEAN ABSOLUTE ERROR OF 0.12 PPM. THE 5-FOLD CROSS VALIDATION IS CARRIED ON THE SAME TRAINING SET SHUFFLED RANDOMLY WITH THE SAME RANDOM SEED.........................................................86

FIGURE 2.25: PREDICTED $^1$H CHEMICAL SHIFT VALUES VS OBSERVED $^1$H CHEMICAL SHIFTS VALUES ON THE TRAINING DATASET SET USING RFR. TESTING ON THE TRAINING SET GAVE AN MAE OF 0.03 PPM.........................................................................................................................87

FIGURE 2.26: PREDICTED $^1$H CHEMICAL SHIFT VALUES VS OBSERVED $^1$H CHEMICAL SHIFT VALUES ON THE TRAINING DATASET SET USING THE XGBOOST REGRESSOR. TESTING ON THE TRAINING SET GAVE A MEAN ABSOLUTE ERROR OF 0.02 PPM. OVERFITTING IS MORE VISIBLE COMPARED TO THE RESULTS SHOW IN FIGURE 2.25. ....................................................................88

FIGURE 2.27: THE SMILES STRING FOR BENZALDEHYDE (HMDB0006115) WAS PROVIDED IN THE INPUT SECTION TOGETHER WITH THE SELECTED SOLVENT OPTION. CHEMAXON'S JCHEM CONVERTS THE SMILES STRING INTO A 2D STRUCTURE......................................................90

FIGURE 2.28: THE PREDICTED $^1$H CHEMICAL SHIFT VALUES FROM THE INPUT IN FIGURE 2.27 TOGETHER WITH THE 3D STRUCTURE OF THE MOLECULE. .....................................................90

FIGURE 2.29: A SCATTER PLOT OF THE CORRELATION BETWEEN THE OBSERVED VS PREDICTED $^{1}$H CHEMICAL SHIFTS IN THE FIRST HOLDOUT DATASET, FOR NMRPRED (WHICH USES THE RFR). ....................92

FIGURE 2.30: A SCATTER PLOT OF THE OBSERVED VS PREDICTED $^{1}$H CHEMICAL SHIFTS IN THE FIRST HOLDOUT DATASET, USING PREDICTIONS FROM MNOVA. ....................93

FIGURE 2.31: A SCATTER PLOT OF THE OBSERVED VS PREDICTED $^{1}$H CHEMICAL SHIFTS IN THE FIRST HOLDOUT DATASET USING PREDICTIONS FROM NMRSHIFTDB2. ....................94

FIGURE 2.32: THE OBSERVED $^{1}$H CHEMICAL SHIFT VALUE OF HYDROGEN ATOM #14 FOR BENZALDEHYDE (HMDB0006115) AND THE PREDICTED $^{1}$H CHEMICAL SHIFT VALUES WITH DIFFERENT PREDICTORS. ....................96

FIGURE 2.33: THE OBSERVED $^{1}$H CHEMICAL SHIFT VALUE OF HYDROGEN ATOM #19, #18 IN MEVALONIC ACID (HMDB0000227 AND THE PREDICTED $^{1}$H CHEMICAL SHIFT VALUES FOR DIFFERENT CHEMICAL SHIFT PREDICTORS. ATOMS #18 AND #19 ARE DIASTEREOTOPIC. .........96

FIGURE 2.34: A COMPARISON OF THE MEAN ABSOLUTE ERRORS (MAES) FOR EACH OF THE 36 MOLECULES IN THE FIRST HOLDOUT DATASET AMONG THE DIFFERENT PREDICTORS (NMRPRED, MNOVA, NMRSHIFTDB2 AND THE DFT PREDICTOR). ....................97

FIGURE 2.35: A COMPARISON OF MEAN ABSOLUTE ERRORS FOR EACH OF THE 36 MOLECULES IN THE HOLDOUT DATASET AMONG EACH OF THE FOUR PREDICTORS: NMRPRED, MNOVA, NMRSHIFTDB2 AND THE DFT PREDICTOR. ....................98

Figure 2.36: A comparison of the prediction of the diastereotopic hydrogens among the NMRPred, MNOVA, NMRShiftDB2 and DFT predictors. Red highlights indicate that the diastereotopic hydrogens were not detected. .............................99

Figure 2.37: Performance comparison between 1st and 2nd holdout datasets among the four predictors: NMRPred, MNOVA, NMRShiftDB2 and the DFT predictor. 100

Figure 2.38: The class distribution of chemicals seen in the two holdout datasets compared to the NMRPred training dataset. ............................................................102

Figure 2.39: The three NP-MRD molecules (NP0006813, NP0040444 and NP0035870) and the variation in the molecular structures in the training dataset of the same subclass compounds. .........................................................................................103

Figure 2.40: The simple linear equation that can be used to predict the $^{1}H$ chemical shift values of hydrogen atoms for molecules dissolved in CDCl3 relative to those dissolved in water. .........................................................................................105

Figure 2.41: The simple linear equation that can be used to predict the $^{1}H$ chemical shift values of hydrogen atoms of molecules dissolved in DMSO relative to those dissolved in water. .........................................................................................105

Figure 3.1: A table of $^{13}C$ chemical shift assignments for L-Isoleucine (HMDB00172) in the HMDB. The numbers in the peak column refer to carbon atom positions drawn in same image. 119

Figure 3.2: The structure of L-Isoleucine (HMDB00172) as downloaded from the HMDB server with atom numbering generated via MarvinSketch. 120

Figure 3.3: The final chemical shift assignment file for L-Isoleucine (HMDB00172) using manual remapping. As shown here, one must begin by mapping the atoms from Figure 3.1 and Figure 3.2, then one must replace the Peak column in Figure 3.1 with the atomic positions from Figure 3.2. Atomic features for the ML model must therefore be calculated from the structure in Figure 3.2 since the structure from Figure 3.1 exists only as an image file, not as an SDF file. 121

Figure 3.4: The SDF file for Thioacetamide (bmse000781) from the BMRB (left side) and the same structure from PubChem (right side). 122

Figure 3.5: The SDF file for Barbituric acid (bmse000346) from the BMRB (left side) and the same structure from PubChem (right side). 123

Figure 3.6: The aldehyde carbon chemical shift values in D,L-Glyceraldehyde (bmse000225) and in D-Ribulose 5-phosphate (bmse000278) in the collected training dataset (top left corner) and the predicted chemical shift values by MNOVA for the same compound's aldehyde carbons. 124

Figure 3. 7: $^{13}$C chemical shift ranges for various functional groups in organic compounds. 125

Figure 3.8: An illustration of carbon bond hybridization (left side) and the meaning of α, β and δ substituents (right side). 127

Figure 3.9: Chemical structure of chorismic acid (bmse000075). If the targeted carbon atom # is 6 (C:6), the zero, one and two bond neighborhood atoms are circled with red, yellow and green color. As an example, C:6 is not a carboxylic (-COOH) carbon. One bond away from C:6, the carboxylic acid appears 1 time and two bonds away, it appears 0 times. 129

Figure 3.10: A schematic diagram illustration the process of finding the best parameter for various ML models using 5-fold cross validation and selecting the best model using the test/holdout data. 134

Figure 3.11: Scatter plots comparing the true vs. predicted values on the training dataset for the four algorithms. The GBR and SVR models appear to be the most well learned models. The chart below the graph images, shows test and training error ratios for each model. 136

Figure 3.12: A bar graph showing the performance of each of the 4 ML models for the holdout dataset. The Gradient Boost Regressor, which is represented by the green bar, shows the best performance. The chart below the graph shows performance of the four models for both the 5-fold cross validation and on the holdout dataset. 139

Figure 3.13: A bar graph showing the performance of the four ML models on the 2nd holdout dataset. The Gradient Boost Regressor, indicated by the green bar, shows the best performance. The performance of the other three algorithms changed slightly compared to the results shown in Figure 3.13. The chart below

THE GRAPH SHOWS THE PERFORMANCE COMPARISON OF THE FOUR MODELS BETWEEN THE TWO HOLDOUT DATASETS. 140

FIGURE 3.14: THE SMILES STRING FOR BETA-ALANINE (HMDB0000056) WAS PROVIDED IN THE INPUT TOGETHER WITH A SELECTED CHEMICAL SHIFT REFERENCE OPTION. CHEMAXON'S JCHEM CONVERTS THE SMILES STRING INTO A 2D STRUCTURE. 141

FIGURE 3.15: THE PREDICTED $^{13}$C CHEMICAL SHIFT VALUES FROM THE INPUT IN FIGURE 3.15 TOGETHER WITH THE 3D STRUCTURE OF THE MOLECULE. 142

FIGURE 3.16: SCATTER PLOTS SHOWING THE EXPERIMENTALLY MEASURED $^{13}$C CHEMICAL SHIFTS VERSUS PREDICTED $^{13}$C CHEMICAL SHIFTS FOR NMRPRED-CARBON (TOP LEFT CORNER), MNOVA (TOP RIGHT CORNER) AND NMRSHIFTDB2 (LOWER LEFT CORNER), RESPECTIVELY. THE UPPER CHART AT THE LOWER RIGHT CORNER SHOWS THE CHEMICAL SUBCLASSES FOR THE COMPOUNDS WHERE THE OUTLIERS WERE IDENTIFIED. THE LOWER CHART AT THE LOWER RIGHT CORNER SHOWS THE EXPERIMENTALLY MEASURED AND PREDICTED CHEMICAL $^{13}$C SHIFT VALUES FOR THE OUTLIERS BY NMRPRED-CARBON, MNOVA AND NMRSHIFTDB2. 144

FIGURE 3.17: THE SUBCLASS DISTRIBUTION GRAPH OF CHEMICALS SEEN IN THE TWO HOLDOUT DATASETS COMPARED TO THE NMRPRED-CARBON TRAINING DATASET. THE CHART AT THE BOTTOM SHOWS HOW THE SUBCLASSES ARE GIVEN A NUMBERED LABEL. 145

FIGURE 3.18: THE PREDICTION PERFORMANCE OF INDIVIDUAL COMPOUNDS AS GENERATED BY NMRPRED-CARBON, MNOVA AND NMRSHIFTDB2. THE GREEN COLOR AND THE RED COLOR INDICATE THE BEST AND THE WORST PREDICTIONS BY NMRPRED-CARBON, RESPECTIVELY. 146

FIGURE 3.19: A COMPARISON OF THE MEAN ABSOLUTE ERRORS (MAES) FOR EACH OF THE 65 MOLECULES IN THE FIRST HOLDOUT DATASET FOR THE DIFFERENT PREDICTORS (NMRPRED-CARBON, MNOVA, NMRSHIFTDB2). 148

FIGURE 3.20: PERFORMANCE COMPARISON BETWEEN THE 1ST AND 2ND HOLDOUT DATASETS AMONG THE FOUR PREDICTORS: NMRPRED-CARBON, MNOVA, NMRSHIFTDB2 AND THE DFT PREDICTOR. 148

FIGURE 3.21: THE CHEMICAL SUBCLASS DISTRIBUTION FOR THE NMRPRED-CARBON TRAINING DATASET AS WELL AS THE 1ST HOLDOUT DATASET AND THE 2ND HOLDOUT DATASET (FROM NP-MRD) AS INDICATED BY MEMBERSHIP IN CHEMICAL SUBCLASSES. 149

# Chapter 1: Introduction

NMR, **N**uclear **M**agnetic **R**esonance spectroscopy (1) is a spectroscopic technique that measures the absorbance of radio frequency (RF) radiation that occurs when molecules are placed within strong magnetic fields. NMR spectroscopy leads to the generation of NMR spectra that consist of many sharp peaks (called resonances) spanning a range of frequencies that reflect the interaction of applied RF energy and the nuclei within the molecule(s) being studied. By measuring the positions, intensities and fine structure of the resonance peaks in an NMR spectrum, the structure of molecules can be determined. NMR spectroscopy has been used by chemists for more than 70 years (2) to characterize the atomic structure and dynamics of molecules. NMR is widely considered to be the gold standard for determining the structure of small organic molecules – both natural and synthetic. NMR is also used to determine the structure of macromolecules such as peptides, proteins and nucleic acids (3) and has found applications in other areas including metabolomics (4), food analysis (5), pharmaceutical formulation (6) and petroleum product assessment (7). Millions of NMR spectra are collected each and every day by 10s of thousands of NMR spectrometers located around the world.

NMR spectroscopy requires the use of specially designed instruments called NMR spectrometers. An NMR spectrometer consists of a powerful, refrigerator-sized superconducting magnet connected to a radio transmitter/receiver (called a transceiver) and a computer or spectral recording device (Figure 1.1). An NMR magnet is very powerful. In fact, NMR magnets are among the strongest magnets found anywhere in the world. Many NMR magnets have a field strength on the order of 12-14 Tesla, which is strong enough to lift a city bus weighing 15,000 kilograms. The most powerful NMR magnets ever made have a field strength of 28 Tesla (8). An NMR magnet is always cylindrical in shape and has a central bore (usually about 5 cm wide) running through its middle. The bore is where the NMR sample is placed (in a specially designed glass tube) along with the NMR probe (which houses the sample and contains the RF electronics to excite and measure the nuclei of interest). To keep an NMR's superconducting magnet cooled to near absolute zero, the magnet must be surrounded by a bath of liquid helium (at -269 $^{o}$C) which is then surrounded by a layer of vacuum insulation which is then surrounded by a bath of liquid nitrogen (at –196 $^{o}$C) which is surrounded by more insulation and encased in a metal container. This layered cooling and insulation infrastructure forms a giant metal thermos-like bottle around the magnet (see Figure 1.2 and Figure 1.3)

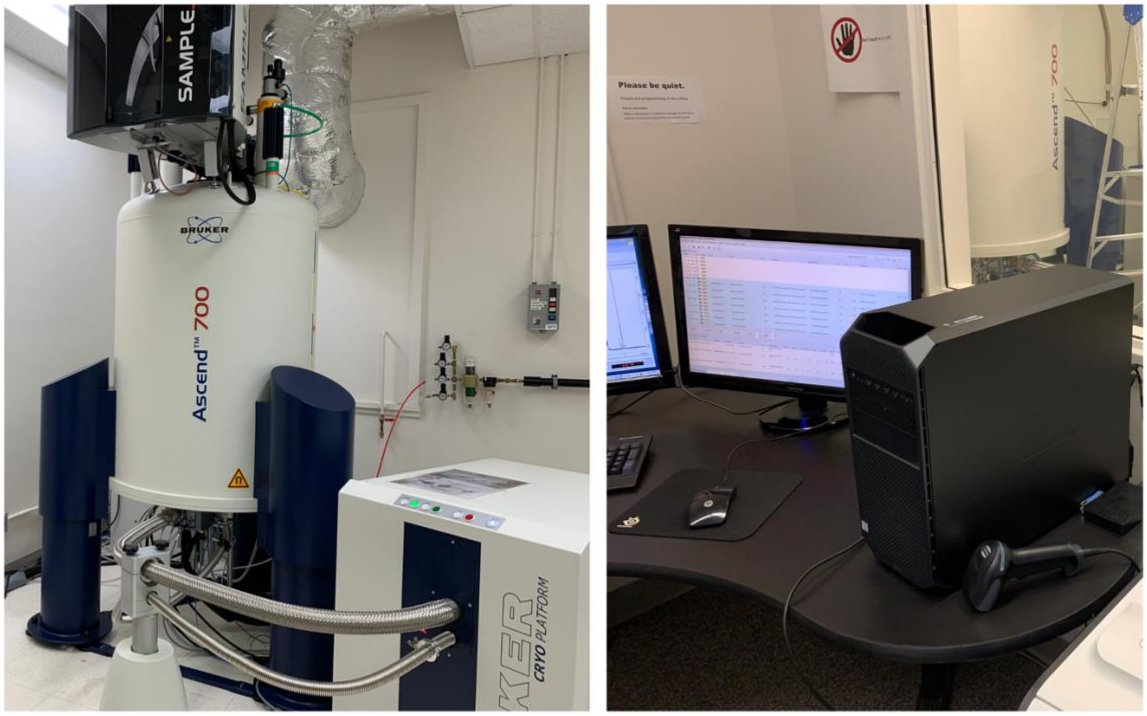


Figure 1.1: An NMR spectrometer (left side) and its computer terminal (right side)

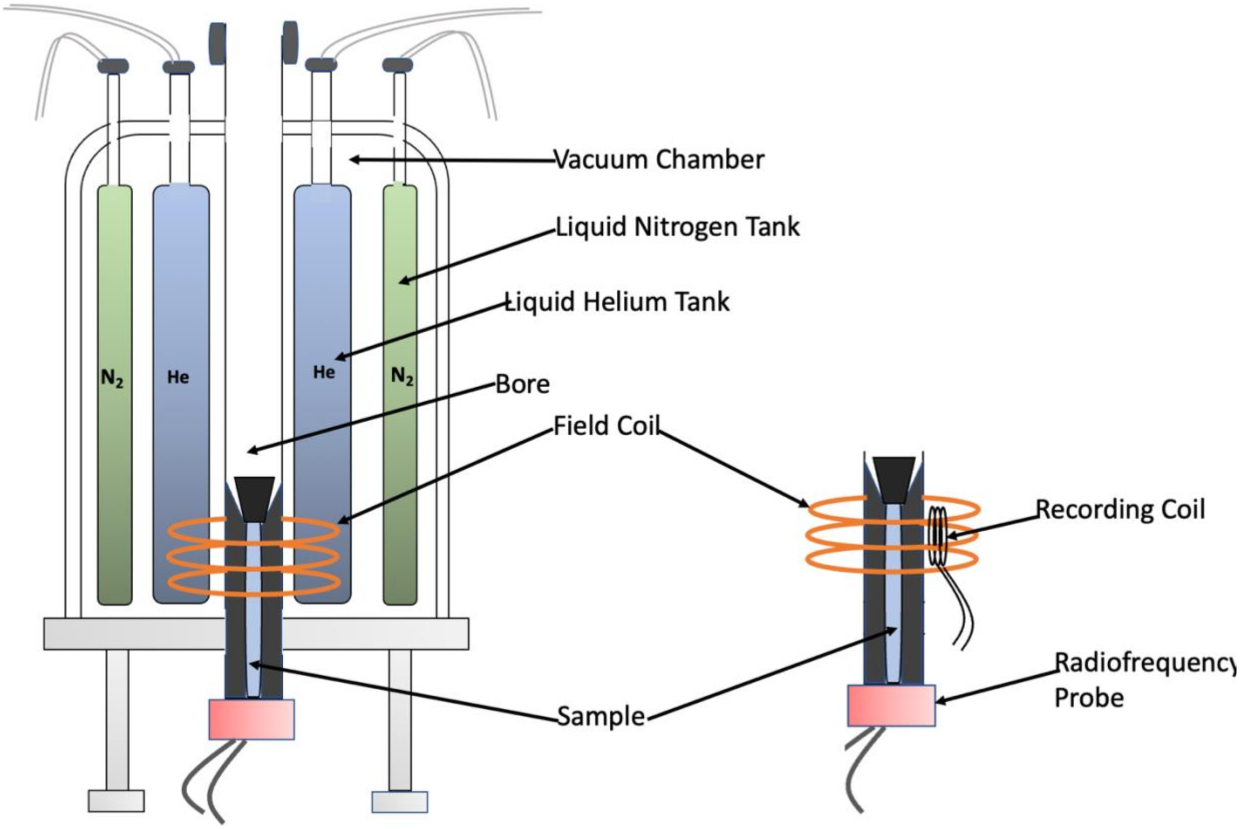


Figure 1.2: A schematic diagram illustrating the main components an NMR spectrometer, including the superconducting magnet (on the left) and the RF probe (on the right).

Inside an NMR magnet bore sits the NMR probe, which is connected externally to a series of RF generators, amplifiers, receivers and computers. The NMR probe is designed to accommodate a narrow glass test tube that is typically 5 mm wide and about 15-20 cm long. An NMR probe consists of a saddle-shaped RF coil (attached to electronics) that wraps partially around the glass test tube to produce a magnetic field in a specific horizontal direction (see Figure 1.2). This glass test tube or NMR sample tube contains the sample of interest.  Most NMR samples consist of materials (chemicals, compounds, powders, biological materials) dissolved in a liquid solvent. Common solvents used in NMR are chloroform, dimethyl sulfoxide, methanol, acetic acid, acetone, acetonitrile, benzene, methylene chloride, pyridine and water. Most of these NMR solvents must be deuterated (as will be explained later). Typically, the sample volume used in NMR

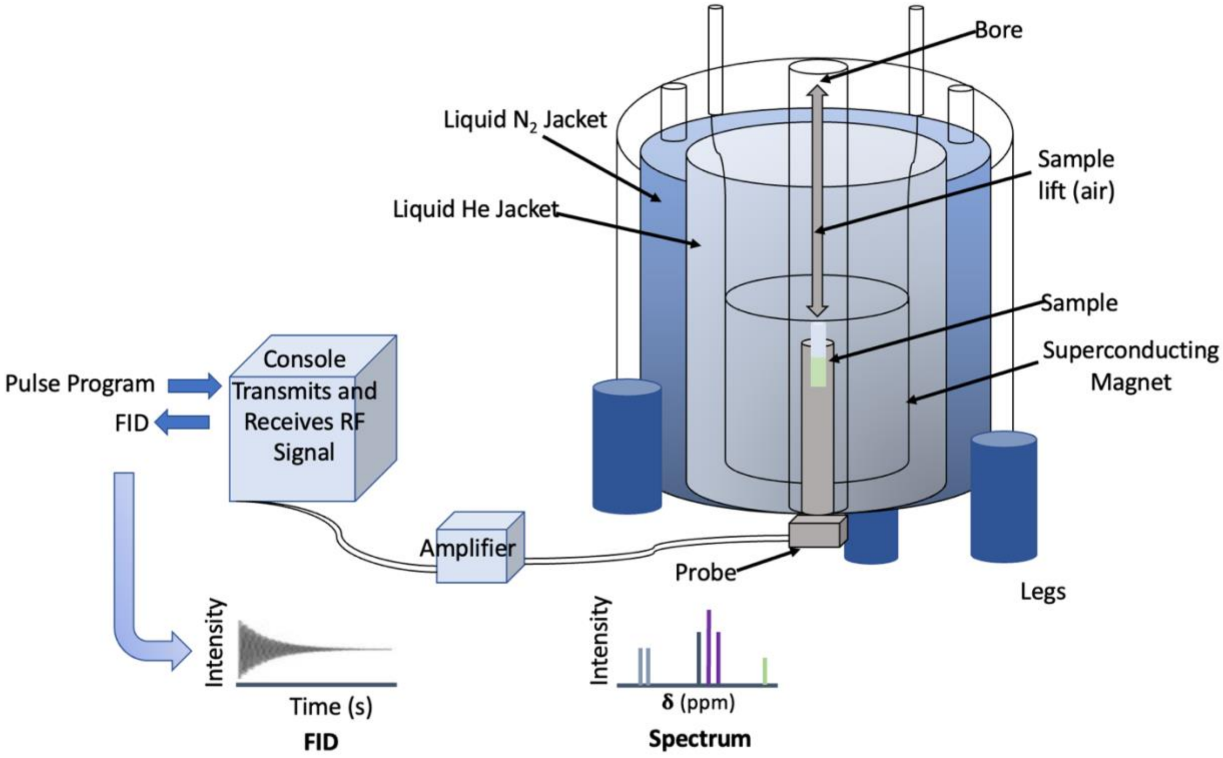


Figure 1.3: A schematic diagram of an NMR spectrometer illustrating the magnet assembly (right), the probe (inside the magnet), the amplifier and console/transceiver. The measured FID is shown on the lower left along with the resulting Fourier transformed NMR spectrum.

is between 250-500 μL and the sample only occupies 2-3 cm inside the tube. When a sample is placed in an NMR spectrometer, an NMR spectrum can be acquired. NMR spectra are collected by sending a short, strong RF pulse through the NMR probe and measuring the RF absorption or response that occurs in the sample. The RF signal is detected by the probe coil and recorded as an oscillating signal that spans several seconds, called a Free Induction Decay (FID). The resulting NMR spectrum, which is obtained by converting the time-dependent data in the FID into a frequency-dependent data file using Fourier transformation, usually consists of a series of sharp peaks that appear at specific frequencies that are very close (within a few parts per million or ppm) to the frequency of the RF excitation energy (Figure 1.3).

In NMR spectroscopy, it is well known that different nuclei absorb at different radio frequencies. Hydrogen ($^1$H) atoms/nuclei absorb at much higher frequencies than carbon ($^{13}$C) atoms, which, in turn, absorb at higher frequencies than deuterium atoms ($^2$H) or nitrogen atoms ($^{15}$N). Likewise, not all nuclei are NMR active (such as $^{12}$C or $^{16}$O). The abundance of certain isotopes also affects the intensity of NMR signals. In organic molecules, $^1$H is very abundant, whereas $^{13}$C and $^{15}$N are very rare. As a result, NMR samples must often be isotopically enriched with $^{13}$C or $^{15}$N to ensure that the signal is strong enough to be detected. The use of solvents that are deuterated ($^2$H) in NMR ensures that the normal $^1$H signal of the compound of interest is not overwhelmed or totally overlapped by the solvent signal. This is because deuterium ($^2$H) resonates at a much different frequency than $^1$H.

The positions of the absorption frequencies or "resonances" that are seen in an NMR spectrum are called chemical shifts. NMR chemical shifts are very sensitive to the electronic environment surrounding each nucleus and can provide a great deal of information about a molecule's covalent and non-covalent structure. Chemical shifts can be reported as a frequency value or as a relative frequency value. Today, most chemical shifts are usually measured as the difference between the resonant frequency of a nucleus and that of a defined reference or reference material (dissolved in the sample of interest), relative to the reference's resonant frequency. This quantity is expressed by $\delta$ (chemical shift) and is reported in parts per million (ppm).

$$\delta = \frac{f - f_{ref}}{f_{ref}} \times 10^6 \tag{1.1}$$

In the equation 1.1, $f$ is the resonant frequency of the nucleus of interest and $f_{ref}$ is the resonant frequency of the reference material. In NMR spectroscopy, the reference material is often TMS (tetramethylsilane – for organic solvents) or DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid – for water), both of which are organo-silicon compounds that have a chemical shift of the attached methylsilane groups formally defined as 0.00 ppm.

NMR spectra are also characterized by a "fine structure" characterized by peak clusters (doublets, triplets, etc.) that correspond to scalar or J-couplings between adjacent (geminal) or nearby (vicinal) atoms. The intensity of the peaks in an NMR spectrum is also informative as the intensities correspond to the numbers of atoms found at a given resonance frequency or a given chemical shift. The intensity and shape of an NMR peak is also affected by the interactions between nearby atoms (via nuclear Overhauser effects or NOEs) and the liquid lattice (relaxation effects) – but these will not be discussed here.

An example of a hydrogen ($^1$H) NMR spectrum for a simple molecule (ethanol, $CH_3CH_2OH$) is shown in Figure 1.4, As seen here, there are three major resonances or peaks in ethanol at three different frequencies (or chemical shifts). These correspond to the three kinds of hydrogen atoms in ethanol, the methyl hydrogens ($CH_3$), the methylene hydrogens ($CH_2$) and the hydroxyl hydrogens (OH). These peaks have some fine structure due to the scalar or J-couplings between neighboring hydrogens. For instance, the spin of the Hs in the -$CH_2$- group creates three split peaks (a triplet) in the signal corresponding to the methyl group (-$CH_3$). Similarly, the three Hs in the methyl group create four split peaks (a quartet) in the -$CH_2$- group signal.

Relative to most other kinds of spectroscopy, NMR is unique in the degree of resolution (sharpness of peaks) that it offers. The narrow peaks in NMR are actually a consequence of the rapid molecular tumbling that happens when molecules are dissolved in a solution. This random tumbling slows the magnetic relaxation times ($T_1$ and $T_2$) and sharpens the energy transitions. This exceptional spectral resolution makes NMR much better than UV, fluorescence or infrared (IR) spectroscopy for extracting detailed atomic or molecular information from chemical compounds. Furthermore, the behavior of nuclei under all kinds of RF excitation conditions is remarkably predictable making NMR very amenable to specially designed RF pulse experiments which allow magnetization to be manipulated and transferred around a molecular almost at will. These “pulse sequence” experiments or RF manipulations lie at the heart of modern NMR spectroscopy (9).

## Principles of the NMR Phenomenon

The principles behind the NMR phenomenon are very complex and took many decades for physicists to understand and develop the appropriate theoretical understanding. Therefore, a detailed explanation of the theory of NMR is beyond the scope of this thesis. Instead, I will provide a shorter, more simplistic explanation. The key to explaining NMR lies in understanding the detailed structure and dynamics of atoms and molecules. Molecules consist of atoms that are covalently bonded together. Atoms consist of nuclei (consisting of protons and neutrons) surrounded by electrons or electron orbitals. Electrons are negatively charged while the protons inside the atomic nuclei are positively charged. Both electrons and protons have a spin. Spin is a quantum property, but it can also be viewed classically or semi-classically as an actual spinning phenomenon when trying to understand NMR. If we imagine protons as tiny positively charged spheres spinning on their own axis, this spinning charge creates a magnetic field. As a result, each proton and each nucleus in a molecule behaves like a microscopic magnet.

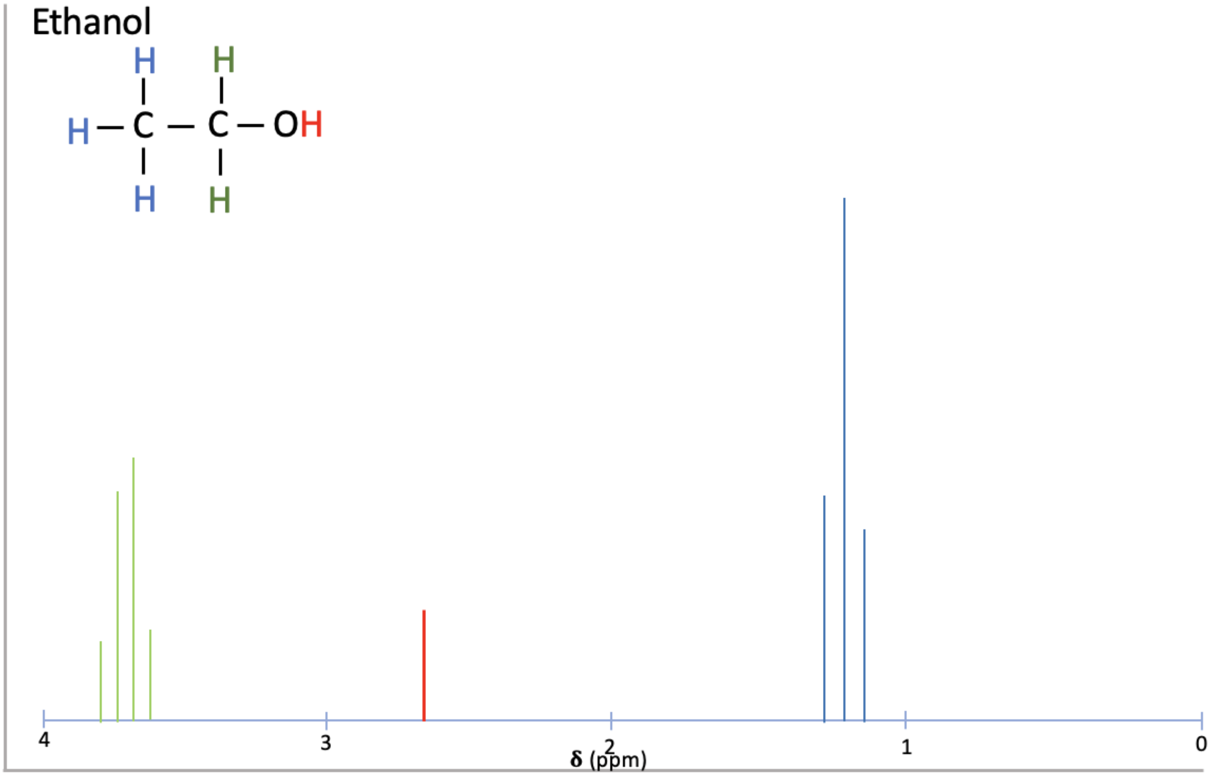


Figure 1.4: $^{1}$H NMR spectrum of ethanol. The methyl (-CH3) and methylene (-CH2-) groups couple with each other which results in a triplet multiplet for the -CH3 signal and a quartet for the -CH2-. The -OH group does not couple with any other protons and as a result the signal for the -OH is singlet.

Without the presence of any external magnetic field, the orientation of these nuclear spins (which are like spinning tops suspended in space) are random, with some spinning clockwise (up) and some spinning counter clockwise (down). But when the molecules are placed in a strong external magnetic field (such as the one provided by an NMR superconducting magnet), the nuclei start to align with the magnetic field and they begin to spin not only about their own axes, but also around the magnetic field. This extra (slower) spinning motion is called precession. It's the same phenomenon seen when a spinning top starts to lose its momentum, starts wobbling and appears ready to fall over. In NMR, the frequency of the nuclear precession is proportional to the strength of the magnetic field. A nucleus can either spin parallel to the magnetic field or in the opposite (antiparallel) direction to the external magnetic field. The spins that are parallel to the external magnetic field are at lower energy whereas the antiparallel spins are at higher energy. In a sample with trillions of molecules, you will have trillions of spinning nuclei -- called a population of spins. Even with a very powerful magnet turned on, the spin populations occupying these two states (up or down) are only slightly different, with slightly more than half of spins in the parallel direction and slightly less than half in the antiparallel direction.

When we send radio frequency (RF) pulses to the sample through the NMR probe and its saddle shaped coil, we can perturb the nuclear spins. This is because the RF pulse has a weak oscillating magnetic field (recall that RF energy is electromagnetic in nature). If the frequency of these RF pulses matches the frequency of the nuclear precisions, a resonance or an RF absorption event occurs. This leads to a transition (or flipping) of nuclear spins from a lower energy state to higher energy state. Classically, this would lead to a flip of the spinning top, which over time would lead to the spinning top to slowly lose its angular momentum and wobble until the top flips back down to the lower energy state. This flipping and wobbling by trillions of nuclear spins all at once, creates an RF signal. This RF signal is detected by the probe coil and recorded as a Free Induction Decay (FID). An FID looks like a graph depicting the up and down motion of a decaying oscillation or a collection of decaying oscillations that weaken over a few seconds (Figure 1.5). This time-dependent FID contains valuable frequency information which can be recovered by Fourier transformation, which converts a time dependent signal to a frequency dependent signal (10). The resulting frequency dependent signal is called an NMR spectrum and it is characterized by a spectral image that contains peaks at different frequency positions corresponding to chemical shifts. This is the classical NMR spectrum seen in Figure 1.4 and 1.6.

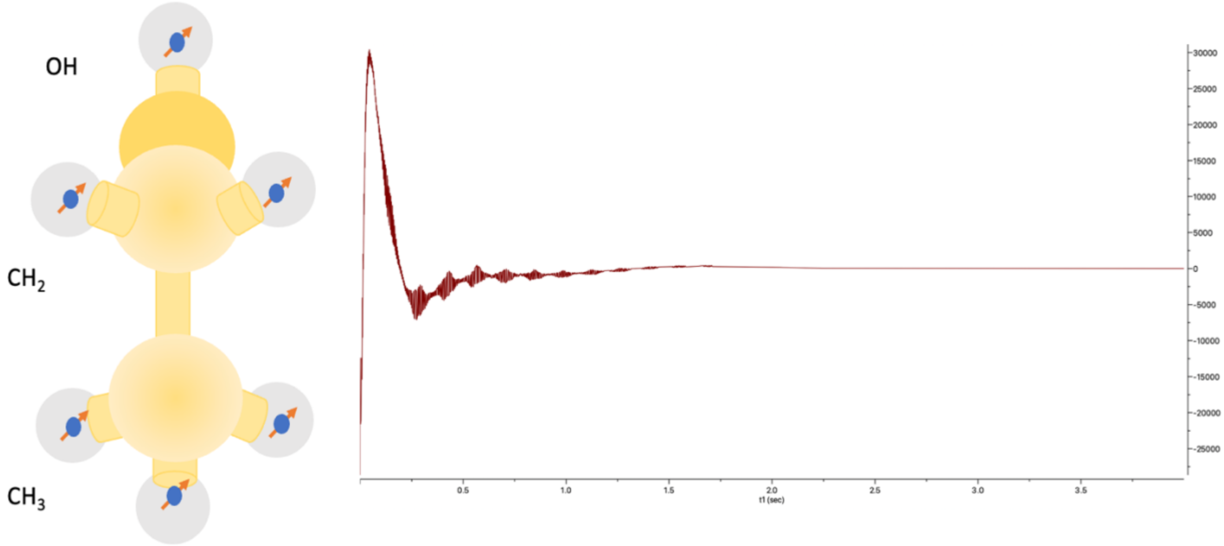


Figure 1.5: An image illustrating the spin of the protons in ethanol (left side). The right side shows the Free Induction Decay (FID) plotted against the time axis, arising from the spinning protons after they have been perturbed by an external RF signal.

As noted earlier, only certain nuclei are NMR compatible or NMR active. Atoms or isotopes that have an even number of protons and an even number of neutrons (such as $^{12}C$ or $^{16}O$) are not NMR active, while atoms or isotopes with an odd number of protons and neutrons (such as $^{1}H$, $^{2}H$, $^{13}C$, $^{15}N$, etc.) are NMR active. The charge distribution around the nucleus also affects the quality of the NMR spectrum. Those nuclei that have a spherical charge distribution (such as $^{1}H$,

$^{13}C$ and $^{19}F$) tend to have a spin of ½ and generate "good" NMR spectra with sharp peaks, while those that have non-spherical charge distributions (such as $^{2}H$ and $^{14}N$) tend to have a spin of 1 or 3/2 and do not generate very good NMR spectra (characterized by broad peaks). Among the different nuclei routinely detected by NMR spectroscopists, $^{1}H$ (proton), $^{13}C$ (carbon), $^{19}F$ (fluorine) and $^{15}N$ (nitrogen) are the most common ones. Among these, $^{1}H$ and $^{13}C$ are the most widely used, especially in the field of organic chemistry. This is because >99% of organic molecules contain hydrogen and/or carbon atoms.

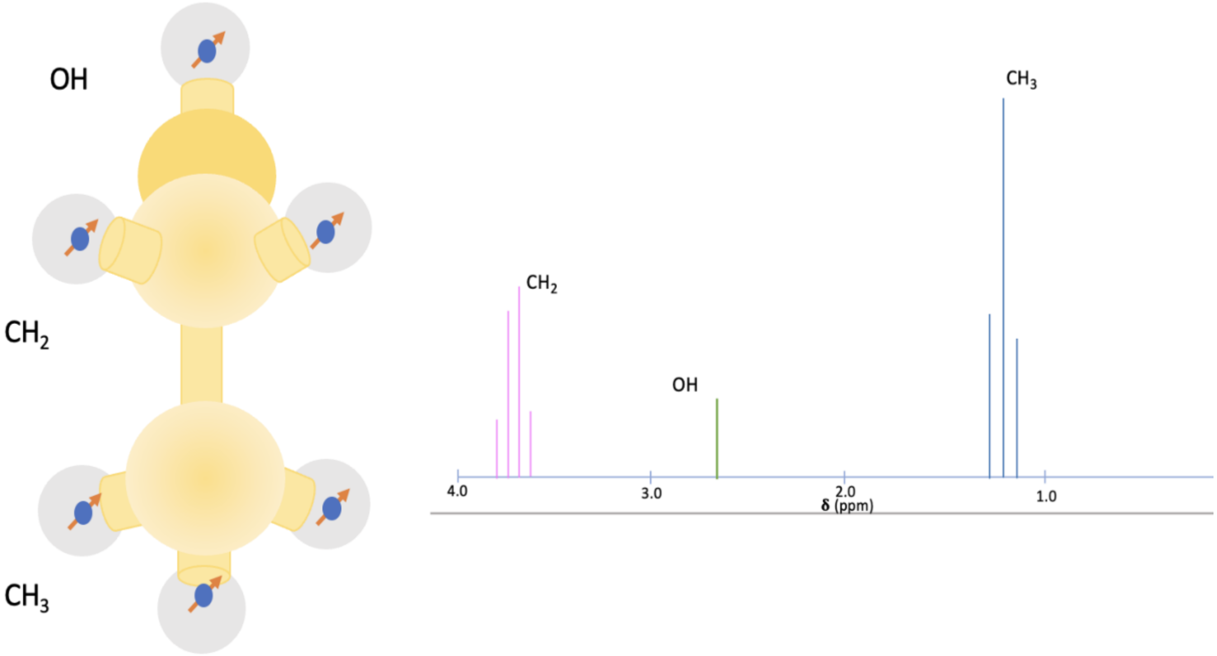


Figure 1.6: An example of a Fourier transformed NMR spectrum arising from the hydrogens (protons) in ethanol. This spectrum is characterized by sharp peaks plotted against their resonance frequencies (chemical shifts).

## Structure Elucidation by NMR

As noted earlier, NMR is widely considered to be the gold standard for determining the structure of small organic molecules – both natural and synthetic. This is because NMR spectra are characterized by sharp, well-defined peaks that can be directly associated to specific atoms within a given molecule. These peaks correspond to the chemical shifts. Chemical shifts can often be assigned to specific atoms or atomic groups in the molecule of interest. In NMR, a chemical or a molecule is considered "assigned" or solved when all the chemical shifts (NMR peaks in the spectrum) are assigned to all the NMR-detectable atoms in the molecule. Chemical shifts are often called the "mileposts of NMR". They serve as reference points to help NMR spectroscopists map out atomic positions, identify chemical groups and determine molecular structures. Over the last few decades, chemical shifts have been used by chemists to successfully determine or confirm the covalent structure of hundreds of thousands of organic molecules. Not only are the chemical shifts sensitive to the type and character of nearby atoms but chemical shifts are also remarkably consistent or "predictive" for different chemical groups or chemical environments. This sensitivity and behavioral consistency has allowed chemists to produce well-defined chemical shift "principles" that allow them to deduce the identity of key chemical groups and thereby determine the precise structures of many small molecules.

To see how this is done, let us go through an example of determining the structure of an "unknown" compound with the formula $C_9H_{10}O_2$. Using this example we can understand how, by looking only into the chemical shifts, detailed information about a molecular structure can be obtained. Figure 1.7 shows the one-dimensional proton or $^1$H NMR spectrum for a compound with this chemical formula. From this $^1$H NMR spectrum we can see there are 4 different peaks or peak clusters, which means there are 4 different types of hydrogen nuclei present. In $^1$H NMR, the values of $^1$H chemical shifts are usually within the range from 0.0 ppm to 12.00 ppm, depending on what particular proton is attached to which type of heavy atom, such as a carbon, nitrogen, oxygen atom, and how that proton is attached. These are called functional groups. Figure 1.8 shows the range of $^1$H chemical shifts values for many common functional groups seen in natural products and synthetic organic molecules. To determine the molecular structure of this "unknown" molecule, the next step is to check the area beneath the peaks. The peak area or peak intensity reveals how many hydrogens are associated with a particular type of hydrogen atom. Peak area determination is normally performed by NMR-specific software and the relative area determined by the software for each peak is shown above (Figure 1.7). So, Peak-A has one hydrogen, Peak-B has five hydrogens, while Peak-C and Peak-D each have two hydrogens. Next, if we look at the value of the chemical shifts in Figure 1.7 and compare with the possible chemical shift values in Figure 1.8, Peak-A is approximately at 11.6

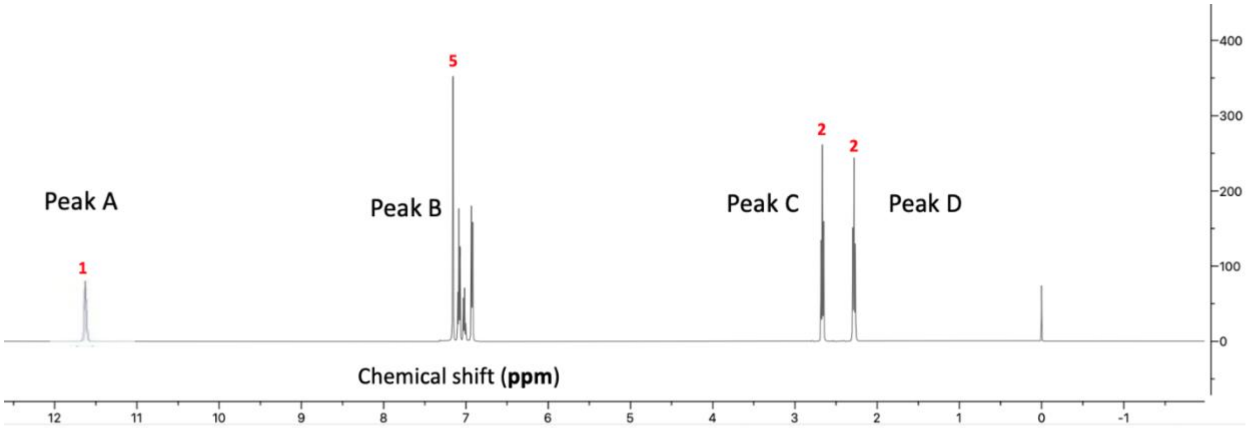


Figure 1.7: An example of $^1$H NMR spectrum for a compound with the chemical formula: $C_9H_{10}O_2$.

ppm which indicates that it likely corresponds to a proton attached to a carboxyl functional group, thus indicating there are two oxygens and one carbon in this “unknown” chemical formula. Peak-B is at 7.4 ppm which indicates, according to Figure 1.8, that these protons are attached to an aromatic group, which would suggest the presence of a benzene ring consisting of six carbon atoms in the molecular formula. But as there are just five hydrogens in Peak-B, we would have to assume that within the six-carbon aromatic group there are five attached hydrogens and remaining carbon has a different (non-hydrogen) substituent attached. Peak-C and Peak-D are located at around 2.7 ppm and 2.3 ppm, respectively. They likely indicate the presence of C-H functional groups. According to the peak areas, these peaks have two hydrogens, which means there are two -CH2- groups in the chemical formula. We can also assume that the two carbons from Peak-C and

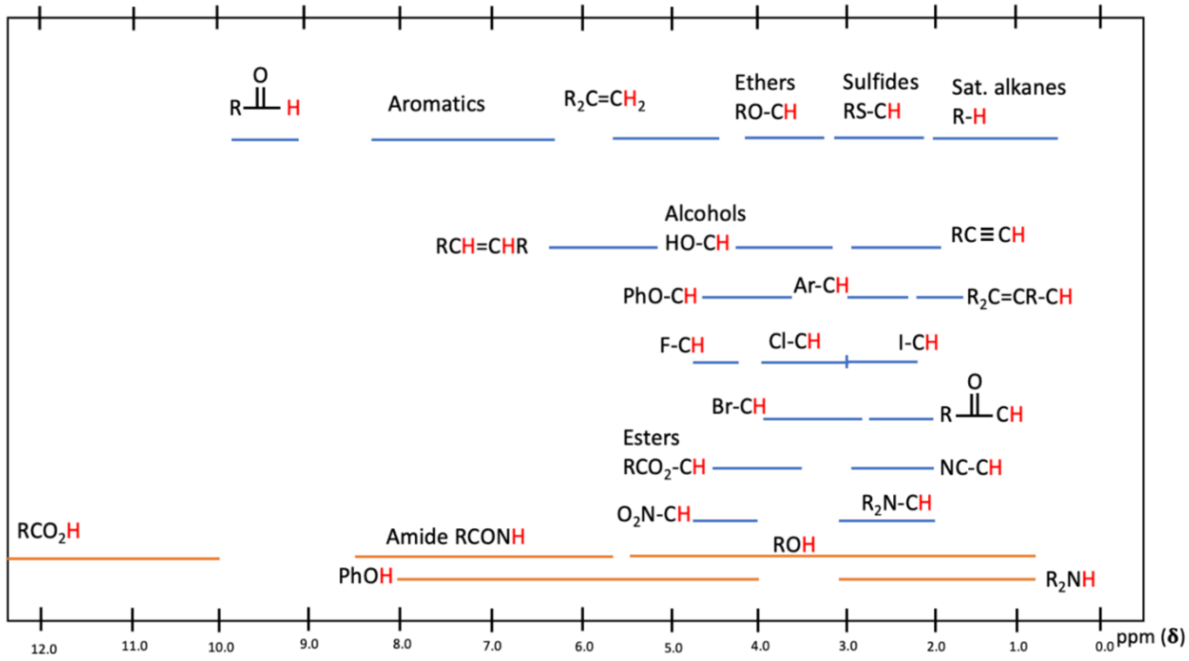


Figure 1.8: Approximate range of proton chemical shift values for some common functional groups.

Peak-D are internal carbons. Because if they were at the ends of the molecule, they would need three hydrogens (indicating the presence of $CH_3$ groups). From this data on both chemical shifts and integrated peak areas, we can rationalize the structure or substructures of a total of 9 carbons, 10 hydrogens and 2 oxygens. The individual substructures that we get from each peak are shown in Figure 1.9. Finally, if we put all the above observations together, the only meaningful molecular structure that we obtain 3-Phenyl-propionic acid, which is shown in Figure 1.10.

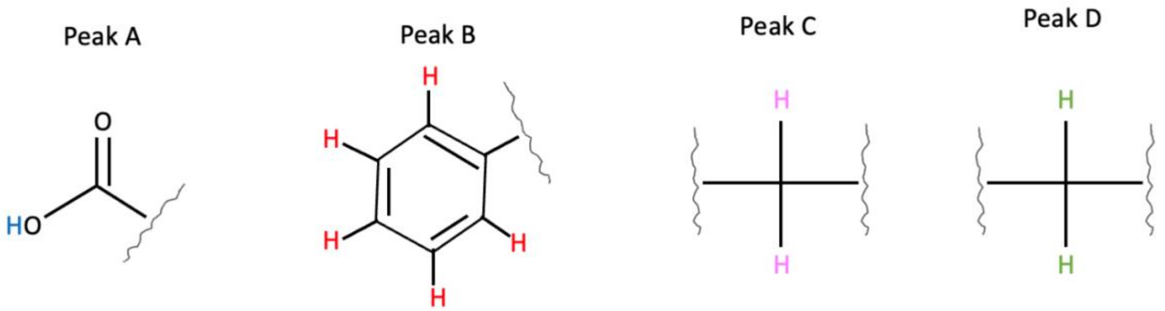


Figure 1.9: Individual possible substructure structures derived from the NMR peaks in [Figure 1.7].

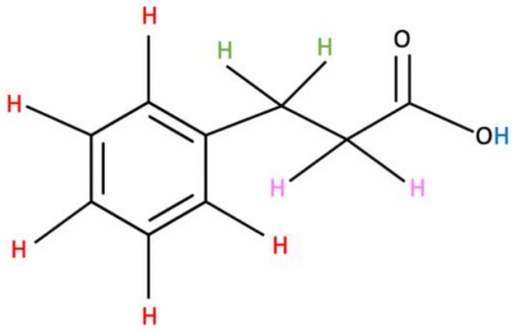


Figure 1.10: The derived structure (3-Phenyl-propionic acid) from the $^1$H NMR spectrum in [Figure 1.7].

Determining a chemical structure by analysing NMR peak positions and peak areas can be quite tedious and is often very challenging. It's a bit like solving a jigsaw puzzle. Because so many organic molecules have already had their structures solved by NMR and their assignments determined by NMR spectroscopists, a more common trend in NMR is to identify a compound or assign its spectra by matching the observed $^{1}$H and/or $^{13}$C NMR spectra (or chemical shifts) with the $^{1}$H and/or $^{13}$C NMR spectra in a spectral library. A spectral library consists of molecular structures, their NMR spectra and the corresponding NMR assignments. By performing simple spectral similarity or chemical shift matching to the spectral library, it is often possible to identify/assign a known compound or identify a structurally similar compound – which makes the assignment process many times faster and infinitely easier.

As a result, many NMR spectral libraries that contain NMR data for small molecules have been built. These include the Biological Magnetic Resonance Databank (BMRB) (11), NMRShiftDB2 (12), the Spectral Database System (SDBS) (13), and the Natural Products Magnetic Resonance Database (NP-MRD) (14). In addition, several commercial NMR spectral libraries have been developed including Advanced Chemistry Development (ACD/Labs) and the Wiley spectral database collection. These spectral libraries contain $^{1}$H and/or $^{13}$C NMR data for thousands of molecules collected over a range of NMR spectrometer frequencies. These libraries are widely used in the fields of synthetic chemistry, food chemistry, metabolomics and natural product chemistry.

While NMR spectral libraries are useful for many different fields of chemistry or analytical chemistry, they often only cover a tiny fraction of the known set of compounds being studied. For instance, in the field of metabolomics, fewer than 1000 compounds with high quality NMR spectra have been collected and deposited into the Human Metabolome Database (15). This compares to the >250,000 chemicals that are in the HMDB (which translates to <0.5% compound coverage). Similarly, the number of experimentally assigned NMR spectra in the NP-MRD is <20,000 whereas the number of known natural products in the NP-MRD is >300,000 (which translates to <7% coverage coverage). With the ever-increasing number of known human metabolites or known natural products being identified, collecting experimental NMR data on each of these compounds and completing their assignments is an almost impossible task. If we could accurately predict NMR chemical shifts or NMR spectra and avoid the tedious work of sample preparation and NMR spectral collection/assignment, this would save an immense amount of laboratory time. Indeed, an accurate NMR spectral prediction system would allow NMR spectroscopists to rapidly build an enormous library of predicted NMR spectra that could be readily used for the identification (and quantification) of compounds in almost any sample. The development of an accurate NMR spectral prediction system is the main motivation behind my thesis.

## Chemical Shift Prediction

There are four different kinds of approaches to predict NMR chemical shifts from structural data (16–18). These include: 1) quantum *mechanical ab initio* techniques; 2) rule based or classical physics techniques; 3) database or look-up methods and 4) machine learning based methods. I will briefly review all four approaches.

## *Ab initio* Techniques

*Ab initio* or quantum mechanical approaches do not rely on empirical knowledge. Instead, they use quantum theory to model the electron density and electron probability distributions of atoms and molecules. This allows one to directly calculate chemical shifts based on the electron effects on the observed nuclei. The best results for *ab initio* chemical shift calculation are obtained using the Density Functional Theory (DFT) technique. DFT is a computational quantum mechanical modelling method that is widely used in physics, chemistry and materials science to investigate the electronic structure of many-body systems, specifically atoms and molecules. Using DFT, the properties of a many-electron system (such as a molecule) can be determined by using functionals, i.e., functions of another function. In the case of DFT, these are functionals of the spatially dependent electron density. The performance of a DFT calculation depends on what functionals and basis sets are used. DFT calculations have become quite routine with the release of many freeware packages and commercial packages such as Quantum Espresso, VASP, LAMMPS, Gaussian, GAMESS and Schrodinger.  A detailed discussion of DFT and the theory behind it is beyond the scope of this thesis, however an excellent review regarding DFT and chemical shift calculation are available (19). DFT is capable of providing chemical shift prediction results that are reasonably close to experimental values, with RMSEs (root mean square errors) of 0.2-0.4 ppm for $^{1}$H shifts and 3-5 ppm for $^{13}$C shifts (20, 21). Unfortunately, the time required for performing a DFT calculation, even for one small organic structure is very long, approximately 3-24 hours (depending on the computer speed and memory configuration). It can be substantially more if the structure is very large (>50 atoms). Further improvements in DFT accuracy require larger basis sets which require even more compute time. The speed of chemical shift prediction is a very important criterion, especially if one is trying to calculate chemical shifts for 100's of thousands of molecules. As a result, there have been increasing efforts by researchers to develop more rapid methods to counter the long calculation times required for DFT methods, without compromising the prediction accuracy.

## Rule-based Techniques

Rule-based methods are empirical approaches that use observed correlations or observed trends in chemical shifts along with additive rules to estimate which neighboring atoms or functional groups change the electron density (and hence chemical shift) of the atom of interest. Early examples of this approach include hand-made, manually implementable rules developed by Shoolery et. al. (16, 18). These authors published a set of additive rules to calculate the $^{13}$C chemical shifts of methylene groups. The main idea is that every atom has a basic chemical shift value and the observed chemical shift value for a given atom of interest can be predicted by adding up the basic shifts of neighboring substituent atoms -- which might be several bonds away. Using this concept, Grant and Paul (22) extended the ideas of Shoolery and developed a manual method for predicting the $^{13}$C chemical shifts in linear alkanes. Since then, many more extensions of this rule-based or additive approach for chemical shift calculation have been developed, enabling the prediction of $^{13}$C chemical shifts for many different classes of organic compounds.  These approaches have also been applied to the prediction of $^{1}$H chemical shifts. For $^{1}$H chemical shift calculation, contributions from the closest (alpha) neighbors generally only need to be counted, while for $^{13}$C NMR chemical shifts substituents from the alpha, beta, gamma and delta position must typically be counted. So, to predict $^{13}C$ chemical shifts manually, a comparatively large number neighbor effects needed to be summed up, which can lead to errors, even for simple compounds. To reduce the errors and accelerate the speed of manual calculation, a computer program was developed for the estimation of the $^{13}C$ chemical shifts in 1990 (23). This early $^{13}C$ chemical shift predictor took a chemical structure as a linear chain as its input. From this input a connection table was built with the information about the atom type, the type of the central atom (the atom to have its chemical shift predicted), the types of immediate atom neighbors and the type of bonds associated with the central atoms. In this program, each atom was given a number that encodes all the connectivity information. Next, for every single bonded carbon atom, the chemical shift increments of its neighboring atoms were added to the base chemical shift value, to calculate the predicted chemical shift value. Unfortunately, this early prediction program was limited to handling single bond acyclic compounds with noncyclic functional groups. Later, an improved version of a computer program for predicting $^{13}C$ chemical shifts was proposed by Andras and Erno, who introduced an extended set of additivity increments (24, 25). This new program, called $^{13}$CShift, had the capability to select the appropriate additivity rules automatically for each carbon atom in the submitted molecule. This program also had the flexibility to extend the rule parameters, and to add new rules. In their assessment of the program, the authors predicted approximately 168,900 carbon chemical shifts and compared the predicted values with their experimental chemical shift values. The average error was -0.29 ppm with a standard deviation of 5.5 ppm. While the performance was generally good, this model suffered from the uncertainty of when it will work correctly and when it will fail (24). Within a few years, the same authors developed a new program using additivity rules to predict $^{1}H$ NMR chemical shifts (26, 27), but again, like other additivity models, these models also suffered from predictive uncertainty in terms of when it will work and when it will fail (26). Because of their high level of uncertainty and the limited applicability of additive rules to work for more exotic structures, work on rule-based methods for chemical shift prediction has largely stopped.

## Database Approaches

Another method for predicting NMR chemical shifts is to use database algorithms. In this method, a large database of chemical structures and their associated experimental chemical shift assignments is compiled. Within the database, each atom in each structure is described using a set of features/descriptors which reflect the characteristics of the chemical environment where the atom exists. To predict the chemical shifts of a new molecule, the structure is queried in the database using its descriptors and the database is searched for exactly matching or similar structures. When similar structures are found, the predicted chemical shifts are returned as the weighted average value of the experimental chemical shift values corresponding to the found similar structures. To improve the performance, most databases approaches encode information about the atomic environment in the database. The most popular method for encoding atomic environment information is the Hierarchical Ordered Spherical description of Environment coding method or the HOSE code. HOSE coding was first developed by W. Bremser (28) in 1978. Since it was first introduced, HOSE coding has become the gold standard for empirical NMR chemical shift prediction. HOSE coding describes the spherical environment of every atom in a molecule and complete ring systems. The HOSE description is symbolic and priority rules are predefined.

The description of an atom is done via progressively larger spheres surrounding the atom of interest. Starting with a given atom the topological descriptions are captured based on what is contained in the first sphere (usually about 1.4 -1.5 Angstroms in radius). The substituents of the central atom found in the first sphere, are then collected in a predefined sequence and converted to a symbolic notation. The sphere is then enlarged to be about 3 Angstroms in radius and more topological and connectivity information is collected. This is repeated for a third sphere that is about 4.5 Angstroms in radius – See Figure 1.11. Usually, descriptions are collected up to at least three spheres for HOSE coding. Through the calculated HOSE codes a database is created that has the spherical descriptions of all the atoms in each molecule stored and linked with the observed chemical shift values. By searching such a HOSE encoded database, chemical shifts can be predicted for molecules that share only partial structural similarity. If multiple matches are found in the database, a mean value of all the linked chemical shift values is calculated and returned. The quality of a HOSE code chemical shift prediction strongly depends on the structural diversity and the size of the database. The HOSE approach was first used successfully for chemical shift prediction by Steinbeck et al. in 2003 (12). To make the approach more appealing, they developed an open source and open content research database for chemical structures and chemical shift assignments called NMRShiftDB. NMRShiftDB also provided an openly accessible HOSE-coded-based chemical shift prediction tool. This HOSE-code predictor has the capacity to predict chemical shifts for $^{1}$H, $^{13}$C atoms in most molecules.

The NMRShiftDB algorithm works as follows: If the atom of interest has several matched HOSE codes, the final prediction result is provided as the average of the linked chemical shift

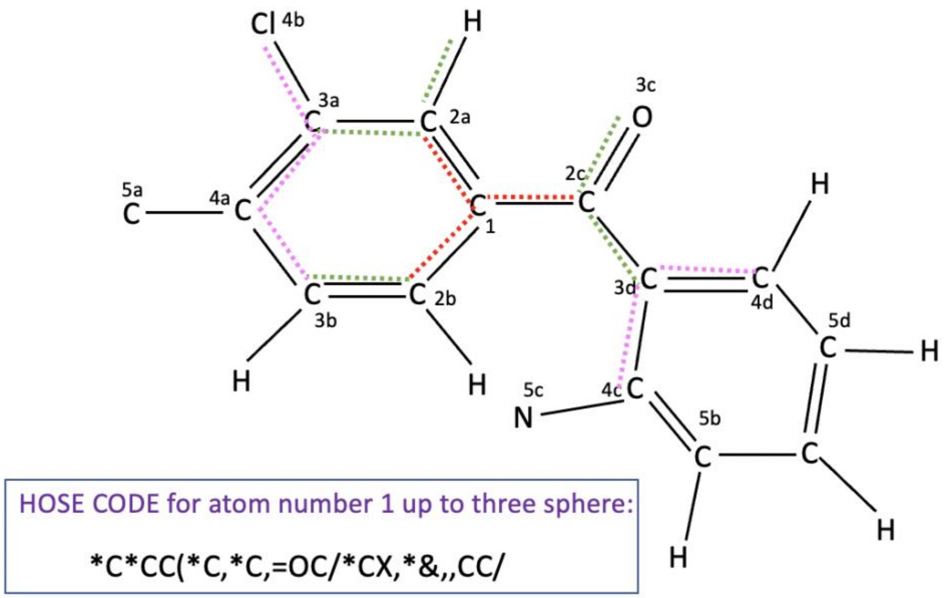
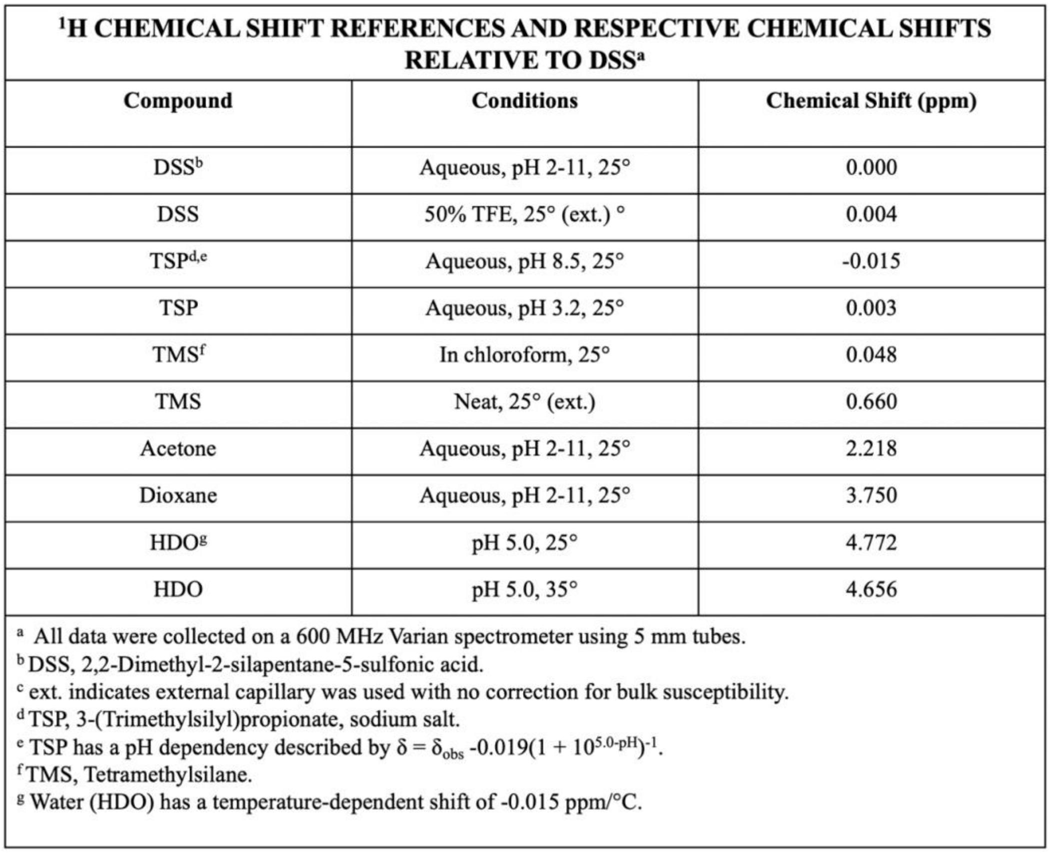

Cl 4b
H
3a
2a
3c
O
2c
H
5a
C
4a
1
3d
4d
5d
H
3b
2b
H
H
5c
N
4c
5b
H
H
HOSE CODE for atom number 1 up to three sphere:
*C*CC(*C,*C,=OC/*CX,*&,,CC/

Figure 1.11: The HOSE CODE for atom number 1 up to three spheres. The dashed lines represent the path generated for different spheres from atom number 1. The 1$^{st}$, 2$^{nd}$ and 3$^{rd}$ spheres from atom number 1 are represented by red, green and purple dashed lines, respectively.

values for each of the matched HOSE codes. When the number of the matching codes becomes 10 or more, the smallest and largest chemical shifts values are considered as the boundaries of the confidence limit. If the number of matching HOSE codes is insufficient to predict the chemical shift or the confidence limit, the search for the HOSE codes is decremented by one sphere until any suitable value is found.

Unfortunately, the early NMRShiftDB or HOSE code method did not include stereo descriptions of molecules. To address this limitation, Kuhn et al. (29) proposed an extended version of HOSE encoding which was stereo-aware. Kuhn's team implemented this approach in the next version of NMRShiftDB, called NMRShiftDB2. The stereo-aware version of this HOSE code method improved the MAE by 0.70 ppm in the prediction of $^{13}$C chemical shifts and by 0.04 ppm in the prediction of $^{1}$H chemical shifts. Though this type of chemical shift prediction tool is relatively slow, it is very popular. Indeed, the NMRShiftDB2 chemical shift predictor is one of the most popular tools for chemical shift prediction in the field of chemistry. However, one of the major drawbacks of the NMRShiftDB2 prediction tool is, it does not consider solvent effects nor does it consider the effects of using different chemical shift reference reagents, which was discussed by Wishart et al (30, 31)and is described in more detail in Table 1.1 and Table 1.2. The

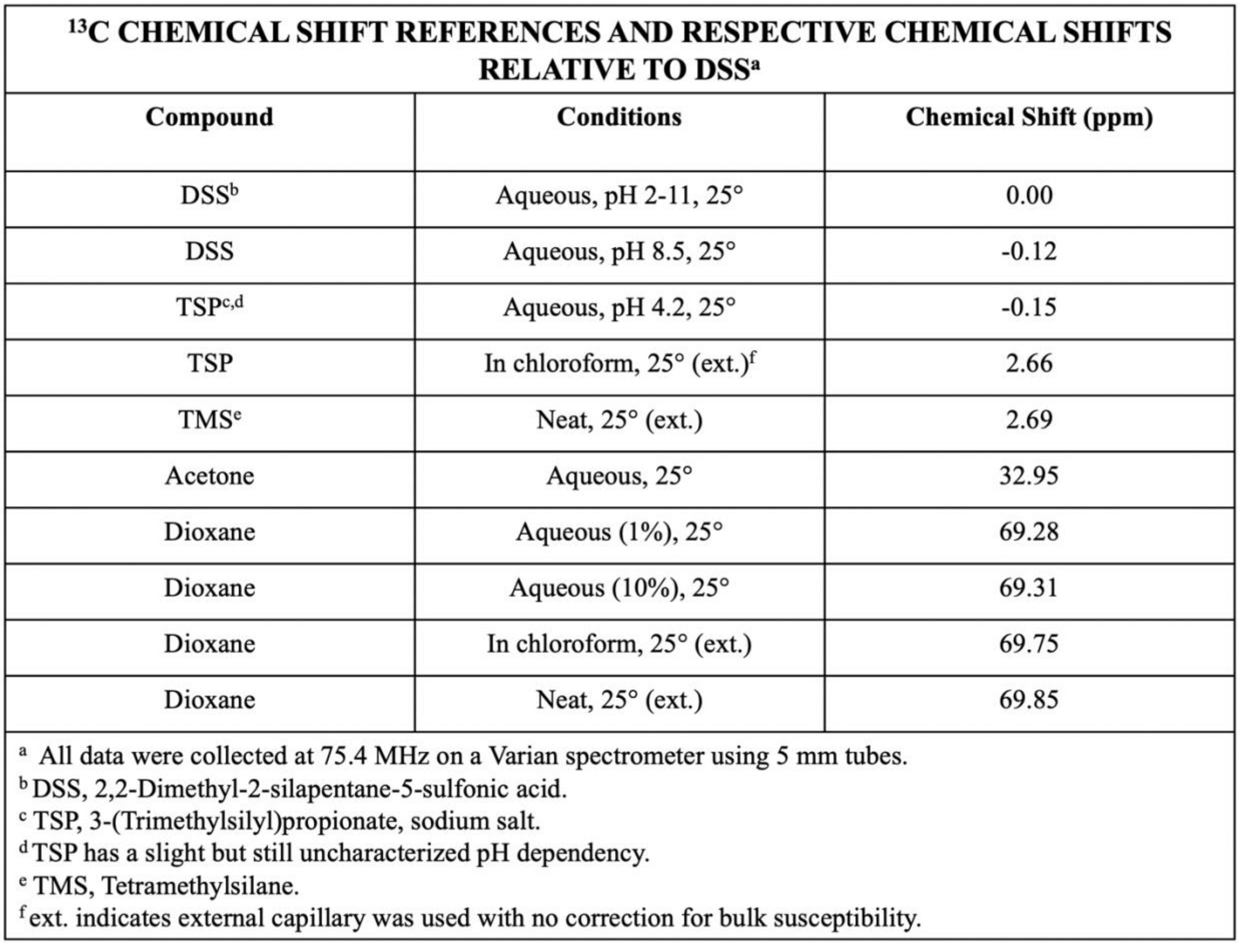


**$^{1}$H CHEMICAL SHIFT REFERENCES AND RESPECTIVE CHEMICAL SHIFTS RELATIVE TO DSS[a]**

| Compound | Conditions | Chemical Shift (ppm) |
|---|---|---|
| DSS[b] | Aqueous, pH 2-11, 25° | 0.000 |
| DSS | 50% TFE, 25° (ext.) ° | 0.004 |
| TSP[d,e] | Aqueous, pH 8.5, 25° | -0.015 |
| TSP | Aqueous, pH 3.2, 25° | 0.003 |
| TMS[f] | In chloroform, 25° | 0.048 |
| TMS | Neat, 25° (ext.) | 0.660 |
| Acetone | Aqueous, pH 2-11, 25° | 2.218 |
| Dioxane | Aqueous, pH 2-11, 25° | 3.750 |
| HDO[g] | pH 5.0, 25° | 4.772 |
| HDO | pH 5.0, 35° | 4.656 |

[a] All data were collected on a 600 MHz Varian spectrometer using 5 mm tubes.
[b] DSS, 2,2-Dimethyl-2-silapentane-5-sulfonic acid.
[c] ext. indicates external capillary was used with no correction for bulk susceptibility.
[d] TSP, 3-(Trimethylsilyl)propionate, sodium salt.
[e] TSP has a pH dependency described by $\delta = \delta_{obs} -0.019(1 + 10^{5.0\text{-pH}})^{-1}$.
[f] TMS, Tetramethylsilane.
[g] Water (HDO) has a temperature-dependent shift of -0.015 ppm/°C.

Table 1.1: $^{1}$H chemical shift references and respective chemical shift differences relative to DSS. Reference: Wishart et al. (30).



$^{13}$C CHEMICAL SHIFT REFERENCES AND RESPECTIVE CHEMICAL SHIFTS RELATIVE TO DSS[a]

| Compound | Conditions | Chemical Shift (ppm) |
|---|---|---|
| DSS[b] | Aqueous, pH 2-11, 25° | 0.00 |
| DSS | Aqueous, pH 8.5, 25° | -0.12 |
| TSP[c,d] | Aqueous, pH 4.2, 25° | -0.15 |
| TSP | In chloroform, 25° (ext.)[f] | 2.66 |
| TMS[e] | Neat, 25° (ext.) | 2.69 |
| Acetone | Aqueous, 25° | 32.95 |
| Dioxane | Aqueous (1%), 25° | 69.28 |
| Dioxane | Aqueous (10%), 25° | 69.31 |
| Dioxane | In chloroform, 25° (ext.) | 69.75 |
| Dioxane | Neat, 25° (ext.) | 69.85 |

[a] All data were collected at 75.4 MHz on a Varian spectrometer using 5 mm tubes.
[b] DSS, 2,2-Dimethyl-2-silapentane-5-sulfonic acid.
[c] TSP, 3-(Trimethylsilyl)propionate, sodium salt.
[d] TSP has a slight but still uncharacterized pH dependency.
[e] TMS, Tetramethylsilane.
[f] ext. indicates external capillary was used with no correction for bulk susceptibility.

Table 1.2: $^{13}$C chemical shift references and respective chemical shift differences relative to DSS. Reference: Wishart et al (30).

type of chemical shift standards/references and solvent used in NMR play a critical role in the determination of chemical shift values. Unfortunately, it is very common to publish NMR data without mentioning the solvent used or the chemical shift referencing methods. Wishart et al (30, 31), noted that because of these inconsistencies between NMR laboratories, a random error between 0.05-0.15 ppm was common for $^{1}$H chemical shift data, while for carbon and nitrogen chemical shift data, a ~2.0 ppm error is often present. While a 0.05 ppm error in $^{1}$H chemical shift data is not that significant, an error of 2.0 ppm is quite large and not acceptable. Table 1.1 and Table 1.2 show how chemical shift values must be adjusted among different chemical shift reference standards.

## Machine Learning Methods

The latest methods to be used for NMR chemical shift prediction are based on machine learning (ML) methods. Machine learning is a branch of artificial intelligence (AI) that enables a computer to learn automatically from past data or past experiences. ML can be used to identify patterns and to predict results without being explicitly programmed to do so (32, 33). ML gathers information directly from the provided data and learns from it. The more input data or input samples provided and depending on the performance task, the more diversity in the dataset, the better the learning and the better the performance. ML has become a very powerful problem-solving technique for various fields such as computational finance (credit scoring); image processing or computer vision (facial recognition, motion tracking etc.); aerospace or automotive production (predictive maintenance) as well as natural language processing, voice recognition, computational biology, drug discovery, DNA sequencing etc. The use of ML in chemistry is not new. ML has been successfully utilized in computational chemistry to predict molecular properties (logP, pKa), chemical reactions (BioTransformer (34)), to construct quantitative structure activity relationships (35), to predict mass spectra (36–38) and to predict NMR spectra or NMR chemical shifts (39). Before discussing how ML is being used to predict NMR chemical shifts, it is worthwhile to discuss how ML works.

ML involves five different steps. These include: 1) data collection; 2) data cleansing and analysis; 3) data partitioning; 4) model training and performance evaluation and finally 5) deployment. These steps are discussed separately below:

**Data Collection in ML**

The first step in any ML activity almost always involves data collection. This includes collecting and measuring information from different sources in such a way that it makes sense for the targeted problem. Data can be collected in numerical, categorical, ordinal, time series, and/or text forms (40). Numerical data are quantitative with data points being exact numbers (such as temperature, housing prices, etc.). Categorical data describe characteristics of the data (color, gender etc.). Categorical data can have a numerical value too but without any mathematical meaning (male is represented with 1 and female with 2). Ordinal data are a type of categorical data that are ordered or ranked in some particular way (cold, warm, hot). Ordinal can also be a mix of numeric and categorical data, consisting of quantitative data split into groups and represented by different categories. For example, $150K-250K price houses can be categorized as low-priced houses, houses with price ranges between $250K-450K can be medium-priced houses and houses with price >$450K can be grouped as high-priced houses. When a sequence of numbers is collected at regular intervals over a period of time, it is called time series data. An as example, measuring the average number of houses sold each day throughout the last 50 days would constitute time series data. The difference between numerical and time series data is that the numerical data doesn’t have any time ordering whereas the time series data has some implied ordering. Lastly, text type data, consists of words or sentences, that may correspond to, for example, opinions from a survey. Collecting the correct data and the correct type of data is very important as it allows one to identify recurrent patterns through data analysis. Once those recurrent patterns are evident it is possible to build predictive models using ML algorithms that would look for these trends, learn from them and predict future trends. Information rich, error free data is crucial for building a high performing ML predictor or model.

### Data Cleansing and Analysis in ML

The next step in ML is data cleansing and analysis (40, 41). In this step, the collected data are inspected, transformed and irrelevant data are removed. This initial stage of data analysis is intended to allow ML model developers to more fully explore their data and clean it. This can be done by examining several rows of the sample set, checking basic statistics and fixing null or missing numerical values (through imputation or data extrapolation). Converting the data set into some kind of graphical representation is often very useful especially where the data anomalies or patterns are not clear from numerical statistics. Data visualization (through heat maps, correlation maps or principal component analysis) helps ML model developers identify data outliers or check the effects of any major data changes. After the data cleansing step has been completed, the next stage involves transforming the data and then determining which features might be useful in training the ML model. For data compatibility it is often necessary to do specific data transformations. An example of a data transformation is converting non-numeric features into numeric types. To ensure optimal performance of the model, certain types of transformations often have to be done, such as lower-casing of text data, normalizing numeric features, etc. Lastly, feature importance assessments or feature selection calculations must be done to help in model training. The overall goal of feature selection or feature importance assessment is to infer meaningful insights about the data and make the data suitable for use in various ML algorithms. Feature selection can be done manually or automatically. In this step a decision must be made regarding which ML algorithms are compatible for the collected dataset and the specific feature set.

### Data Partitioning in ML

In the data partitioning phase, the dataset is divided into subgroups. In ML usually two groups are created. One data grouping is used to train the model and the other part is used to test the model. The basic rule in ML is to split the dataset and assign 80% of the dataset for training and 20% for testing the model, respectively. There is no specific set of rules on how the dataset should be partitioned. Usually, random sampling or stratified random sampling methods are used in most ML data partitioning processes. In random sampling, the ML modelling process is protected from bias arising from specific or intrinsic data characteristics. However, random sampling may inadvertently cause an uneven distribution of the data in the training and testing datasets. Stratified random sampling is another random sampling process but this approach ensures that the data is properly distributed between training and testing sets.

### Model Training and Performance Evaluation in ML

In ML the training dataset is used to train the model, i.e., to let the algorithm learn from the data. This process allows the one to tune different parameters in the ML algorithm, develop the models and compare the performance between different models. In many cases the training process involves some kind of N-fold cross validation of the training data set.  This reduces the chances of overtraining. Once the ML model is trained, the performance of the trained model is measured by how it handles completely new or never-before-seen observations. Typically, the testing (or hold-out) dataset is used for this purpose. Ideally the performance of the model (through different measures of accuracy or sensitivity or correlation) for both the training and the testing (hold-out) data set should be within 5% of each other. After the training and performance evaluation are complete, the final model must be saved in a particular file format and deployed in a production or live environment. A diagram illustration the basic workflow in the ML process is shown in Fig 1.12.

### Types of Machine Learning Algorithms

There are four general types of ML methods: supervised, unsupervised, semi-supervised and reinforcement learning. Supervised ML works on a labeled dataset. That is, for input x there is a mapped output y. In the training dataset for supervised learning, one needs the correct pairs of input x and expected output label y. In supervised learning this method eventually learns to provide reasonably accurate guesses of the output by taking just the input alone without the output label (33). Supervised learning models are subdivided into classification and regression methods. Classification methods predict categorial outputs e.g., “yes” or “no”, “true” or “false” etc. Some examples of popular classification ML algorithms are Artificial Neural Networks (ANN), random forest (RF) algorithms, decision trees (DT), logistic regression (LR) methods, and support vector

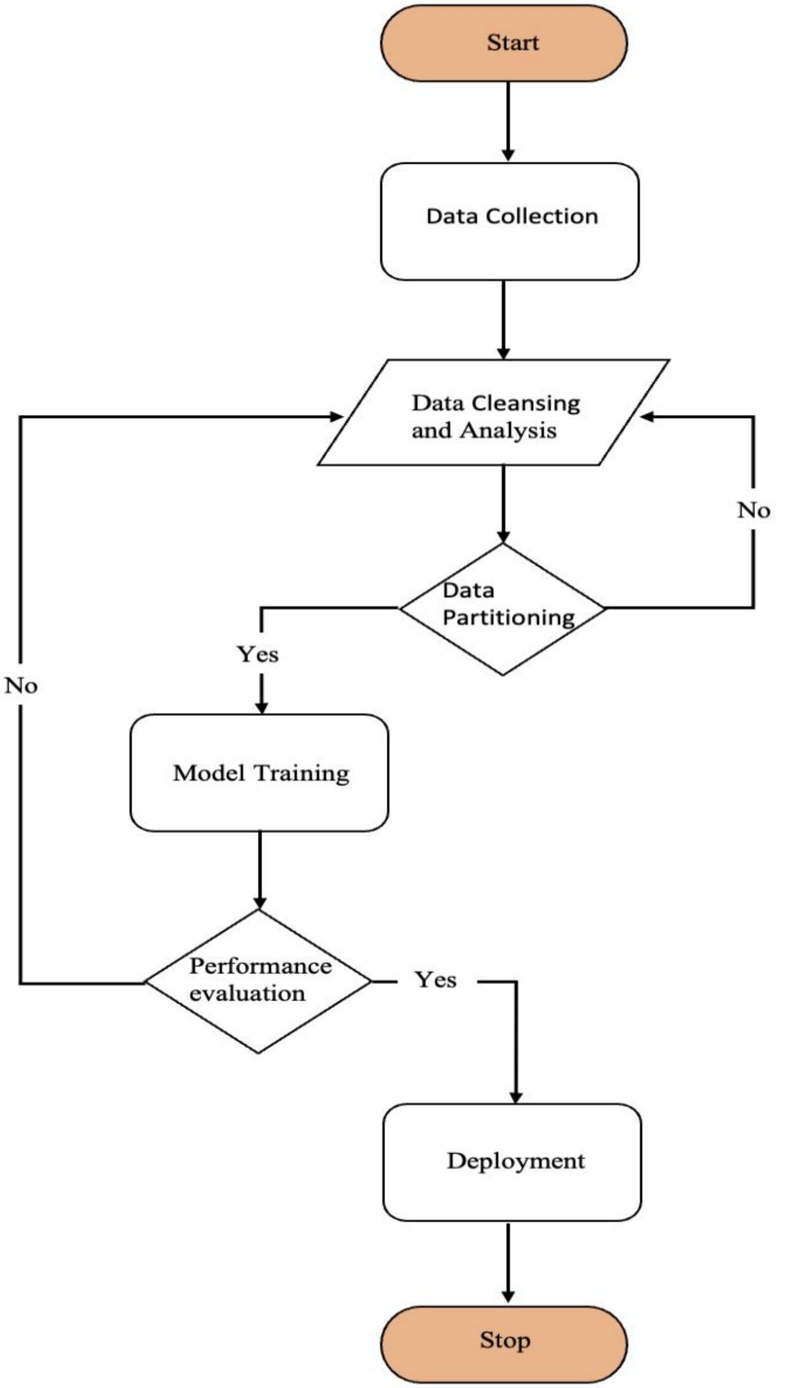


Figure 1.12: A basic flow chart outlining the Machine Learning process.

machine (33) models (SVMs). Regression algorithms handle continuous output variables and try to predict numbers or numerical data instead of categories. Common examples of regression algorithms are simple linear regression, multivariate regression, decision tree regression, and lasso regression. Note also that RF and SVM regression is also possible.

Unsupervised ML is different than supervised ML. In unsupervised ML, the training data are not associated with any output labels. Depending on the patterns, structures, similarities, and differences in the training dataset, an unsupervised ML algorithm decides to divide the dataset into groups and learns from those groupings how to predict the output without any supervision. Unsupervised algorithms are subdivided into two categories: clustering and association. Unsupervised clustering algorithms will group the dataset into clusters based on certain selected parameters and predict the output by deciding which type of groups the input data falls into. Examples of unsupervised clustering algorithms are the K-means clustering algorithm, the mean-shift algorithm, DBSCAN, etc. Unsupervised association algorithms find the dependency of various data items and then map the associated variables to create the input files. Examples of unsupervised association algorithms are the Apriori algorithm, the Eclat algorithm, and the FP-growth algorithm.

Semi-supervised algorithms can use both labeled and unlabeled data. In other words, semi-supervised ML is a combination of supervised and unsupervised ML. Lastly, reinforcement learning is a type of ML algorithm that uses a feedback-based process and learns from experience only. Almost all ML methods I will mention in this thesis are supervised ML methods.

## Machine Learning Algorithms in the Prediction of NMR Chemical Shifts

ML approaches to predict NMR chemical shifts are becoming increasingly popular. They are being used particularly in the prediction of $^{1}$H and $^{13}$C chemical shift prediction. Artificial Neural Networks (ANN) have become very popular in the prediction of NMR chemical shifts. ANNs were introduced in the field of chemistry beginning in the early 1990s, which is very close to the time when the first publication on $^{13}$C chemical shift prediction using ANNs was published (42–44). These papers used ANNs to predict $^{13}$C chemical shifts in monosubstituted benzenes. Since then, ANNs have been adapted to predict $^{1}$H and $^{13}$C NMR chemical shifts for other organic molecules including alkanes, alkenes, substituted benzenes, cyclohexanes and so on. However, these early ANNs were not much different in their capabilities from the rule-based systems developed in the 1960s and 1970s. To make ANN chemical shift prediction more general, Meiler et. al. (45) in 2000, trained an ANN model consisting of 40K molecules with 526,565 $^{13}$C chemical shifts atoms. The model consisted of 9 individually trained NN models with 9 types of carbons and 368 atomic features/descriptors. To mitigate solvent and chemical shift referencing effects, Meiler only selected experimental chemical shift values measured in $CDCl_3$/or $CCl_4$ and only chose those compounds that were referenced to tetramethylsilane (TMS) as the internal standard. Meiler's new ANN outperformed the rule-based methods in terms of accuracy. It was also 1000 times faster than the database search method for chemical shift prediction. Based on the test dataset which had more than 15,000 carbon atoms, the Meiler ANN model showed a mean $^{13}$C chemical shift deviation of 1.80 ppm and a standard deviation of 2.10 ppm.

An improved version of Meiler's ANN was released later (46) by adding more atomic descriptors to describe the carbon atom environment more completely. The new model was trained with the same previous ~15K carbons. The mean deviation for the improved ANN model fell to 1.60 ppm (compared to 1.80 ppm with the previous ANN). This new method was compared to the HOSE code-based prediction using another independent dataset of 100 molecules. The ANN based model showed a standard deviation of 2.7 ppm compared to the SD of 2.6 ppm with the HOSE code-based predictor. Therefore, this improved ANN model had a performance that was almost as good as the HOSE code prediction model.

In 2002, DeSousa et al. (39) used counter propagation neural networks (CPNNs) to predict $^1$H chemical shifts via ML. DeSousa et. al. used topological, physicochemical and geometrical descriptors to ensure the robustness of their model. The study was done on four types of protons separately: protons in aromatic systems, protons in the pi non-aromatic systems, protons in a rigid aliphatic system and protons in a non-rigid aliphatic system. Chloroform ($CDCl_3$) was the solvent for all experimental $^1$H chemical shifts used. The dataset was restricted to protons attached to carbons only and only those molecules with C, H, N, O, S, F, Cl, Br and I in their molecular formula. The authors tested their trained model on 259 test cases and obtained a Mean Absolute Error (MAE) of 0.25 ppm between the experimental and predicted $^1$H shifts. Additionally, the model was able to differentiate between stereoisomeric protons in nonaromatic pi and rigid aliphatic systems. Overall, this ML model was able to provide a similar quality of chemical shift prediction results as found in the best available commercial software packages. However, it is important to note that the model was trained on a very small dataset, consisting of only 744 protons. Given the small size of both the training and testing data set, doubts may be expressed about the generality of the model.

To improve the robustness of their original model, Binev and DeSousa (47) used an ensemble Feed Forward Neural Network (FFNN) with the same training set. In this ML method, a set of networks are independently trained with the same training dataset but are built with different configurations. Each of the constructed networks can provide different results for the same example. An average result from the individual network's predictions provides a final prediction result for the example. These authors optimized the selection of descriptors, the number of the neurons in the hidden layer and the size of the FFNN ensembles. Using a larger test dataset consisting of 952 test samples, these authors found a significant improvement. In particular, they obtained a $^{1}$H chemical shift MAE of 0.29 ppm, whereas with the previous model, the $^{1}$H chemical shift MAE was 0.36 ppm. To further improve the model, they expanded the data set with more experimental chemical shift data and used the previously trained ensembles of FFNNs (48). In this later study they used an Associative Neural Network (ASNN) where the new dataset served as a memory for the ASSN. This led to a significant performance improvement on the 952 test samples, achieving a $^{1}$H MAE of 0.19 ppm.

In 2008 Kuhn et. al. (49), assessed several different machine learning methods for predicting $^{1}$H NMR chemical shifts. A number of algorithms were tested, including multivariate linear regression, support vector machines (SVM), decision trees and random forest techniques to create a $^{1}$H chemical shift predictor. They used the nmrshiftdb2 data set as their training/testing set. At the time, this database had nearly 18K $^{1}$H experimental chemical shifts. The protons in this data set were divided into four classes: protons with an aromatic system, protons in a non-aromatic pi system, and protons in rigid and non-rigid aliphatic systems. The percentage of proton shifts in the training dataset for each proton class was 21%, 7%, 27% and 45%, respectively. CDKit (Chemistry

Development Kit) (50, 51) was used to calculate the properties or descriptors of the atoms. In total 246 atomic descriptors were used. Interestingly, Kuhn et al. found that HOSE code methods performed better than all other machine learning algorithms. Random forests and j48 decision trees algorithms performed the best among the ML methods, especially when both categorical and numerical features were used.

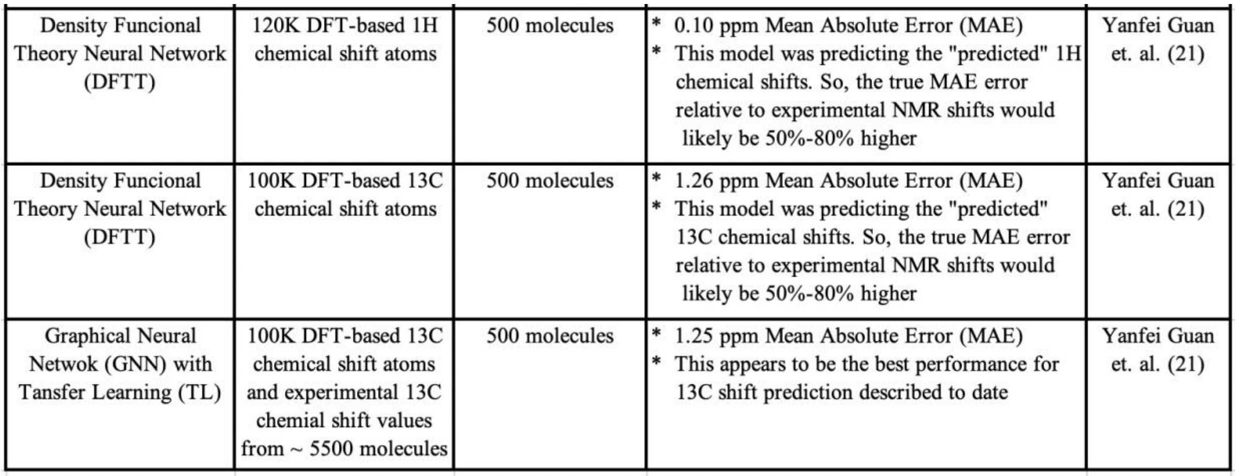
Recently Jonas and Khun (52) tried one of the most popular ML methods, graphical neural network (GNNs), to predict both $^{1}$H and $^{13}$C chemical shifts. They used ~32K molecules with an average molecular size of 29 atoms. They compared their test results with the standard HOSE code methods and found that their GNN either tied or outperformed the HOSE code method. Unfortunately, the training and testing dataset used in this study ignored solvent effects. As noted before this can alter experimental $^{1}$H and $^{13}$C chemical shift values. These authors also ignored stereochemical and geometry-specific effects in their experiment. In addition to the work of Kuhn et. al. (52), another study was conducted by Kwon et. al. (53) in 2020 using message passing neural networks (MPNNs). MPNNs can reduce the time and space complexity that GNNs tend to suffer from. Furthermore, instead of using explicit hydrogen atoms in their model, they used implicit hydrogen atoms as node features of adjacent nodes. As a result, their edge representation is more flexible than that of Khun's. This allowed them to add more features to make the molecular graphical representation more informative. As with most other ML studies the training/testing dataset was taken from nmrshiftdb2. The performance of this MPNN model was measured against the standard HOSE code and Kuhn's GNN model (52). These authors found that their MPNN model performed better than these other two models. In particular the MPNN model achieved an MAE 0.22 ppm for $^{1}$H chemical shift prediction whereas the HOSE code and GNN models only achieved an MAE 0.33 ppm and 0.28 ppm, respectively. For $^{13}C$ chemical shift prediction, the MPNN model achieved an MAE of 1.36 ppm and whereas the other models had MAEs of 2.85 ppm and 1.43 ppm. Even though this MPNN model shows some promising results, the authors noted that it suffers from inconsistency in terms of handling different solvent and temperature conditions.

In 2021, Yanfei Guan et. al. (21) tried a new ML approach called transfer learning (TL). In this using strategy, they developed a GNN model with DFT calculated chemical shift data to predict $^{1}H$ chemical shifts and then applied TL to predict experimentally measured $^{13}C$ chemical shifts. They named this model DFTNN (density functional theory neural network). This approach nicely bypasses the problems of collecting and fixing/cleaning a large dataset with a large number of assignment errors, partially assigned structures, or incomplete spectral data. To construct the DFT set they turned to the nmrshiftdb2 database, which has 43K molecules, then they selected 20K neutral organic molecules whose molecular weight was less than 500 Daltons. After that molecular weight filter was applied, they selected those molecules that appeared to be computationally manageable and which exhibited good structural diversity. The final dataset that they used contained 8K molecules. These 2D structures were then converted into 3D structures and their chemical shifts were calculated using a standard DFT based method. The resulting dataset produced 120K DFT-based $^{1}H$ chemical shift assignments and 100K DFT-based $^{13}C$ chemical shift assignments. The authors then trained a GNN model and used 500 held-out structures as a test set to measure the performance of their model. The resulting model produced an MAE of 0.10 ppm for $^{1}H$ chemical shifts and an MAE of 1.26 ppm for $^{13}C$ chemical shifts. Note that these models were only able to "predict" $^{1}H$ and $^{13}C$ shifts, so the true MAE error relative to experimental NMR shifts would likely be 50%-80% higher. Nevertheless, they felt their model was sufficiently well trained to be able to learn how to predict experimental NMR shifts. Therefore, they refined their GNN model by using TL on actual experimental NMR data. To do so, they took ~5500 molecules with experimental shifts from the nmrshiftdb2 database and held out 500 molecules testing the TL model. They tested their model only on $^{13}$C data because nmrshiftdb2 data doesn't have solvent information for all the compounds and it is well known that $^{1}$H chemical shifts are very sensitive to solvent. This TL model is called ExpNN-dft and it achieved an MAE of 1.25 ppm on experimental $^{13}$C NMR data. This appears to be the best performance for $^{13}$C shift prediction described to date.

This TL model has a few limitations, however. For instance, it does not process molecules with formal charges and is limited to handling molecules with the elements C, H, N, O, S, P, F, Cl. Furthermore, this tool requires 3D structures as input and it must generate conformers of a molecule by using Merck molecular force field (MMFF), so, to some extent, the accuracy of the model is controlled by the quality of the MMFF 3D structure generation tool. Also, this tool is limited to calculating chemical shifts for those molecules with less than 50 heavy atoms. Users must also specify stereochemistry manually where appropriate. Moreover, the authors did not compare the results with other models like HOSE code-based models and other commercially available prediction tools. Table 1.3 compares the performance of the various ML-based predictors for $^{1}$H and $^{13}$C chemical shift prediction that were described here.

In 2021, Jonas et al. (54) conducted a review on different approaches of NMR chemical shift predictions and concluded that HOSE code and ML based techniques perform comparably. From my own literature review, I found some promising ML based predictors for $^{1}$H and $^{13}$C NMR chemical shifts but none of them quite offers the complete package in terms of handling solvent effects, pH effects, reference compound effects or molecules with formal charges.

| Machine learning algorithm | Number of the molecules/ samples used in the training dataset | Number of the molecules/sample used in the test dataset | Results | Reference |
|---|---|---|---|---|
| ANN | 40K molecules with 526,565 13C chemical shifts atoms | ~1K molecule with >15K 13C chemical shift atoms atoms | * Mean deviation of 1.80 ppm<br>* Standard deviation of 2.10 ppm<br>* 1000X faster than the database search method | Meiler et. al. (45) |
| ANN | 40K molecules with 526,565 13C chemical shifts atoms | ~1K molecule with >15K 13C chemical shift atoms | * Mean deviation of 1.60 ppm<br>* Better than the model by Meiler et. Al. (45) | Meiler et. al. (46) |
| Counter Propagation Neural Networks (CPNNs) | 744 1H chemical shifts atoms | 259 1H chemical shifts atoms | * 0.25 ppm Mean Absolute Error (MAE) | DeSousa et. al. (39) |
| Feed Forward Neural Network (FFNN) | 744 1H chemical shifts atoms | 259 1H chemical shifts atoms | *. 0.24 ppm Mean Absolute Error (MAE)<br>* Better than the model by DeSousa et. al. (39) | Binev and DeSousa (47) |
| Counter Propagation Neural Networks (CPNNs) | 744 1H chemical shifts atoms | 952 1H chemical shifts atoms | *. 0.36 ppm Mean Absolute Error (MAE) | DeSousa et. al. (39) |
| Feed Forward Neural Network (FFNN) | 744 1H chemical shifts atoms | 952 1H chemical shifts atoms | *. 0.29 ppm Mean Absolute Error (MAE)<br>*. Better than the model by DeSousa et. al. (39) | Binev and DeSousa (47) |
| Associative Neural Network (ASNN) | 5631 1H chemical shifts atoms | 952 1H chemical shifts atoms | * 0.19 ppm Mean Absolute Error (MAE)<br>* Better than the model by Binev and DeSousa (47) | DeSousa et. al. (48) |
| Message Passing Neural Network (MPNN) | 1H chemical shift atoms from nmrshiftdb2 databse. Exact number is not available | 1H chemical shift atoms from nmrshiftdb2 databse. Exact number is not available | * 0.22 ppm Mean Absolute Error (MAE)<br>* Better than HOSE CODE based model<br>* Better than the model by Jonas and Khun (52) | Kwon et. al. (53) |
| Message Passing Neural Network (MPNN) | 13C chemical shift atoms from nmrshiftdb2 databse. Exact number is not available | 13C chemical shift atoms from nmrshiftdb2 databse. Exact number is not available | * 1.36 ppm Mean Absolute Error (MAE)<br>* Better than HOSE CODE based model<br>* Better than the model by Jonas and Khun (52) | Kwon et. al. (53) |



| | | | | |
|---|---|---|---|---|
| Density Funcional Theory Neural Network (DFTT) | 120K DFT-based 1H chemical shift atoms | 500 molecules | * 0.10 ppm Mean Absolute Error (MAE)<br>* This model was predicting the "predicted" 1H chemical shifts. So, the true MAE error relative to experimental NMR shifts would likely be 50%-80% higher | Yanfei Guan et. al. (21) |
| Density Funcional Theory Neural Network (DFTT) | 100K DFT-based 13C chemical shift atoms | 500 molecules | * 1.26 ppm Mean Absolute Error (MAE)<br>* This model was predicting the "predicted" 13C chemical shifts. So, the true MAE error relative to experimental NMR shifts would likely be 50%-80% higher | Yanfei Guan et. al. (21) |
| Graphical Neural Netwok (GNN) with Tansfer Learning (TL) | 100K DFT-based 13C chemical shift atoms and experimental 13C chemial shift values from ~ 5500 molecules | 500 molecules | * 1.25 ppm Mean Absolute Error (MAE)<br>* This appears to be the best performance for 13C shift prediction described to date | Yanfei Guan et. al. (21) |

Table 1.3: A table illustrating the performance of various ML based predictors for $^{1}$H and $^{13}$C chemical shift prediction.

## Thesis Hypothesis and Aims

The lack of large reference libraries containing experimentally measured $^{1}$H and $^{13}$C NMR chemical shifts combined with the experimental challenges of acquiring experimental NMR data means that there is a strong need to develop computational tools that can accurately predict $^{1}$H and $^{13}$C NMR chemical shifts. My literature review indicates that there are several promising approaches for predicting $^{1}$H and $^{13}$C chemical shifts for small molecules, but that none of the existing methods achieves sufficient accuracy or handles solvent effects, pH effects, reference compound effects or molecules with formal charges. *I hypothesize that it is possible to develop ML-based methods that can accurately predict $^{1}$H and $^{13}$C chemical shifts of small molecules with an MAE of <0.20 ppm for $^{1}$H shifts and an MAE of <2 ppm for $^{13}$C shifts*. To test this hypothesis, I will explore the use of support vector machine algorithms, random forest algorithms along with the Gradient Boost regressor, XGBoost regressor and CatBoost regressor to develop appropriate ML predictive models. I will also use experimental NMR data sets that have been carefully curated and partitioned to ensure that the NMR data was collected in the same solvents and have been consistently referenced using DSS or other IUPAC-approved standards.

## Thesis Outline

This thesis will describe my efforts to test the above hypothesis. Chapter 1 (this chapter) provides the background to the topic, including an introduction to NMR, to structure determination by NMR, to the theory behind NMR and NMR chemical shifts and to the current methods used to predict or calculate $^1$H and $^{13}$C NMR chemical shifts. Chapter 2 will focus on the application of machine learning to the prediction of $^1$H chemical shifts of small organic molecules. In this chapter, I will discuss the collection/curation of the NMR training and testing data, the ML algorithms that I tested and the results that I achieved.  Chapter 3 will describe my efforts to apply machine learning to the prediction of $^{13}$C chemical shifts of small molecules. I will describe the collection and curation of the NMR training/testing data, the ML algorithms that I tested and the results that I achieved. Chapter 4 is the last and concluding chapter for this thesis. In this final chapter I will discuss the successes and failures that I faced and describe some potential solutions that may further improve the performance of my ML algorithms.

# Chapter 2: Application of Machine Learning to the Prediction of $^1$H Chemical Shifts of Small Organic Molecules

## Introduction

Chemical shifts serve as the reference points for NMR. They help NMR spectroscopists map out atomic positions, reveal the identity of key chemical groups and ultimately help NMR spectroscopists determine the atomic structures of many small organic molecules. In the field of organic chemistry, $^1$H and $^{13}$C NMR chemical shifts are the most widely used types of chemical shifts in the elucidation of chemical structures. This is because >99% of organic molecules contain hydrogen and/or carbon atoms. The very high natural abundance of $^1$H, combined with the exceptional sensitivity of $^1$H nuclei (due to $^1$H's high gyromagnetic ratio) make $^1$H chemical shifts particularly easy to measure – at least relative to $^{13}$C chemical shifts. In addition, $^1$H chemical shifts are exquisitely sensitive to subtle molecular bonding and structural or geometric effects, which means that $^1$H shifts can provide a tremendous amount of information about molecular structure (pairwise $^1$H proximity, $^1$H bonding, $^1$H geometry). Because of their importance in chemical structure elucidation by NMR, their ready availability in chemical shift databases and their utility in interpreting chemical structures, I decided to focus initially on the prediction of $^1$H chemical shifts using machine learning (ML).

In this chapter I will first provide the formal problem definition. Then I will briefly discuss the terminology and the methodology. I will then go into more detail regarding the collection/curation procedure used in obtaining the experimental NMR $^1$H chemical shift data, including both the training and testing data. Then I will discuss the ML algorithms that I tested, how the performance of each ML model was measured. and the results that I achieved. Finally, I will discuss the comparison of my results with other published or readily available $^1$H chemical shift prediction methods (both academic and commercial).

## Problem Definition

Simply stated, the problem to be solved is to *take a single chemical structure of a small molecule expressed as a SMILES string and to accurately predict the $^1$H chemical shifts of all the observable hydrogen atoms in that molecule in different NMR solvents.* The entire problem can be further decomposed into four separate tasks. The first task is to convert SMILES strings into usable 3D molecular structures with correct chemical geometry. The second task is to add hydrogen atoms to the appropriate heavy atoms in the generated molecular structure for a given NMR solvent. The third task is to develop a method that uses the generated (or available) 3D chemical structure to determine the $^1$H chemical shifts for all visible or detectable hydrogen atoms in the chosen molecule. The fourth task is to modify the predicted chemical shifts to match those seen in the chosen NMR solvent. In the following paragraphs I will briefly explain the terminology used in this problem definition.

## Terminology and Background

SMILES (Simplified Molecular-input Line-entry System) is computer-compatible method or a line notation for describing the structure of chemical species that uses combinations of ASCII character strings. By combining letters (for atomic symbols) with non-alphanumeric characters (bond symbols) chemical structures can be easily represented as text strings. The structures that are generated using SMILES are hydrogen-suppressed, which means that the molecules are represented without hydrogens or H symbols. For instance, the structure for ethanol ($CH_3CH_2OH$) can be represented by the following SMILES string:

1) ***CCO***

Typically, a number of equally valid SMILES strings can be written for the same molecule. For example, CCO, OCC and C(O)C all specify the structure of ethanol. The SMILES specification was initiated by David Weininger at the USEPA Mid-Continent Ecology Division Laboratory in Duluth Minnesota in the 1980s (55). A SMILES string can be converted to a 3D chemical structure that is commonly represented in chemistry as a Structure Data File or an SDF. An SDF is an XML file (Extensible Markup Language) file that describes the atomic positions and bonding patterns of a molecule with specific x,y,z coordinates for specific atoms. SDF files can be generated from SMILES strings using a chemistry package called RDKit (56). RDKit is an open access collection of cheminformatics and machine-learning software written in C++ and Python that has been under development since 2011 (57). RDKit uses a large collection of known or pre-defined bond lengths and bond geometry rules to ensure that all generated chemical structures are geometrically and structurally correct. Any given chemical structure written in SDF can also be decorated with hydrogen atoms using functions written within RDKit. Likewise, based on various rules that are appropriate to the behavior of hydrogen atoms in a given NMR solvent, hydrogen atoms can be removed through atom-specific commands supplied by RDKit. For instance, hydrogen atoms connected to nitrogen or oxygen atoms are known to be rapidly "lost" when compounds are dissolved in water. As a result, they are not visible in the $^1$H NMR spectra of molecules dissolved in water. On the other hand, hydrogen atoms connected to nitrogen or oxygen atoms will remain attached when compounds are dissolved in chloroform. As a result, they are visible in the $^1$H NMR spectra of molecules dissolved in chloroform. Once the geometrically correct atomic structure of a given molecule has been generated, it is then possible to use this structural information to predict the $^1$H chemical shifts. These predictions are normally based on their atomic positions, geometry and bonding patterns (molecular and atomic features). However, $^1$H chemical shifts are also sensitive to the solvent and so systematic corrections to predicted $^1$H chemical shifts must often have to be applied to correct for different “solvent effects”.

## Methodological Outline

The use of ML methods to predict $^1$H chemical shifts requires large collections of accurately generated chemical structures (with correct placement of all H atoms) and accurate, experimentally assigned $^1$H chemical shifts. These structure/shift collections also must have consistent atomic numbering schemes and information about which NMR solvent was used to collect the experimental. A number of databases exist which contain small molecule structures, $^1$H chemical shift assignments and NMR solvent data. These include the HMDB (15), BioMagResBank (11) NMRShiftDB (12) and NP-MRD (14). Compiling, checking and cleaning experimentally collected NMR data proved to be particularly challenging and details regarding the collection and cleaning of this NMR training/testing data are provided in this chapter. Since all $^1$H chemical shifts values are real numbers, the problem of chemical shift prediction via ML requires training and assessment using ML-based regression methods. It also requires the selection of appropriate atomic, molecular and structural features to optimize the performance of the chosen regressor. The feature selection process is particularly important and will be described, in more detail, later in this chapter. In addition to identifying the optimal set of features for $^1$H chemical shift prediction we also explored several popular ML regression algorithms. Different regressors can often yield different results and so it is important to assess different regressors. We evaluated a Support Vector Regressor (SVR), a Random Forest Regressor (RFR), an Extreme Gradient Boosting Regressor (XGBoostRegressor or XGBR) and a Categorial Boosting Regressor (CatBoostRegressor). The comparative performance of these regressors is described as are the performance of these ML models relative to several well-regarding chemical shift predictors.

## Performance Evaluation Metric

When performing regression learning and regression analysis it is important to choose an appropriate evaluation metric. With regression learning, one can use several performance metrics including correlation coefficients, standard deviation or mean absolute error (MAE). We chose MAE to assess our regressor's performance. The higher the MAE, the less accurate the algorithm The MAE is expressed by the equation

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left|y_j - \hat{y}_j\right| \tag{2.1}$$

Where $y_j$ is the chemical shift of the j$^{th}$ sample (atom) in the dataset and $\hat{y}_j$ is the predicted chemical shift of the j$^{th}$ sample (atom) in the dataset. And *N* is the total number of hydrogen atoms in the dataset.

## The Initial $^{1}$H NMR Chemical Shift Dataset

As discussed in the previous chapter, there are several large, publicly accessible NMR chemical shift libraries consisting of experimentally measured $^{1}$H and $^{13}$C chemical shifts for small molecules. The quality of these databases is quite variable as not all are particularly consistent in tracking or storing important experimental information such as solvent, pH, temperature, charge status or chemical shift reference compounds. In cases where this information was catalogued we found tremendous variability in the solvents, pH, chemical shift references and temperatures. Furthermore, there is no consistent or universal method for matching atom numbers to specific chemical shifts. This lack of consistency and uniformity made the data collection and curation process quite challenging.

Based on the quality and coverage available among the various NMR chemical shift databases we decided to work with just three chemical shift libraries: 1) the HMDB, 2) the BMRB and 3) the GISSMO library. The HMDB (Human Metabolome Database) (15) is a comprehensive, high-quality, freely available online database of the small molecule metabolites found in the human body. It contains 768 experimentally collected $^{1}$H NMR spectra for 768 compounds. We found the experimental NMR data and $^{1}$H chemical shift assignments were of very high quality and almost all were collected in a single solvent -- water. The second chemical shift library we used was the Biological Magnetic Resonance Databank (BMRB) (11). The BMRB compiles experimental NMR chemical shift data for both small molecules and large (protein) molecules. It contains over 1000 biological small molecules with assigned $^{1}$H and $^{13}$C chemical shifts at multiple spectrometer frequencies. We found the experimental NMR data and $^{1}$H chemical shift assignments in the

BMRB were of high quality (a few assignment errors were evident) and almost all chemical shifts were collected in a single solvent – water. The third chemical shift library we chose was the Guided Ideographic Spin System Model Optimization (GISSMO) (58) library. The GISSMO database contains about 1000 small molecules and small molecule fragments with assigned or chemical shifts for $^{1}$H. Almost all the chemical shifts in GISSMO were collected in water. For all chemical shifts used in these databases, the chemical shift reference was set to 0.00 ppm using DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) and the pH value was generally reported as being between 7.0-7.4.

## The Training Dataset

Machine learning requires the use of both training and validation (or holdout) datasets. The training dataset is typically a high quality, "gold standard" dataset containing the expected input and the desired output. More simply, it is the dataset that is used to learn the predictive model. The training dataset consisted of 577 molecules with complete 3D structures (with attached protons) and fully assigned $^{1}$H chemical shifts. 430 of these molecules were obtained from the HMDB library. These 430 molecules had a total of 3333 experimentally measured $^{1}$H chemical shift values. Another 103 molecules were obtained from the BMRB library, which corresponded to 508 experimentally measured $^{1}$H chemical shifts. The last set of 44 molecules was collected from the GISSMO library which contributed 366 experimentally measured $^{1}$H chemical shifts. Altogether our training dataset consisted of 4207 experimentally measured $^{1}$H chemical shift values from 577 diverse molecules. These 577 molecules had an average molecular weight of 162, with the smallest molecule having a molecular weight of 31 Daltons and the largest having a molecular weight of 566 Daltons. $^{1}$H chemical shifts in the training dataset were collected in water and referenced to DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid). In assembling the training data set we made sure to include a structurally diverse range of molecules including organic acids, alcohols, amino acids and nucleotides. Note that most of the molecules chosen were relatively water soluble and had a biological origin (microbial, plant or animal). The bias towards natural products was deliberate as we are primarily interested in predicting $^{1}$H chemical shifts for metabolites and other naturally occurring chemicals.

## The Holdout Dataset

To measure the performance of the different trained ML models for $^{1}$H chemical shift prediction we also had to assemble a holdout dataset. Like the training dataset, the holdout dataset is typically a high quality, "gold standard" dataset containing the expected input and the desired output. More simply, the holdout set is a dataset that has not previously been seen by the ML model which is used to test the predictive performance of model.  This means that the dataset was neither used to train the ML model nor selected with any prior knowledge or bias. We compiled two sets of holdout chemical shifts. Our first holdout dataset consisted of 36 structurally diverse molecules chosen at random from the HMDB, BMRB or GISSMO, each of which was dissolved in water and each of which was referenced to DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid). These 36 molecules had a total of 272 experimentally measured $^{1}$H chemical shifts. The average molecular weight of these 36 molecules was 156 Daltons, with the lowest molecular weight being 78 Daltons and the highest being 307 Daltons. The second holdout dataset consisted of 22 organic compounds that were chosen at random from the NP-MRD database. These 22 compounds had a total of 442 experimentally determined $^{1}H$ chemical shifts. All 22 of these compounds were dissolved in deuterated chloroform ($CDCl_3$) and referenced to tetramethylsilane (TMS). These solvent and chemical shift reference conditions are obviously different than those in the first holdout set. Therefore, to bring the chemical shift data in-line with what is reported for compounds dissolved in water and referenced to DSS we had to make some chemical shift adjustments. Based on data provided by Wishart et al. (30, 31), we adjusted all TMS referenced $^{1}H$ chemical shifts in the second holdout set to match DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) referenced $^{1}H$ chemical shifts. Furthermore, because $CDCl_3$ has a different polarity and hydrogen bonding character than water, we also had to adjust the reported $^{1}H$ chemical shifts to match those reported in water, using the solvent scaling equation mentioned at the end of this chapter. For the molecules in this second hold-out set, the average molecular weight was 306 Daltons, with the lowest molecular weight being 224 Daltons and the highest molecular weight being 429 Daltons.

## Atom and Chemical Shift Labeling

A persistent problem with chemical shift assignments is that there is no standard or consistent way to label which atoms are assigned to which $^{1}H$ chemical shifts. Typically, chemical shift assignments are presented visually with atom labels marked on an image of the structure and the chemical shifts are presented separately in a table with the corresponding atom labels from the structural image. In other words, the chemical image provides a chemical shift “map” that associates numbered (or lettered) heavy atoms to atoms bearing hydrogen atoms. While this visual approach to structural or chemical shift mapping works well for humans, it is not computer readable. Figure 2.1 shows an example of a typical text file found in many chemical shift databases that contains the chemical shift values for a compound called 2-isopropylmalic acid. As seen in this file, a table lists a numbered set of carbon atoms along with the associated $^{1}$H chemical shifts for the hydrogen atoms attached to those carbons. These carbon atom numbers correspond to those in Figure 2.2. However, this atom ordering or atom numbering varies tremendously from one structure to another structure, as there is currently no agreed-upon standard atom-numbering scheme. We will use Marvin Sketch (59, 60) to illustrate an example of the problems associated with atom labeling in NMR. Marvin Sketch is among the most popular chemical structure rendering and structure editing tools in use today. Marvin Sketch, like many other structure drawing tools starts the atom numbering for a given molecule with the first atom that the user draws. So if someone starts drawing the chemical structure of 2-isopropylmalic acid beginning with the oxygen atom bearing the alcoholic hydroxyl group, that oxygen will be numbered as atom #1. On the other hand, if someone else drew the same molecule beginning with a carbon atom, that carbon atom will be numbered as atom #1. Therefore, the atom numbering of most molecules drawn with commercial software tools varies depending on how it was drawn by each user. When we analyzed the structures and assignments in the HMDB library, we found (as expected) the molecular structures did not have the same pattern of numbering. Furthermore, the numbering system that was originally used was saved as an image in the HMDB, but the structure's SDF files were not saved at the same time. As a result, the SDF files that could be downloaded from the HMDB had a completely different numbering system. Figure 2.3 shows how the same structure with different atom numbering schemes was found in HMDB library.

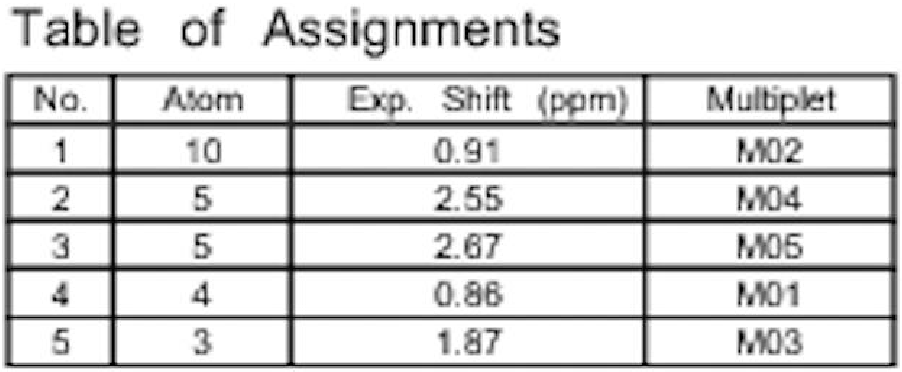
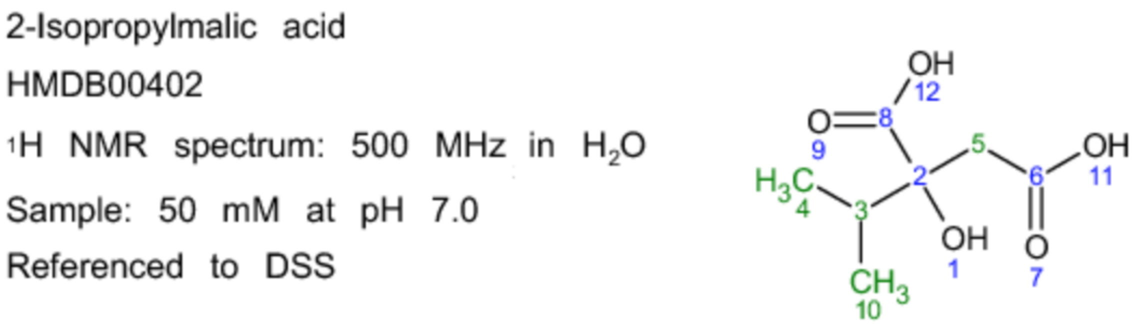


Table of Assignments

| No. | Atom | Exp. Shift (ppm) | Multiplet |
|---|---|---|---|
| 1 | 10 | 0.91 | M02 |
| 2 | 5 | 2.55 | M04 |
| 3 | 5 | 2.67 | M05 |
| 4 | 4 | 0.86 | M01 |
| 5 | 3 | 1.87 | M03 |

Figure 2.1: A table of $^1H$ chemical shift assignments for hydrogen atoms connected to carbon atoms in 2-Isopropylmalic acid (HMDB00402). The numbers in the Atom column refer to carbon atom positions drawn in Figure 2.2.



2-Isopropylmalic acid

HMDB00402

$^1H$ NMR spectrum: 500 MHz in $H_2O$

Sample: 50 mM at pH 7.0

Referenced to DSS

Figure 2.2: The atom-numbered structural image of 2-Isopropylmalic acid (HMDB00402) with heavy atoms numbered in the figure. These numbers are used to map the measured $^1H$ chemical shifts assignments in Figure 2.1.

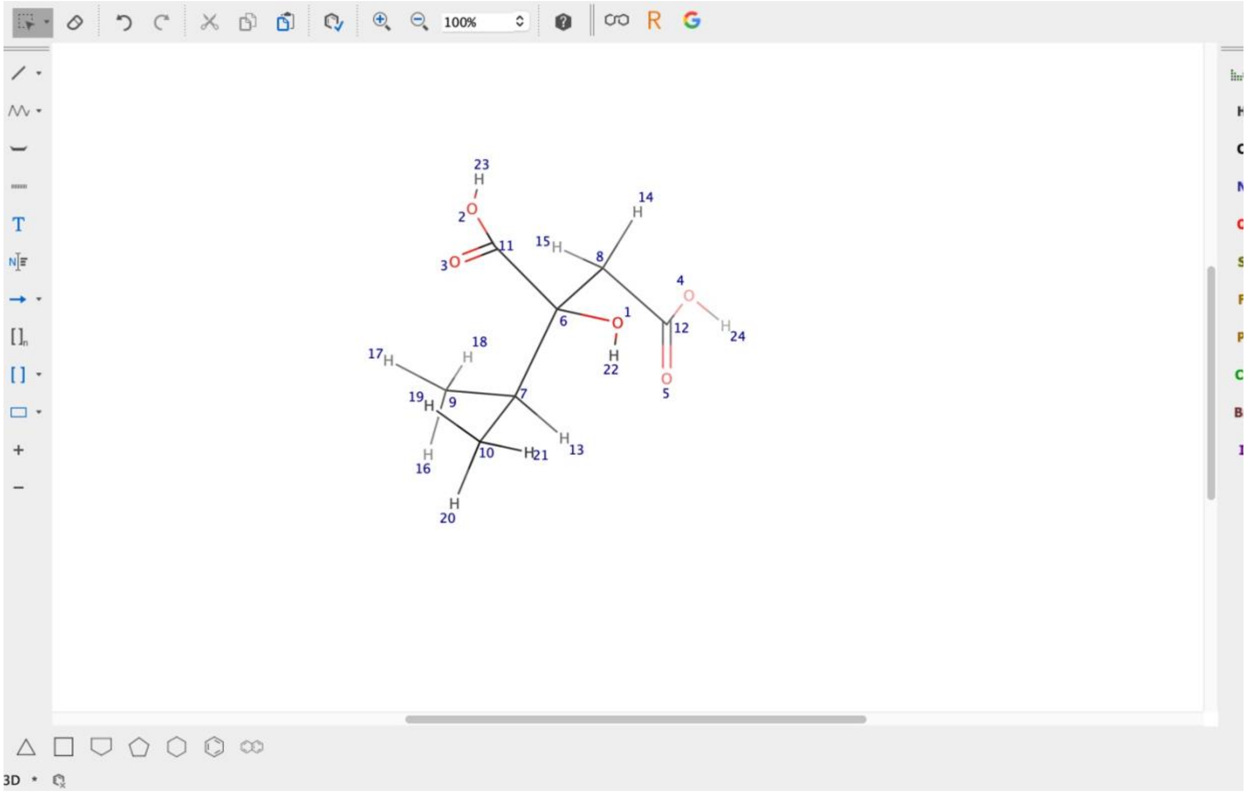


Figure 2.3: The structure of 2-Isopropylmalic acid (HMDB00402) as downloaded from the HMDB server with atom numbering generated via Marvin Sketch.

In order to rectify the problem, we had to manually map the original (PDF or PNG) image of the structure and its atom numbering scheme saved in the HMDB, to the SDF structure files downloaded from HMDB. Another unexpected problem that emerged with the HMDB files was the fact that not all SDF structure files were consistent. We found that some of the molecular structure files for some chemicals were as rendered as "flat" two dimensional structures whereas others were rendered as proper three-dimensional structures.

To overcome these problems, we used a program called Atom Label Assignment Tool using InChI String (ALATIS) (61). ALATIS produces a robust 3D molecular structure and a consistent atom numbering scheme in a stable, repeatable fashion. Using ALATIS, we converted all the structure files from the HMDB into three-dimensional SDF structure files with consistent atom numbering. Next, using Marvin Sketch from ChemAxon, we rotated the structure around different axes to align it with the original PNG or PDF structure image posted in the HMDB. For each molecule, we manually mapped the two atom number schemes to each other by looking at their images side by side. We then manually changed the atom numbers (where the chemical shifts were assigned) in the chemical shift assignment files. Figure 2.4 shows the final chemical shift assignments text file after implementing the above-described procedure. With the chemical shifts properly aligned to the atom numbers represented in the SDF file, we were able to properly calculate all the atomic features used for our ML models. As might be expected, the atom-remapping process was quite time consuming.

In the course of conducting this atom remapping we found a number of problems. For instance, some molecules did not have the same number of $^1$H chemical shifts as H atoms (excluding degenerate shifts seen in methyl groups). In these cases, we simply discarded those molecules from the dataset. We also found a number of duplicate molecules. To remove the duplicate molecules from our datasets, we first converted all structures in our training and holdout datasets into InChI (International Chemical Identifier) strings using RDKit. InChi is a textual identifier for chemical substances created to provide a uniform method of encoding molecular information. We then compared the InChIs to each other. If any common InChI was found, the duplicate compound corresponding to that redundant InChi was eliminated. We also found several

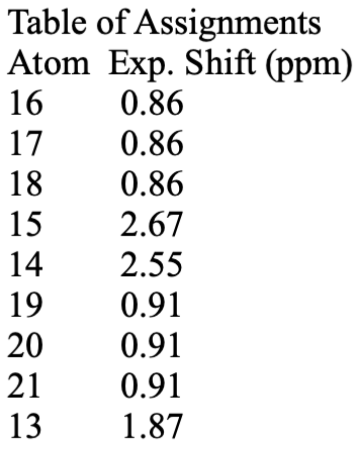


Table of Assignments

| Atom | Exp. Shift (ppm) |
|---|---|
| 16 | 0.86 |
| 17 | 0.86 |
| 18 | 0.86 |
| 15 | 2.67 |
| 14 | 2.55 |
| 19 | 0.91 |
| 20 | 0.91 |
| 21 | 0.91 |
| 13 | 1.87 |

Figure 2.4: The final chemical shift assignment file for 2-Isopropylmalic acid (HMDB00402) using the process of manual atom. As shown here, one must begin by mapping the atoms of Figure 2.2 and Figure 2.3, then one must replace the Atom column in Figure 2.1 with the atomic positions from Figure 2.3. Atomic features for the ML model must therefore calculated from the structure in Figure 2.3 since the structure from Figure 2.2 exists only as an image file, not an SDF file.

errors in the molecular SDF files. For instance, we found that some molecular isomers had identical SDF files. One such example involved Erythritol (HMDB0002994) and D-Threitol (HMDB00041336) (Figure 2.5). To correct this error, we downloaded the correct SDF files from PubChem (62).

After completing the structure "cleaning" and remediation process we then manually checked all the $^1$H chemical shift assignments for all the molecules in the data set. In this checking and correction procedure, we used a commercial program called MNOVA (MestReNovA) (63). MNOVA is a popular NMR data analysis package which offers a full selection of software tools for processing and visualising high-resolution NMR spectra. We used MNOVA-predicted chemical shifts to identify manually assigned chemical shifts that seemed unusual or questionable. If the difference between the MNOVA predicted shift and the observed/reported shifts was >1.0 ppm for any hydrogen atom in a given molecule, we manually rechecked those assignments and made appropriate corrections if errors were found. If we could not rationalize the difference, we discarded that entry. We also used information from the Reich $^1$H chemical shift database (64) to cross check the experimentally reported $^1$H NMR chemical shift values against those predicted based on their known positions within molecules. Additionally, we used the BMRB database to compare reported $^1$H chemical shift assignments against those reported in the HMDB database (where structural overlaps occurred). This also helped correct mis-assigned chemical shifts. To further assess the chemical shift assignments, several NMR experts were also involved.

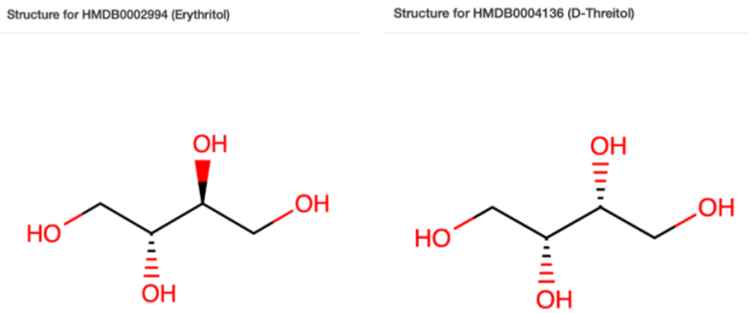
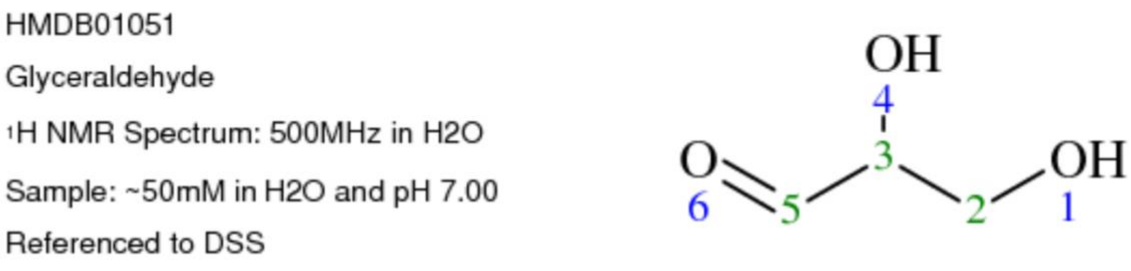

Structure for HMDB0002994 (Erythritol)
Structure for HMDB0004136 (D-Threitol)
OH
HO
OH
OH
OH
HO
OH
OH

Figure 2.5: Erythritol (left side) and D-Threitol (right side). They both had same molecular SDF files.

Here we will show two examples of how incorrect chemical shift assignments were identified and corrected in the training dataset. The first example is Glyceraldehyde (HMDB HMDB01051). Figure 2.6 shows the chemical structure of Glyceraldehyde and atom numbering scheme stored in the HMDB. Figure 2.7 (right side) shows the PNG image file of the assigned chemical shifts for Glyceraldehyde as stored in the HMDB and the corresponding mapped atom numbers using Marvin Sketch from its SDF file.



HMDB01051

Glyceraldehyde

$^1$H NMR Spectrum: 500MHz in H2O

Sample: ~50mM in H2O and pH 7.00

Referenced to DSS

Figure 2.6: Molecular structure of Glyceraldehyde (HMDB01051) in the stored image file.

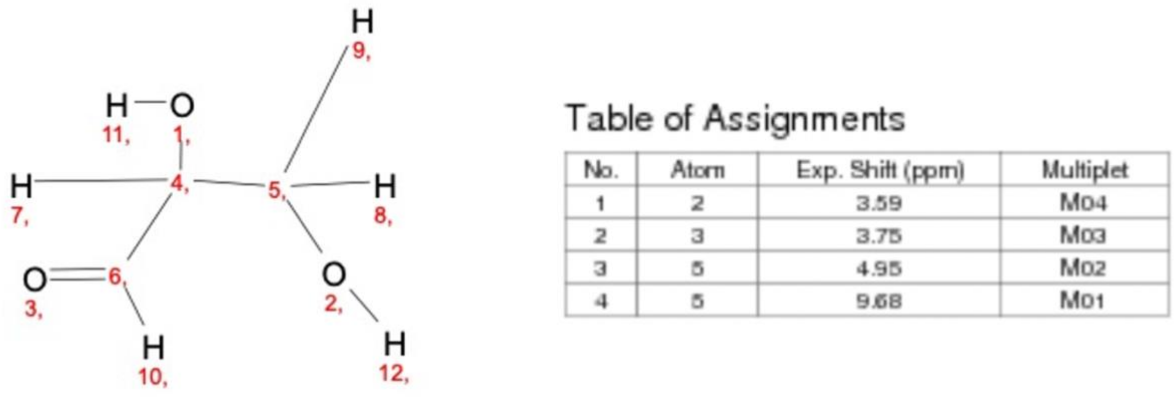
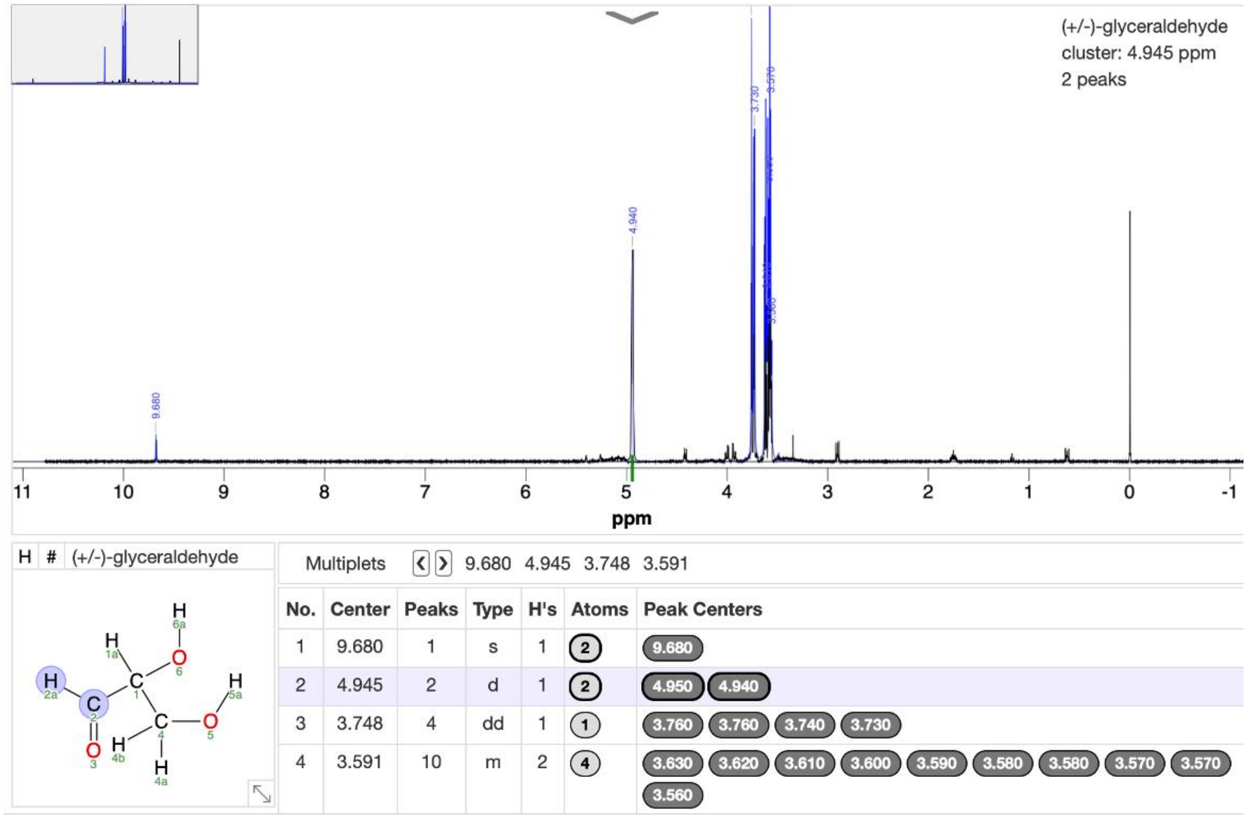


Table of Assignments

| No. | Atom | Exp. Shift (ppm) | Multiplet |
|---|---|---|---|
| 1 | 2 | 3.59 | M04 |
| 2 | 3 | 3.75 | M03 |
| 3 | 5 | 4.95 | M02 |
| 4 | 5 | 9.68 | M01 |

Figure 2.7: Molecular structure of Glyceraldehyde (HMDB01051) drawn using Marvin Sketch (left side). The table of chemical shift assignments corresponds to the atom numbering in Figure 2.6 (right side).

If we compare the structure in Figure 2.6 with that in Figure 2.7, we can see that the carbon atom 5 in Figure 2.6 matches with the carbon atom 6 in the SDF file in Figure 2.7 (left side). Thus, the number of the attached hydrogen atom with this carbon is 10. We mentioned previously that in the image file of HMDB's chemical shift assignment table, the chemical shift of a hydrogen is normally mapped to the carbon atom that it is bonded to. From the table in Figure 2.7 (right side), we see that the chemical shift assignment for atom number 5, has two values: 4.95 ppm and 9.68 ppm. This is not physically possible and indicates some ambiguity in the chemical shift assignment (Figure 2.8).



(+/-)-glyceraldehyde
cluster: 4.945 ppm
2 peaks

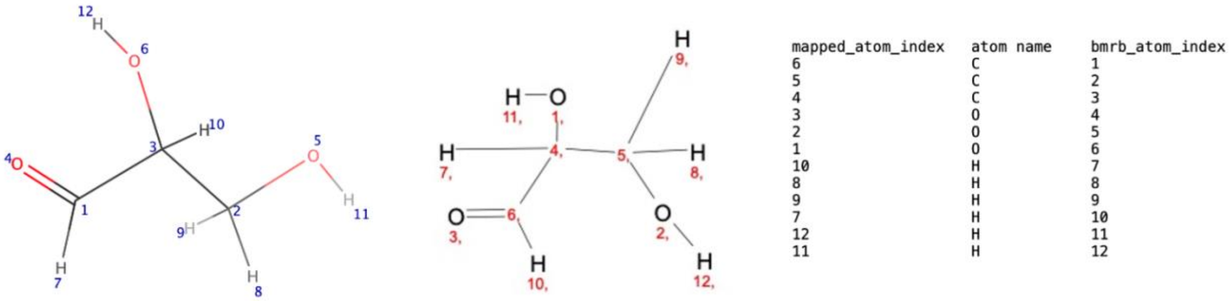
H # (+/-)-glyceraldehyde

Multiplets 9.680 4.945 3.748 3.591

| No. | Center | Peaks | Type | H's | Atoms | Peak Centers |
|---|---|---|---|---|---|---|
| 1 | 9.680 | 1 | s | 1 | 2 | 9.680 |
| 2 | 4.945 | 2 | d | 1 | 2 | 4.950 4.940 |
| 3 | 3.748 | 4 | dd | 1 | 1 | 3.760 3.760 3.740 3.730 |
| 4 | 3.591 | 10 | m | 2 | 4 | 3.630 3.620 3.610 3.600 3.590 3.580 3.580 3.570 3.570 3.560 |

Figure 2.8: The spectrum of Glyceraldehyde (HMDB01051) as displayed in HMDB's Jpectra viewer. The auto assignment function could not confirm the correct value of that hydrogen atom 2.

To investigate this problem further, we checked the BMRB database entry for Glyceraldehyde. The BMRB ID for Glyceraldehyde is bmse000298. As might be expected, the atom numbering system used in the BMRB was different than that of the HMDB (Figure 2.9 left and middle image). As a result, we had to perform another atom number mapping. Figure 2.9 (right most image) shows how the atom number mapping was done between the HMDB SDF file and the BMRB SDF file. Here we see that hydrogen atom number 10 in HMDB maps to atom number 7 in the BMRB. In the BMRB we found that the assigned chemical shift value for that hydrogen atom was 3.583 ppm. We tried to run MNOVA to see what it would predict for the chemical shift of atom number 10.



| mapped_atom_index | atom name | bmrb_atom_index |
|---|---|---|
| 6 | C | 1 |
| 5 | C | 2 |
| 4 | C | 3 |
| 3 | O | 4 |
| 2 | O | 5 |
| 1 | O | 6 |
| 10 | H | 7 |
| 8 | H | 8 |
| 9 | H | 9 |
| 7 | H | 10 |
| 12 | H | 11 |
| 11 | H | 12 |

Figure 2.9: The atom numbering difference between the BMRB (left image) and the atom numbering system the HMDB (middle image) for Glyceraldehyde. Atom number mapping between the HMDB SDF file and the BMRB SDF file for Glyceraldehyde (right image).

Unfortunately, MNOVA could not assign the chemical shift value to hydrogen atom 10 (Figure 2.12). We then analyzed the observed NMR spectrum and noticed there was a small peak at ~9.6 ppm. The observed J-coupling for that peak indicated that it should correspond to hydrogen atom 10. We also looked into the Reich chemical shift database and found that aldehyde $^{1}$H chemical shifts typically are between ~9.3 ppm to ~10 ppm. However, we can see from Figure 2.8, the peak at ~4.9 ppm is much more intense than the peak at ~9.68ppm. After discussing issue with NMR experts, we learned that when aldehydes are dissolved in water, the aldehyde can convert to an alcohol through a reversible equilibrium with a hydrate (geminal-diol or gem-diol) (Figure 2.10).

As a result, Glyceraldehyde can exist in two structural states, one low abundance state as an aldehyde and one higher abundance state as a diol. Therefore, to solve this issue, we created two SDF files for Glyceraldehyde, one with only the HC=O (aldehyde) group and another with $HC(OH)_2$. The structures for the two SDF files are shown in Figure 2.11. For the SDF file containing the diol, the chemical shift for atom number 8 (the left side of the Figure 2.11)  is 4.95 ppm. On the other hand, for the SDF file containing the aldehyde (atom number 10) on the right side of the Figure 2.11 has a chemical shift of 9.68 ppm. Thus, we were able to correct the ambiguity in the NMR chemical shift assignment for Glyceraldehyde by creating two molecules with two separate types of hydrogen atoms.

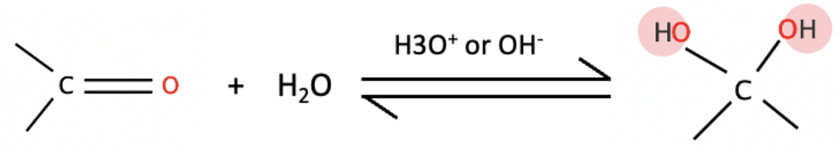


Figure 2.10: Carbonyl function of aldehydes and ketones creating a reversible equilibrium with a hydrate with the present of acid or base. The aldehyde converts to a diol and vice versa.

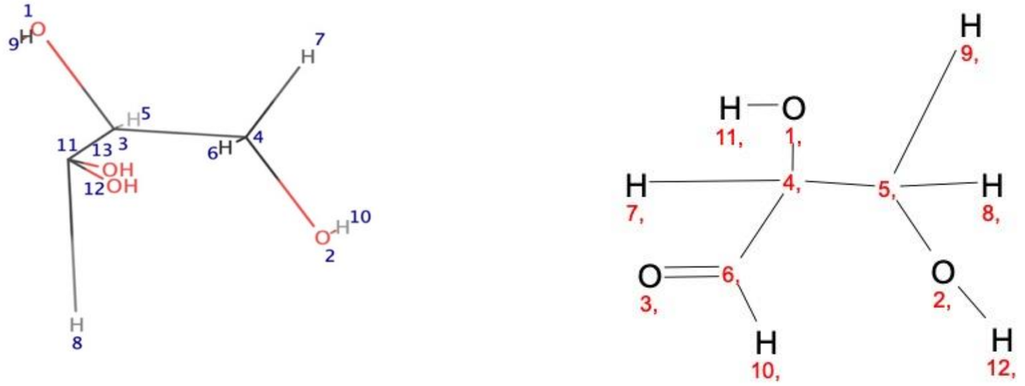


Figure 2.11: Structure of the SDF containing HC(OH)2 (left side) and the structure of the SDF containing HC=O (right side).

Another example where a chemical shift correction was required involved Pipecolic acid (HMDB0000070 - Figure 2.13). Figure 2.14 shows the image file and the chemical shift assignments table along with the atom numbering system in the molecular SDF file for Pipecolic acid. As can be seen in Figure 2.13, carbon atom number 3 corresponds to the carbon atom number 7 in Figure 2.14 (left side) and thus we found the attached hydrogen atoms with this carbon were assigned atom numbers 15 and 16. If we closely look into the chemical shift assignments table in Figure 2.14 (right side), we can see that carbon atom number 3 has three different chemical shift values, whereas there should only be two chemical shift values. As with the previous example, we conducted the same curation steps (MNOVA prediction, BMRB comparison, analysis of the NMR

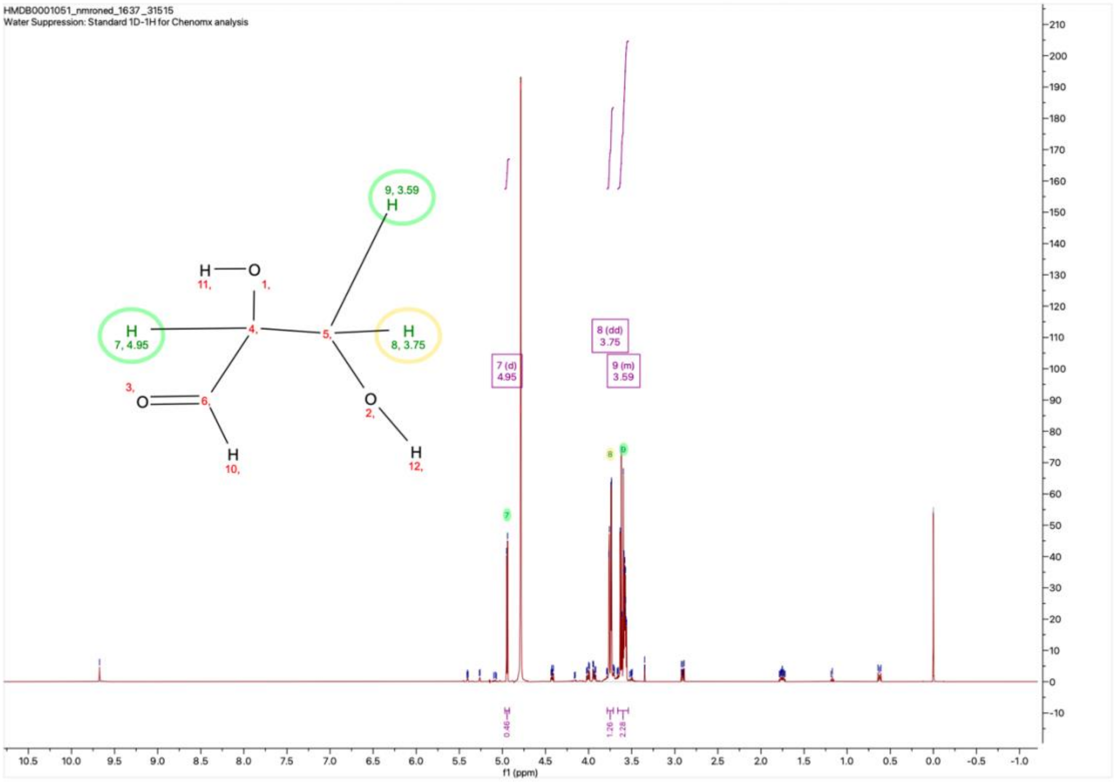
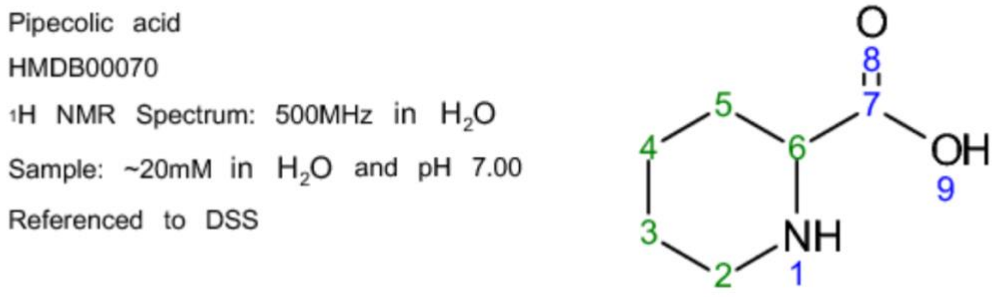


Figure 2.12: MNOVA could not assign the chemical shift value for hydrogen atom 10 from the provided spectrum.

spectrum, consulting NMR experts) and found out that the first two chemical shift assignments indicated the chemical shift values for hydrogen atom 15 and 16 in Figure 2.14. The third chemical shift value (2.99 ppm) should have been associated with the hydrogen atoms bonded with carbon atom number 2 (Figure 2.13). Thus the 2.99 ppm chemical shift should be assigned to either hydrogen atom number 17 or 18 in Figure 2.14 (right side).



Pipecolic acid

HMDB00070

1H NMR Spectrum: 500MHz in $H_2O$

Sample: ~20mM in $H_2O$ and pH 7.00

Referenced to DSS

Figure 2.13: Chemical structure of Pipecolic Acid (HMDB00070) and atom numbering scheme in the HMDB image file.

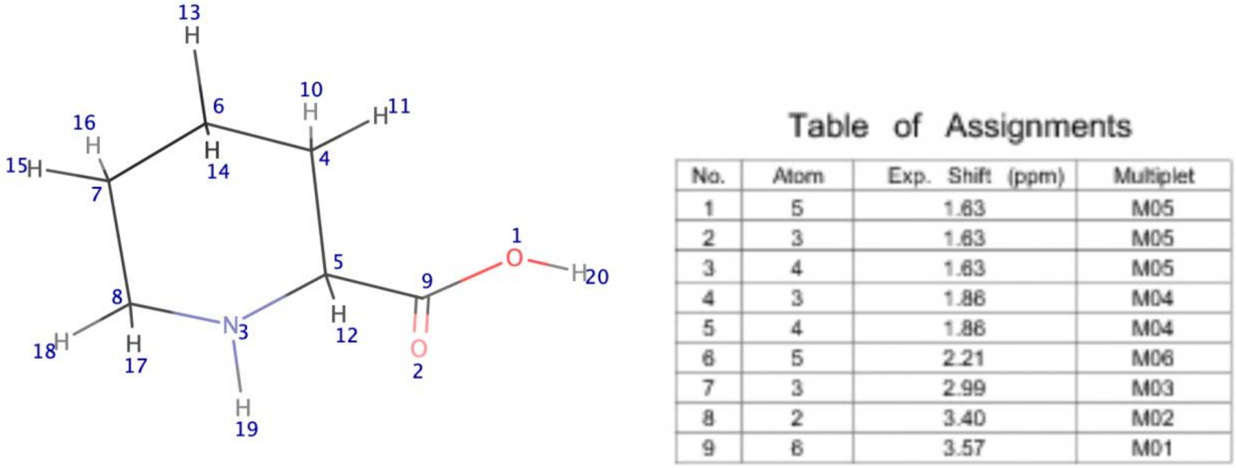


Table of Assignments

| No. | Atom | Exp. Shift (ppm) | Multiplet |
|---|---|---|---|
| 1 | 5 | 1.63 | M05 |
| 2 | 3 | 1.63 | M05 |
| 3 | 4 | 1.63 | M05 |
| 4 | 3 | 1.86 | M04 |
| 5 | 4 | 1.86 | M04 |
| 6 | 5 | 2.21 | M06 |
| 7 | 3 | 2.99 | M03 |
| 8 | 2 | 3.40 | M02 |
| 9 | 6 | 3.57 | M01 |

Figure 2.14: Atom numbering system for Pipecolic Acid in the molecular SDF file (left side) and the image file for the chemical shift assignments table (right side) corresponding to the Figure 2.13.

## Correcting Diastereotopic Proton Assignments

Another challenge we encountered in remediating $^1H$ chemical shift assignments involved the correction of diastereotopic $^1H$ chemical shift assignments. Diastereotopic protons are pairs of hydrogen atoms attached to the same heavy atom in a molecule containing at least one chiral center. They often have distinct $^1H$ chemical shifts. These kinds of protons often belong to a $CH_2$ group located in a chiral molecule, although diastereotopic protons can also be found in achiral compounds. A chiral molecule is a molecule that cannot be superimposed on its mirror image. On the other hand, an achiral compound can be superimposed on its mirror image. Achiral molecules either have a symmetry plane or a symmetry centre.  To understand diastereotopic protons a little better, let us look at the alkene, 1,1-Dimethylethylene (Figure 2.15 left side). We can see the symmetry plane for this molecule in Figure 2.15 on the right side. Since the two olefinic protons (Ha, $H_b$) are equivalent from this molecule's mirror plane, their $^1H$ chemical shifts are the same.

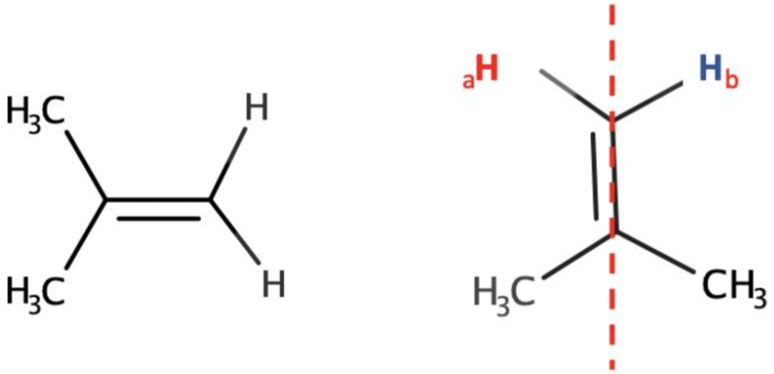


Figure 2.15: 1,1-Dimethylethylen (left side). Two olefinic protons **aH** = **H**$_b$ have identical chemical shifts in the mirror plane of symmetry (right side)

The molecule loses its symmetry if we alter one of the methyl ($CH_3$) groups by changing the carbon to a nitrogen or adding another functional group. The olefinic protons no longer have the same chemical environment since they are no longer identical. Now that there are two olefinic protons in this new asymmetric molecule (Figure 2.16), they are diastereotopic and exhibit different chemical shift values. But it is difficult to say which chemical shift value is associated to Ha or $H_b$. This is because the chemical shift value can flip between the two protons (Figure 2.17). To incorporate this diastereotopic "ambiguity" into our dataset, we calculated the prochirality of all diastereotopic hydrogen atoms. Prochiral molecules are those that can go from being achiral to being chiral in just one step. Prochirality, then, is the quality of an achiral molecule that allows for a single-step transition to chirality. We calculated the prochirality property using the prochirality function in RDKit. We tagged each of the prochiral hydrogens as "1" and "2" using this function. Between the two chemical shift values of the hydrogens in each $CH_2$ group, the higher chemical shift value was assigned to the hydrogen atom that had the prochiral tag "2". Similarly, the lower chemical shift value was assigned to the hydrogen atom that had the prochiral tag "1". A total of ~900 $^1$Hs in our dataset had a prochiral tag and we rearranged the chemical shift values for those hydrogens using our prochiral tag approach.

As part of our remediation effort we also found that for some molecules that had cis and trans isomers, the deposited NMR spectrum did not match with the correct isomeric molecule. Isomers are compounds that contain exactly the same number of atoms, i.e., they have exactly the same molecular formula, but differ from each other by the way in which the atoms are arranged. In chemistry a *cis* isomer is defined as an isomer in which two comparable atoms or groups of atoms are on the same side of a double bond. While *trans*-isomers are defined as isomers in which two similar atoms or groups of atoms are opposed to one another along a double bond. Figure 2.18 shows the example of the *cis* and *trans* variants of Dimethylethylene. Although *cis* and *trans*

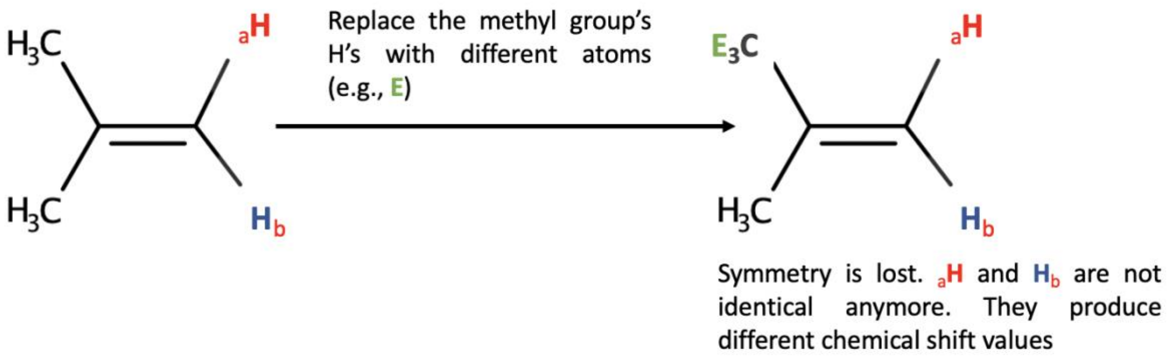

$H_3C$
$_aH$
$H_3C$
$H_b$
Replace the methyl group's H's with different atoms (e.g., E)
$E_3C$
$_aH$
$H_3C$
$H_b$
Symmetry is lost. $_aH$ and $H_b$ are not identical anymore. They produce different chemical shift values

Figure 2.16: After replacing the methyl group's hydrogens with a different heavy atom E, the molecule is no longer symmetrical.

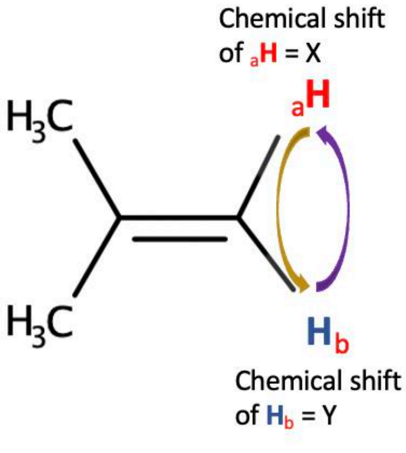


Figure 2.17: If $_a$**H** has chemical shift value **X** and **H**$_b$ has **Y**, those values can switch between each other.

isomers share the same molecular weight and formula, there are several clear differences between them. Several examples of incorrect *cis/trans* spectra of such isomers were identified in the HMDB training set, including 2-Octenoic acid (HMDB0000392) and Cinnamic acid (HMDB0000567). For these molecules each of the SDF files depicted a *cis* isomer whereas the corresponding 1H NMR spectra were for the *trans* isomer. We manually corrected all incorrect *cis/trans* isomers in our dataset.

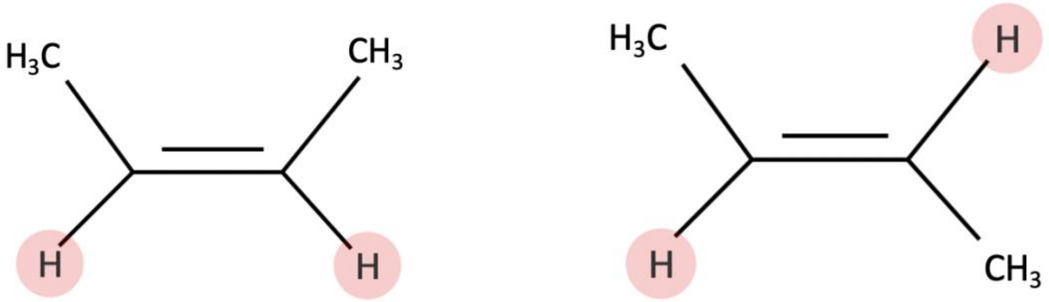


Figure 2.18: Two hydrogen atoms (or the two methyl: $CH_3$ groups) in Dimethylethylene are on the same side of the double bond indicating a cis-Dimethylethylene (left side). When the methyl groups are on the opposite site of the double bond this indicates a trans-Dimethylethylene (right side).

## Feature Identification

In machine learning (ML), a feature is a measurable property or characteristic of a phenomenon. Choosing informative, discriminating and independent features is a crucial element in developing or training effective ML algorithms in pattern recognition, classification and regression. Since our central objective is to accurately predict $^1$H chemical shifts from chemical structures it was essential that we include known features that have been previously determined (by physicists and chemists) to have a major impact on $^1$H chemical shifts. Based on a review of the literature we found that some of the atomic or molecular factors that have an impact on $^1$H chemical shift values are: 1) inductive effects (65–67), 2) van der Waals interactions (68), 3) anisotropic effects (68), and 4) hydrogen bonding effects (69). The inductive effect is defined as the effect on the electron density in one portion of a molecule due to electron-withdrawing or electron-donating groups elsewhere in the molecule. The inductive effect is often described by the electronegativity of specific atom or an adjacent atom. The more electronegative a heavy atom is, the greater the desheilding effect on the attached proton is. An electronegative atom draws an electron from the hydrogen atom, leaving the hydrogen atom with less electron density surrounding its nucleus. This leaves the hydrogen atom's nucleus more susceptible to the effects of the external magnetic field, a phenomenon known as the desheilding effect. This desheilding effect shifts the resonance frequency of that hydrogen atom to a lower value (downfield), thereby producing a higher chemical shift value. With the van der Waals interaction, there is a desheilding effect too. The van der Waals potential is an interaction between non-bonded atoms that has both a short-range repulsive force and a long range (weak) attractive force. The repulsive van der Waals interactions are associated with desheilding effects with $^1$H chemical shifts, while the attractive interactions produce smaller (and opposing) shielding effects with $^1$H chemical shifts. Shielding effects increase the electron density around a nucleus, shifting the resonance frequency higher (upfield), thereby producing a lower chemical shift value. Van der Waals effects are most evident when a molecule is sterically overcrowded. In a sterically crowded or sterically hindered molecule, the electron cloud of the bulky group tends to repel the electron cloud of the nearby protons. As a result, these protons are more exposed (desheilded), and their $^1$H chemical shift value increases. The next most significant factor in determining a $^1$H chemical shift value is the anisotropic effect. The word "anisotropic" means "non-uniform". So magnetic anisotropy means that there is a non-uniform magnetic field that typically arises as a result of the non-uniform electron distribution arising from pi bonds. Many chemicals exhibit varying patterns of electron distribution around their nucleus. Anisotropy can create both shielding and desheilding effects on a proton. The anisotropic effect is more prominent in molecules with double bonds (i.e., pi bonds) such as alkenes, alkynes, aromatic molecules and ketones/aldehydes. Anisotropic effects give rise to a phenomenon called “ring current” shifts which lead to a higher (more downfield) chemical shift for protons attached to aromatic rings.

Chemical shifts can also depend on the presence of hydrogen bonds within or around a given molecule. Hydrogen bonds are non-covalent bonds that may exist between hydrogen bond donors and hydrogen bond acceptors. A hydrogen bond donor contains the hydrogen atom which participates in the hydrogen bond (OH, NH, etc.) whereas the hydrogen bond acceptor contains lone electron pairs (C=O) that attract the hydrogen atom. The stronger the bonding, the more downfield the chemical shift for the hydrogen atom that is participating in the hydrogen bond. Hydrogen bond effects are seen in peptides or arise through solvent interactions

Given their importance in $^1$H chemical shift determination we tried to incorporate as many of the above factors into our feature set. These features include geometric features, physicochemical descriptors and topological descriptors that describe the structure of the molecule, the geometry of the molecule and the character of the bonds and atoms that make up the molecule. One example of a geometrical feature is The Radical Distribution Function (RDF). The RDF describes the probability distribution to find the center of a particle in a given position at a radial distance “r” from the center of a reference sphere. Examples of physicochemical features or descriptors that are important for chemical shift calculations are the partial atomic charge of a proton, the effective polarizability of a proton, and atomic electronegativities. Examples of topological descriptors are numerical representations of information regarding the size, shape, branching, presence of heteroatoms, and different bonds in molecules.

For each molecule in our training and testing dataset, the three-dimensional SDF files were used to generate a set of appropriate atomic features. This was done using the Chemistry Development Kit (CDKit). CDKit is a Java-based cheminformatics software package developed by Steinbeck et. al. (70). CDKit has a library called QSAR, that enables the rapid calculation of many atomic and molecular properties relevant for chemical shift calculation. In particular, the QSAR library generates values for 30 atomic descriptors. These atomic features encompassed nearly all of the atomic traits that would be expected to have an impact on $^{1}$H chemical shift values. In addition to CDKit there is another commonly used cheminformatics package called RDKit (56). However, the number of descriptors in RDKit was much less (only 17 atomic descriptors) and these overlapped with the 30 features in the CDKit library. The atomic features that the CDKit library can calculate include the hybridization state of an atom; its atomic valence; the covalent radius; the Van der waals radius; the atom’s position in the periodic table; the number of non-hydrogen substituents attached to an atom; the effective polarizability of an heavy atom;  an atom's "resistance" to a change in its atomic charge as well as its capacity to delocalize charges (known as the “inductive atomic hardness” and the “inductive atomic softness” of an atom, respectively); the ability of an atom with lone pair electrons to ionize; the connectivity to an aromatic system or conjugated system; the partial charges of atoms in pi bonds (if any); the partial charges of atoms in sigma bonds (if any); the total partial charge; the electronegativity; the proton affinity; and the radical distribution functions (described earlier). Figure 2.19 shows a list of the 30 atomic features generated by CDKit.

In addition to these atom-specific or target-atom features, we also took into account the influence of nearby atoms, which is known to influence $^{1}$H chemical shift values. We carried out an experiment by assessing the quality of the $^{1}$H shift predictions of the target hydrogen by incorporating information from the first, second, third, and fourth closest neighbours. We discovered that the effect of the closest three atoms produced the best results for $^{1}$H chemical shift prediction. The closest neighbours of the target hydrogen atom were calculated by measuring

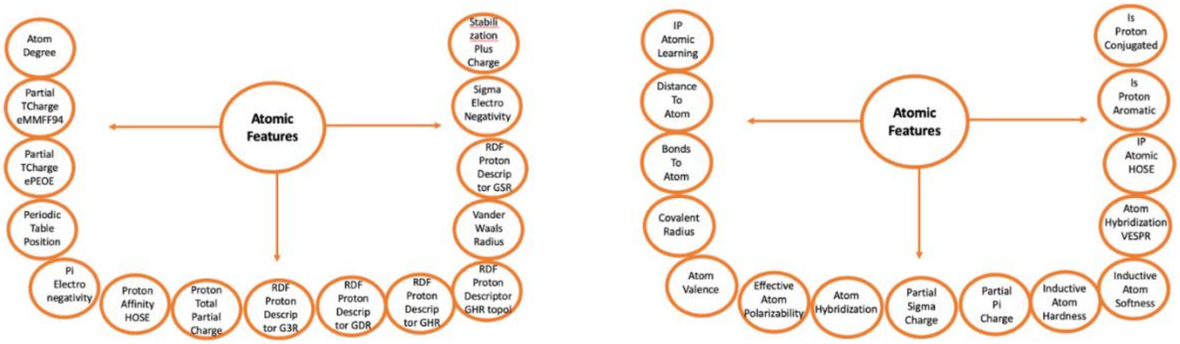


Figure 2.19: Atomic features that can be calculated from a molecular structure using the CDK package. All values are numeric.

their spatial distance (using the x,y,z coordinates). Therefore, to describe the environment surrounding every target hydrogen atom, we used 28 features to describe the target hydrogen atom and 30 features to describe each of the 3 spatially nearest atoms (for a total of 118 features). The hybridization state and valence state were the two features we disregarded for the target hydrogen atom. This is due to the fact that hydrogen does not hybridize, and its valence will never change.

For the ML algorithm, these atomic features had to appear in a specific order. The feature set for the targeted hydrogen atom (28 features) had to be listed first, followed by the 30 features for the closest atom, followed by the 30 features of the 2nd closest atom, and finally followed by the 30 features from the 3rd closest atom. For example, in Figure 2.20, if the atom of interest is #10, the nearest three atoms are #3, #4, and #6.

Because an atom's chemical shift can also be affected by the presence of different functional groups (Figure 1.8 in Chapter 1), we also included 89 types of chemical functional groups or their particular chemical substructures. Table 2.1 shows the list of the chemical functional groups using SMART strings. These functional groups were used to annotate the target hydrogen atom's molecular neighborhood. This neighborhood included functional groups up to four bonds away. To describe this molecular neighborhood property, we counted how many times each functional group was present in the four-bond neighborhood (Figure 2.20 right side). This four-bond neighborhood was determined after some trial-and-error assessment on the influence of the functional group effects from one, two, three, or four bonds away. Therefore, our feature set included 118 atomic features and 89 neighborhood descriptors. In addition, we added three more descriptors to the feature set, to account for the chirality of the molecule. These included the chirality and prochirality of the target hydrogen atom as well as the spatial distance from a chiral centre of that target hydrogen. Table 2.2 shows how the feature space was constructed for each instance. All features are numeric.

| Name | SMART Strings | Name | SMART Strings | Name | SMART Strings |
|---|---|---|---|---|---|
| Carbonyl | [$([CX3]=[OX1]);!$([CX3](=[OX1])[OX2H1])] | Guanidine | [CX3](=N)(N)N | Glucose | C1([CX4H2][OX2H1])C([OH])C([OH])C([OH])CO1 |
| Carbonyl, aromatic | a[$([CX3]=[OX1]);!$([CX3](=[OX1])[OX2H1])] | double bonded oxygen | [OX1] | Ribose | C1([CX4H2][OX2H1])C([OH])C([OH])C([OH])O1 |
| Carboxylate anion | [CX3](=[OX1])([-OX1]) | Phosphoric Acid | [PX4](=[OX1])(-[OX2])(-[OX2])(-[OX2]) | Ribose | O1[CH]([OH])CCCC1 |
| Carboxyl | [CX3](=[OX1])[OX2H1] | Phosphate | [PX4](=[OX1])([-O])([OX2])([OX2]) | Ribose | O1[CH]([OH])CCC1 |
| aromatic carboxyl | [a[CX3](=[OX1])[OX2H1] | Pyridine | [cR1]1[+nR1][cR1][cR1][cR1][cR1]1 | Cyclohexene | [CR1]1=[CR1][CR1][CR1][CR1][CR1]1 |
| Formic Acid | [CX4H1][CX3](=[OX1])[OX2H1] | Pyridine | [cR1]1[nR1][cR1][cR1][cR1][cR1]1 | Cyclohexane | [CR1]1[CR1][CR1][CR1][CR1][CR1]1 |
| Amide | [CX4H1][NX3H1][CX3](=[OX1]) | Pyridone | [cR1]1(=O)[nR1][cR1][cR1][cR1][cR1]1 | Cyclohexanone | [CR1]1(=O)[CR1][CR1][CR1][CR1][CR1]1 |
| Amide | [CX4H2][CX3](=[OX1])[NX3H1] | Pyrimidine | [cR1]1[cR1][nR1][cR1][nR1][cR1]1 | Quinoline | c12cccnc1cccc2 |
| Aldehyde | [CX3H1]=[OX1] | Quinazoline | [cR2]1[cR1][nR1][cR1][nR1][cR2]1 | Quinolinone | c12c(=O)ccnc1cccc2 |
| aromatic aldehyde | a[CX3H1]=[OX1] | Nucleoside | [cR1]1[cR1][nR1][cR1](=O)[nR1][cR1]1([NX3H2]) | Purine | c12cncnc1ncn2 |
| Olefinic | [CX3;R0]=[CX3;R0] | Uracil | [cR1]1(=O)[nR1][cR1][cR1][cR1](=O)[nR1]1 | Nucleoside | c12c(=O)ncnc1ncn2 |
| Alkyne | [CX2]#[CX2] | Geminal diol | [CX4H1]([OX2H1])([OX2H1]) | Naphthalene | c12ccccc1cccc2 |
| Halogens | [Fl,Cl,Br,I] | Benzene | [cR1]1[cR1][cR1][cR1][cR1][cR1]1 | Nitro | [[+NX3](=[OX1])[-O] |
| aromatic halogen | a[Fl,Cl,Br,I] | Nitrile | C#N | Phoshoester | C[OX2]P |
| Methine | [CX4H1] | Pyrimidine | [cR1]1[nR1][cR1](=O)[nR1][cR1][cR1]1 | phosphonic Acid | [PX4](=[OX1])([OX2H])([OX2H) |
| Methylene | [CX4H2] | Imidazole | [cR1]1[cR1][nR1H0][cR1][nR1H0]1 | Ether | [CX4][O][CX4] |
| aromatic methyl | a[CX4H3] | Nucleoside | c12c(=O)ncnc1ncn2 | Sulphur Oxide | [$([SX4](=[OX1])=[OX1]);!$([SX4](=[OX1])(=[OX1])([OX2]-))] |
| Amine | [CX4H3]N | Imidazole | [nH1]1[cR1][nR1][cR1][cR1]1 | Quinone | [CR1]1(=[OX1])[CR1]=[CR1][CR1](=[OX1])[CR1]=[CR1]1 |
| Methyl bonded with 6 membered ring with one ionized nitrogen | [CX4H3][+nX3r6] | Tetrahydrofuran | [CR1]1[CR1][CR1][CR1][O]1 | Sulphide | C[SX2;R0]C |

| | | | | | |
|---|---|---|---|---|---|
| Alcohol | [$(C[OX2H1]);!$([CX3](=[OX1])[OX2H1])] | Furan-2,3,4(5H)-trione | [CR1]1(=[OX1])[CR1](=[OX1])[CR1](=[OX1])[CR1][O]1 | Primary Amine | [CH2][NX3H2] |
| aromatic alcohol | [$(c[OX2H1]);!$([CX3](=[OX1])[OX2H1])] | Ribosel | [CR1]1(C)[CR1]([OX2H1])[CR1]([OX2H1])[CR1][O]1 | Sulphide | S[CX4H3] |
| Nitrogen | [NX3H0] | Oxane | [CR1]1[CR1][CR1][CR1][CR1][O]1 | Amine | [+N][CX4H3] |
| Secondary Amine | [NX3H1] | Lactone | [CR1]1(=O)[CR1][CR1][CR1][CR1][O]1 | Methyl | [CX4H3] |
| Primary Amine | [NX3H2] | Oxolane | [cR1]1[cR1][cR1][cR1][oR1]1 | Pteridine | c12cncnc1nccn2 |
| Ammonia | [NX3H3] | Pyrrolidine | [CR1]1[NR1][CR1][CR1][CR1]1 | Aromatic ring | [aR1] |
| Sulphonic | [SX4](=[OX1])(=[OX1])([OX2]) | Ester | C(=[OX1])[OX2;H0] | Aliphatic ring | [AR1] |
| Name | SMART Strings | Name | SMART Strings | Name | SMART Strings |
| Amide | [CX3](=[OX1])[NX3H2] | Thiazole | [cR1]1[+nR1][cR1][cR1][sR1]1 | Thiazole | c1scnc1 |
| Secondary Amine | [CX4H2][NX3H1] | 2,3,3a,4,6,6a-Hexahydro-1H-thieno[3,4-d]imidazole-2-one | C12CSCC1NC(=O)N2 | | |
| Amide | [NX3H1][CX3]=[OX1] | Hetarocyclic ring (with sulphur atom) | [SR1] | | |
| Nitrosamine | [NX3][NX2](=[OX1]) | Six membered ring with one hetaro atom (nitrogen atom) | [nX2r6] | | |
| Amio Acid | [CX4]([NX3H2])C(=[OX1])[OX2H] | five membered ring with one hetaro atom (nitrogen atom) | [nX2r5] | | |

Table 2.1: The list of the chemical functional groups and chemical substructures (written in SMART strings) that were used to annotate the feature space.

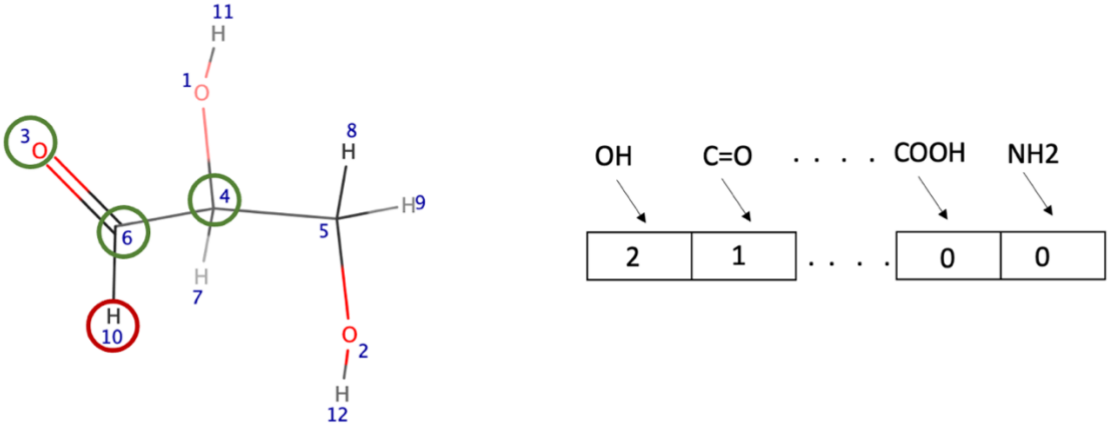


Figure 2.20: If the atom of interest is #10, the nearest three atoms are #3, #4, and #6 (left side). If the atom of interest is #10 (right side), then looking at atoms up to four bonds away we can see that an OH group appeared 2 times, a C=O group appeared 1 time, while COOH, and NH2 groups never appeared and so on.

| Dataset Instances | **Feature Space** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 28 features of the targated hydrogen atom | 30 features of the 1st nearest atom | 30 features of the 2nd nearest atom | 30 features of the 3rd nearest atom | Presence of 89 types of chemical structure | Chirality of the molecule | Prochirality of the targated hydrogen atom | Distance from the chiral center to the targated hydrogen atom |
| 1 | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. |
| 2 | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. |
| ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. |
| 4207 | ….. | ….. | ….. | ….. | ….. | ….. | ….. | ….. |

Table 2.2: The sequence of all features, used to train the different ML models.

## Methodology

To evaluate which ML algorithm would yield the best results for $^1$H chemical shift prediction, we trained and tested four different regression algorithms: a Support Vector Regressor (SVR), a Random Forest Regressor (RFR), an Extreme Gradient Boosting Regressor (XGBoostRegressor or XGBR) and a Categorial Boosting Regressor (CatBoostRegressor).

The training and validation methodology was performed using standard cross-validation methods with independent training and testing (hold-out) datasets. For each of the four regressor algorithms we tuned the hyper parameters using internal cross validation (5-fold) over the training dataset. Once the best model with the tuned hyperparameter was found, the model was trained on the entire training dataset of 4207 experimentally measured $^1$H chemical shift values from 577 molecules. Then the trained model was used to predict the $^1$H chemical shift values for two different holdout datasets, one consisted of 272 and the other set consisted of 442 $^1$H chemical shift values. The holdout datasets were used to measure how the model behaved against previously unseen data. Since our hold-out dataset was relatively small, we chose the best performing model using the internal cross-validation method. For both inner and outer cross validation steps, we used 5-fold cross validation by shuffling the training dataset randomly (using a random number seed). Once the best ML model was identified among the 4 regressors, the performance of that ML model was compared (using the MAE) against the results of two popular $^1$H chemical shift predictors (that use machine learning), namely MNOVA and NMRshiftDB2 as well as $^1$H chemical shifts calculated using quantum mechanical (density functional theory or DFT) methods. However, before describing the results of this experiment, I will briefly provide a high-level explanation of each of the regressor algorithms used in our study.

## Support Vector Regressor

Support Vector Regression (SVR) (71) is a supervised machine learning algorithm that can be used to predict discrete or numeric values (such as chemical shifts) through regression. Regression is a statistical technique that relates a dependent variable ($^1$H chemical shift, in this case) to one or more independent (explanatory) variables. As with any supervised ML system, labeled training data must be provided. Support Vector Regression uses the same principle as Support Vector Machines (SVMs). The basic idea behind SVR is to find the best fit line between observed data and predicted data. In SVR, the best fit line is the hyperplane that fits the maximum number of points. The Support Vector Machine, or SVM, is the supervised learning model upon which SVR is built. It is one of the best-known ML models and has been applied to classification problems when classes cannot be separated linearly. More specifically, an SVM (72) is a discriminative classifier that takes features from labeled examples and applies kernel transformations on them to produce a hyperplane that separates a class from other classes. By using support vectors that are training instances close to the hyperplane with high influence, the SVM optimizes the hyperplane, separating the classes as much as possible. The same idea of hyperplane optimization applies to Support Vector Regression. In other words, SVR employs SVMs to solve regression issues. SVR is ideal for addressing regression problems when numerical data points cannot be regressed or fit via simple linear or polynomial functions.

## Random Forest Regressor (RFR)

Random Forest (73) is a supervised ensemble learning technique for classification and regression. The idea behind ensemble learning is built on the observation that a group of people with varying levels of expertise in a given field can come up with a solution that is often superior to that of a single expert. The goal of ensemble learning techniques is to improve the efficacy of decision making or classification by combining several machine learning (ML) algorithms together. Random Forest (RF) employs ensemble learning by assembling multiple decision trees (adding trees together to make a forest) in such a manner to provide an output which is the consensus on the best solution to the problem. Decision trees use a succession of true/false inquiries regarding the components of a data set to arrive at an answer. In the Random Forest algorithm multiple decision trees are constructed concurrently using the "bagging" technique from random bootstrap samples of the data set and features. Because of the randomness that is used in the assembly of these decision trees, bias is less likely to occur because individual trees have minimal correlations with one another. The issue of overfitting, which happens when a model integrates too much "noise" in the training data (making bad decisions as a result) is further mitigated by the existence of many trees. Subsets of the training data are randomly sampled by each tree in a random forest. These smaller data sets are then fit to the model, and the predictions are combined. Through replacement sampling, it is possible to employ many instances of the same data again and again, leading to decision trees that have diverse decision-making properties. Even when a significant portion of the data is missing, Random Forest is quite robust and it appear to manage missing values and retain good accuracy. The RF approach virtually reduces model overfitting due to the "majority rules" output. RF is a useful tool for dimensionality reduction since it can handle very large data sets with hundreds of input variables. Random Forest Regression (RFR) is a supervised learning algorithm that uses RF learning for regression instead of classification. RFR is ideal for solving regression problems when numerical data points cannot be regressed or fit via simple linear or polynomial functions.

## XGBoost Regressor

Extreme Gradient Boosting or XGBoost (74) is a distributed, scalable gradient-boosted decision tree (GBDT) machine learning model. It is considered to be one of the best ML methods for regression, classification, and ranking. XGBoost is based on several well-known ML concepts, including supervised machine learning, decision trees, ensemble learning, and gradient boosting. Supervised machine learning employs algorithms to train a model to detect patterns in a dataset containing both labels and features, and then uses the trained model to predict the labels on the features for a new dataset. While similar in concept but differently implemented, a decision tree algorithm produces a model that predicts the label by analysing a tree of if-then-else true/false feature questions and estimating the minimum number of questions required to assess the probability of making the right choice. On the other hand, a Gradient Boosting Decision Tree (GBDT) is a decision tree ensemble learning approach for classification and regression that is similar to random forest. The concept behind "gradient boosting" is to "boost" or improve a single weak model by fusing it with a number of additional weak models in order to produce a model that is stronger when taken as a whole. Gradient boosting formalizes the process of additively creating weak models using gradient descent optimization over an objective function. To reduce errors, gradient boosting begins with group of shallow decision trees that are iteratively trained by GBDTs, with each iteration using the error residuals of the prior model to fit the new model. The weighted average of all the tree predictions represents the final projection. The XGBoost algorithm uses the same concept of GBDT but the trees are constructed in parallel as opposed to sequentially (which is done in GBDT). XGBoost employs a level-wise approach, scanning over gradient values and assessing the quality of splits at each potential split in the training set using partial sums. XGBoost is a scalable and an extremely accurate implementation of gradient boosting. It was created primarily to enhance the performance and computational speed of ML models. In recent years, XGBoost has seen a substantial increase in popularity. Like RFR, XGBoost regression is ideal for solving regression problems when numerical data points cannot be regressed or fit via simple linear or polynomial functions.

## CatBoost Regressor

CatBoost or Category Boost is another variant of Gradient Boosting algorithm that works with categorical data rather than numerical data. Numeric data can be converted or binned into categorical data. To employ CatBoost for $^1$H NMR chemical shift prediction it is necessary to do this numeric to categorical conversion. CatBoost is a very recent open-source machine learning method that works similarly to XGBoost but with a slightly different strategy. CatBoost provides a nice method of managing categorical data, which reduces the amount of categorical feature translation. Unlike XGBoost, CatBoost uses symmetric decision trees. Symmetric trees, also known as balanced trees, are decision trees where the splitting condition is consistent for every node at every depth of the tree. This implies that the splitting condition must produce the lowest loss over all nodes of the same depth. This provides faster computation and evaluation along with greater control against overfitting. On the other hand, XGBoost produces asymmetric trees which means that the splitting conditions for each node within the same depth can vary. CatBoost also differs from XGBoost in the type of boosting method used. CatBoost offers tremendous versatility in how it handles heterogeneous, sparse, and categorical data while still supporting quick training times and already adjusted hyperparameters. CatBoost regression is ideal for solving regression problems when numerical data points cannot be regressed or fit via simple linear or polynomial functions.

## Coding Details

Python and a variety of Python libraries, including the scikit-learn library were used to develop and test all the ML models described herein. Specifically, Python “2.7.0” was initially used, which was later replaced with Python "3.7.0". For all ML applications we utilized the Python scikit-learn library (version "1.0.2”). In addition, another Python library, RDKit (versions "2019_03" and "2020.09"), was used for a variety of cheminformatics tasks, including turning SMILES strings into 3D structures and matching SMILES strings to detect the existence of different functional groups within a target molecule. RDKit was also used to calculate chirality, prochirality, and other molecular properties. As previously mentioned, we used the Java-based CDKit to determine the 30 sets atomic features for our chemical shift calculations. To simplify operations, we used Java to construct a separate “.jar” file that could be called from Python to obtain the values of the various atomic descriptors. We conducted all our model development (training, testing, evaluating, etc.) on an Ubuntu 20.04 machine running the Linux 5.4.0-131generic kernel on the University of

Alberta's (UofA) Cybera server. The server was equipped with 8 Intel Xeon E5-2630 v3 @ 2.40 Gz CPUs and had a total of 32 GB of RAM.

The process of selecting an optimal set of hyperparameters for a given ML model and for a given dataset is known as "hyperparameter optimization" or tuning. Each of the algorithms we tested included a number of hyperparameters. One of the benefits of using sckit-learn is that majority of the hyperparameters can be handled by using the default settings. However, in the course of this work we also discovered that the most crucial parameters for optimizing tree-based algorithms are the number of trees (n_estimators) and its maximum depth (max_depth). Likewise, for SVMs, the regularization parameter (C) and the selection of the appropriate kernel are most important hyperparameters. We discovered that the model would frequently overfit if the "rbf," kernel was used, so we stuck with the linear kernel and experimented with several values of C to find the best SVM hyperparameters.

An important component of any ML training procedure involves cross-validation (CV). CV is particularly useful as a method to assess ML model performance. It allows one to use training data more effectively to perform tasks such as parameter adjustment without running the danger of data leakage, (a situation where the model gains access to knowledge that it otherwise shouldn't have). In order to implement CV, we must first divide our dataset into training and validation sets. The model is then trained on the training set and validated on the validation set, thereby allowing us to test several models without utilizing our hold-out (or test) set, which should only be used after we have selected our model. In ML it is normal practise to use CV to compare the performance of several different ML models. However, when the same CV is used to assess the performance of the ML models as well as to select the optimum (hyper)parameters, the problem data leakage can occur (Figure 2.21). This is because the models have effectively "seen" the test data while optimizing the hyperparameters.

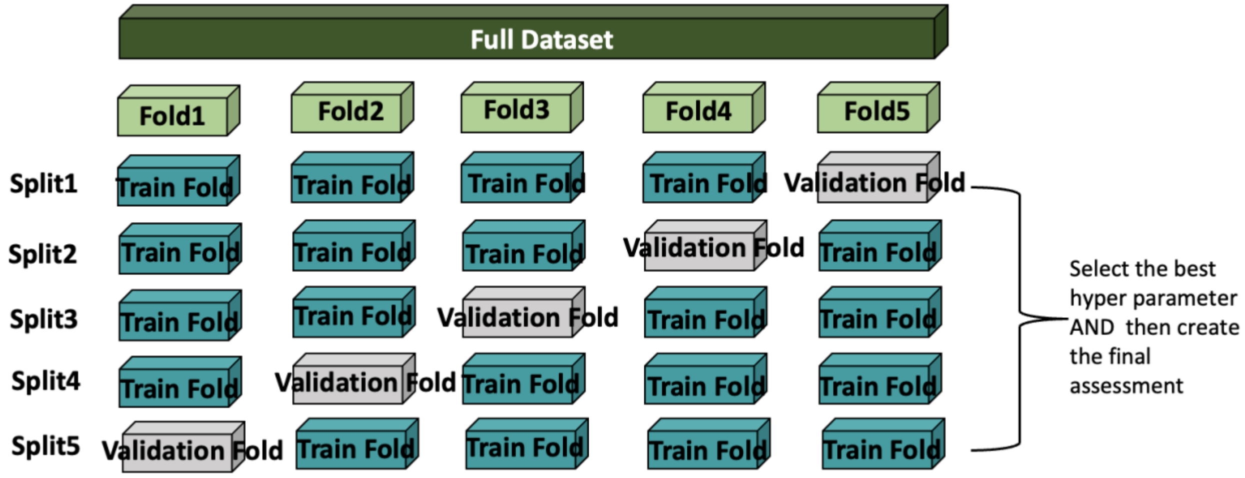


Figure 2.21: An optimistically skewed estimate of the model performance results from combining hyper-parameter adjustment and model evaluation in the same CV loop.

Use of the nested cross-fold technique is one way to prevent the problem of data leakage. In nested CV, the entire training dataset is partitioned into k-folds. Each k-1 fold is separated once more into a j-fold around the outer loop of the k-fold. Inside the so-called inner loop of the j-fold cross fold, the hyperparameters are adjusted and the model's performance is assessed using the fold data from the k-fold CV. A schematic version of the process is depicted in Figure 2.22. As previously noted, our hold-out dataset was relatively small; hence, we used an internal cross validation approach to more completely assess the model performance. The training dataset (4207 samples) for our experiment was divided into k = 5 folds (the outer loop). The outer loop had training samples of 3365, 3365, 3366, 3366 and 3366 $^{1}$H chemical shifts for each of the k = 5 folds. The outside test sample consisted of 842, 842, 842, 841, and 841 $^{1}$H chemical shifts, respectively. Each set of outer train samples were split once more into a set of inner j = 5 folds. Figure 2.23 displays the sample distribution for each outer and inner loop. To avoid data leakage, we split the dataset or created folds based on molecular identities instead of the $^1$H chemical shift labels.

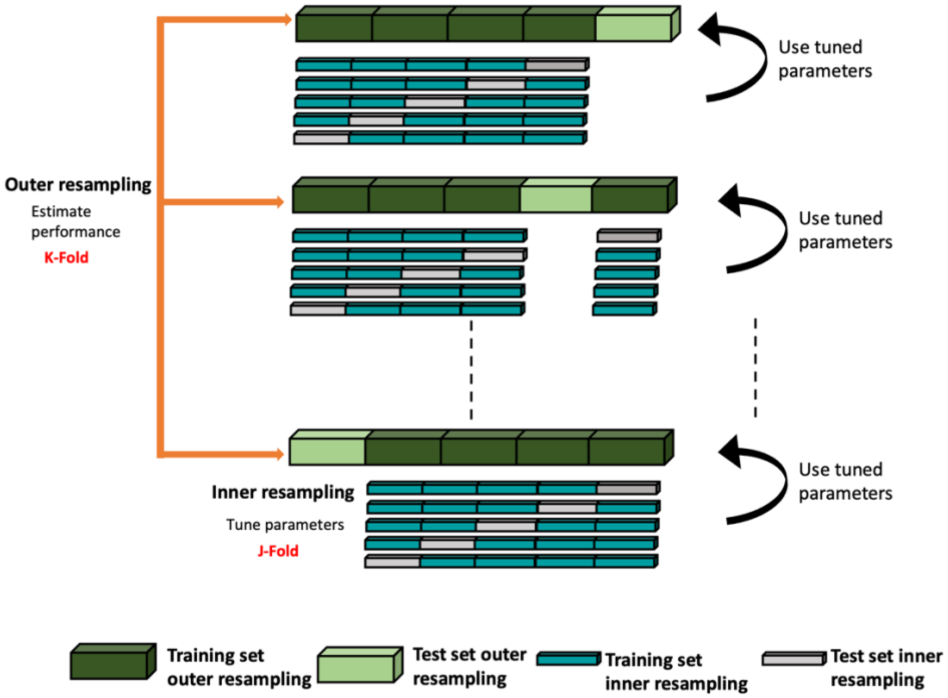
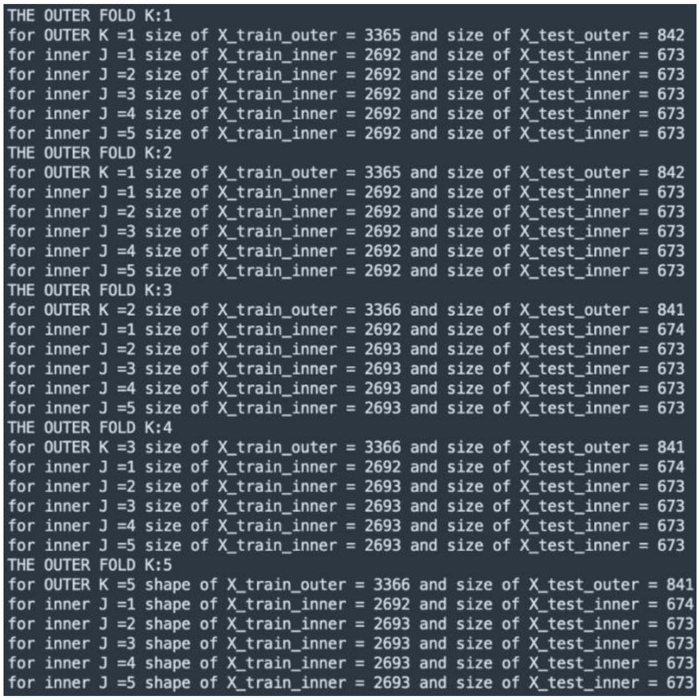


Figure 2.22: An explanation of Nested CV. The outer loop is used to estimate the model error while the inner loop is used to search for hyper-parameters.



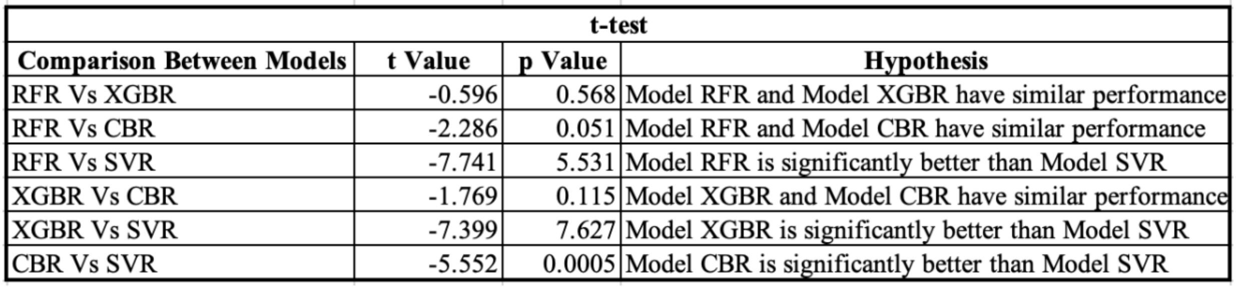
```
THE OUTER FOLD K:1
for OUTER K =1 size of X_train_outer = 3365 and size of X_test_outer = 842
for inner J =1 size of X_train_inner = 2692 and size of X_test_inner = 673
for inner J =2 size of X_train_inner = 2692 and size of X_test_inner = 673
for inner J =3 size of X_train_inner = 2692 and size of X_test_inner = 673
for inner J =4 size of X_train_inner = 2692 and size of X_test_inner = 673
for inner J =5 size of X_train_inner = 2692 and size of X_test_inner = 673
THE OUTER FOLD K:2
for OUTER K =1 size of X_train_outer = 3365 and size of X_test_outer = 842
for inner J =1 size of X_train_inner = 2692 and size of X_test_inner = 673
for inner J =2 size of X_train_inner = 2692 and size of X_test_inner = 673
for inner J =3 size of X_train_inner = 2692 and size of X_test_inner = 673
for inner J =4 size of X_train_inner = 2692 and size of X_test_inner = 673
for inner J =5 size of X_train_inner = 2692 and size of X_test_inner = 673
THE OUTER FOLD K:3
for OUTER K =2 size of X_train_outer = 3366 and size of X_test_outer = 841
for inner J =1 size of X_train_inner = 2692 and size of X_test_inner = 674
for inner J =2 size of X_train_inner = 2693 and size of X_test_inner = 673
for inner J =3 size of X_train_inner = 2693 and size of X_test_inner = 673
for inner J =4 size of X_train_inner = 2693 and size of X_test_inner = 673
for inner J =5 size of X_train_inner = 2693 and size of X_test_inner = 673
THE OUTER FOLD K:4
for OUTER K =3 size of X_train_outer = 3366 and size of X_test_outer = 841
for inner J =1 size of X_train_inner = 2692 and size of X_test_inner = 674
for inner J =2 size of X_train_inner = 2693 and size of X_test_inner = 673
for inner J =3 size of X_train_inner = 2693 and size of X_test_inner = 673
for inner J =4 size of X_train_inner = 2693 and size of X_test_inner = 673
for inner J =5 size of X_train_inner = 2693 and size of X_test_inner = 673
THE OUTER FOLD K:5
for OUTER K =5 shape of X_train_outer = 3366 and size of X_test_outer = 841
for inner J =1 shape of X_train_inner = 2692 and size of X_test_inner = 674
for inner J =2 shape of X_train_inner = 2693 and size of X_test_inner = 673
for inner J =3 shape of X_train_inner = 2693 and size of X_test_inner = 673
for inner J =4 shape of X_train_inner = 2693 and size of X_test_inner = 673
for inner J =5 shape of X_train_inner = 2693 and size of X_test_inner = 673
```

Figure 2.23: The size of the data samples in the inner and outer loops of the nested CV as used in our experiment.

## Results and Discussion

After implementing each of the 4 regressors on the UofA Cybera cluster, a series of performance tests was conducted with the training dataset. During the 5-fold cross validation training stage, it was found that both the Random Forest Regressor (RFR) and the XGBoost Regressor performed better than the other two algorithms. In particular, the RFR and XGBoost produced an MAE of 0.12 ppm and 0.13 ppm (respectively) over the completes set of 4207 experimentally measured $^1$H chemical shift values. The Support Vector Regressor (SVR) had an MAE of 0.20 ppm whereas the CatBoost Regressor produced an MAE of 0.14 ppm. Figure 2.24 shows the comparison of 5-fold cross validation errors among the 4 different algorithms. As seen here, the RFR's MAE of 0.12 ppm exceeds the performance of all other algorithms we tested.

Even though the RFR outperformed the other regression algorithms, the XGBoost regressor, as well as CBR, showed a somewhat similar performance to that of RFR. I performed a paired t-test to see if there was a statistically significant difference between these models (Table 2.3). RFR, XGBR and CBR performed similarly based on the results of the t-test but much better than SVR. Next we checked how both models performed over previously unseen data, we used both the RFR and XGBoost regressor to predict the $^1$H chemical shifts in the first holdout dataset (which consisted of 272 $^1$H chemical shifts from 36 molecules). This test showed both the XGBoost Regressor and RFR achieved the same MAE of 0.11 ppm with standard deviation of 0.19 ppm and 0.18 ppm, respectively . To further distinguish between the two models, we analyzed the "Train Error", which is the measured error on the training dataset, for both models. We found that XGBoost Regressor's MAE was 0.02 ppm, which was a little lower than the RFR "Train Error (which had an MAE of 0.03ppm). This indicates that the XGBoost Regressor overfits a little more than the RFR model.



| t-test | | | |
|---|---|---|---|
| **Comparison Between Models** | **t Value** | **p Value** | **Hypothesis** |
| RFR Vs XGBR | -0.596 | 0.568 | Model RFR and Model XGBR have similar performance |
| RFR Vs CBR | -2.286 | 0.051 | Model RFR and Model CBR have similar performance |
| RFR Vs SVR | -7.741 | 5.531 | Model RFR is significantly better than Model SVR |
| XGBR Vs CBR | -1.769 | 0.115 | Model XGBR and Model CBR have similar performance |
| XGBR Vs SVR | -7.399 | 7.627 | Model XGBR is significantly better than Model SVR |
| CBR Vs SVR | -5.552 | 0.0005 | Model CBR is significantly better than Model SVR |

Table 2. 3: Paired t-test result among RFR, XGBR, CBR and SVR models.

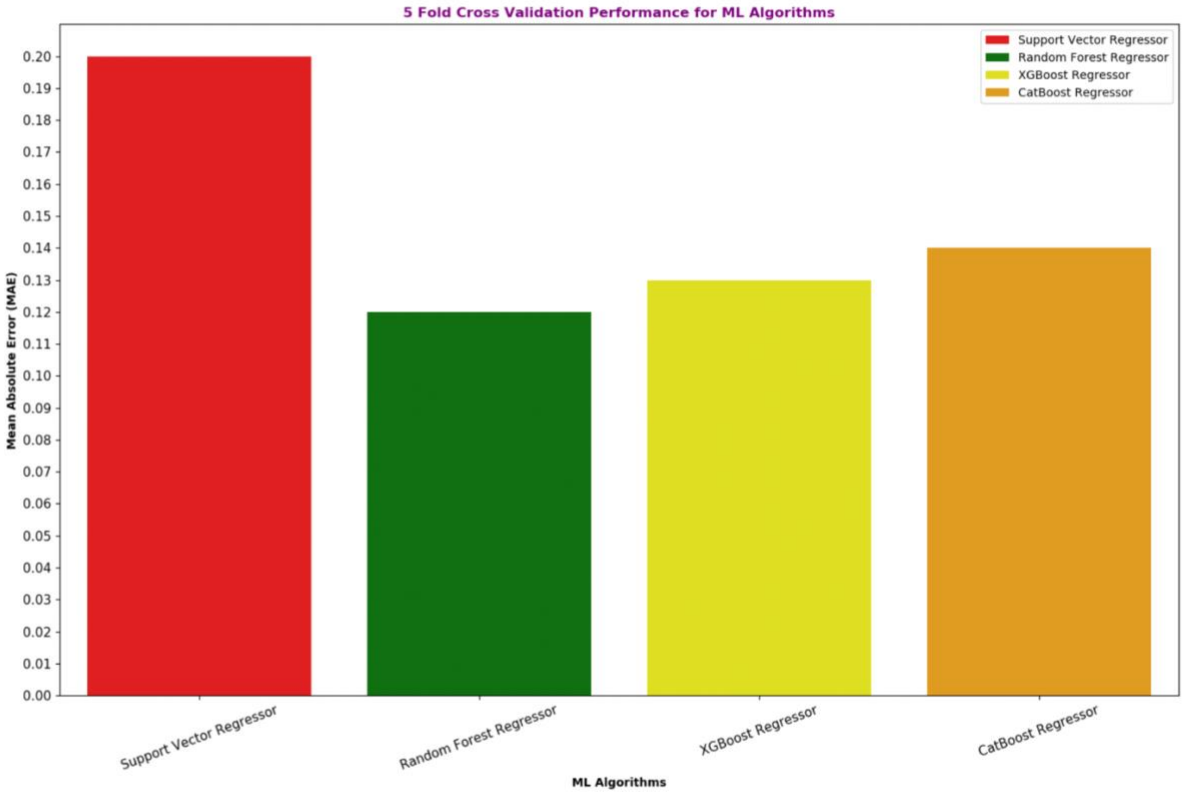


Figure 2.24: Comparison of the 5-fold cross validation performance for all 4 algorithms. Random Forest does the best with the smallest mean absolute error of 0.12 ppm. The 5-fold cross validation is carried on the same training set shuffled randomly with the same random seed.

Figures 2.25 and 2.26 show the plots for the predicted $^{1}$H chemical shift values vs the experimentally measured $^{1}$H chemical shifts values on the training dataset set using the RF and XGBoost Regressors. We also measured the Pearson correlation index or coefficient of determination ($R^2$) between observed and predicted $^{1}$H chemical shifts for both models. When assessing the effectiveness of a machine learning model based on regression, the $R^2$ value is a crucial indicator. It is determined by calculating the variation in the predictions that the dataset can explain. An $R^2$ of 1 indicates that the model is perfect, and when it is zero, the model is no better than a random guess. For the 5-fold cross validation (training) stage, the $R^2$ score was 0.99 for both

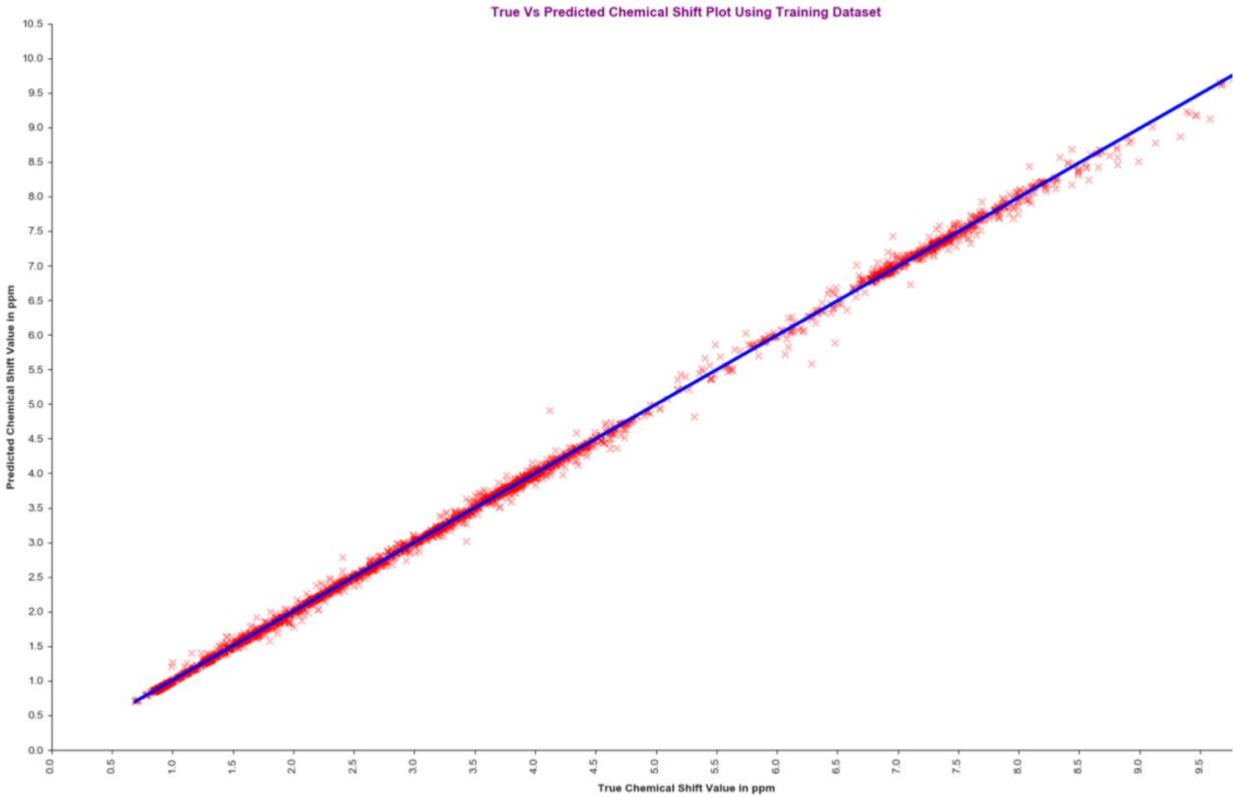


Figure 2.25: Predicted $^1$H chemical shift values vs observed $^1$H chemical shifts values on the training dataset set using RFR. Testing on the training set gave an MAE of 0.03 ppm.

the RFR and XGBoost Regressor, which did not change when test was performed on the first holdout dataset. Out of curiosity, we conducted the same analysis on the third best algorithm, the CatBoost Regressor. Even though the "Train Error" measurement for the CatBoost Regressor indicated it did not overfit and even though it produced an MAE of 0.14 ppm on the 5-fold cross validation test, when the CatBoost Regressor was applied to the holdout dataset, its performance dropped significantly, yielding a very high MAE of 1.19 ppm with a standard deviation of 1.32 ppm. The second holdout dataset was also evaluated using the RFR. The second holdout dataset consisted of 442 $^1$H chemical shifts derived from 22 randomly selected compounds in the NP-MRD database. This dataset yielded an MAE of 0.36 ppm with a standard deviation of  0.56 ppm and an

$R^2$ value of 0.92, which was significantly worse than the result achieved for the first holdout dataset. The reasons for this significant drop in performance are discussed in the Discussion section. We also predicted $^1$H chemical shifts for this 2nd set of holdout dataset using the 2nd and 3rd best predictors which were the XGBoost and CatBoost Regressors. These predictors had MAEs of 1.03 ppm and 1.01 ppm, respectively with standard deviations of 1.83 ppm and 1.82 ppm respectively. Given their poor overall performance, we did not analyze these predictors any further.

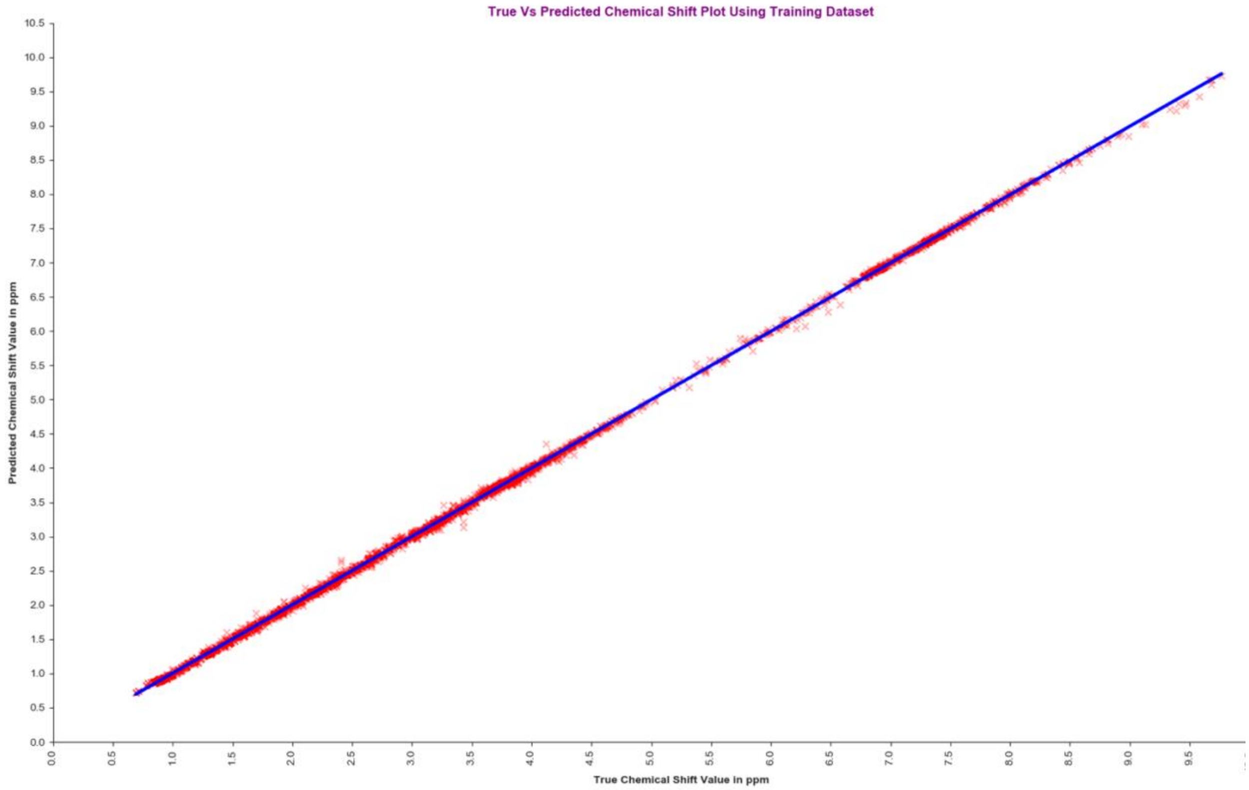
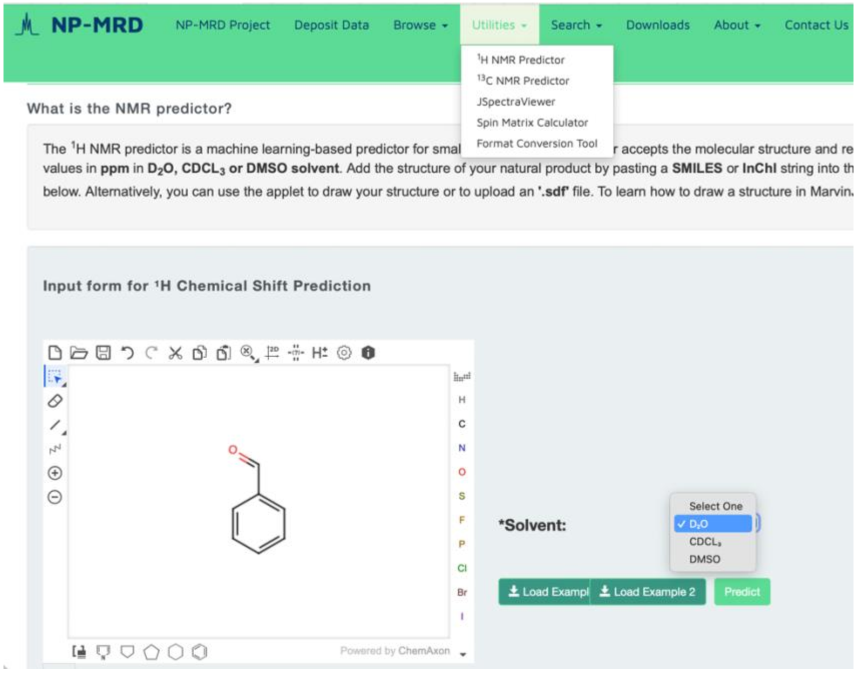


Figure 2.26: Predicted $^1$H chemical shift values vs observed $^1$H chemical shift values on the training dataset set using the XGBoost Regressor. Testing on the training set gave a mean absolute error of 0.02 ppm. Overfitting is more visible compared to the results show in Figure 2.25.

Even though the performance of the RFR was uneven across the two holdout datasets, we ultimately selected the RFR regressor as our best ML model. This model was used  as the "final" predictor for our chemical shift prediction program called NMRPred.

NMRPred represents the complete NMR chemical shift predictor package. NMRPred accepts SMILES data, converts the SMILES string to a 3D SDF file with atomic coordinates including the attached hydrogens (via RDKit), calculates the atomic feature sets using CDKit and then calculates the $^{1}$H NMR shifts using the RFR. Currently NMRPred is integrated with NP-MRD website under the utility tool known as "$^{1}$H NMRPredictor". The public version of NMRPred will be available on GitHub with a README file in the link https://github.com/zsayeeda/NMR_Prediction.git. A sample input for the NP-MRD database $^{1}$H Chemical Shift Predictor ("$^{1}$H NMRPredictor"), which employs NMRPred, is shown in Figure 2.27. The input is the SMILES formula (O=CC1=CC=CC=C1) corresponding to Benzaldehyde (HMDB0006115), which has been pasted in the input field. The structure is displayed via ChemAxon's JChem software which displays the chemical structure in a standard 2D format. Users must select the solvent from the pull-down options listed under "Solvent". The generated 3D structure of Benzaldehyde and NMRPred's predicted $^{1}$H chemical shift values for Benzaldehyde are shown in Figure 2.28.



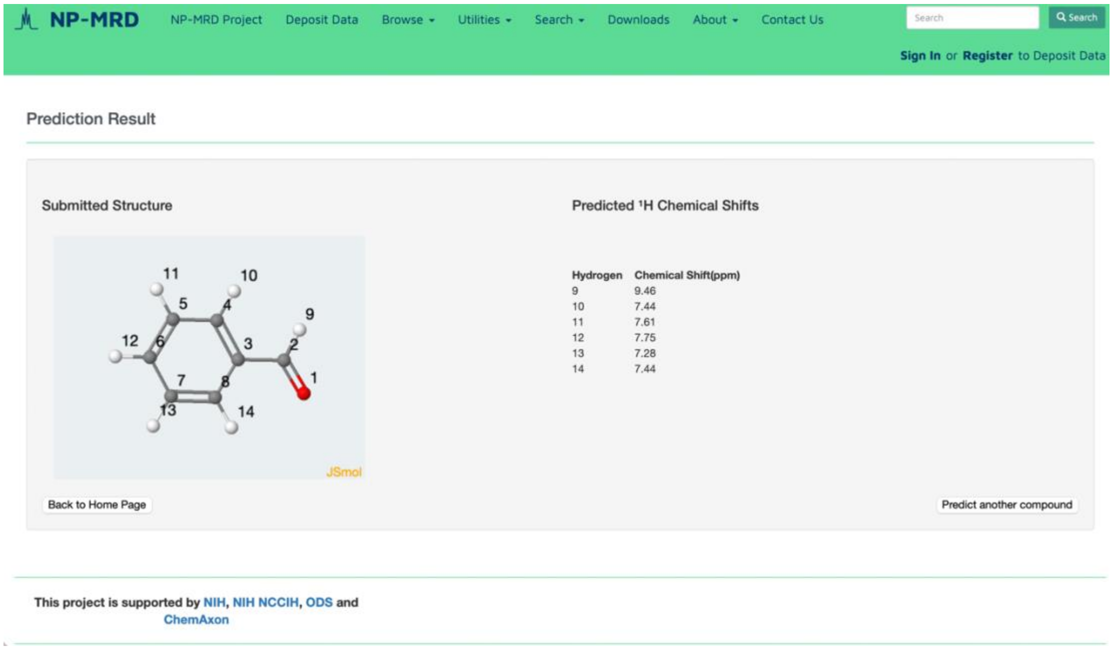
What is the NMR predictor?

The $^1$H NMR predictor is a machine learning-based predictor for smal r accepts the molecular structure and re values in **ppm** in **$D_2O$, $CDCL_3$ or DMSO solvent**. Add the structure of your natural product by pasting a **SMILES** or **InChI** string into th below. Alternatively, you can use the applet to draw your structure or to upload an **'.sdf'** file. To learn how to draw a structure in Marvin.

Input form for $^1$H Chemical Shift Prediction

***Solvent:** Select One / ✓ $D_2O$ / $CDCL_3$ / DMSO

Figure 2.27: The SMILES string for Benzaldehyde (HMDB0006115) was provided in the input section together with the selected solvent option. ChemAxon's JChem converts the SMILES string into a 2D structure.



Prediction Result

Submitted Structure

Predicted $^1$H Chemical Shifts

| Hydrogen | Chemical Shift(ppm) |
|---|---|
| 9 | 9.46 |
| 10 | 7.44 |
| 11 | 7.61 |
| 12 | 7.75 |
| 13 | 7.28 |
| 14 | 7.44 |

This project is supported by NIH, NIH NCCIH, ODS and ChemAxon

Figure 2.28: The predicted $^1$H chemical shift values from the input in Figure 2.27 together with the 3D structure of the molecule.

## Performance Comparison Against Popular Methods

We compared the performance of NMRPred with several popular $^1$H chemical shift predictors, including MNOVA (63), NMRShiftDB2 (12) and DFT based calculations (75) performed by NWChem. We evaluated these predictors against the observed $^1$H chemical shift values for both holdout datasets. We found that NMRPred outperformed all three predictors. In particular, NMRPred had an MAE of 0.11 ppm for the first holdout dataset.  On the other hand, MNOVA yielded an MAE of 0.15 ppm, NMRShiftDB2 had an MAE of 0.17 ppm while the NWChem DFT method had an MAE of 0.28 ppm.

Scatter plots have been generated that show the observed $^1$H chemical shift values vs. the predicted $^1$H chemical shift values for the holdout dataset for NMRPred, MNOVA and NMRShiftDB2 in Figure 2.29, 2.30 and 2.31, respectively. The $R^2$ was identical for all 3 models (0.99). We chose not to provide a scatter plot for the DFT predictions because it displayed the worse performance across the board. Looking more closely at the scatter plot for the observed vs. predicted $^1$H chemical shifts in the first holdout dataset for NMRPred (Figure 2.29) we see that it has two major outliers. At the top right corner of the plot, the most visible outlier belongs to an aldehyde proton for Benzaldehyde (HMDB0006115). MNOVA, NMRShiftDB2 and the DFT method predicted this aldehyde chemical shift much more accurately than NMRPred. The true chemical shift value for this proton is 9.93 ppm. NMRPred predicted that this aldehyde hydrogen's chemical shift would be 9.04 ppm whereas MNOVA, NMRShiftDB2 and DFT predicted its shift to be 9.95 ppm, 9.94 ppm and 9.88 ppm, respectively This difference is illustrated in Figure 2.32. We found that the reason for this discrepancy was likely due to undertraining in the original training dataset, In particular, we only had 10-12 compounds with aldehyde hydrogens in the training dataset and their chemical shift values range from 7.92-9.69 ppm. Based on the modest range of aldehyde proton chemical shifts in the training set and the fact that none of the aldehyde protons were close to aromatic rings, it makes sense that NMRPred would predict a chemical shift value of just 9.04 ppm.

The second major outlier identified in the NMRPred results was for hydrogen atom number 19 in Mevalonic acid (HMDB0000227/HMDB0059629) (Figure 2.33). The true chemical shift

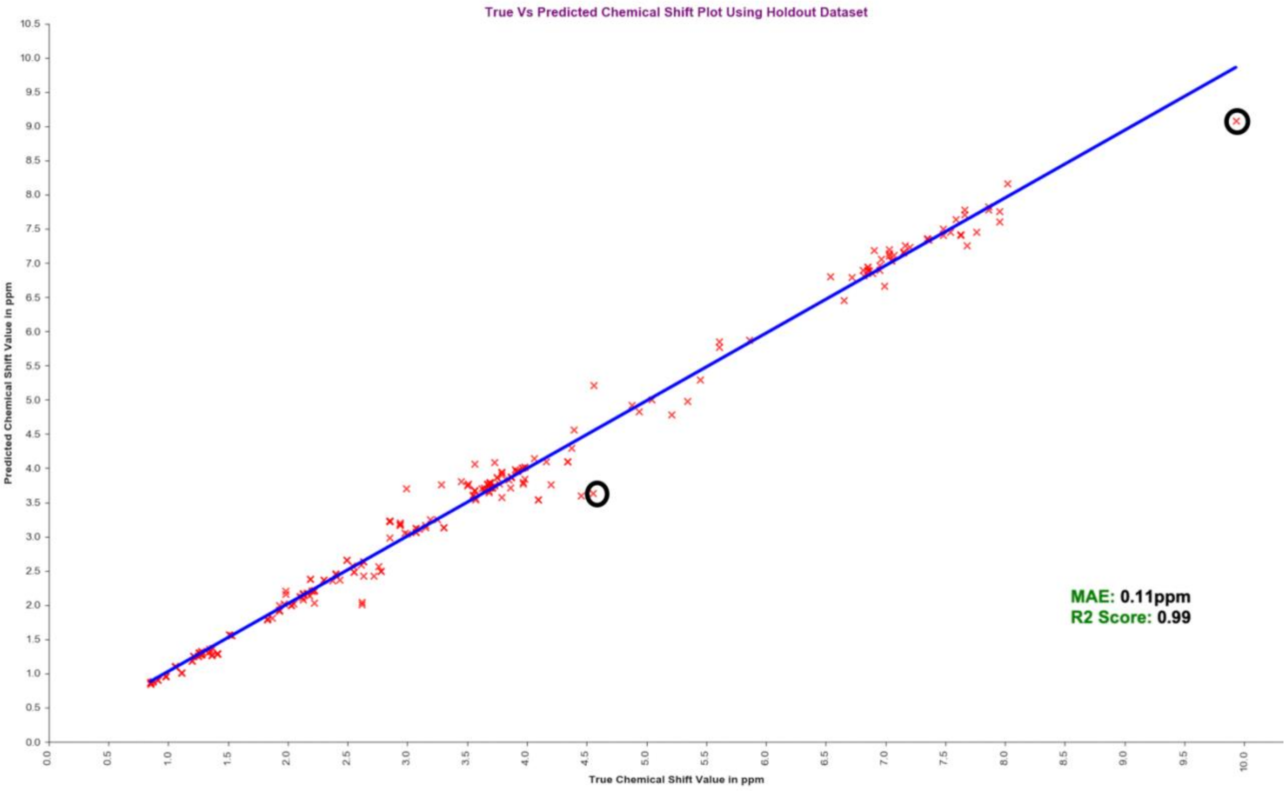


Figure 2.29: A scatter plot of the correlation between the observed vs predicted $^1$H chemical shifts in the first holdout dataset, for NMRPred (which uses the RFR).

value for that atom is 4.55 ppm. The predicted $^1H$ chemical shift values for that atom by NMRPred, MNOVA NMRShiftDB2 and the DFT predictor were 3.63 ppm, 3.65 ppm, 3.85 ppm and 3.90ppm, respectively. In this case, the prediction by the DFT predictor was closer to the true chemical shift value than that of the other predictors. NMRPred and MNOVA predicted values that were also comparatively close. However, all three predictors except for the DFT predictor, were more than 0.70 ppm off the correct value, which still represents a substantial error. This error may arise from undertraining or poor representation of molecules similar to as Mevalonic acid in the training set for all predictors. Interestingly, even though NMRPred did not do as well as the other shift predictors for this compound, it could still differentiate the diastereotopic chemical shift values for

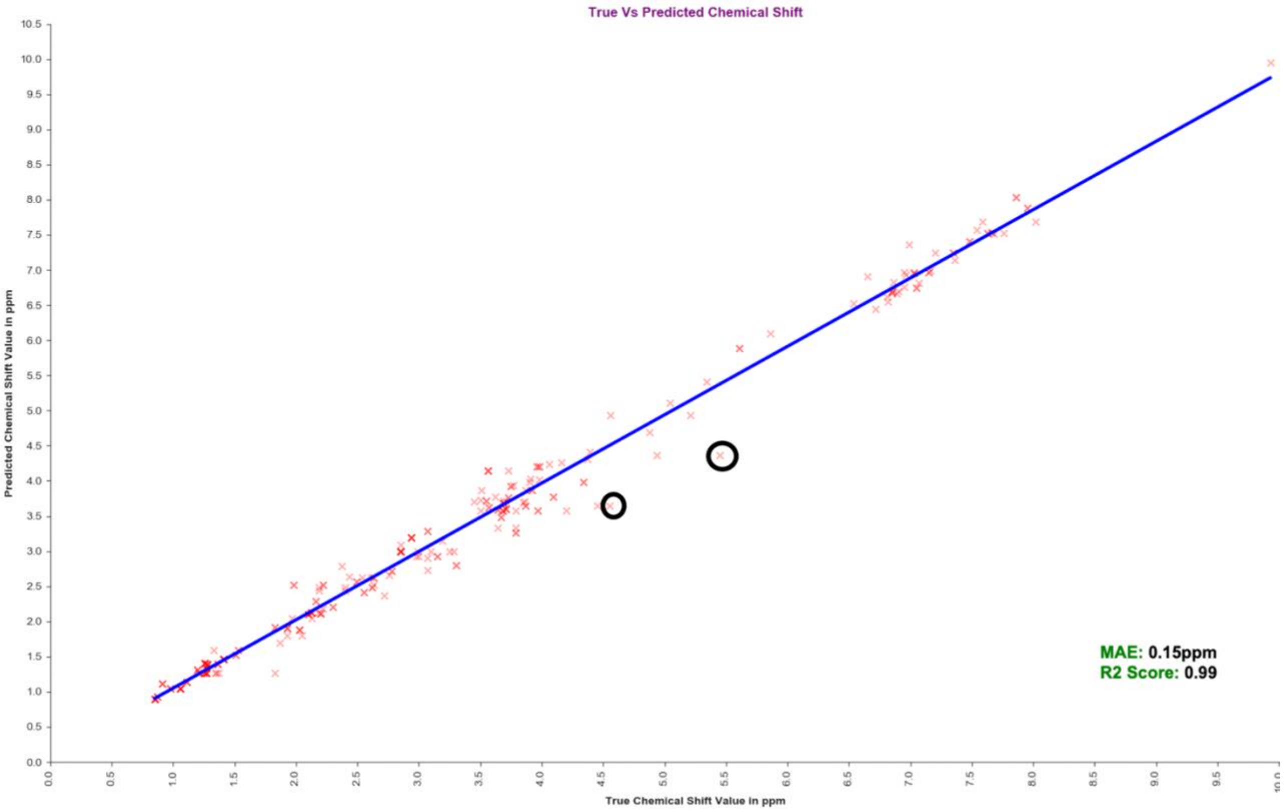


Figure 2.30: A scatter plot of the observed vs predicted $^1H$ chemical shifts in the first holdout dataset, using predictions from MNOVA.

the two CH2 hydrogens (atoms #19 and #18). On the other hand, neither MNOVA nor NMRShiftDB2 were able to predict the diastereotopic properties of these hydrogens. Note that the DFT predictor was able to distinguish between these diastereotopic hydrogens, the accuracy of this prediction was the worst among all predictors (Figure 2.35).

NMRPred was not alone in generating $^1$H chemical shift outliers. Both MNOVA and NMRShiftDB2, (see Figure 2.30 and Figure 2.31) had difficulty predicting the $^1$H chemical shifts

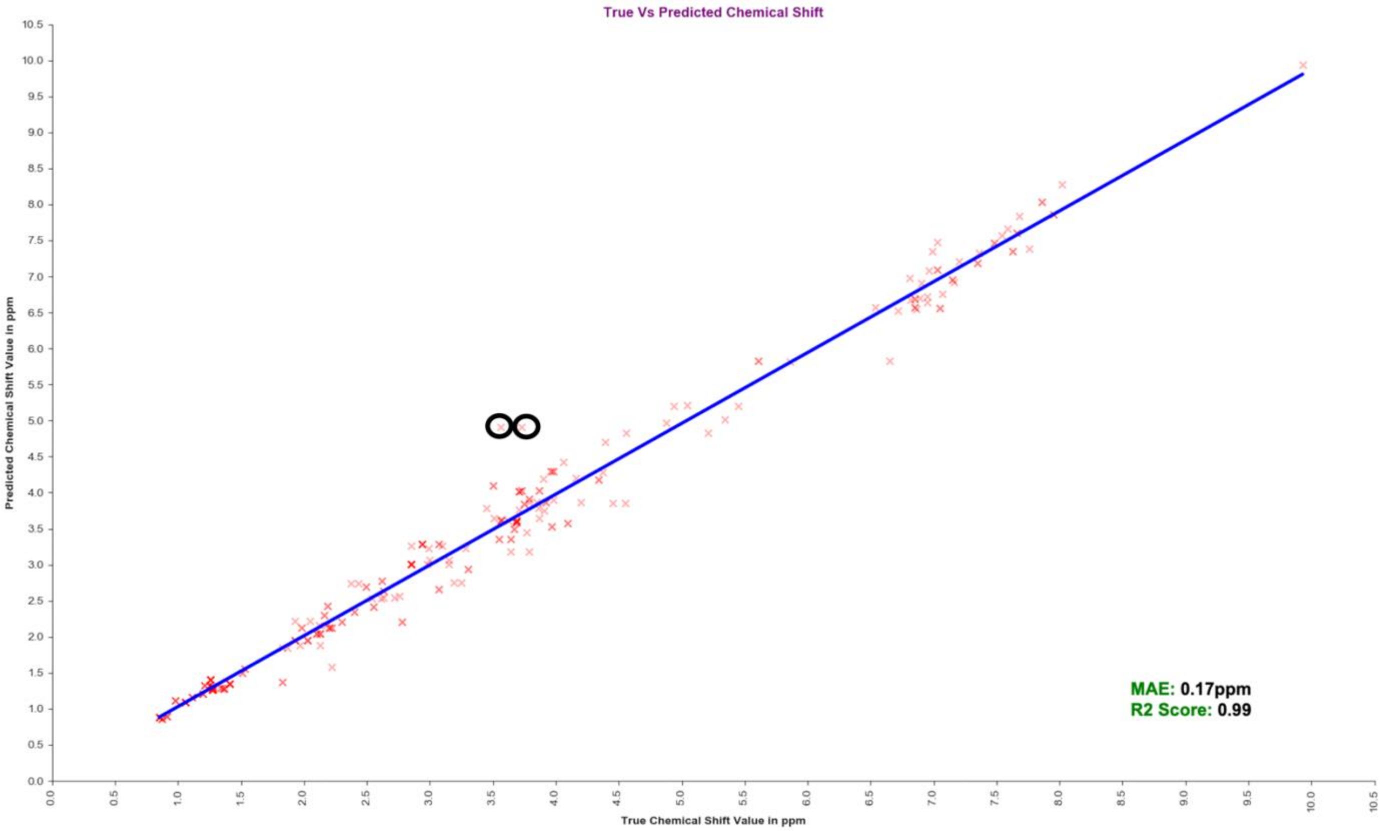


Figure 2.31: A scatter plot of the observed vs predicted $^1$H chemical shifts in the first holdout dataset using predictions from NMRShiftDB2.

of alpha-Muramic and beta-Muramic acid. On the other hand, the $^1$H chemical shifts predicted by NMRPred for these two molecules were quite accurate and could be used to differentiate between these two isomers. We also found that the DFT predicted results could distinguish between these two isomers, but the prediction results for these two molecules with the DFT predictor were not any better than NMRPred, as shown in Figure 2.34.

Figures 2.34 and 2.35 show a comparison of the calculated mean absolute errors (MAEs) for NMRPred, MNOVA, NMRShiftDB2 and the DFT predictor as calculated over all 36 molecules on the first (HMDB) holdout dataset. From those figures, we see that NMRPred had the most accurate $^1$H chemical shift predictions for 26/36 molecules (72.22%), while MNOVA had the most accurate $^1$H chemical shift predictions for 7/36 molecules (19.44%), NMRShiftDB2 had the most accurate $^1$H chemical shift predictions for just 4/36 molecules (11.11%), while the DFT predictor did not have the most accurate $^1$H chemical shift predictions for any of the molecules.

In addition to these assessments of $^1$H chemical shift prediction accuracy, we also analyzed how these four predictors could handle diastereotopic hydrogens. In our first holdout dataset, we identified 50 hydrogens that were diastereotopic. Among these 50 hydrogen atoms, NMRPred was able to predict the correct diastereotopic property for 44/50 (88%) of them. On the other hand, MNOVA could predict that property for only 24/50 (48%) of the hydrogen atoms. Interestingly, NMRShiftDB2 could not predict this property at all. Figure 2.36 shows the results for the diastereotopic predictions among the different predictors.

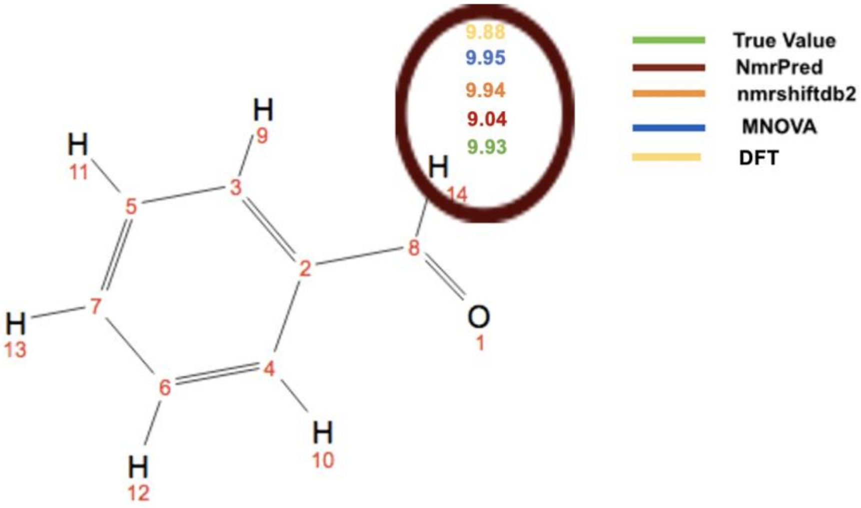


Figure 2.32: The observed $^{1}H$ chemical shift value of hydrogen atom #14 for Benzaldehyde (HMDB0006115) and the predicted $^{1}H$ chemical shift values with different predictors.

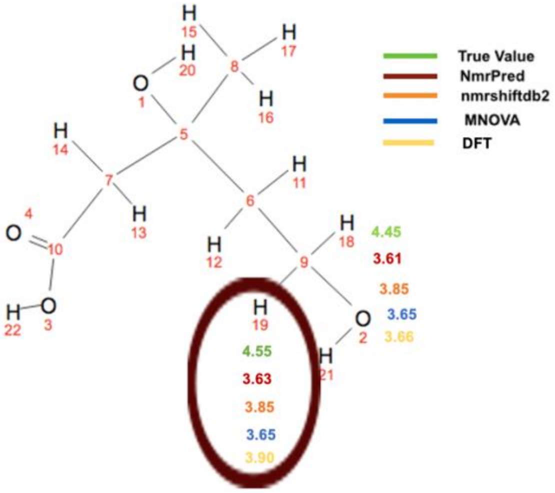


Figure 2.33: The observed $^{1}H$ chemical shift value of hydrogen atom #19, #18 in Mevalonic acid (HMDB0000227 and the predicted $^{1}H$ chemical shift values for different chemical shift predictors. Atoms #18 and #19 are diastereotopic.

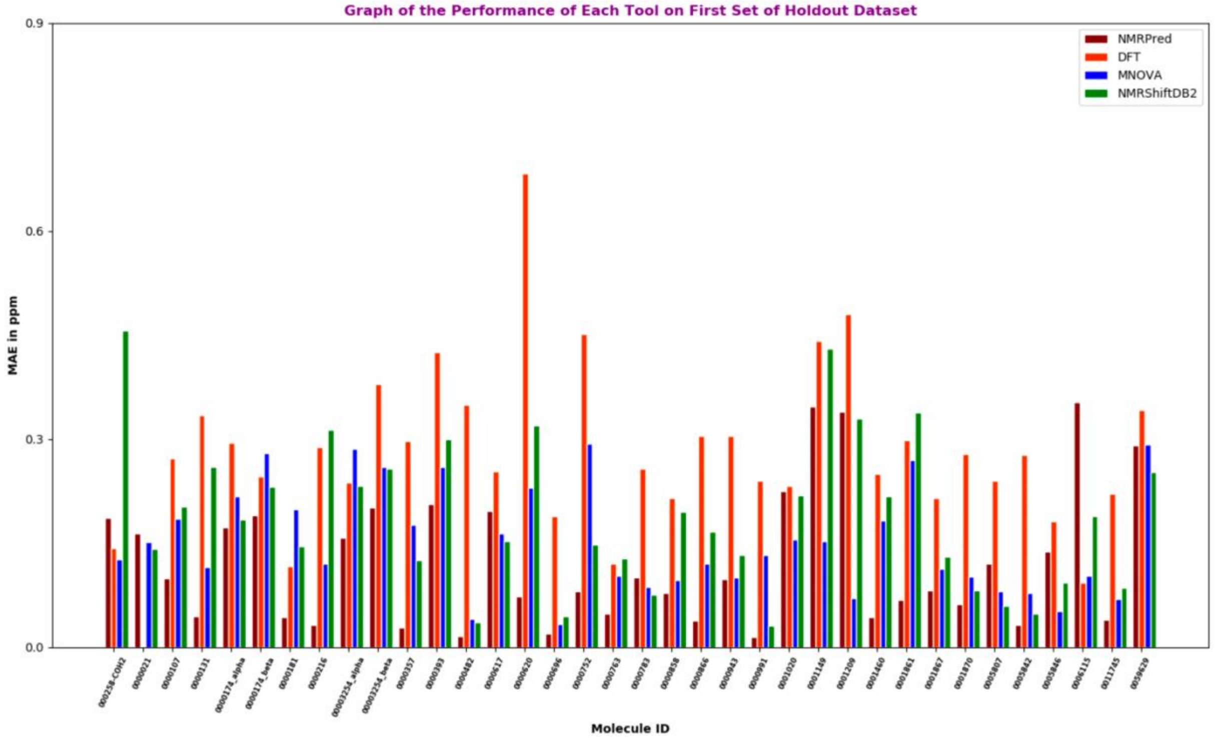
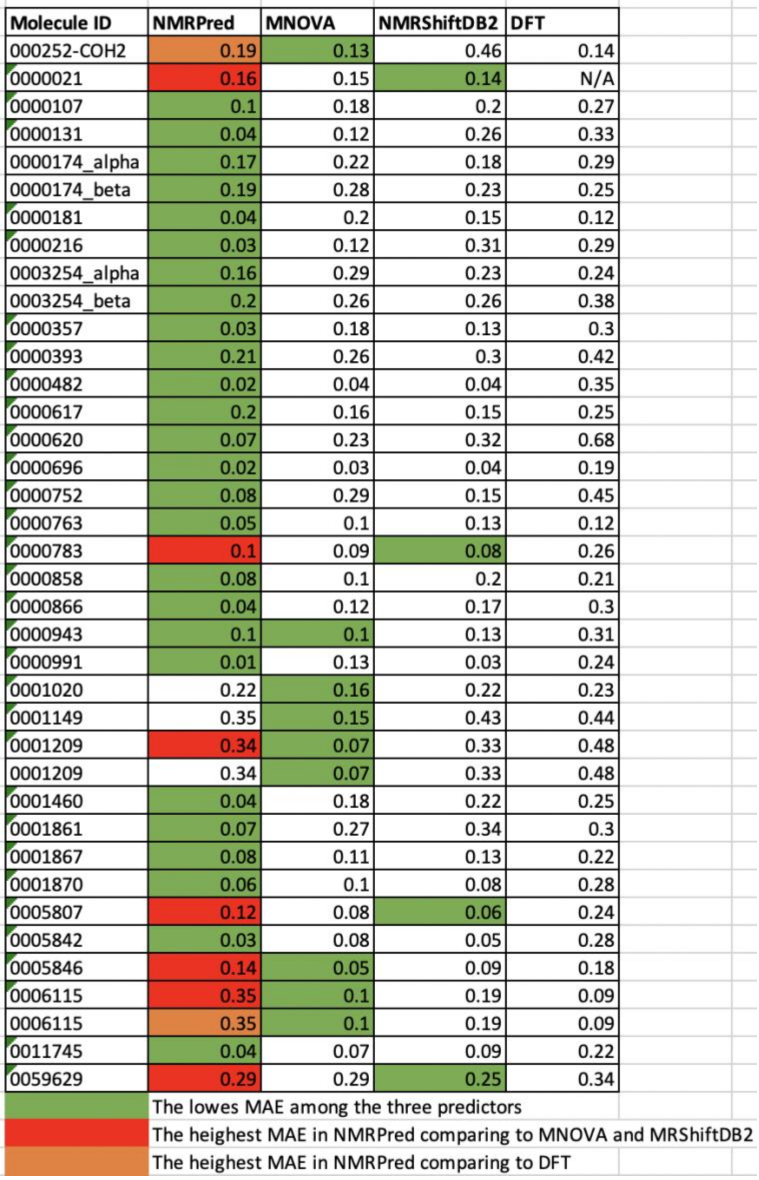


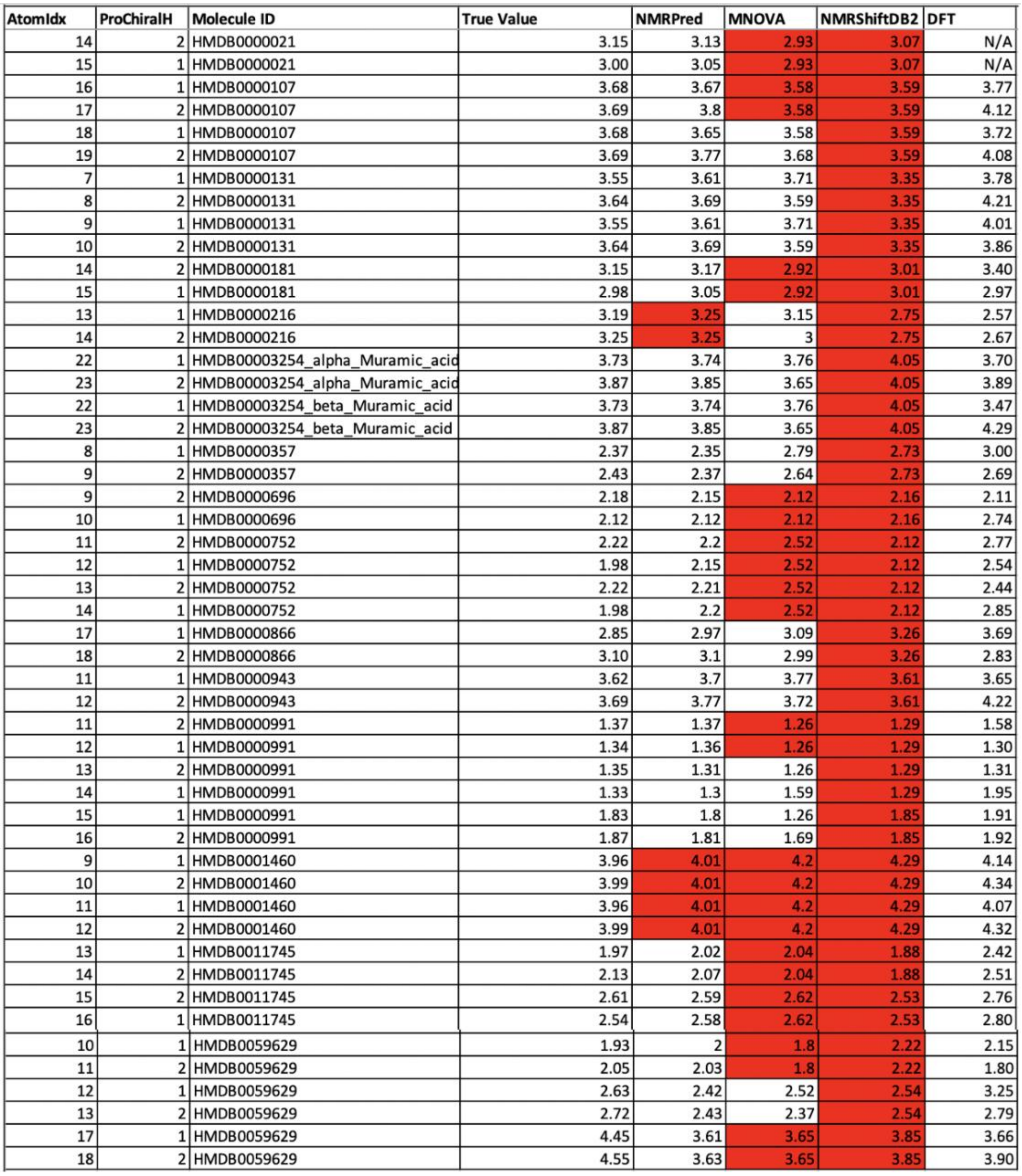
Figure 2.34: A comparison of the mean absolute errors (MAEs) for each of the 36 molecules in the first holdout dataset among the different predictors (NMRPred, MNOVA, NMRShiftDB2 and the DFT predictor).

To further test the performance NMRPred we also evaluated it against a second holdout dataset. This second holdout set consisted of $^1$H chemical shift assignments from the NP-MRD that included 22 molecules with 442 experimental $^1$H chemical shift assignments. NMRPred was evaluated on this second (NP-MRD) holdout dataset and the MAE was determined to be 0.36 ppm. This was substantially higher than the 0.11 ppm MAE for first holdout dataset (which was from HMDB). Furthermore, this prediction performance was found to be much worse than MNOVA (MAE = 0.20 ppm), NMRShiftDB2 (MAE = 0.25 ppm) and DFT (MAE = 0.23 ppm). For this NP-MRD holdout dataset MNOVA performed the best. Figure 2.37 shows the performance comparison between the 1st and 2nd holdout dataset among NMRPred, MNOVA, NMRShiftDB2 and the DFT predictor.



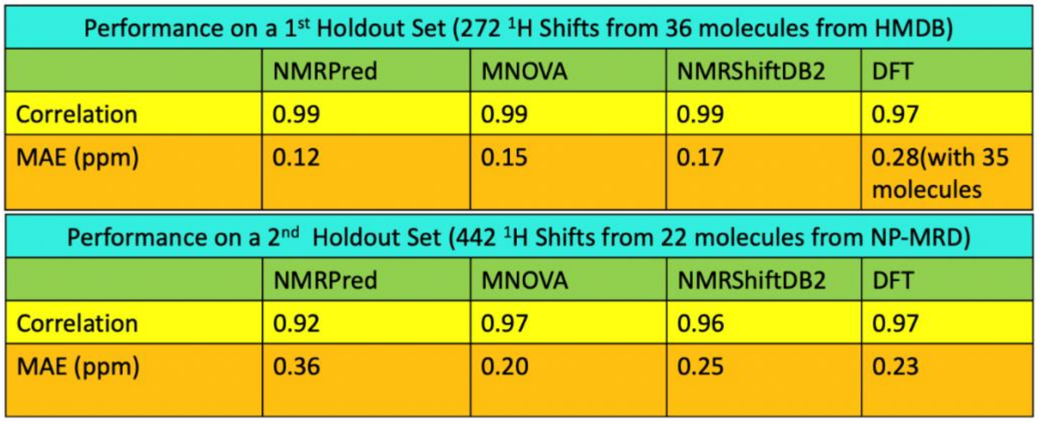
| Molecule ID | NMRPred | MNOVA | NMRShiftDB2 | DFT |
|---|---|---|---|---|
| 000252-COH2 | 0.19 | 0.13 | 0.46 | 0.14 |
| 0000021 | 0.16 | 0.15 | 0.14 | N/A |
| 0000107 | 0.1 | 0.18 | 0.2 | 0.27 |
| 0000131 | 0.04 | 0.12 | 0.26 | 0.33 |
| 0000174_alpha | 0.17 | 0.22 | 0.18 | 0.29 |
| 0000174_beta | 0.19 | 0.28 | 0.23 | 0.25 |
| 0000181 | 0.04 | 0.2 | 0.15 | 0.12 |
| 0000216 | 0.03 | 0.12 | 0.31 | 0.29 |
| 0003254_alpha | 0.16 | 0.29 | 0.23 | 0.24 |
| 0003254_beta | 0.2 | 0.26 | 0.26 | 0.38 |
| 0000357 | 0.03 | 0.18 | 0.13 | 0.3 |
| 0000393 | 0.21 | 0.26 | 0.3 | 0.42 |
| 0000482 | 0.02 | 0.04 | 0.04 | 0.35 |
| 0000617 | 0.2 | 0.16 | 0.15 | 0.25 |
| 0000620 | 0.07 | 0.23 | 0.32 | 0.68 |
| 0000696 | 0.02 | 0.03 | 0.04 | 0.19 |
| 0000752 | 0.08 | 0.29 | 0.15 | 0.45 |
| 0000763 | 0.05 | 0.1 | 0.13 | 0.12 |
| 0000783 | 0.1 | 0.09 | 0.08 | 0.26 |
| 0000858 | 0.08 | 0.1 | 0.2 | 0.21 |
| 0000866 | 0.04 | 0.12 | 0.17 | 0.3 |
| 0000943 | 0.1 | 0.1 | 0.13 | 0.31 |
| 0000991 | 0.01 | 0.13 | 0.03 | 0.24 |
| 0001020 | 0.22 | 0.16 | 0.22 | 0.23 |
| 0001149 | 0.35 | 0.15 | 0.43 | 0.44 |
| 0001209 | 0.34 | 0.07 | 0.33 | 0.48 |
| 0001209 | 0.34 | 0.07 | 0.33 | 0.48 |
| 0001460 | 0.04 | 0.18 | 0.22 | 0.25 |
| 0001861 | 0.07 | 0.27 | 0.34 | 0.3 |
| 0001867 | 0.08 | 0.11 | 0.13 | 0.22 |
| 0001870 | 0.06 | 0.1 | 0.08 | 0.28 |
| 0005807 | 0.12 | 0.08 | 0.06 | 0.24 |
| 0005842 | 0.03 | 0.08 | 0.05 | 0.28 |
| 0005846 | 0.14 | 0.05 | 0.09 | 0.18 |
| 0006115 | 0.35 | 0.1 | 0.19 | 0.09 |
| 0006115 | 0.35 | 0.1 | 0.19 | 0.09 |
| 0011745 | 0.04 | 0.07 | 0.09 | 0.22 |
| 0059629 | 0.29 | 0.29 | 0.25 | 0.34 |
| 🟩 | The lowes MAE among the three predictors | | | |
| 🟥 | The heighest MAE in NMRPred comparing to MNOVA and MRShiftDB2 | | | |
| 🟧 | The heighest MAE in NMRPred comparing to DFT | | | |

Figure 2.35: A comparison of mean absolute errors for each of the 36 molecules in the holdout dataset among each of the four predictors: NMRPred, MNOVA, NMRShiftDB2 and the DFT predictor.



| AtomIdx | ProChiralH | Molecule ID | True Value | NMRPred | MNOVA | NMRShiftDB2 | DFT |
|---|---|---|---|---|---|---|---|
| 14 | 2 | HMDB0000021 | 3.15 | 3.13 | 2.93 | 3.07 | N/A |
| 15 | 1 | HMDB0000021 | 3.00 | 3.05 | 2.93 | 3.07 | N/A |
| 16 | 1 | HMDB0000107 | 3.68 | 3.67 | 3.58 | 3.59 | 3.77 |
| 17 | 2 | HMDB0000107 | 3.69 | 3.8 | 3.58 | 3.59 | 4.12 |
| 18 | 1 | HMDB0000107 | 3.68 | 3.65 | 3.58 | 3.59 | 3.72 |
| 19 | 2 | HMDB0000107 | 3.69 | 3.77 | 3.68 | 3.59 | 4.08 |
| 7 | 1 | HMDB0000131 | 3.55 | 3.61 | 3.71 | 3.35 | 3.78 |
| 8 | 2 | HMDB0000131 | 3.64 | 3.69 | 3.59 | 3.35 | 4.21 |
| 9 | 1 | HMDB0000131 | 3.55 | 3.61 | 3.71 | 3.35 | 4.01 |
| 10 | 2 | HMDB0000131 | 3.64 | 3.69 | 3.59 | 3.35 | 3.86 |
| 14 | 2 | HMDB0000181 | 3.15 | 3.17 | 2.92 | 3.01 | 3.40 |
| 15 | 1 | HMDB0000181 | 2.98 | 3.05 | 2.92 | 3.01 | 2.97 |
| 13 | 1 | HMDB0000216 | 3.19 | 3.25 | 3.15 | 2.75 | 2.57 |
| 14 | 2 | HMDB0000216 | 3.25 | 3.25 | 3 | 2.75 | 2.67 |
| 22 | 1 | HMDB00003254_alpha_Muramic_acid | 3.73 | 3.74 | 3.76 | 4.05 | 3.70 |
| 23 | 2 | HMDB00003254_alpha_Muramic_acid | 3.87 | 3.85 | 3.65 | 4.05 | 3.89 |
| 22 | 1 | HMDB00003254_beta_Muramic_acid | 3.73 | 3.74 | 3.76 | 4.05 | 3.47 |
| 23 | 2 | HMDB00003254_beta_Muramic_acid | 3.87 | 3.85 | 3.65 | 4.05 | 4.29 |
| 8 | 1 | HMDB0000357 | 2.37 | 2.35 | 2.79 | 2.73 | 3.00 |
| 9 | 2 | HMDB0000357 | 2.43 | 2.37 | 2.64 | 2.73 | 2.69 |
| 9 | 2 | HMDB0000696 | 2.18 | 2.15 | 2.12 | 2.16 | 2.11 |
| 10 | 1 | HMDB0000696 | 2.12 | 2.12 | 2.12 | 2.16 | 2.74 |
| 11 | 2 | HMDB0000752 | 2.22 | 2.2 | 2.52 | 2.12 | 2.77 |
| 12 | 1 | HMDB0000752 | 1.98 | 2.15 | 2.52 | 2.12 | 2.54 |
| 13 | 2 | HMDB0000752 | 2.22 | 2.21 | 2.52 | 2.12 | 2.44 |
| 14 | 1 | HMDB0000752 | 1.98 | 2.2 | 2.52 | 2.12 | 2.85 |
| 17 | 1 | HMDB0000866 | 2.85 | 2.97 | 3.09 | 3.26 | 3.69 |
| 18 | 2 | HMDB0000866 | 3.10 | 3.1 | 2.99 | 3.26 | 2.83 |
| 11 | 1 | HMDB0000943 | 3.62 | 3.7 | 3.77 | 3.61 | 3.65 |
| 12 | 2 | HMDB0000943 | 3.69 | 3.77 | 3.72 | 3.61 | 4.22 |
| 11 | 2 | HMDB0000991 | 1.37 | 1.37 | 1.26 | 1.29 | 1.58 |
| 12 | 1 | HMDB0000991 | 1.34 | 1.36 | 1.26 | 1.29 | 1.30 |
| 13 | 2 | HMDB0000991 | 1.35 | 1.31 | 1.26 | 1.29 | 1.31 |
| 14 | 1 | HMDB0000991 | 1.33 | 1.3 | 1.59 | 1.29 | 1.95 |
| 15 | 1 | HMDB0000991 | 1.83 | 1.8 | 1.26 | 1.85 | 1.91 |
| 16 | 2 | HMDB0000991 | 1.87 | 1.81 | 1.69 | 1.85 | 1.92 |
| 9 | 1 | HMDB0001460 | 3.96 | 4.01 | 4.2 | 4.29 | 4.14 |
| 10 | 2 | HMDB0001460 | 3.99 | 4.01 | 4.2 | 4.29 | 4.34 |
| 11 | 1 | HMDB0001460 | 3.96 | 4.01 | 4.2 | 4.29 | 4.07 |
| 12 | 2 | HMDB0001460 | 3.99 | 4.01 | 4.2 | 4.29 | 4.32 |
| 13 | 1 | HMDB0011745 | 1.97 | 2.02 | 2.04 | 1.88 | 2.42 |
| 14 | 2 | HMDB0011745 | 2.13 | 2.07 | 2.04 | 1.88 | 2.51 |
| 15 | 2 | HMDB0011745 | 2.61 | 2.59 | 2.62 | 2.53 | 2.76 |
| 16 | 1 | HMDB0011745 | 2.54 | 2.58 | 2.62 | 2.53 | 2.80 |
| 10 | 1 | HMDB0059629 | 1.93 | 2 | 1.8 | 2.22 | 2.15 |
| 11 | 2 | HMDB0059629 | 2.05 | 2.03 | 1.8 | 2.22 | 1.80 |
| 12 | 1 | HMDB0059629 | 2.63 | 2.42 | 2.52 | 2.54 | 3.25 |
| 13 | 2 | HMDB0059629 | 2.72 | 2.43 | 2.37 | 2.54 | 2.79 |
| 17 | 1 | HMDB0059629 | 4.45 | 3.61 | 3.65 | 3.85 | 3.66 |
| 18 | 2 | HMDB0059629 | 4.55 | 3.63 | 3.65 | 3.85 | 3.90 |

Figure 2.36: A comparison of the prediction of the diastereotopic hydrogens among the NMRPred, MNOVA, NMRShiftDB2 and DFT predictors. Red highlights indicate that the diastereotopic hydrogens were not detected.



| Performance on a 1st Holdout Set (272 $^{1}$H Shifts from 36 molecules from HMDB) | | | | |
|---|---|---|---|---|
| | NMRPred | MNOVA | NMRShiftDB2 | DFT |
| Correlation | 0.99 | 0.99 | 0.99 | 0.97 |
| MAE (ppm) | 0.12 | 0.15 | 0.17 | 0.28(with 35 molecules |

| Performance on a 2nd Holdout Set (442 $^{1}$H Shifts from 22 molecules from NP-MRD) | | | | |
|---|---|---|---|---|
| | NMRPred | MNOVA | NMRShiftDB2 | DFT |
| Correlation | 0.92 | 0.97 | 0.96 | 0.97 |
| MAE (ppm) | 0.36 | 0.20 | 0.25 | 0.23 |

Figure 2.37: Performance comparison between 1st and 2nd holdout datasets among the four predictors: NMRPred, MNOVA, NMRShiftDB2 and the DFT predictor.

The significant drop in performance by NMRPred relative to the other chemical shift predictors suggested that either NMRPred was over-trained or that it was under-trained. Earlier evaluations on the "Train-error" suggested that over-training was modest or unlikely. Given the relatively small training set originally used to train and test NP-MRD, we suspected that under-training was more likely the problem and that the poor performance by NMRPred was due to the fact that it had not seen (or been trained on) many of the chemical structure classes seen in the second (NP-MRD) holdout dataset. To test this hypothesis, we used ClassyFire (76) to quantitatively assess the chemical structure classes seen in NP-MRD's training dataset and the two (HMDB and NP-MRD) holdout datasets. ClassyFire is a computer program that automatically classifies all known chemical compounds into one of more than 4800 different structural categories using chemical structure information. ClassyFire uses the ChemOnt database, which is an extensive, adaptable, and fully calculable chemical taxonomy database, to assign each chemical to a specific chemical superclass, class and subclass. Using ClassyFire we found that our original training dataset contained molecules from 90 different chemical subclasses. The first holdout dataset (with 36 molecules from the HMDB) 34/36 had structures that belonged to at least one of these chemical subclasses. The two exceptions were, one for the compound that belonged to the subclass “Short-chain hydroxy acids and derivatives” and the other for the compound that belonged to “Thiophosphoric acid esters”. On the other hand, for the second holdout dataset (with 22 molecules from the NP-MRD) only 3/22 molecules belonged to chemical subclasses found in the original training dataset. These three molecules were: NP0006813 (Butyl 2,4-dihydroxy-6-methylbenzoate), NP0040444 (Flavalin I), and NP0035870 (9-(3-methylbutanoyl)-8,10-dehydrothymol). These molecules belong to the chemical subclasses known as "Benzoic acids and derivatives", "Carbonyl compounds", and "Cresols”, respectively. The number of molecules belonging to those subclasses in the original NMRPred training dataset was 22, 14, and 5, respectively. We found the MAE for the $^{1}$H chemical shifts for these 3 compounds, as predicted by NMRPred, was 0.15 ppm whereas for MNOVA and NMRShiftDB2, the MAEs were 0.05 ppm and 0.06 ppm. For these three compounds, we anticipated that NMRPred would yield better results than MNOVA and NMRShiftDB2, but it clearly did not. For ~7.11% of the training dataset and ~ 27.27% of the holdout dataset, ClassyFire was unable to identify the chemical subclasses. Figure 2.38 and Table 2.4 show the chemical subclass distribution for the training dataset, the first holdout dataset (from HMDB) and the second holdout dataset (NP-MRD). Additionally, Figure 2.39 shows the above mentioned three NP-MRD molecules and the variation in the molecular structures in the training dataset of the same subclass compounds. The last bar in the graph indicates the total number of the compounds for which chemical subclasses were unknown or for which ClassyFire could not determine. Given this data distribution and the variation in the structures in the training dataset, we can conclude that NMRPred was under-trained and that its poor performance for the

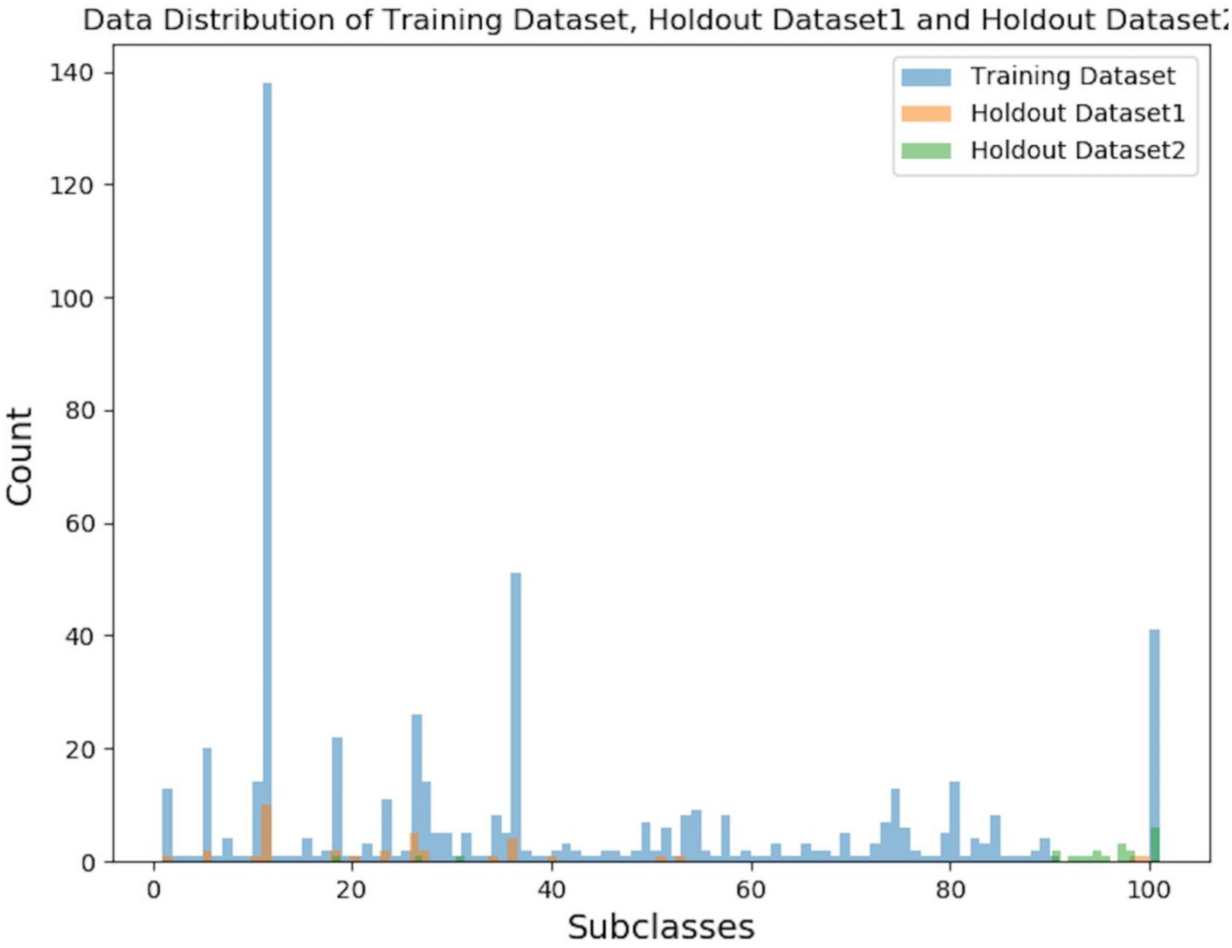


Figure 2.38: The class distribution of chemicals seen in the two holdout datasets compared to the NMRPred training dataset.

NP-MRD (second) holdout dataset was due to the fact that NMRPred had not been trained on any molecules or a sufficient number of molecules belonging to the chemical subclasses seen in the NP-MRD (second) holdout dataset. Given the relatively small training set of molecules and chemical shifts originally used to develop NMRPred, this was not entirely unexpected.

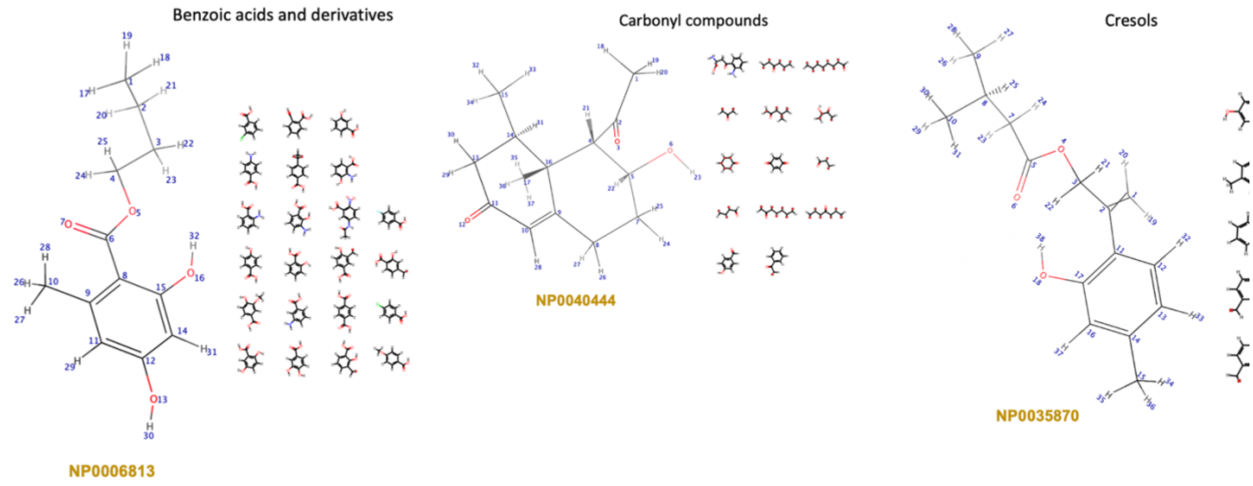


Figure 2.39: The three NP-MRD molecules (NP0006813, NP0040444 and NP0035870) and the variation in the molecular structures in the training dataset of the same subclass compounds.

| Chemical Subclasses | # of Occuarance in Training Dataset | # of Occuarance in Holdout Set1 | # of Occuarance in Holdout Set2 (NP-MRD) |
|---|---|---|---|
| 1-hydroxy-2-unsubstituted benzenoids | 13 | 1 | 0 |
| Alcohols and polyols | 20 | 2 | 0 |
| Amines | 14 | 1 | 0 |
| Amino acids, peptides, and analogues | 138 | 10 | 0 |
| Benzoic acids and derivatives | 22 | 2 | 1 |
| Benzoyl derivatives | 1 | 1 | 0 |
| Beta hydroxy acids and derivatives | 11 | 2 | 0 |
| Carbohydrates and carbohydrate conjugates | 26 | 5 | 0 |
| Carbonyl compounds | 14 | 2 | 1 |
| Cresols | 5 | 0 | 1 |
| Dicarboxylic acids and derivatives | 8 | 1 | 0 |
| Fatty acids and conjugates | 51 | 4 | 0 |
| Furoic acid and derivatives | 2 | 1 | 0 |
| Imidazoles | 6 | 1 | 0 |
| Indoles | 8 | 1 | 0 |
| 1-benzopyrans | 0 | 0 | 2 |
| Acetophenones | 0 | 0 | 1 |
| Benzoquinolines | 0 | 0 | 1 |
| Dihydrobenzophenanthridine alkaloids | 0 | 0 | 1 |
| Diterpenoids | 0 | 0 | 2 |
| O-methylated flavonoids | 0 | 0 | 1 |
| Sesquiterpenoids | 0 | 0 | 3 |
| Terpene lactones | 0 | 0 | 2 |
| Short-chain hydroxy acids and derivatives | 0 | 1 | 0 |
| Thiophosphoric acid esters | 0 | 1 | 0 |
| Unknown | 41 | 0 | 6 |

Table 2.4: The chemical class distribution in the NMRPred training dataset, the 1st holdout dataset (from HMDB) and the 2nd holdout dataset (from NP-MRD) as indicated by membership in chemical subclasses.

## $^1H$ NMR Chemical Shift Predictions in $CDCl_3$ and DMSO

All of our training and testing data for our $^1H$ chemical shift predictors were done using compounds dissolved in $H_2O$. While water is a common solvent used in NMR-based metabolomics, in the world of natural product chemistry, most compounds are dissolved in other solvents, such as chloroform ($CDCl_3$) or dimethylsulfoxide (DMSO). It is also known that different solvents will cause systematic "solvent" shifts (due to anisotropic effects) that will move chemical shifts up or down relative to those measured in water. Likewise, organic solvents tend to prevent hydrogen exchange (unlike water) and so hydrogen atoms from labile hydrogens attached to OH and NH function groups will be visible in the NMR spectrum. To determine the systematic shift arising from $CDCl_3$ and DMSO relative to water, we evaluated the reported $^1H$ chemical shift values of a number of identical compounds dissolved in water, $CDCl_3$, and DMSO. With this information in hand, we were able to identify straightforward linear relationships between the $^1H$ chemical shift values reported in water those reported in $CDCl_3$ as well as the linear relationships between the $^1H$ chemical shift values reported in water those reported in DMSO. These equations and the quality of the fit between the different pairs of $^1H$ chemical shifts are shown in Figure 2.40 and Figure 2.41. These equations have been incorporated into NMRPred to adjust the predicted $^1H$ chemical shift values for molecules dissolved in $CDCl_3$ and DMSO respectively.

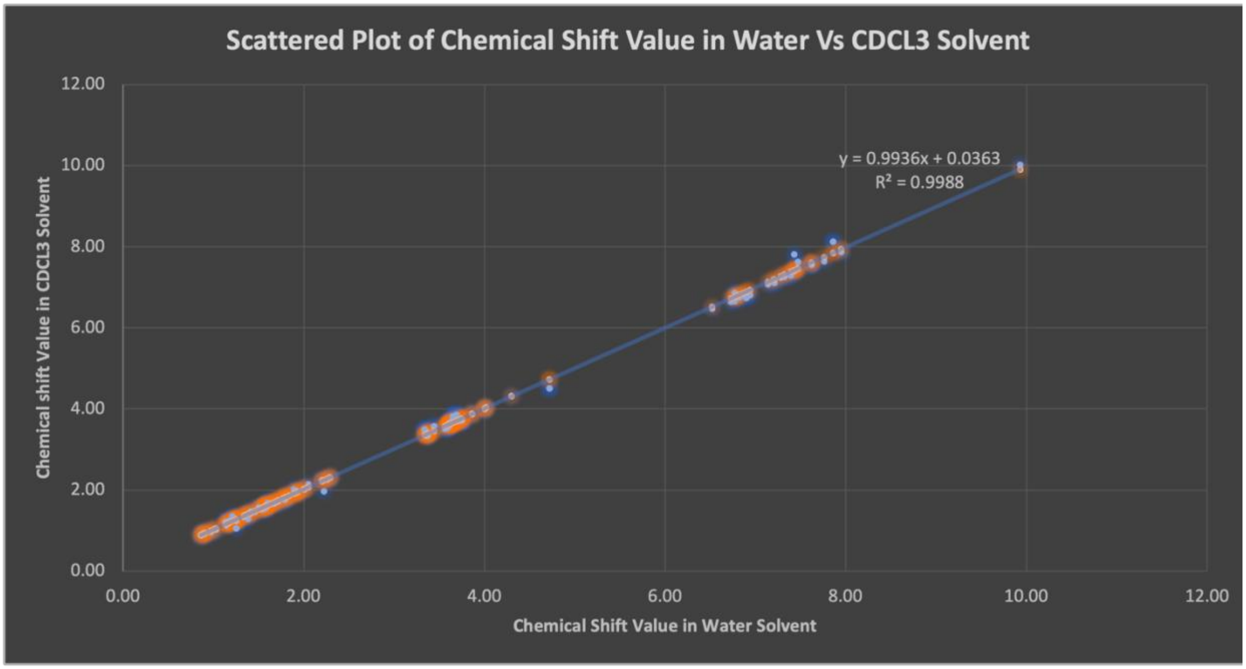


Figure 2.40: The simple linear equation that can be used to predict the $^{1}$H chemical shift values of hydrogen atoms for molecules dissolved in CDCl3 relative to those dissolved in water.

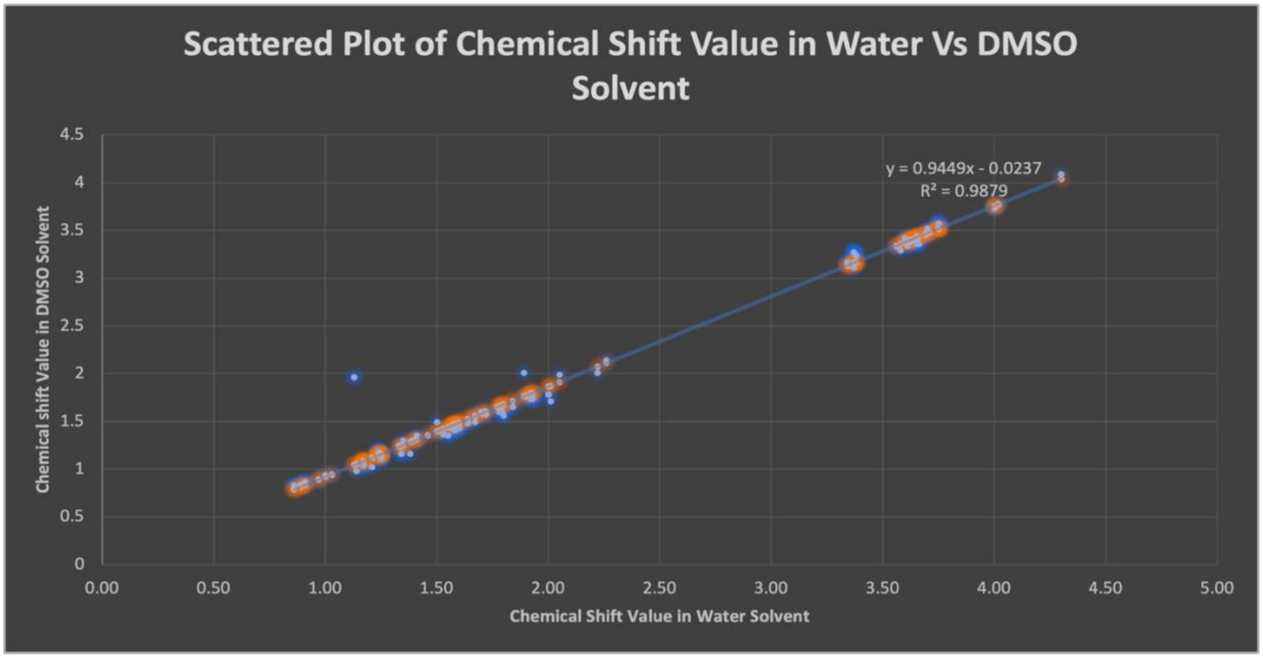


Figure 2.41: The simple linear equation that can be used to predict the $^{1}$H chemical shift values of hydrogen atoms of molecules dissolved in DMSO relative to those dissolved in water.

# Conclusion

As outlined in this chapter I have described how I successfully assembled, curated and cleaned a moderately large database of experimentally acquired $^1$H chemical shifts from several well-known NMR databases including HMDB, BMRB and GISSMO. This extracted dataset required a considerable amount of manual remediation to produce the high-quality dataset needed to train and test my machine learning models. This resulting dataset was then split into a training set (consisting of 577 molecules and 4207 $^1$H chemical shifts) and two different holdout sets (consisting of 36 molecules from the HMDB (holdout set #1) and 22 molecules from the NP-MRD (holdout set #2) molecules with a total of 714 $^1$H chemical shifts). After the $^1$H chemical shift datasets had been prepared (consisting of 3D structures and experimentally assigned $^1$H chemical shifts for each molecule), I used literature-derived data and the CDKit program to calculate an appropriate set of molecular features for $^1$H chemical shift prediction. These features had to capture the key molecular, geometric and atomic properties that are known to contribute to chemical shift effects at both local and distant levels. Intelligent feature selection was able to reduce the initial size of the feature set from thousands to just 210 features for each $^1$H atom under consideration.

I chose four different machine learning (regressor) algorithms to assess their performance in predicting $^1$H chemical shifts. 5-fold internal cross validation was used to optimize the hyperparameters for each of the regressors on the training set and an external 5-fold cross validation was used to evaluate their performance via MAE and $R^2$ calculations. As the size of our dataset was not as large as that used by most other chemical shift predictors, we decided to use nested cross validation to prevent the adverse influence of data leakage. Training error assessment was also done to determine in any of the regressor algorithms were over-trained. The best performing algorithm identified from the external cross validation was then assessed on two holdout datasets. The top performing algorithm was identified as a Random Forest Regressor (RFR). This ML model was incorporated into the chemical shift prediction program called NMRPred.

NMRPred has a number of other functions that allow it to accept a SMILES string for an organic molecule, to generate a 3D chemical structure (in SDF format), to decorate the input molecule with hydrogens, to calculate all the relevant atomic, molecular and geometric features of the input molecule and then to pass these features into the RFR model to generate the $^1$H chemical shifts. NMRPred is also able make solvent chemical shift adjustments (using linear modeling) for both $CDCl_3$ and DMSO.

NMRPred was then evaluated against several commercial or popular $^1$H chemical shift predictors, including MNOVA, NMRShiftDB and a DFT method using one holdout dataset derived from the HMDB and another holdout dataset derived from NP-MRD. The MAE of NMRPred for the HMDB holdout set (holdout set #1) of 36 molecules with 272 experimental hydrogen chemical shift was determined to be 0.11 ppm. This result was found to be better than MNOVA (MAE = 0.15 ppm) NMRShiftDB2 (MAE = 0.17 ppm) and the DFT predictor (MAE = 0.28 ppm). Further comparisons showed that NMRPred showed the best prediction result for 72.2% of the molecules in the HMDB holdout dataset, whereas MNOVA showed the best results for 19.4% of the molecules in the holdout dataset and NMRShiftDB2 showed the best results for just 11.1% of the molecules in the holdout dataset. NMRPred also exhibited superior ability to identify and differentiate diastereotopic protons relative to both MNOVA and NMRShiftDB2.

On the other hand, when NMRPred was run on the NP-MRD holdout dataset (holdout dataset #2) consisting of 22 molecules with 442 experimental $^1$H chemical shifts, the MAE was 0.36 ppm. This was found to be worse than the other predictors, including MNOVA (0.20 ppm), the DFT predictor (0.23 ppm) and NMRShiftDB2 (MAE = 0.25 ppm), even after correcting for solvent differences (many compounds in the NP-MRD holdout set were dissolved in $CDCl_3$). These results were quite disappointing and suggest that NMRPred (and its RFR model) were either overtrained or undertrained. We did a number of evaluations of the RFR model to test for overtraining. It is notable that the "Train error" for the RFR model was just 0.03 ppm while the external cross-validation error was 0.12 ppm (as was the holdout error). This suggests that a small degree of overfitting likely occurred. However, the more significant culprit for NMRPred's poor performance on holdout dataset #2 appears to be undertraining. Based on the structural diversity of the NP-MRD holdout dataset (using ClassyFire's structural classification method) when compared to the structural diversity of the HMDB holdout dataset (holdout dataset #1), it appears that the NMRPred model was undertrained. In other words, too few examples of key functional groups, key chemical properties or key molecular geometries were available in the original HMDB training set to allow it to properly handle the novel structures seen in the NP-MRD holdout dataset (holdout dataset #2). To address this issue of undertraining, the training dataset would have to be made much larger and much more diverse.

It is also notable that feature dimensionality in our training dataset is high. If the training data contains too many features and not enough examples. the model will tend to do very well in the training dataset and exhibit a much worse performance when evaluated on the holdout dataset. We believe our model suffered from at least two problems: 1) insufficient training (due to a training set that was too small); 2) too many features (with too few training samples). Unfortunately, I was unable to find the time to expand the $^{1}$H NMR chemical shift dataset or refine the atomic feature space before the mandatory end-date for my MSc program.

# Chapter 3: Application of Machine Learning to the Prediction of $^{13}C$ Chemical Shifts of Small Organic Molecules

## Introduction

By definition, organic molecules must contain carbon atoms. This fact means that the characterization of carbon atoms is a critical component of any effort aimed at the structural determination or structure description of organic chemicals. However, unlike the situation for hydrogen (as described in Chapter 2), where the most abundant isotope of hydrogen (i.e., $^{1}H$) is NMR active, the most abundant isotope of carbon (i.e., $^{12}C$) is NMR inactive. The NMR active isotope of carbon is $^{13}C$ and, unfortunately, this isotope has a natural abundance of just 1.1%. Furthermore, the gyromagnetic ratio ($\gamma$) for $^{13}C$ is only ¼ that of $^{1}H$. Since the sensitivity of an NMR signal is proportional to the cube of the gyromagnetic ratio, this means that the signal intensity arising from a natural abundance $^{13}C$ NMR resonance is only 0.011/64 or 1/5700 of that of a natural abundance $^{1}H$ NMR resonance. As a result, the collection of natural abundance $^{13}C$ NMR spectra can be very time consuming and often require large amounts of material to get a useful NMR signal. However, by enriching or synthesizing an organic compound with $^{13}C$ instead of $^{12}C$ (a process called isotopic labeling), it is possible to increase the NMR sensitivity of that isotopically labeled compound by more than 90-fold. Furthermore, by using a technique called proton-decoupling (which converts $^{13}C$ multiplets into singlets), it is possible to not only simplify the $^{13}C$ NMR spectra but to enhance the $^{13}C$ resonance signals by a factor of two or more (79).

Likewise, by conducting specially designed NMR experiments (i.e., NMR pulse sequences) that involve detecting the $^{13}C$ nuclei through the attached $^{1}H$ nuclei, it is possible to greatly enhance the $^{13}C$ signal, even at natural abundance. These NMR experiments are called HSQC (heteronuclear single quantum correlation) and HMQC (heteronuclear multiple quantum correlation) experiments (79, 80). As a result, $^{13}C$-$^{1}H$ HMQC and $^{13}C$-$^{1}H$ HSQC NMR experiments have become very common methods for collecting $^{13}C$ chemical shift assignments over the past 20 years. These experiments have led to hundreds of thousands of $^{13}C$ chemical shifts being measured and assigned for tens of thousands of organic molecules.

As a result of this concerted effort to collect and analyze $^{13}C$ chemical shifts, a number of important insights have been gained. In particular, $^{13}C$ chemical shifts have been found to span a much wider range of chemical shift space (from 0 to 220 ppm) than $^{1}H$ chemical shifts (which only span from 0 to 10 ppm). This is likely due to the fact that $^{13}C$ chemical shifts tend to be much more sensitive to the electronic environment of the carbon atom being measured and to the functional groups that are attached to it. As a result, $^{13}C$ NMR spectra are simpler, have less severe problems with overlapping peaks, are more comparable across different magnetic field strengths, and are less susceptible to solvent effects. This makes $^{13}C$ NMR spectra easier to interpret and easier to assign. Indeed, many organic chemists and natural product chemists believe that $^{13}C$ NMR chemical shifts can often provide as much, if not more, information about the structure of a molecule than $^{1}H$ shifts. However, it should also be noted that $^{13}C$ chemical shifts tend to be more sensitive to the chemical shift reference compound (79) that is chosen than $^{1}H$ chemical shifts. In particular, the $^{13}C$ chemical shifts referenced using TMS (tetramethylsilane) can be up to 2.7 ppm different than those $^{13}C$ chemical shifts referenced using DSS (sodium trimethylsilylpropanesulfonate). Nevertheless, given their importance in NMR-based structure analysis and determination we decided to develop a $^{13}$C chemical shift predictor to complement the work on the $^{1}$H chemical shift predictor (described in Chapter 2).

This chapter will describe the development and testing of this machine learning-based $^{13}$C chemical shift predictor. It will describe the collection and curation process used to assemble the $^{13}$C NMR training and testing data, the specific ML algorithms that were tested, how the different $^{13}$C predictors' performance was measured, the results that were achieved for these predictors, and finally an assessment of how the optimal $^{13}$C shift predictor compared with several popular programs for $^{13}$C NMR chemical shift prediction. I will also discuss some of the reasons for the poorer-than-expected performance of my $^{13}$C predictor.

## Problem Definition

My task was defined as follows: Predict the $^{13}$C chemical shifts of all the carbon atoms in a given small molecule using a single chemical structure written as a SMILES string. In addition to this primary task, the predicted $^{13}$C chemical shifts must be adjustable to match different NMR chemical shift reference standards (such as DSS, TSP and TMS). This task was broken down into four separate steps. The first step involved developing a method to transform the input SMILES strings into 3D structures with the proper chemical geometry. The second step involved modifying the generated structure by adding hydrogen atoms to the relevant heavy atoms. The third step involved creating a feature set and an ML-based model that used the 3D structural coordinates and atom types to ascertain the $^{13}$C chemical shifts for each carbon atom in the selected molecule. The fourth step involved adjusting the predicted $^{13}$C chemical shifts to match the $^{13}$C chemical shifts for selected NMR chemical shift reference compounds. The same chemical or cheminformatics terminology and many of the same chemical analysis programs used in Chapter 2 were also used here and so I will refer the reader to Chapter 2 for this information.

## Methodological Outline

Similar to $^1$H chemical shift prediction (as described in Chapter 2), the application of ML algorithms to predict $^{13}$C chemical shifts needs a large database of accurate chemical structures and precisely (experimentally) determined $^{13}$C chemical shifts. Additionally, the atomic numbering systems utilized in such a structure/shift collection must be consistent, and it is necessary to specify which chemical shift reference standard (DSS, TSP or TMS) was used to gather the experimental data. A number of databases exist which contain small molecule structures, $^{13}$C chemical shift assignments, chemical shift reference compounds and NMR solvent data. These include the HMDB (15), BioMagResBank (11) NMRShiftDB (12) and NP-MRD (14). However, compiling, re-referencing and uniformly numbering these chemical shift assignments proved to be particularly difficult as there is little standardization in the field. This chapter provides details on the compilation and cleaning of these NMR training/testing data. Because the values for all $^{13}$C chemical shifts are real numbers, ML-based regression algorithms were used to perform chemical shift prediction. To maximize the effectiveness of each regressor, considerable effort had to be put into the selection of suitable atomic, molecular, and structural properties (or features). The process of feature selection, which is crucial, will also be covered in this chapter. Along with choosing the optimal structural features for $^{13}$C chemical shift prediction, we investigated several popular ML regression techniques. It is important to evaluate different regressors since they frequently produce diverse findings. We tested four different regressors: a Support Vector Regressor (SVR), a Random Forest Regressor (RFR), and Extreme Gradient Boosting Regressor (XGBoostRegressor or XGBR), and a Gradient Boosting Regressor (GBR). The performance of these ML predictors alone and in comparison to several popular chemical shift predictors are discussed.

## Performance Evaluation Metric

As described in Chapter 2 equation 2.1, the performance evaluation metric we chose to evaluate the effectiveness of our regressor was the mean absolute error (MAE). To recall, the MAE is expressed by the equation

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left|y_j - \hat{y}_j\right|$$

Where $y_j$ is the chemical shift of the j<sup>th</sup> sample (atom) in the dataset and $\hat{y}_j$ is the predicted chemical shift of the j<sup>th</sup> sample (atom) in the dataset and *N* is the total number of carbon atoms in the dataset.

## The Initial $^{13}$C NMR Chemical Shift Dataset

There are a number of sizable, openly accessible NMR chemical shift libraries with experimentally determined $^{1}$H and $^{13}$C chemical shifts for small compounds, as was covered in chapters 1 and 2. Unfortunately, not all of these databases are very consistent in keeping track of or saving crucial experimental data like solvent, pH, sample temperature, or chemical shift reference compounds. As a result, the data collection and curation process was quite difficult for assembling a useful $^{13}C$ chemical shift database for training and testing.

Based on their quality and coverage, we chose to work with three public chemical shift databases: 1) the NP-MRD, 2) the BMRB, and 3) the HMDB. The Natural Products Magnetic Resonance Database (NP-MRD) (14) is a comprehensive, open-access electronic resource where the NMR data on natural products, metabolites, and other biologically derived substances can be deposited, distributed, searched for, and retrieved. Using the NP-MRD, we collected 346 experimental $^{13}C$ NMR spectra with fully assigned chemical shifts. TMS (Tetramethylsilane) reference was used to reference nearly all of the chemical shifts from the NP-MRD dataset. The Biological Magnetic Resonance Databank (BMRB) (11) has over 1000 assigned $^{1}H$ and $^{13}C$ chemical shifts at various spectrometer frequencies for small molecules. Although there were a few obvious assignment errors, we generally found that the experimental NMR data and $^{13}C$ chemical shift assignments in the BMRB were of very good quality, and that practically all chemical shifts were referenced to a single reference compound —TMS. The HMDB (Human Metabolome Database) (15), which contains 99 experimentally collected $^{13}C$ NMR spectra and assignments, was found to have high quality data and nearly all of the chemical shifts were referenced to DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid).

## The Training Dataset

After carefully reviewing and checking each of the chemical shift assignments for each of the molecules selected from the three databases, a final set of 318 non-redundant compounds corresponding to 4983 experimentally acquired $^{13}$C chemical shifts was assembled. The checking and reviewing process consisted of two steps: 1) manually confirming if the reported $^{13}$C chemical shifts were consistent with chemical shifts of identical or structurally similar compounds reported elsewhere and 2) manually confirming if the reported $^{13}$C chemical shifts were consistent with predicted chemical shifts generated by MNOVA (a commercial program). The details of this checking and review process are described later. Because the $^{13}$C chemical shifts were collected using both TMS and DSS as chemical shift references, we re-referenced all the reported $^{13}$C shifts to DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) by adjusting the reported shifts, as discussed in Wishart et al. (30, 31).

We then randomly split the entire dataset into two groups consisting of 80% of the compounds (the training dataset) and 20% of the compounds (the holdout dataset). We performed the split at the level of the molecule rather than at the level of the chemical shifts to prevent any data leakage (data leakage is described in chapter 2). After the split, the training dataset consisted of 253 molecules with complete 3D structures (with attached protons) and fully assigned, consistently referenced $^{13}$C chemical shifts. 151 of these molecules were obtained from the NP-MRD library corresponding to a total of 3318 experimentally measured $^{13}$C chemical shift values. Another 53 molecules were obtained from the BMRB, corresponding to a total of 300 experimentally measured $^{13}$C chemical shifts. The last set of 49 molecules was collected from the HMDB which contributed 278 experimentally measured $^{13}$C chemical shifts. Altogether our training dataset consisted of 3896 experimentally measured $^{13}$C chemical shift values from 253 chemically diverse molecules. These $^{13}$C molecules had an average molecular weight of 284, with the smallest molecule having a molecular weight of 46 Daltons and the largest having a molecular weight of 762 Daltons.

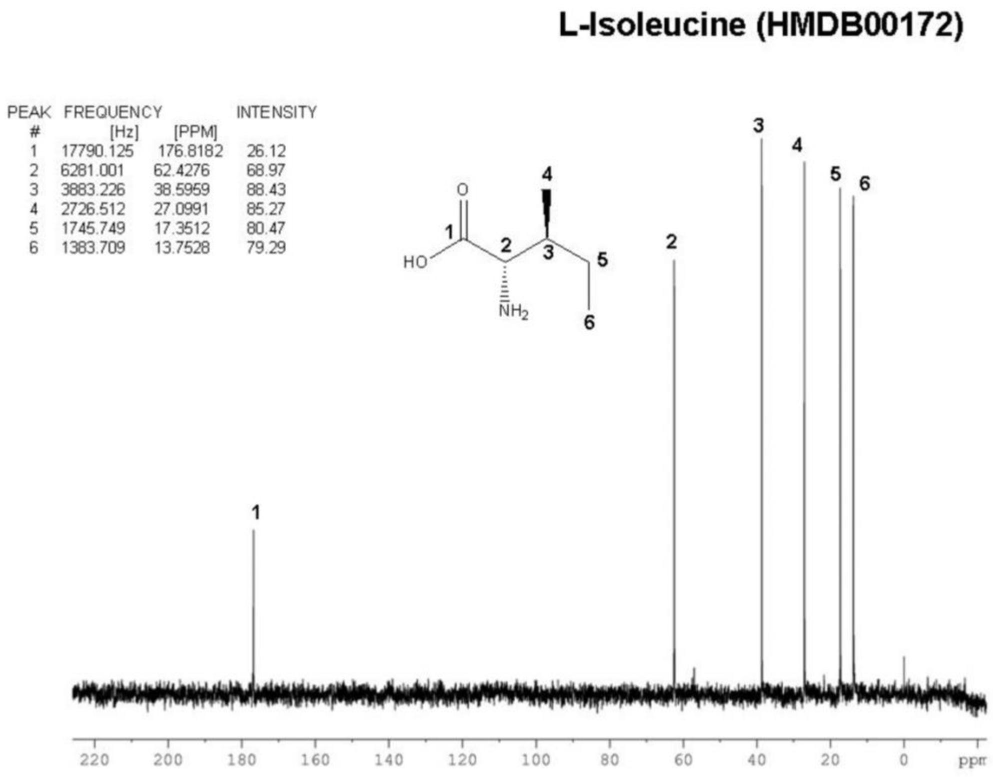
## The Holdout Dataset

A holdout dataset is crucial in order to properly assess how well a given ML model is performing. A holdout dataset is a dataset that the ML model hasn't seen before and is used to test the model's capacity for prediction. The holdout dataset, like the training dataset, must also be a "gold standard" dataset that includes the intended input and expected output. The first holdout dataset we used consisted of 65 molecules corresponding to a total of 1087 experimentally measured $^{13}$C chemical shifts. The average molecular weight of these 65 molecules was 299 Daltons, with the lowest molecular weight being 74 Daltons and the highest being 708 Daltons. We also created a second holdout dataset that consisted of 22 organic compounds that were chosen at random from the NP-MRD database. These 22 compounds had a total of 653 experimentally determined $^{13}$C chemical shifts. Their average molecular weight was 306 Daltons, with the lowest molecular weight being 224 Daltons and the highest molecular weight being 429 Daltons. As noted earlier, we re-referenced all the reported $^{13}$C shifts to DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) by adjusting the reported shifts, as discussed in Wishart et al. (30, 31).

## Atom and Chemical Shift Labeling

The inability to consistently identify which atoms are assigned to which $^{13}$C chemical shifts is a long-standing issue with chemical shift assignments. As a general rule, chemical shifts are typically supplied individually in a table with the matching atom labels from a picture of the structure, and the chemical shift assignments are typically displayed visually with atom labels marked on that structure image. In other words, the chemical picture offers a chemical shift "map" that connects heavy atoms with numbers or letters. Although this visual method of structural or chemical shift mapping is effective for people, computers cannot interpret it. To better appreciate the problem, the $^{13}$C chemical shift values for a compound called L-Isoleucine (HMDB0000172) are displayed in an example text file in Figure 3.1. This sort of display is common in many chemical shift databases. As can be seen in this Figure, a table lists a numbered collection of peaks together with the $^{13}$C chemical shifts that are assigned to each atom. Since there is currently no accepted standard atom-numbering scheme, the atom numbering varies greatly from one structure to another. We will use MarvinSketch (59, 60) to illustrate an example of the problems associated with atom labeling in NMR. As previously mentioned, MarvinSketch, like many other tools for sketching structure, begins the atom numbering for a specific molecule with the first atom that the user draws. Therefore, the carbon atom corresponding to the carbonyl group in Isoleucine will be numbered as atom #1 if someone begins drawing the chemical structure of Isoleucine from that atom. If someone else drew the same molecule starting with an oxygen atom, that oxygen atom would be numbered as atom #1. As a result, depending on how each user drew a particular molecule when using their preferred commercial software tool, the atom numbering changes. When working with the HMDB's collection of $^{13}$C chemical shift assignments we discovered (as predicted) that the molecular structures did not share the same pattern of numbering with the assignments in the HMDB library. This likely arose because the SDF files for the structure were not saved simultaneously with the original chemical shift numbering scheme, which is shown in Figure 3.2.

By manually mapping the original (PDF or PNG) image of the structure and its atom numbering scheme saved in the HMDB to the SDF structure files obtained from HMDB, we were able to fix these issues (as we previously did with the $^{1}H$ chemical shift data). The presence of inconsistent SDF structure files for certain compounds was another issue we discovered in the HMDB database. We were able to fix this issue by using the ALATIS software, as described in Chapter 2. We transformed all of the HMDB's structure files into three-dimensional SDF structure files using the consistent atom numbering generated by ALATIS. Each structure was then rotated along various axes using MarvinSketch from ChemAxon in order to match the original PNG or PDF structure image uploaded in the HMDB. We manually mapped the two different atom



| PEAK # | FREQUENCY [Hz] | [PPM] | INTENSITY |
|---|---|---|---|
| 1 | 17790.125 | 176.8182 | 26.12 |
| 2 | 6281.001 | 62.4276 | 68.97 |
| 3 | 3883.226 | 38.5959 | 88.43 |
| 4 | 2726.512 | 27.0991 | 85.27 |
| 5 | 1745.749 | 17.3512 | 80.47 |
| 6 | 1383.709 | 13.7528 | 79.29 |

Figure 3.1: A table of $^{13}C$ chemical shift assignments for L-Isoleucine (HMDB00172) in the HMDB. The numbers in the peak column refer to carbon atom positions drawn in same image.

numbering schemes for each molecule by comparing the images of the two molecules with each other. The chemical shift assignments were then manually modified. After using the approach, the

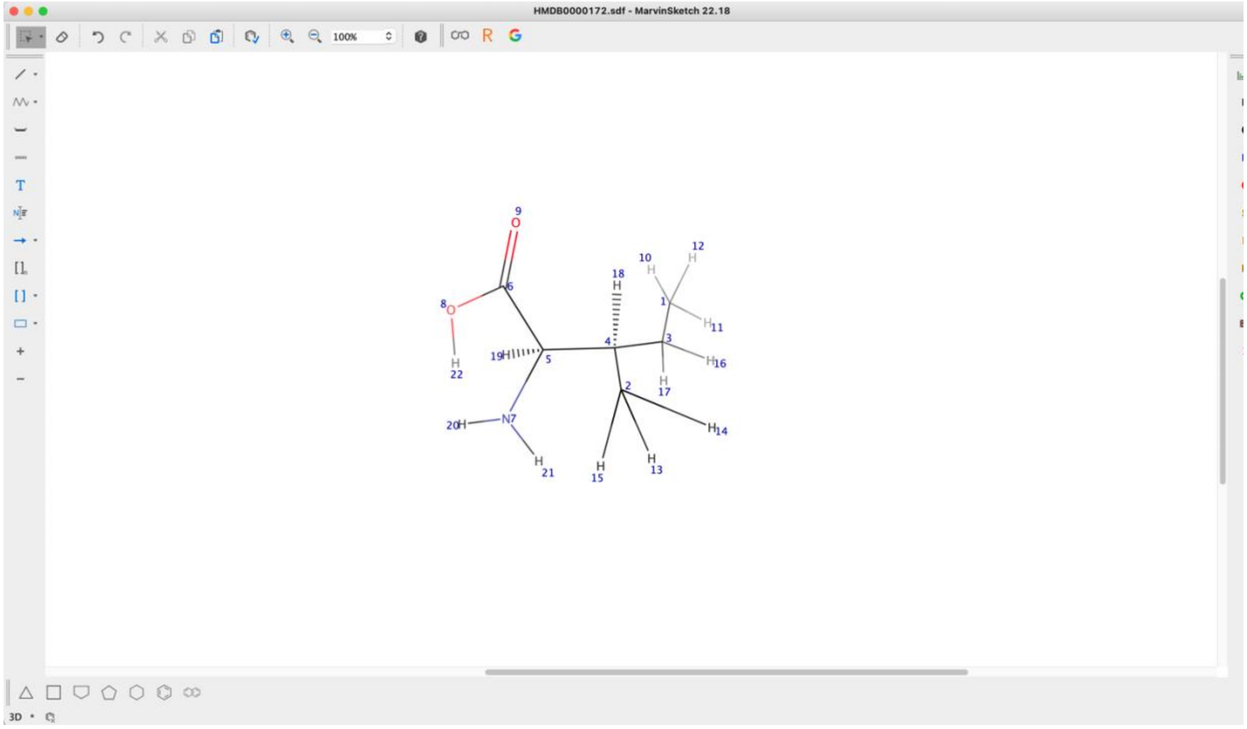


Figure 3.2: The structure of L-Isoleucine (HMDB00172) as downloaded from the HMDB server with atom numbering generated via MarvinSketch.

corrected chemical shift assignments were generated as displayed in Figure 3.3. This realignment and re-mapping procedure ensured that we could correctly and consistently calculate all the atomic features needed for our ML models. The atom remapping process took several weeks of manually intensive work. The identical process also had to be followed for the molecules and assignments collected from the BMRB. Fortunately, the NP-MRD provides well-structured NMR chemical shift data with uniform SDF structure and atom numbering. Thus we were able to use the NP-MRD dataset without having to undertake this laborious manual re-mapping.

We discovered several issues while carrying out this atom remapping process. For instance, we found that some compounds in the initial dataset were actually duplicates of each other. To identify these duplicates, we first used RDkit to transform all of the structures in our training and holdout datasets into InChI (International Chemical Identifier) strings. This allowed us to do some simple text string comparisons to identify and remove the duplicate molecules from those datasets. We also discovered several mistakes in the molecular SDF files. Sometimes the SDF file was not corresponding to the correct structure it had the chemical shift file for. A good illustration of this was Thioacetamide (bmse000781) (Figure 3.4) which are Tautomers, and Barbituric acid (bmse000346) (Figure 3.5). We excluded those compounds from the dataset if we were unable to determine the correct chemical shift values or the correct structure.

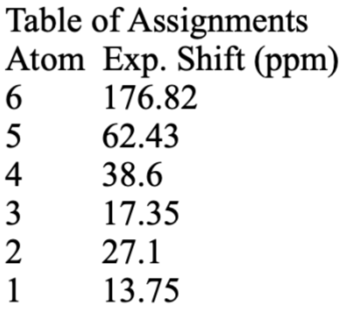


Table of Assignments

| Atom | Exp. Shift (ppm) |
|---|---|
| 6 | 176.82 |
| 5 | 62.43 |
| 4 | 38.6 |
| 3 | 17.35 |
| 2 | 27.1 |
| 1 | 13.75 |

Figure 3.3: The final chemical shift assignment file for L-Isoleucine (HMDB00172) using manual remapping. As shown here, one must begin by mapping the atoms from Figure 3.1 and Figure 3.2, then one must replace the Peak column in Figure 3.1 with the atomic positions from Figure 3.2. Atomic features for the ML model must therefore be calculated from the structure in Figure 3.2 since the structure from Figure 3.1 exists only as an image file, not as an SDF file.

After the chemical shift re-mapping and structure remediation process was finished, we applied a chemical shift “sanity” check to ensure that the reported $^{13}$C chemical shifts were reasonable. We employed two programs, NMRShiftDB2 (12) as well as the commercial program MNOVA (MestRe NovA) (63) to perform this sanity check. Both programs have reasonably accurate $^{13}$C chemical shift predictors. We used both programs to identify problematic assignments. For example, in our training dataset we found that the compounds D,L-Glyceraldehyde (bmse000225) and D-Ribulose 5-phosphate (bmse000278) had aldehyde $^{13}$C chemical shift assignments of 92.46 ppm and 103.82 ppm, respectively (Figure 3.6). According to the range of $^{13}$C chemical shift values in Figure 3.7, ketone and aldehyde carbons should have chemical shift values between 190 -220 ppm. We also calculated the $^{13}$C chemical shift values for the aldehydes for these two compounds using NMRShiftDB2. The predicted values with NMRShiftDB2 for both aldehyde carbon atoms came as 200.84 ppm. We also used MNOVA, which returned chemical shift values of 197.06 ppm and 198.98 ppm, respectively. Such a range of discrepancies led us to further analyze the $^{1}$H chemical shift assignments attached to that aldehyde

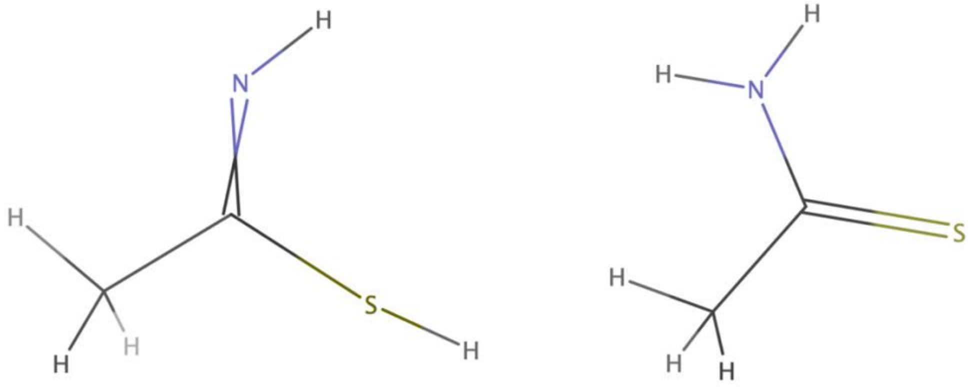


Figure 3.4: The SDF file for Thioacetamide (bmse000781) from the BMRB (left side) and the same structure from PubChem (right side).

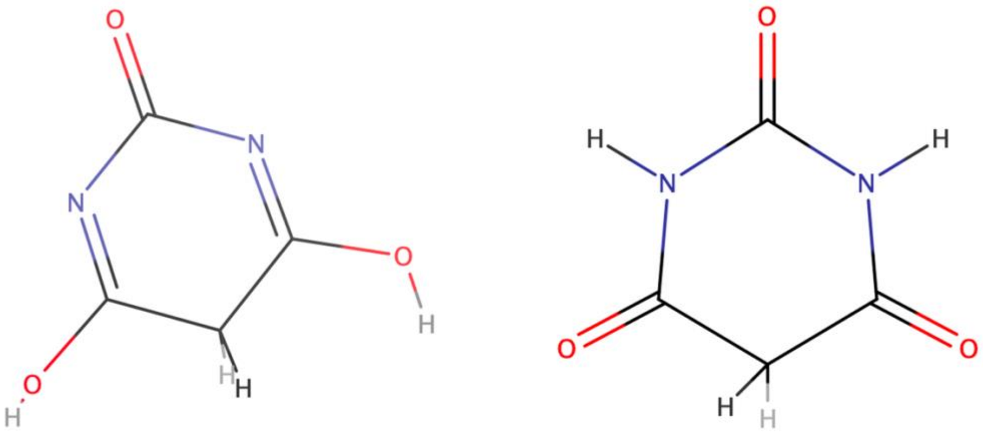
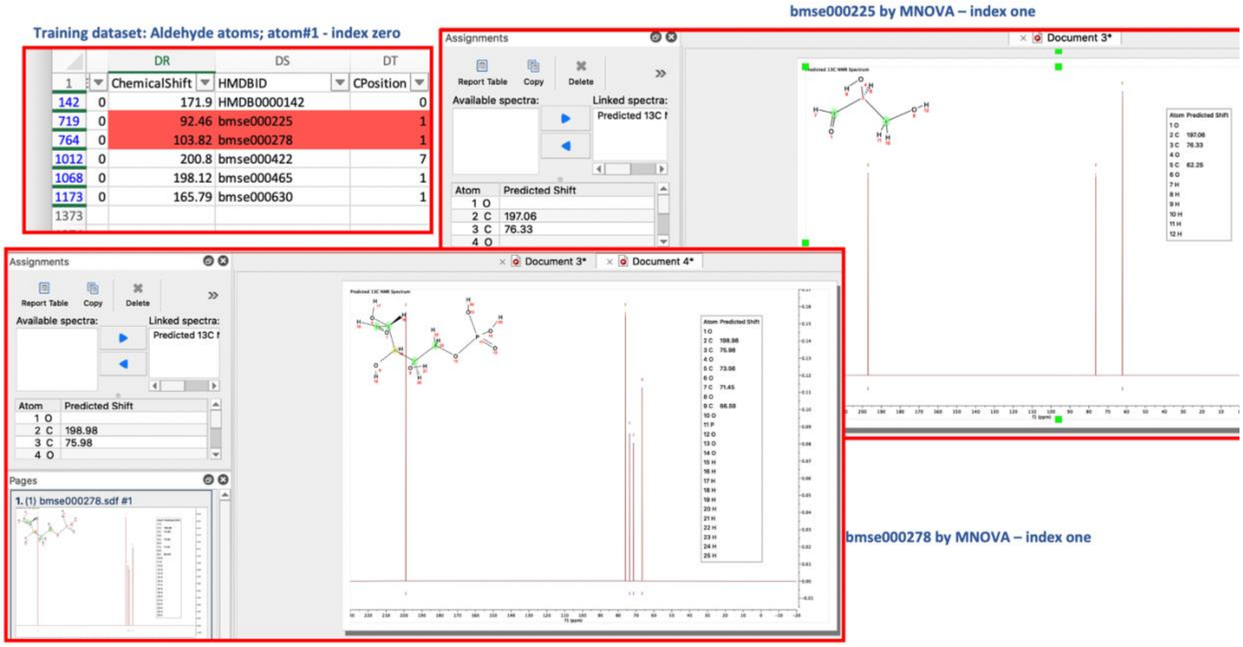


Figure 3.5: The SDF file for Barbituric acid (bmse000346) from the BMRB (left side) and the same structure from PubChem (right side).

carbon. From the two-dimensional $^{1}$H-$^{13}$C HSQC NMR spectrum in the BMRB, we found that the corresponding $^{1}$H aldehyde chemical shifts were given as 4.90 ppm and 5.20 ppm. As discussed in chapter 2, such upfield $^{1}$H chemical shifts for a presumptive aldehyde indicates that aldehyde (HC=O) reacted with water (Figure 2.10) and generated a diol HC(OH)2. In those cases where we



Training dataset: Aldehyde atoms; atom#1 - index zero

|  |  | DR | DS | DT |
|---|---|---|---|---|
| 1 |  | ChemicalShift | HMDBID | CPosition |
| 142 | 0 | 171.9 | HMDB0000142 | 0 |
| 719 | 0 | 92.46 | bmse000225 | 1 |
| 764 | 0 | 103.82 | bmse000278 | 1 |
| 1012 | 0 | 200.8 | bmse000422 | 7 |
| 1068 | 0 | 198.12 | bmse000465 | 1 |
| 1173 | 0 | 165.79 | bmse000630 | 1 |
| 1373 |  |  |  |  |

bmse000225 by MNOVA – index one

| Atom | Predicted Shift |
|---|---|
| 1 O |  |
| 2 C | 197.06 |
| 3 C | 76.33 |
| 4 O |  |
| 5 C | 62.26 |
| 6 O |  |
| 7 H |  |
| 8 H |  |
| 9 H |  |
| 10 H |  |
| 11 H |  |
| 12 H |  |

bmse000278 by MNOVA – index one

| Atom | Predicted Shift |
|---|---|
| 1 O |  |
| 2 C | 198.98 |
| 3 C | 75.98 |
| 4 O |  |
| 5 C | 73.56 |
| 6 O |  |
| 7 C | 71.45 |
| 8 O |  |
| 9 C | 66.59 |
| 10 O |  |
| 11 P |  |
| 12 O |  |
| 13 O |  |
| 14 O |  |
| 15 H |  |
| 16 H |  |
| 17 H |  |
| 18 H |  |
| 19 H |  |
| 20 H |  |
| 21 H |  |
| 22 H |  |
| 23 H |  |
| 24 H |  |
| 25 H |  |

Figure 3.6: The aldehyde carbon chemical shift values in D,L-Glyceraldehyde (bmse000225) and in D-Ribulose 5-phosphate (bmse000278) in the collected training dataset (top left corner) and the predicted chemical shift values by MNOVA for the same compound's aldehyde carbons.

were able to determine the chemical modifications leading to the chemical shift discrepancy, we created two structure (SDF) files and generated new assignment files, otherwise, we discarded these problematic molecules.

As a further check to avoid potential mis-assignments we compared the observed $^{13}$C assignments with the predicted $^{13}$C assignments. In particular, if the discrepancy between the NMRShiftDB2 predicted chemical shift values and the observed/reported shifts was >4.0 ppm for any carbon atom in a particular molecule, we manually rechecked those assignments and made the necessary modifications. We eliminated an entry if we were unable to explain the discrepancy. We also used information from the Reich $^{13}$C chemical shift database (64) to cross check the experimentally reported $^{13}$C NMR chemical shift values against those predicted based on their known positions within molecules.

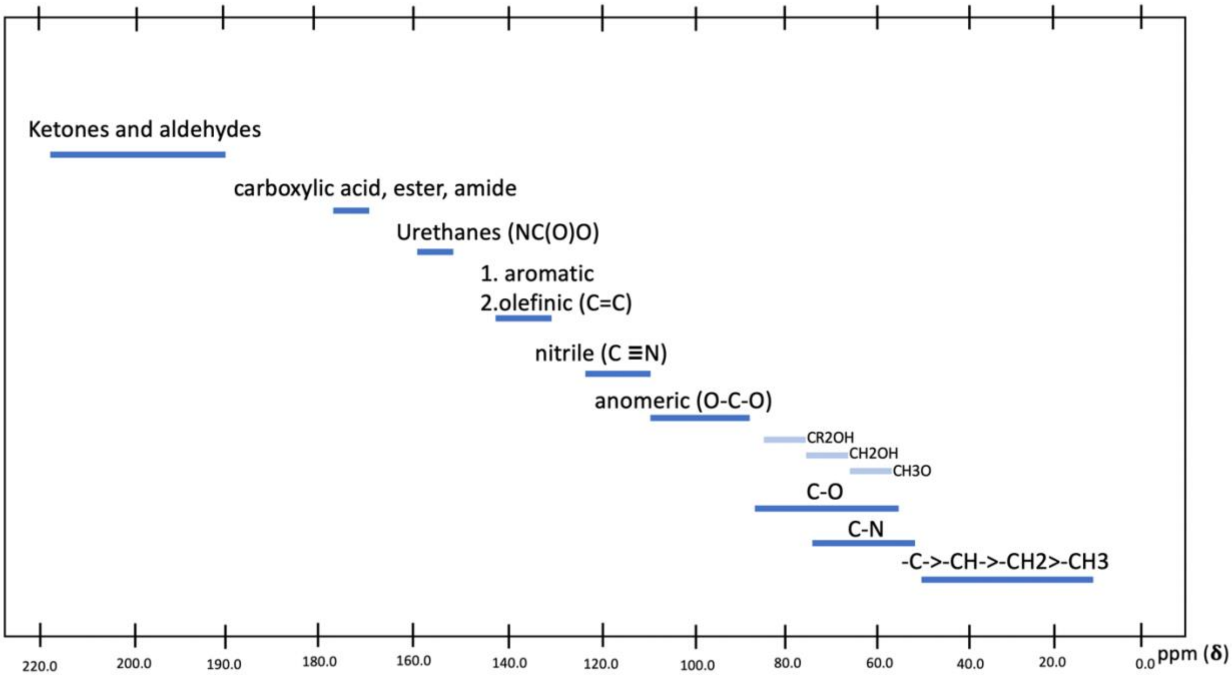


Figure 3. 7: $^{13}$C chemical shift ranges for various functional groups in organic compounds.

## Feature Identification

A feature in machine learning (ML) is a measurable attribute or quality of a phenomenon. When creating or training efficient ML algorithms for pattern recognition, classification, and regression, choosing informative, discriminating, and independent features is a key step. For our task of predicting $^{13}$C chemical shifts, it was crucial that we incorporate characteristics that have already been proven (by physicists and chemists) to have a significant influence on $^{13}$C chemical shifts.

Based on a survey of the $^{13}$C chemical shift literature and other resources, we found that the bond geometry around the carbon atom (i.e., the type of bond hybridization), the electronegativity of the other atoms bonded to the carbon of interest and other substituent effects were the key parameters that had the most influence over $^{13}$C chemical shift values (81).

The hybridization of carbon atoms is typically understood to mean "carbon geometry." In chemistry, bond hybridization involves combining two atomic orbitals to create a new type of bonding orbital called hybridized orbitals. The three types of hybridized bonds that carbon atoms can form are sp3, sp2, and sp (Figure 3.8 left side). The sp3 hybridization is characterized by a single bond (say to a hydrogen atom) leading to a tetrahedral configuration where the carbon atom is bonded to four other atoms as might be found in methane. The sp2 hybridization leads to a trigonal planar arrangement where the carbon atom is bonded to two sp3 (single bonds) and one double bond (sp2) as might be found in acetone, where the sp2 bond is between carbon and oxygen. When a carbon atom forms a triple bond (say with nitrogen), an sp bond is formed. This is a configuration seen with hydrogen cyanide, leading to a linear arrangement of atoms. The level or type of bond hybridization affects the bond strength, the bond lengths and the overall molecular geometry of carbon-containing molecules (82). In terms of chemical shift, the sp3 carbons are the most upfield (0–70 ppm), while the sp2 and sp3 carbons are further downfield at 100–150 ppm. Note that carbonyl-type sp2 carbons resonate at 160–220 ppm, while acetylene-type sp (triple bond) carbons resonate at 210–220 ppm (83). The effects on $^{13}$C chemical shifts are most pronounced for substituent modifications at the alpha (one bond away), beta (two bonds away), and gamma (three bonds away) positions associated with the carbon atom of interest (Figure 3.8 right side). The electronegativity of the bonded atom is the primary factor affecting the

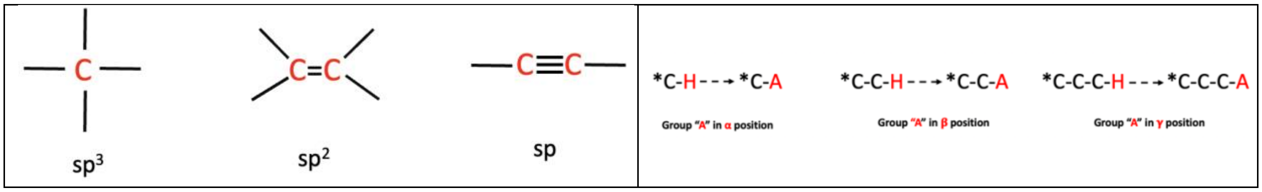


Figure 3.8: An illustration of carbon bond hybridization (left side) and the meaning of α, β and δ substituents (right side).

majority of α-substituent effects. The more electronegative the α-substituent is, the more downfield the chemical shift. For the atoms in the second row of the periodic table, the electronegativity rule functions reasonably well. However, there is a "heavy atom" effect that can override the electronegativity rule. In comparison to a carbon in an analogous saturated alkane, the $^{13}$C shifts of carbons connected to double bonds are altered comparatively little. While carbonyl substituents do induce considerable downfield shifts, other triple-bonded substituents, such as acetylene and nitrile groups, unexpectedly cause large upfield shifts. As a general rule, nearly all of the $^{13}$C chemical shifts arising from β-effects lead to downfield shifts, while upfield shifts arise from γ-effects (except for organometallic substituents) (84). Double and triple bonds exhibit the same α, β and γ-effects on carbon chemical shifts. In addition to the substituent/neighboring functional group effect, $^{13}$C chemical shifts are also influenced by proximity to conjugated ring systems (85).

Because of their significance in determining $^{13}$C chemical shifts, we made an effort to include as many of the aforementioned factors in our feature set as possible. The three-dimensional SDF files were utilised to construct a set of the relevant atomic characteristics for each molecule in our training and testing datasets. This was done using the Python-based cheminformatics package called RDKit (56). The atomic features we considered for each target carbon atom were: the number of not-H substituents connected to the carbon; the carbon atom's bond hybridization state; it’s Gasteiger charge; whether the carbon is part of an aromatic system or a conjugated system; the size of the ring (if part of an aromatic system); the total number of H atoms attached to that carbon; it’s electronegativity; and the atomic numbers of the attached atoms. All the atomic features were calculated using RDKit expect for the electronegativity property (which is not supported by RDKit). The electronegativity property was calculated using CDKit, a Java-based cheminformatics software package developed by Steinbeck et. al. (70).

An atom’s chemical shift can also be affected by the presence of different functional groups, so we also included 71 types of chemical functional groups or their particular chemical substructures (Table 3.1 shows the list of the chemical functional groups and particular chemical substructures encoded as SMART strings) to annotate the target carbon atom’s molecular neighborhood. These functional groups and chemical substructures were identified from literature reviews, other online resources and by analyzing the outliers while training our models. The molecular neighborhood included functional groups up to two bonds away (α and β effects). To describe this molecular neighborhood property, we first determined whether the carbon atom of interest belongs to a particular functional group (which we called the zero bond neighborhood), then we determined how many times each functional group was present one-bond away and finally how many times each functional group was present two-bonds away (Figure: 3.9). This two-bond neighborhood was determined after some trial-and-error assessment on the influence of different functional group effects from zero, one, two and three bonds away. The feature set for the targeted carbon atom (12 features) was listed first, followed by the 55 features for the zero bond neighborhood, followed by the 71 features of the one bond neighborhood, and finally the 71 features of the two bond neighborhood (Figure 3.10). Therefore, our feature set included 12 atomic features and 55 + (71 X 2) = 197 neighborhood descriptors. In all cases the features were numeric.

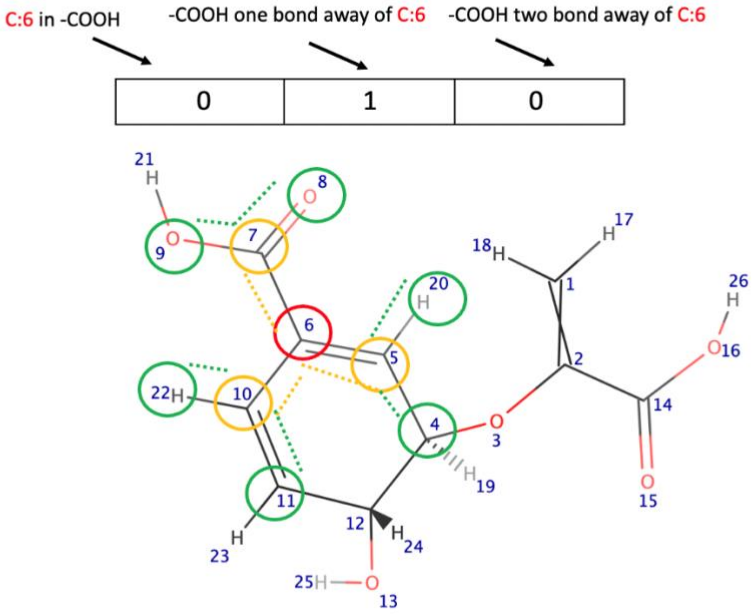
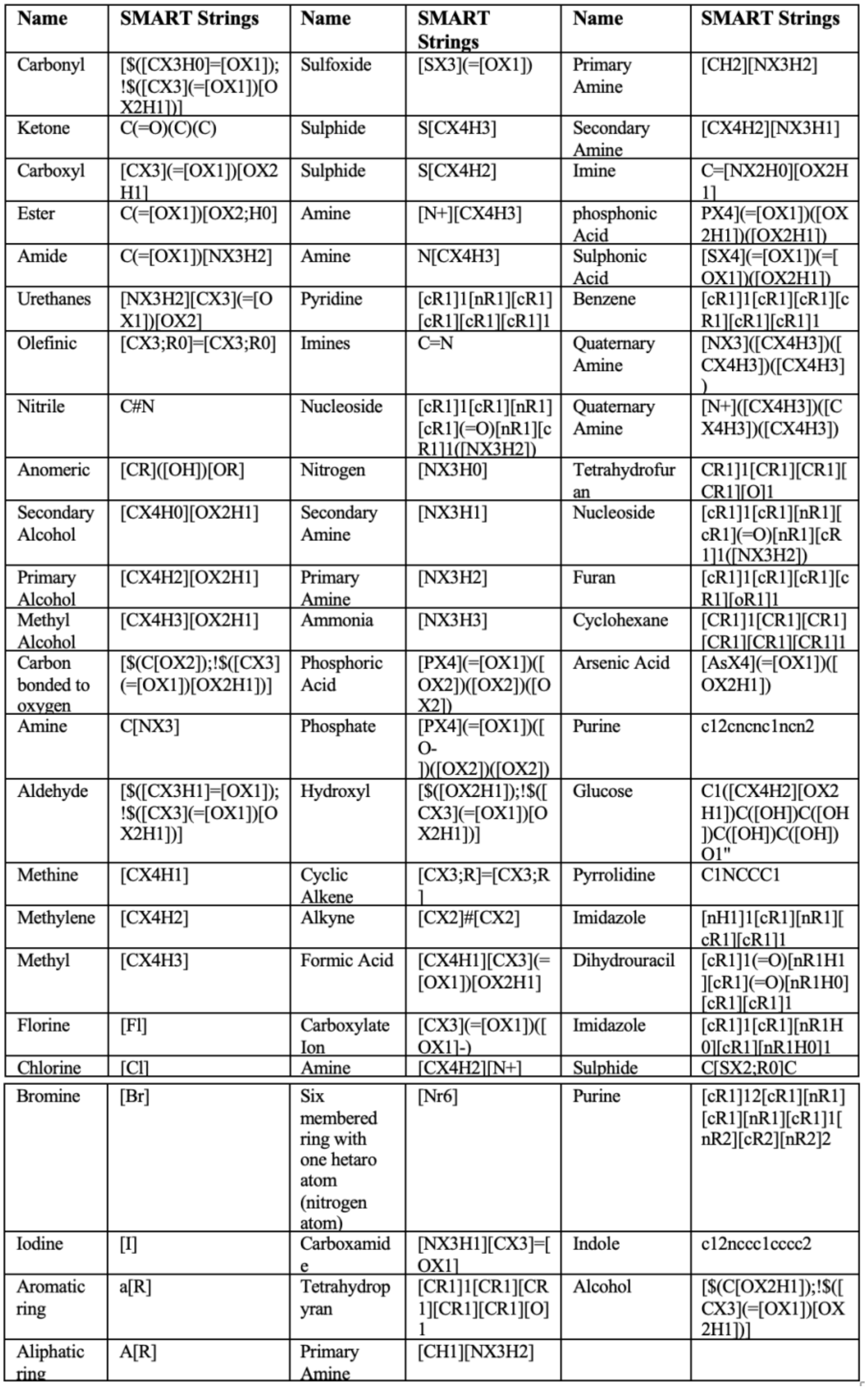


Figure 3.9: Chemical structure of chorismic acid (bmse000075). If the targeted carbon atom # is 6 (C:6), the zero, one and two bond neighborhood atoms are circled with red, yellow and green color. As an example, C:6 is not a carboxylic (-COOH) carbon. One bond away from C:6, the carboxylic acid appears 1 time and two bonds away, it appears 0 times.



| Name | SMART Strings | Name | SMART Strings | Name | SMART Strings |
|---|---|---|---|---|---|
| Carbonyl | [$([CX3H0]=[OX1]);!$([CX3](=[OX1])[OX2H1])] | Sulfoxide | [SX3](=[OX1]) | Primary Amine | [CH2][NX3H2] |
| Ketone | C(=O)(C)(C) | Sulphide | S[CX4H3] | Secondary Amine | [CX4H2][NX3H1] |
| Carboxyl | [CX3](=[OX1])[OX2H1] | Sulphide | S[CX4H2] | Imine | C=[NX2H0][OX2H1] |
| Ester | C(=[OX1])[OX2;H0] | Amine | [N+][CX4H3] | phosphonic Acid | PX4](=[OX1])([OX2H1])([OX2H1]) |
| Amide | C(=[OX1])[NX3H2] | Amine | N[CX4H3] | Sulphonic Acid | [SX4](=[OX1])(=[OX1])([OX2H1]) |
| Urethanes | [NX3H2][CX3](=[OX1])[OX2] | Pyridine | [cR1]1[nR1][cR1][cR1][cR1][cR1]1 | Benzene | [cR1]1[cR1][cR1][cR1][cR1][cR1]1 |
| Olefinic | [CX3;R0]=[CX3;R0] | Imines | C=N | Quaternary Amine | [NX3]([CX4H3])([CX4H3])([CX4H3]) |
| Nitrile | C#N | Nucleoside | [cR1]1[cR1][nR1][cR1](=O)[nR1][cR1]1([NX3H2]) | Quaternary Amine | [N+]([CX4H3])([CX4H3])([CX4H3]) |
| Anomeric | [CR]([OH])[OR] | Nitrogen | [NX3H0] | Tetrahydrofuran | CR1]1[CR1][CR1][CR1][O]1 |
| Secondary Alcohol | [CX4H0][OX2H1] | Secondary Amine | [NX3H1] | Nucleoside | [cR1]1[cR1][nR1][cR1](=O)[nR1][cR1]1([NX3H2]) |
| Primary Alcohol | [CX4H2][OX2H1] | Primary Amine | [NX3H2] | Furan | [cR1]1[cR1][cR1][cR1][oR1]1 |
| Methyl Alcohol | [CX4H3][OX2H1] | Ammonia | [NX3H3] | Cyclohexane | [CR1]1[CR1][CR1][CR1][CR1][CR1]1 |
| Carbon bonded to oxygen | [$(C[OX2]);!$([CX3](=[OX1])[OX2H1])] | Phosphoric Acid | [PX4](=[OX1])([OX2])([OX2])([OX2]) | Arsenic Acid | [AsX4](=[OX1])([OX2H1]) |
| Amine | C[NX3] | Phosphate | [PX4](=[OX1])([O-])([OX2])([OX2]) | Purine | c12cncnc1ncn2 |
| Aldehyde | [$([CX3H1]=[OX1]);!$([CX3](=[OX1])[OX2H1])] | Hydroxyl | [$([OX2H1]);!$([CX3](=[OX1])[OX2H1])] | Glucose | C1([CX4H2][OX2H1])C([OH])C([OH])C([OH])C([OH])O1" |
| Methine | [CX4H1] | Cyclic Alkene | [CX3;R]=[CX3;R] | Pyrrolidine | C1NCCC1 |
| Methylene | [CX4H2] | Alkyne | [CX2]#[CX2] | Imidazole | [nH1]1[cR1][nR1][cR1][cR1]1 |
| Methyl | [CX4H3] | Formic Acid | [CX4H1][CX3](=[OX1])[OX2H1] | Dihydrouracil | [cR1]1(=O)[nR1H1][cR1](=O)[nR1H0][cR1][cR1]1 |
| Florine | [Fl] | Carboxylate Ion | [CX3](=[OX1])([OX1]-) | Imidazole | [cR1]1[cR1][nR1H0][cR1][nR1H0]1 |
| Chlorine | [Cl] | Amine | [CX4H2][N+] | Sulphide | C[SX2;R0]C |
| Bromine | [Br] | Six membered ring with one hetaro atom (nitrogen atom) | [Nr6] | Purine | [cR1]12[cR1][nR1][cR1][nR1][cR1]1[nR2][cR2][nR2]2 |
| Iodine | [I] | Carboxamide | [NX3H1][CX3]=[OX1] | Indole | c12nccc1cccc2 |
| Aromatic ring | a[R] | Tetrahydropyran | [CR1]1[CR1][CR1][CR1][CR1][O]1 | Alcohol | [$(C[OX2H1]);!$([CX3](=[OX1])[OX2H1])] |
| Aliphatic ring | A[R] | Primary Amine | [CH1][NX3H2] | | |

Table 3.1: The list of the chemical functional groups and particular chemical substructures considered in the different molecular neighborhood written out as SMART strings.

## Methodology

We trained and tested four different regression algorithms: a Support Vector Regressor (SVR), a Random Forest Regressor (RFR), an Extreme Gradient Boosting Regressor (XGBoostRegressor or XGBR), and a Gradient Boosting Regressor (GBR) in order to determine which machine learning algorithm would produce the best $^{13}$C chemical shift prediction results.

The training and validation process was performed using standard cross-validation methods with training and testing (hold-out) datasets. For each of the four regressor algorithms we tuned the hyperparameters using 5-fold cross validation over the training dataset. Once the best hyperparameters for the different models were found, the models were trained on the entire training dataset of 3896 experimentally measured $^{13}$C chemical shift values from the collection of 253 molecules. Then the trained models were used to predict the $^{13}$C chemical shift values for the holdout dataset, which consisted of 1087 $^{13}$C chemical shift values from 65 molecules. Finally, based on the mean absolute error (MAE) calculated from the holdout dataset, the best model was chosen. We further assessed our final model by comparing its $^{13}$C chemical shift prediction performance against the results of two well-known $^{13}$C chemical shift predictors (that use machine learning), namely MNOVA and NMRshiftDB2. We also compared our predictor against $^{13}$C chemical shifts calculated using quantum mechanical (density functional theory or DFT) methods. This was done using a second holdout dataset from the $^{1}$H chemical shift prediction experiment, which was composed of 652 $^{13}$C experimental chemical shifts from 22 molecules from the NP-MRD database. A high-level explanation of each of the regressor algorithms used in our study of $^{13}$C chemical shift prediction was provided in the previous chapter.

# Coding Details

All of the ML models discussed here were created and tested using Python and a range of Python libraries, including the scikit-learn package. In particular, Python version "3.7.0" was used for most input and output coding. The Python scikit-learn library, version "1.0.2," was used for all machine learning applications. Additionally, the Python-based RDKit (versions "2020.09" and "2022.03.5") was employed for a number of cheminformatics function, such as converting SMILES strings into 3D structures and matching SMILES strings to find distinct functional groups within a target molecule. As noted earlier, we calculated the electronegativity feature for our chemical shift calculations using the Java-based CDKit. To make things easier, we built a separate ".jar" file in Java that the Python program could use to get the value of this important atomic descriptor. The University of Alberta's (UofA) Cybera server was used for all of our model development (training, testing, assessing, etc.). The system used an Ubuntu 20.04 computer running the Linux 5.4.0-131generic kernel. The server contained a total of 32 GB of RAM and 8 Intel Xeon E5-2630 v3 CPUs running at 2.40 GHz.

As noted in Chapter 2, "hyperparameter optimization" or "tuning" is a process of choosing the ideal collection of parameters for a specific ML model and a specific dataset. Each of the algorithms we investigated had a variety of hyperparameters. Using sckit-learn has the advantage that most hyperparameters can be handled with the default settings. However, throughout the course of work, we also learned that the number of trees (n estimators) and the tree's maximum depth (max depth) are the most important parameters for optimizing tree-based algorithms. Similarly, the most crucial hyperparameters for SVMs are the regularisation parameter (C) and the choice of the proper kernel. We discovered that the "rbf" kernel would frequently cause the model to overfit, so we stayed with the linear kernel and tried a range of C values to determine the optimal SVM hyperparameters.

Cross-validation (CV) is a very important step in every ML training process. When evaluating the effectiveness of ML models, CV assessment is extremely helpful. It enables more efficient use of training data for operations such as hyperparameter adjustment without running the risk of data leakage (a circumstance in which the model has access to information that it otherwise shouldn't normally have). In the CV process, a test/hold-out set is initially created from the entire data set, which to be used for the model's final evaluation. This hold-out selection must be done before implementing CV. The remaining data, or everything but the hold-out set, is divided into K number of folds (subsets). After that, the CV process involves repeatedly evaluating each the folds, using one of the K folds as the validation set and the remaining folds as the training set, in each iteration. Each fold is utilised as a validation set over the entire K fold repetition process. The procedure for selecting the best model using a 5-fold CV process appears in Figure 3.10. We can gain a more realistic idea of how well our model might perform on data that it has never seen before by training and validating the model K times on various subsets of the same training data. To better understand the model's performance in a K-fold CV, for each combination of hyperparameters, the process scores the model after each iteration, then computes the average over all iterations.

The entire dataset (4983 samples from 319 molecules) in our experiment was divided into approximately 80% (training) and 20% (testing). The split was made based on the molecules rather than on samples to prevent any data leakage. Specifically, 80% (3896 samples with 253 molecules) of the entire dataset served as the training set and the 20% (1087 samples with 65 molecules) of

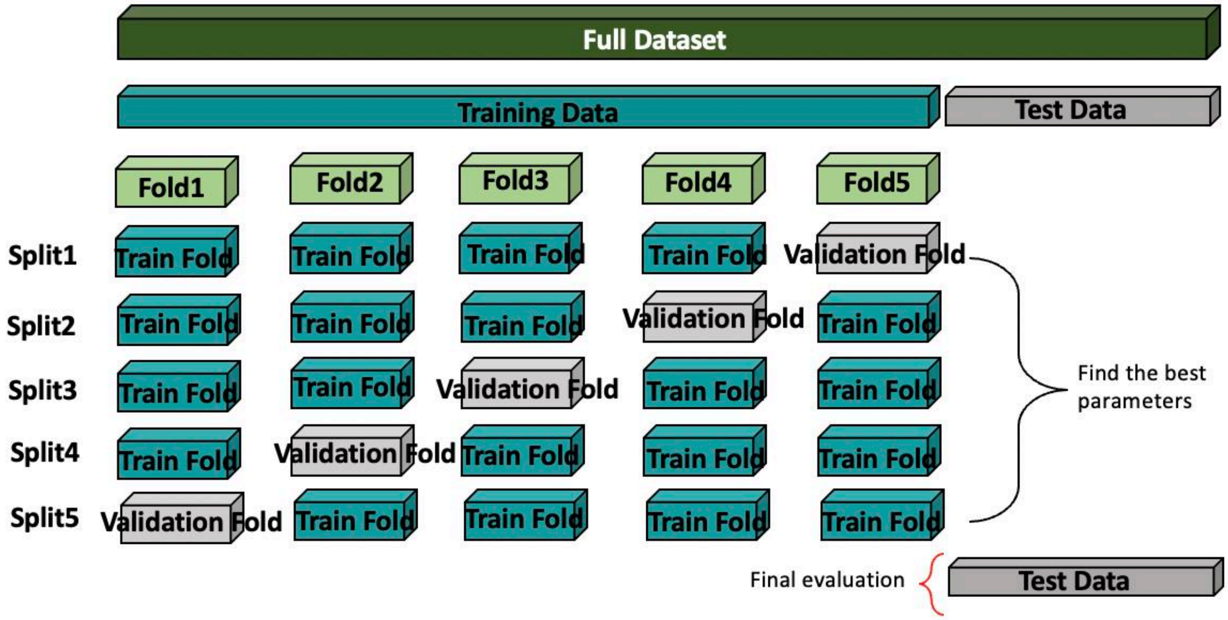


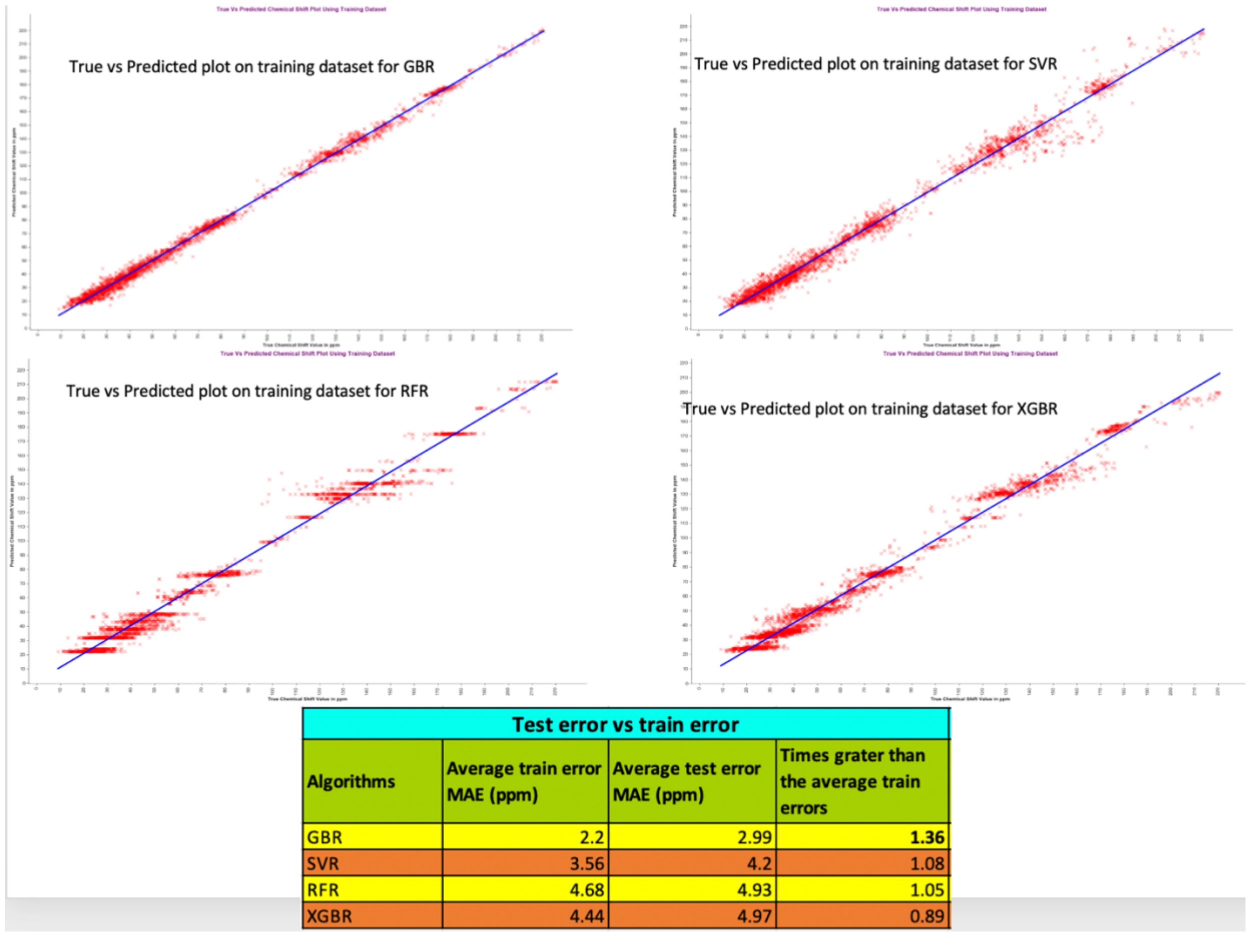
Figure 3.10: A schematic diagram illustration the process of finding the best parameter for various ML models using 5-fold cross validation and selecting the best model using the test/holdout data.

the entire dataset was put aside as a test/holdout set. The holdout set was never used in model training. The training dataset was divided into k = 5 folds and used for hype parameter tuning for the four different regression algorithms. In the k = 5 loop, the number of the training samples was 3269, 3042, 3079, 3088 and 3106 respectively and the number of the validation samples was 627, 854, 817, 808 and 790, respectively.

## Results and Discussion

A series of performance tests were carried out using the training dataset for each of the four regressors as implemented on the UofA Cybera cluster. The average MAE for the test set (using the best hyperparameters for each model) was 2.99 pm for the Gradient Boost Regressor (GBR), 4.20 ppm for the Support Vector Regressor (SVR), 4.93 ppm for the Random Forest Regressor (RFR), and 4.97 ppm for the Extreme Gradient Boosting Regressor (XGBR). These MAEs were determined during the 5-fold cross validation with the training dataset. We also measured the Pearson correlation index or coefficient of determination ($R^2$) between observed and predicted $^{13}C$ chemical shifts for all 4 ML models. The correlation index is determined by calculating the variation in the predictions that the dataset can explain. An $R^2$ of 1 indicates that the model is perfect, while an $R^2$ of zero, indicates the model is no better than a random guess. The average $R^2$ score for GBR was 0.992, for SVR was 0.985 while for the other two regressors it was 0.981. In addition to calculating the average test errors, we checked the average training errors to ensure that none of the models was overfitting. The best-fitting models had average training errors of 2.20 ppm, 3.56 ppm, 4.68 ppm, and 4.44 ppm for the GBR, SVR, RFR and XGBR models, respectively. When we compared the training error and the test error, we found that the average test errors for the GBR was 1.36 times greater than the average training error. Meanwhile the average test error for the SVR was 1.18 times greater than the average training error, and the average test errors for the RFR and XGBR were 1.05 and 1.12 times greater than their average training error, respectively. All four algorithms appeared to have a respectable test to training error ratios, indicating that the models were not overfitting. However, relative to RFR, XGBR and SVR, it appears that the GBR model had the highest ratio which may indicate a slightly more overfitted model. We also analyzed the true vs predicted plot on the training errors for each of the algorithms (Figure 3.11). Although the training and test error gaps for the RFR, XGBR, and SVR models were smaller than those in GBR, it is obvious from these plots that the true vs. predicted values did not exactly match up with the best-fit line for these three predictors. However, the alignment is noticeably better for GBR. Overall, we can conclude that the GBR model performed better than RFR, XGBR, and SVR. It also appears that the SVR's learning quality was generally respectable.

Once the models with the best hyperparameters were found from the 5-fold cross validation process, each of the four models were trained with the entire training dataset (3896 samples). The



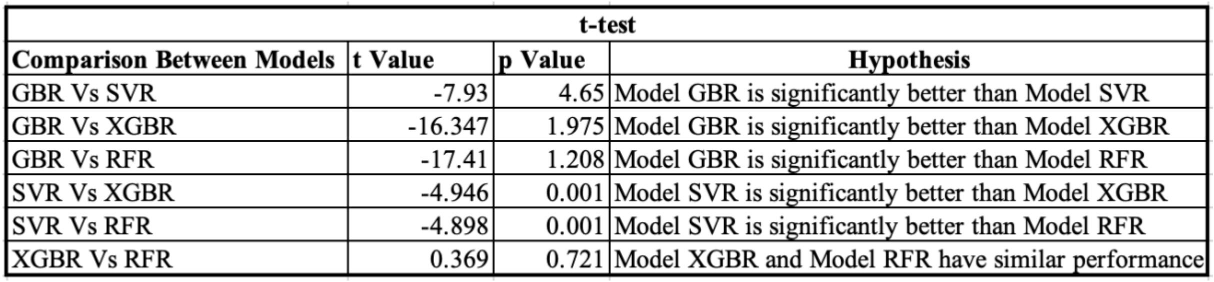
| Test error vs train error | | | |
|---|---|---|---|
| Algorithms | Average train error MAE (ppm) | Average test error MAE (ppm) | Times grater than the average train errors |
| GBR | 2.2 | 2.99 | **1.36** |
| SVR | 3.56 | 4.2 | 1.08 |
| RFR | 4.68 | 4.93 | 1.05 |
| XGBR | 4.44 | 4.97 | 0.89 |

Figure 3.11: Scatter plots comparing the true vs. predicted values on the training dataset for the four algorithms. The GBR and SVR models appear to be the most well learned models. The chart below the graph images, shows test and training error ratios for each model.

trained models were then evaluated against the holdout dataset of 1087 $^{13}$C chemical shifts to select the winning model. In this final model evaluation step, it was found that the GBR performed the best relative to all other models. In particular, the GBR produced an MAE of 2.94 ppm with a standard deviation of 4.20 ppm. The $R^2$ score for this best model was 0.993. The SVR had an MAE of 3.93 ppm with a standard deviation of 5.82 ppm and an $R^2$ score 0.986 while the RFR had the MAE of 4.91 ppm with a standard deviation of 6.67 ppm and an $R^2$ score 0.982. The XGBR was the worst and it produced an MAE of 4.86 ppm with a standard deviation of 6.68 ppm and an $R^2$ of 0.982. The performance evaluation of these four methods is summarised in Figure 3.12. As can be observed, the GBR's MAE of 2.94 ppm (using previously unseen data) outperforms all other algorithms we examined. SVR was the 2nd best model while the XGBR and RFR were the 3rd and 4th best performers, respectively.

To further assess the performance of these predictors, we again used a second holdout dataset (another group of previously unseen data). This 2nd set of holdout dataset was composed of 652 $^{13}$C experimentally measured chemical shifts from 22 molecules found in the NP-MRD database. In this set of unseen data, the GBR model again showed the best performance with an MAE of 6.29 ppm, a standard deviation of 8.65 ppm and an $R^2$ of 0.971. XGBR had the 2nd best performance with an MAE of 8.06 ppm, a standard deviation of. 11.03 ppm and an $R^2$ of 0.956. The worst performer on the 2nd holdout dataset was the SVR model which had an MAE of 9.51 ppm with a standard deviation of 13.58 ppm and an $R^2$ of just 0.936. The RFR achieved an MAE of 8.13 ppm with a standard deviation of 10.94 ppm and an $R^2$ of 0.952. Overall, the GBR remained as the best predictor even though its performance dropped significantly on the 2nd holdout dataset. The reasons for this significant drop in performance are discussed later in this chapter. A comparison of the performance of all four algorithms on the 1st and 2nd holdout datasets is shown in the Figure 3.13.

Furthermore, we performed a paired t-test to see if there was a significant difference between these models (Table 3.2). Based on the t-test result, the GBR model is better than the SVR and XGBR and RFR models. The next performer is SVR while XGBR and RFR performed similarly.

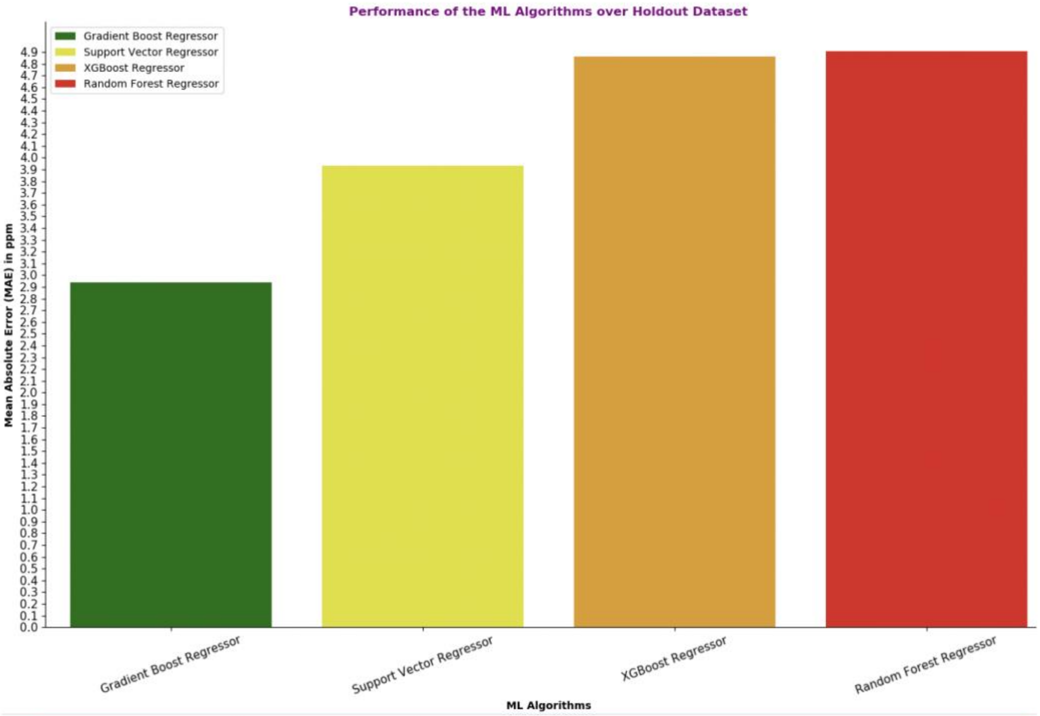


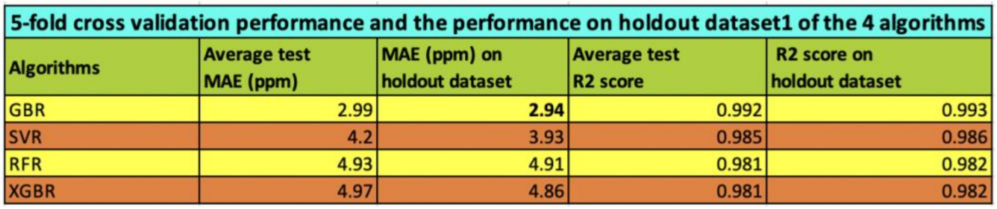
| t-test | | | |
|---|---|---|---|
| **Comparison Between Models** | **t Value** | **p Value** | **Hypothesis** |
| GBR Vs SVR | -7.93 | 4.65 | Model GBR is significantly better than Model SVR |
| GBR Vs XGBR | -16.347 | 1.975 | Model GBR is significantly better than Model XGBR |
| GBR Vs RFR | -17.41 | 1.208 | Model GBR is significantly better than Model RFR |
| SVR Vs XGBR | -4.946 | 0.001 | Model SVR is significantly better than Model XGBR |
| SVR Vs RFR | -4.898 | 0.001 | Model SVR is significantly better than Model RFR |
| XGBR Vs RFR | 0.369 | 0.721 | Model XGBR and Model RFR have similar performance |

Table 3. 2: Paired t-test result among GBR, SVR, XGBR and RFR models

Based on the GBR's performance on the two holdout datasets, we decided that it was the best ML model to use as the "final" predictor for our NMRPred $^{13}C$ chemical shift prediction algorithm.

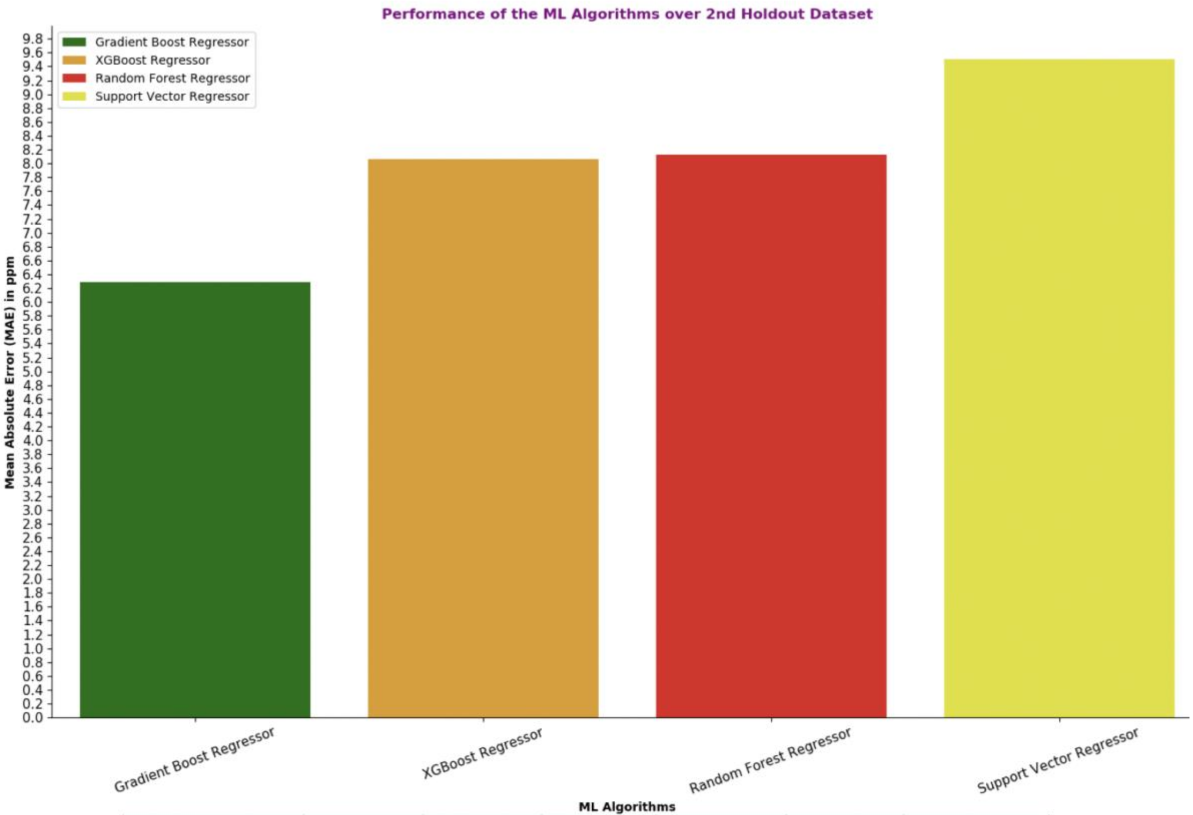
NMRPred-Carbon is the name we chose for the complete NMR chemical shift predictor package. NMRPred-Carbon accepts SMILES data, converts the SMILES string to a 3D SDF file with atomic coordinates and calculates the $^{13}C$ chemical shifts for all $^{13}C$ atoms in the molecule NMRPred-Carbon is an all-inclusive NMR chemical shift prediction toolkit. As part of its calculation process, NMRPred-Carbon calculates the atomic feature sets using RDKit and CDKit, and then uses the Gradient Boost Regressor (GBR) to calculate the $^{13}C$ chemical shifts. The utility tool on the NP-MRD website known as "$^{13}$C NMRPred-Carbonictor" currently includes NMRPred-Carbon. A public version of NMRPred-Carbon for $^{13}$C shift prediction will be available on GitHub with a README file in the link https://github.com/zsayeeda/NMR_Prediction.git. A sample input for the $^{13}$C Chemical Shift Predictor ($^{13}$C NMRPred-Carbonictor"), on NMRPred-Carbon, is shown in Figure 3.14. The input is the SMILES formula (NCCC(=O)O) corresponding to a compound

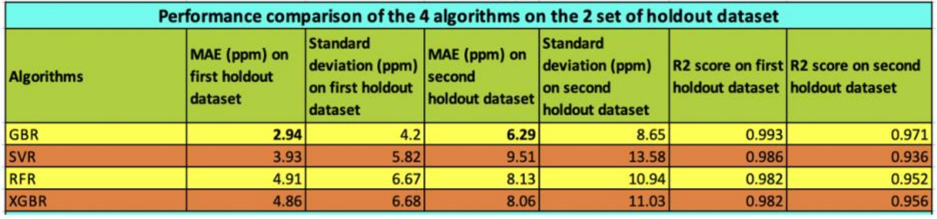



5-fold cross validation performance and the performance on holdout dataset1 of the 4 algorithms

| Algorithms | Average test MAE (ppm) | MAE (ppm) on holdout dataset | Average test R2 score | R2 score on holdout dataset |
|---|---|---|---|---|
| GBR | 2.99 | **2.94** | 0.992 | 0.993 |
| SVR | 4.2 | 3.93 | 0.985 | 0.986 |
| RFR | 4.93 | 4.91 | 0.981 | 0.982 |
| XGBR | 4.97 | 4.86 | 0.981 | 0.982 |

Figure 3.12: A bar graph showing the performance of each of the 4 ML models for the holdout dataset. The Gradient Boost Regressor, which is represented by the green bar, shows the best performance. The chart below the graph shows performance of the four models for both the 5-fold cross validation and on the holdout dataset.





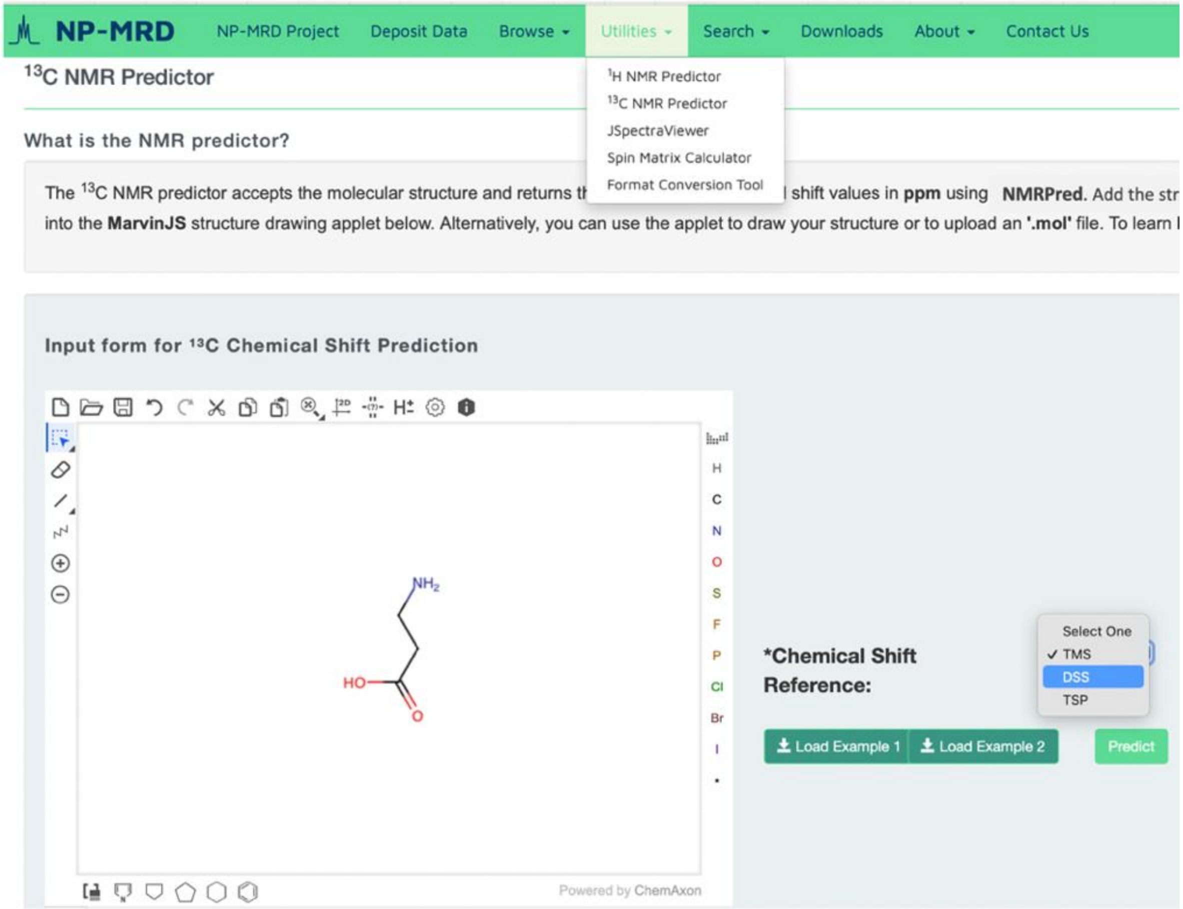
Performance comparison of the 4 algorithms on the 2 set of holdout dataset

| Algorithms | MAE (ppm) on first holdout dataset | Standard deviation (ppm) on first holdout dataset | MAE (ppm) on second holdout dataset | Standard deviation (ppm) on second holdout dataset | R2 score on first holdout dataset | R2 score on second holdout dataset |
|---|---|---|---|---|---|---|
| GBR | **2.94** | 4.2 | **6.29** | 8.65 | 0.993 | 0.971 |
| SVR | 3.93 | 5.82 | 9.51 | 13.58 | 0.986 | 0.936 |
| RFR | 4.91 | 6.67 | 8.13 | 10.94 | 0.982 | 0.952 |
| XGBR | 4.86 | 6.68 | 8.06 | 11.03 | 0.982 | 0.956 |

Figure 3.13: A bar graph showing the performance of the four ML models on the 2nd holdout dataset. The Gradient Boost Regressor, indicated by the green bar, shows the best performance. The performance of the other three algorithms changed slightly compared to the results shown in Figure 3.13. The chart below the graph shows the performance comparison of the four models between the two holdout datasets.

known as beta-Alanine (HMDB0000056), which has been pasted in the input field. ChemAxon's JChem program is used to depict the structure, which does so in a conventional 2D style. Users must choose the chemical reference substance from the pull-down menus displayed under

"Chemical Shift Reference". NMRPred-Carbon first predicts all carbon chemical shifts using DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) as a reference. The predictor then modifies the chemical shift values in accordance with the chosen chemical shift reference (i.e., either DSS, TMS, or TSP), based on the discussions in (30, 31). The generated 3D structure of beta-Alanine and NMRPred-Carbon's predicted $^{13}C$ chemical shift values for beta-Alanine are shown in Figure 3.15.



$^{13}C$ NMR Predictor

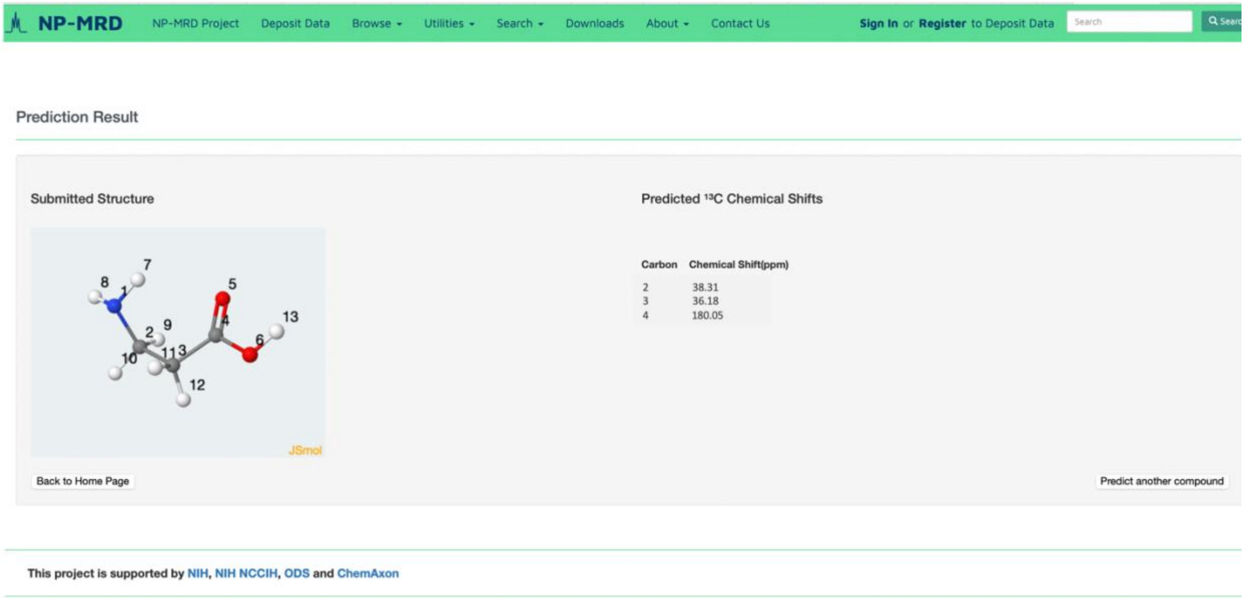
What is the NMR predictor?

The $^{13}C$ NMR predictor accepts the molecular structure and returns th shift values in **ppm** using **NMRPred**. Add the str into the **MarvinJS** structure drawing applet below. Alternatively, you can use the applet to draw your structure or to upload an **'.mol'** file. To learn l

Input form for $^{13}C$ Chemical Shift Prediction

*Chemical Shift Reference: DSS

Figure 3.14: The SMILES string for beta-Alanine (HMDB0000056) was provided in the input together with a selected chemical shift reference option. ChemAxon's JChem converts the SMILES string into a 2D structure.



Prediction Result

Submitted Structure

Predicted $^{13}C$ Chemical Shifts

| Carbon | Chemical Shift(ppm) |
|---|---|
| 2 | 38.31 |
| 3 | 36.18 |
| 4 | 180.05 |

This project is supported by NIH, NIH NCCIH, ODS and ChemAxon

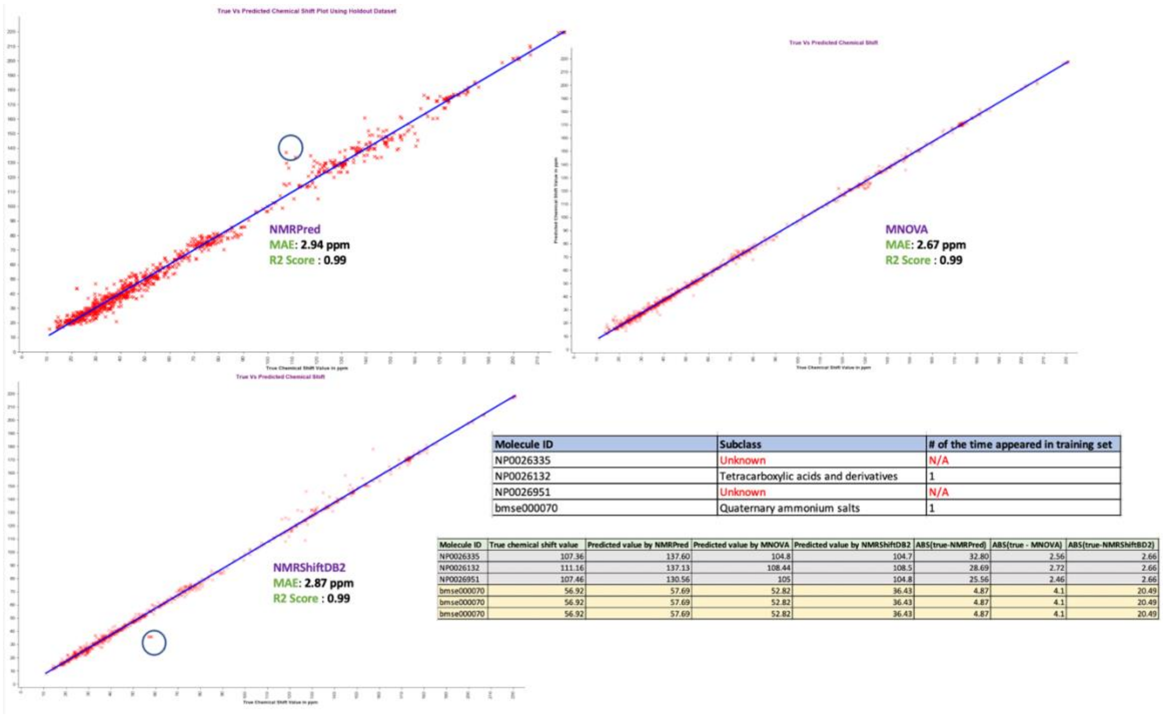
Figure 3.15: The predicted $^{13}C$ chemical shift values from the input in Figure 3.15 together with the 3D structure of the molecule.

## Performance Comparison Against Popular Methods

We evaluated the performance of NMRPred-Carbon in comparison to a several well-known $^{13}C$ chemical shift predictors, including MNOVA (63), NMRShiftDB2 (12) and DFT based calculations (75) performed by NWChem. With the exception of the DFT-based technique for the first holdout dataset, we compared these predictors to the observed $^{13}C$ chemical shift values for both holdout datasets. We found that even though NMRPred-Carbon did not perform better than MNOVA or the NMRShiftDB2 predictors, it performed nearly as well as them for the first holdout dataset. In particular, the first holdout dataset NMRPred-Carbon had an MAE of 2.94 ppm while NMRShiftDB2 had an MAE of 2.87 ppm and MNOVA had an MAE of 2.67 ppm. As seen by these results, MNOVA performed the best of the three algorithms.

Scatter plots were generated to show the observed $^{13}C$ chemical shift values vs. the predicted $^{13}C$ chemical shift values for the holdout dataset for NMRPred-Carbon, MNOVA and NMRShiftDB2 in Figure 3.16. The $R^2$ was identical for all 3 models (0.99). Looking more closely at the scatter plot for NMRPred-Carbon, we see that it had three major outliers for Neodactyloquinone (NP0026335), 1-Homoacevaltrate (NP0026132) and 1beta,16:15,16-diepoxy-cis-ent-cleroda-12,14-dien-18alpha,6alpha-olide (NP0026951). Among these three outliers, for NP0026335 and NP0026951, the chemical subclasses were unknown and one NP0026132 was from the chemical subclass of "Tetracarboxylic acids and derivatives". We looked into the training dataset that this class appeared only one time. The difference between true chemical shift value and the predicted chemical shift value provided by NMRPred-Carbon for NP0026335, NP0026132 and NP0026132 were 30.24 ppm, 25.97 ppm and 23.10 ppm, respectively. Whereas for MNOVA the reported differences were just 2.56 ppm, 2.72 ppm, and 2.46 ppm, respectively. For NMRShiftDB2 the reported differences were 2.66 ppm for all the three cases. NMRPred-Carbon was not alone in generating outliers. From the scattered plot for NMRShiftDB2, we found three major outliers which were not present in NMRPred-Carbon nor in MNOVA. Two of these outliers belonged to Betaine Aldehyde (bmse000070). NMRShiftDB2 predicted these chemical shifts to be 36.43 ppm which was 20.49 ppm higher than the true chemical shift values. On the other hand, the difference between the true chemical shift values and the predicted chemical shift values was 4.1 ppm and 4.87 ppm for MNOVA and NMRPred-Carbon, respectively.

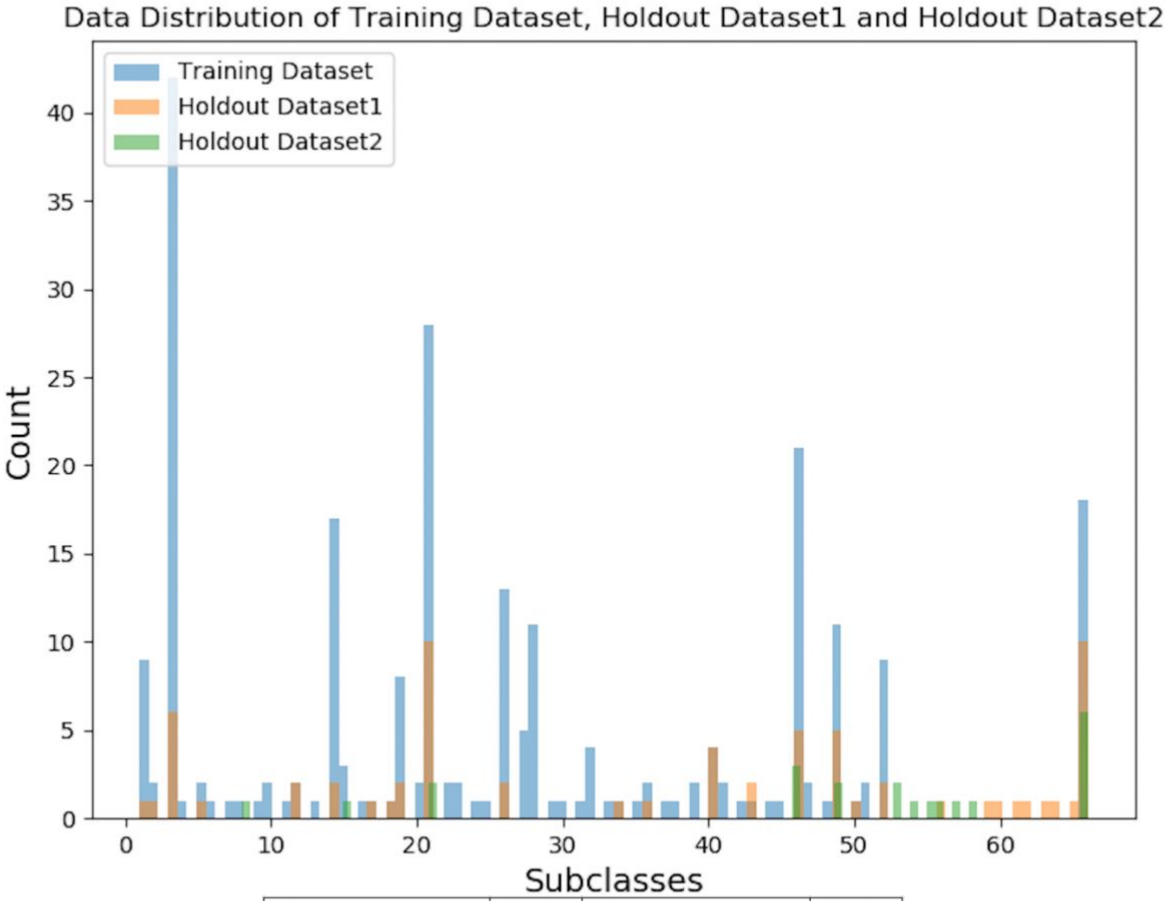
To get a better understanding behind these outliers, we measured the dataset distribution by assessing each compound's chemical subclasses using ClassyFire (76). Figure 3.17 shows the subclasses distribution between the training dataset and the holdout dataset. The last bar in the graph indicates the total number of the compounds for which chemical subclasses were unknown or for which ClassyFire could not detect. Moreover, we compared each compound's prediction performances shown in Figure 3.18 and Figure 3.19. Even though the data distribution for the predicted compounds compared to the training set distribution was not particularly well matched, we observed that for 83% of the compounds, NMRPred-Carbon's predicted values were better than any of the other predictors whereas for 10% of the molecules it performed the worst. Given these factors, it seems that NMRPred-Carbon actually performed quite well on the 1$^{st}$ holdout dataset. As mentioned earlier, to further test the performance NMRPred-Carbon we also evaluated it against

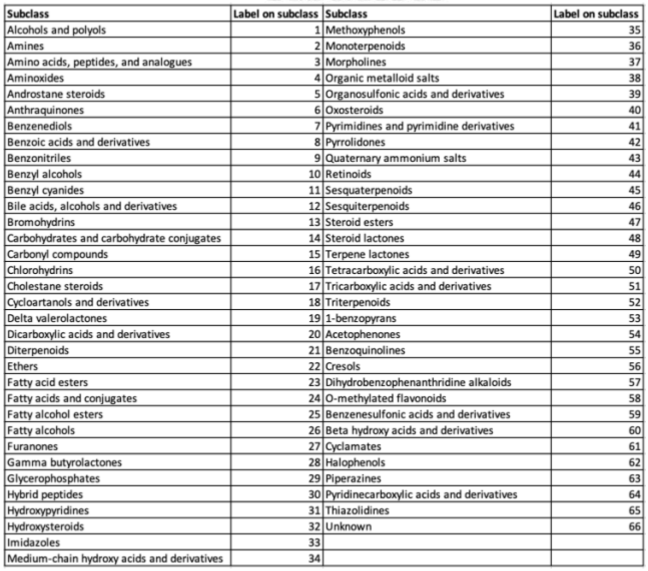


| Molecule ID | Subclass | # of the time appeared in training set |
|---|---|---|
| NP0026335 | Unknown | N/A |
| NP0026132 | Tetracarboxylic acids and derivatives | 1 |
| NP0026951 | Unknown | N/A |
| bmse000070 | Quaternary ammonium salts | 1 |

| Molecule ID | True chemical shift value | Predicted value by NMRPred | Predicted value by MNOVA | Predicted value by NMRShiftDB2 | ABS(true-NMRPred) | ABS(true - MNOVA) | ABS(true-NMRShiftBD2) |
|---|---|---|---|---|---|---|---|
| NP0026335 | 107.36 | 137.60 | 104.8 | 104.7 | 32.80 | 2.56 | 2.66 |
| NP0026132 | 111.16 | 137.13 | 108.44 | 108.5 | 28.69 | 2.72 | 2.66 |
| NP0026951 | 107.46 | 130.56 | 105 | 104.8 | 25.56 | 2.46 | 2.66 |
| bmse000070 | 56.92 | 57.69 | 52.82 | 36.43 | 4.87 | 4.1 | 20.49 |
| bmse000070 | 56.92 | 57.69 | 52.82 | 36.43 | 4.87 | 4.1 | 20.49 |
| bmse000070 | 56.92 | 57.69 | 52.82 | 36.43 | 4.87 | 4.1 | 20.49 |

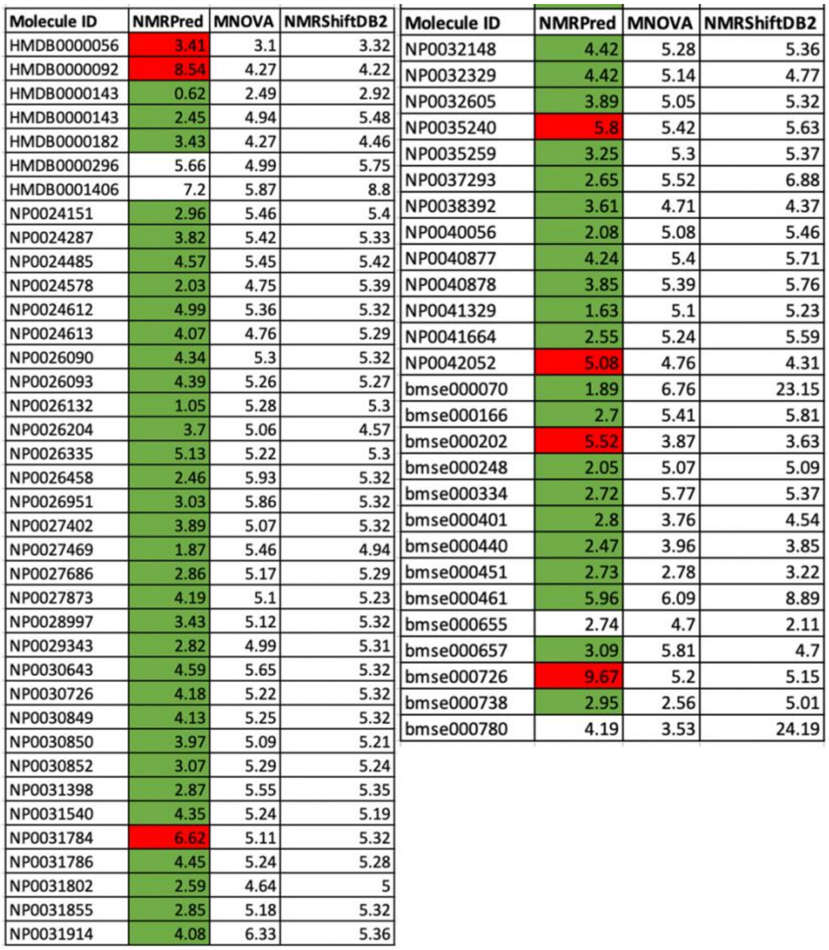
Figure 3.16: Scatter plots showing the experimentally measured $^{13}C$ chemical shifts versus predicted $^{13}C$ chemical shifts for NMRPred-Carbon (top left corner), MNOVA (top right corner) and NMRShiftDB2 (lower left corner), respectively. The upper chart at the lower right corner shows the chemical subclasses for the compounds where the outliers were identified. The lower chart at the lower right corner shows the experimentally measured and predicted chemical $^{13}C$ shift values for the outliers by NMRPred-Carbon, MNOVA and NMRShiftDB2.





| Subclass | Label on subclass | Subclass | Label on subclass |
|---|---|---|---|
| Alcohols and polyols | 1 | Methoxyphenols | 35 |
| Amines | 2 | Monoterpenoids | 36 |
| Amino acids, peptides, and analogues | 3 | Morpholines | 37 |
| Aminoxides | 4 | Organic metalloid salts | 38 |
| Androstane steroids | 5 | Organosulfonic acids and derivatives | 39 |
| Anthraquinones | 6 | Oxosteroids | 40 |
| Benzenediols | 7 | Pyrimidines and pyrimidine derivatives | 41 |
| Benzoic acids and derivatives | 8 | Pyrrolidones | 42 |
| Benzonitriles | 9 | Quaternary ammonium salts | 43 |
| Benzyl alcohols | 10 | Retinoids | 44 |
| Benzyl cyanides | 11 | Sesquaterpenoids | 45 |
| Bile acids, alcohols and derivatives | 12 | Sesquiterpenoids | 46 |
| Bromohydrins | 13 | Steroid esters | 47 |
| Carbohydrates and carbohydrate conjugates | 14 | Steroid lactones | 48 |
| Carbonyl compounds | 15 | Terpene lactones | 49 |
| Chlorohydrins | 16 | Tetracarboxylic acids and derivatives | 50 |
| Cholestane steroids | 17 | Tricarboxylic acids and derivatives | 51 |
| Cycloartanols and derivatives | 18 | Triterpenoids | 52 |
| Delta valerolactones | 19 | 1-benzopyrans | 53 |
| Dicarboxylic acids and derivatives | 20 | Acetophenones | 54 |
| Diterpenoids | 21 | Benzoquinolines | 55 |
| Ethers | 22 | Cresols | 56 |
| Fatty acid esters | 23 | Dihydrobenzophenanthridine alkaloids | 57 |
| Fatty acids and conjugates | 24 | O-methylated flavonoids | 58 |
| Fatty alcohol esters | 25 | Benzenesulfonic acids and derivatives | 59 |
| Fatty alcohols | 26 | Beta hydroxy acids and derivatives | 60 |
| Furanones | 27 | Cyclamates | 61 |
| Gamma butyrolactones | 28 | Halophenols | 62 |
| Glycerophosphates | 29 | Piperazines | 63 |
| Hybrid peptides | 30 | Pyridinecarboxylic acids and derivatives | 64 |
| Hydroxypyridines | 31 | Thiazolidines | 65 |
| Hydroxysteroids | 32 | Unknown | 66 |
| Imidazoles | 33 | | |
| Medium-chain hydroxy acids and derivatives | 34 | | |

Figure 3.17: The subclass distribution graph of chemicals seen in the two holdout datasets compared to the NMRPred-Carbon training dataset. The chart at the bottom shows how the subclasses are given a numbered label.



| Molecule ID | NMRPred | MNOVA | NMRShiftDB2 |
|---|---|---|---|
| HMDB0000056 | 3.41 | 3.1 | 3.32 |
| HMDB0000092 | 8.54 | 4.27 | 4.22 |
| HMDB0000143 | 0.62 | 2.49 | 2.92 |
| HMDB0000143 | 2.45 | 4.94 | 5.48 |
| HMDB0000182 | 3.43 | 4.27 | 4.46 |
| HMDB0000296 | 5.66 | 4.99 | 5.75 |
| HMDB0001406 | 7.2 | 5.87 | 8.8 |
| NP0024151 | 2.96 | 5.46 | 5.4 |
| NP0024287 | 3.82 | 5.42 | 5.33 |
| NP0024485 | 4.57 | 5.45 | 5.42 |
| NP0024578 | 2.03 | 4.75 | 5.39 |
| NP0024612 | 4.99 | 5.36 | 5.32 |
| NP0024613 | 4.07 | 4.76 | 5.29 |
| NP0026090 | 4.34 | 5.3 | 5.32 |
| NP0026093 | 4.39 | 5.26 | 5.27 |
| NP0026132 | 1.05 | 5.28 | 5.3 |
| NP0026204 | 3.7 | 5.06 | 4.57 |
| NP0026335 | 5.13 | 5.22 | 5.3 |
| NP0026458 | 2.46 | 5.93 | 5.32 |
| NP0026951 | 3.03 | 5.86 | 5.32 |
| NP0027402 | 3.89 | 5.07 | 5.32 |
| NP0027469 | 1.87 | 5.46 | 4.94 |
| NP0027686 | 2.86 | 5.17 | 5.29 |
| NP0027873 | 4.19 | 5.1 | 5.23 |
| NP0028997 | 3.43 | 5.12 | 5.32 |
| NP0029343 | 2.82 | 4.99 | 5.31 |
| NP0030643 | 4.59 | 5.65 | 5.32 |
| NP0030726 | 4.18 | 5.22 | 5.32 |
| NP0030849 | 4.13 | 5.25 | 5.32 |
| NP0030850 | 3.97 | 5.09 | 5.21 |
| NP0030852 | 3.07 | 5.29 | 5.24 |
| NP0031398 | 2.87 | 5.55 | 5.35 |
| NP0031540 | 4.35 | 5.24 | 5.19 |
| NP0031784 | 6.62 | 5.11 | 5.32 |
| NP0031786 | 4.45 | 5.24 | 5.28 |
| NP0031802 | 2.59 | 4.64 | 5 |
| NP0031855 | 2.85 | 5.18 | 5.32 |
| NP0031914 | 4.08 | 6.33 | 5.36 |
| NP0032148 | 4.42 | 5.28 | 5.36 |
| NP0032329 | 4.42 | 5.14 | 4.77 |
| NP0032605 | 3.89 | 5.05 | 5.32 |
| NP0035240 | 5.8 | 5.42 | 5.63 |
| NP0035259 | 3.25 | 5.3 | 5.37 |
| NP0037293 | 2.65 | 5.52 | 6.88 |
| NP0038392 | 3.61 | 4.71 | 4.37 |
| NP0040056 | 2.08 | 5.08 | 5.46 |
| NP0040877 | 4.24 | 5.4 | 5.71 |
| NP0040878 | 3.85 | 5.39 | 5.76 |
| NP0041329 | 1.63 | 5.1 | 5.23 |
| NP0041664 | 2.55 | 5.24 | 5.59 |
| NP0042052 | 5.08 | 4.76 | 4.31 |
| bmse000070 | 1.89 | 6.76 | 23.15 |
| bmse000166 | 2.7 | 5.41 | 5.81 |
| bmse000202 | 5.52 | 3.87 | 3.63 |
| bmse000248 | 2.05 | 5.07 | 5.09 |
| bmse000334 | 2.72 | 5.77 | 5.37 |
| bmse000401 | 2.8 | 3.76 | 4.54 |
| bmse000440 | 2.47 | 3.96 | 3.85 |
| bmse000451 | 2.73 | 2.78 | 3.22 |
| bmse000461 | 5.96 | 6.09 | 8.89 |
| bmse000655 | 2.74 | 4.7 | 2.11 |
| bmse000657 | 3.09 | 5.81 | 4.7 |
| bmse000726 | 9.67 | 5.2 | 5.15 |
| bmse000738 | 2.95 | 2.56 | 5.01 |
| bmse000780 | 4.19 | 3.53 | 24.19 |

Figure 3.18: The prediction performance of individual compounds as generated by NMRPred-Carbon, MNOVA and NMRShiftDB2. The green color and the red color indicate the best and the worst predictions by NMRPred-Carbon, respectively.

a second holdout dataset. This second holdout set consisted of 652 $^{13}$C experimental chemical shifts from 22 molecules from the NP-MRD database, which was also used in the $^{1}$H chemical shift prediction experiment (see Chapter 2). On this second holdout dataset, NMRPred-Carbon was assessed, and the MAE was found to be 6.29 ppm. This was significantly greater than the first holdout dataset's 2.94 ppm MAE. Additionally, NMRPred-Carbon performed significantly worse than MNOVA (MAE = 2.87 ppm), DFT (MAE = 3.08 ppm), and NMRShiftDB2 (MAE = 4.02 ppm) on this second holdout set. Overall, MNOVA had the best results for this second holdout dataset. Figure 2.20 shows the performance comparison between the 1st and 2nd holdout datasets for NMRPred-Carbon, MNOVA, NMRShiftDB2 and the DFT predictor. The large decline in NMRPred-Carbon's performance in comparison to other chemical shift predictors indicates that either NMRPred-Carbon was overtrained or that it was undertrained. In our earlier analyses of the "Training-error," overtraining was deemed to be moderate or unlikely. Given NMRPred-Carbon's relatively small training set, we hypothesised that under-training was more likely the issue and that NMRPred-Carbon's poor performance on the second holdout set was caused by the fact that it had not encountered (or been trained on) many of the chemical structure classes present in the second holdout dataset. To test this hypothesis, we again used ClassyFire (76) to quantitatively assess the chemical structure subclasses seen in the training dataset and the 2nd holdout dataset (Figure 3.17). For ~7% of the training dataset, ~ 28% of the first holdout dataset, and ~45% of the second holdout dataset, ClassyFire was unable to identify the chemical subclasses. ClassyFire indicated that compounds from 52 different chemical subclasses were included in our initial training sample. Even though only 50/65 molecules (~77%%) in the first holdout dataset had structures that corresponded to at least one of the training dataset's chemical categories, our predictor nevertheless outperformed all other predictors. On the other hand, for the second holdout dataset (with 22 molecules from the NP-MRD) only 9/22 (~41%) of the molecules belonged to chemical subclasses found in the original training dataset. Figure 3.21 shows a summary of this chemical class distribution. While the overlap of previously seen structure classes is not profoundly different

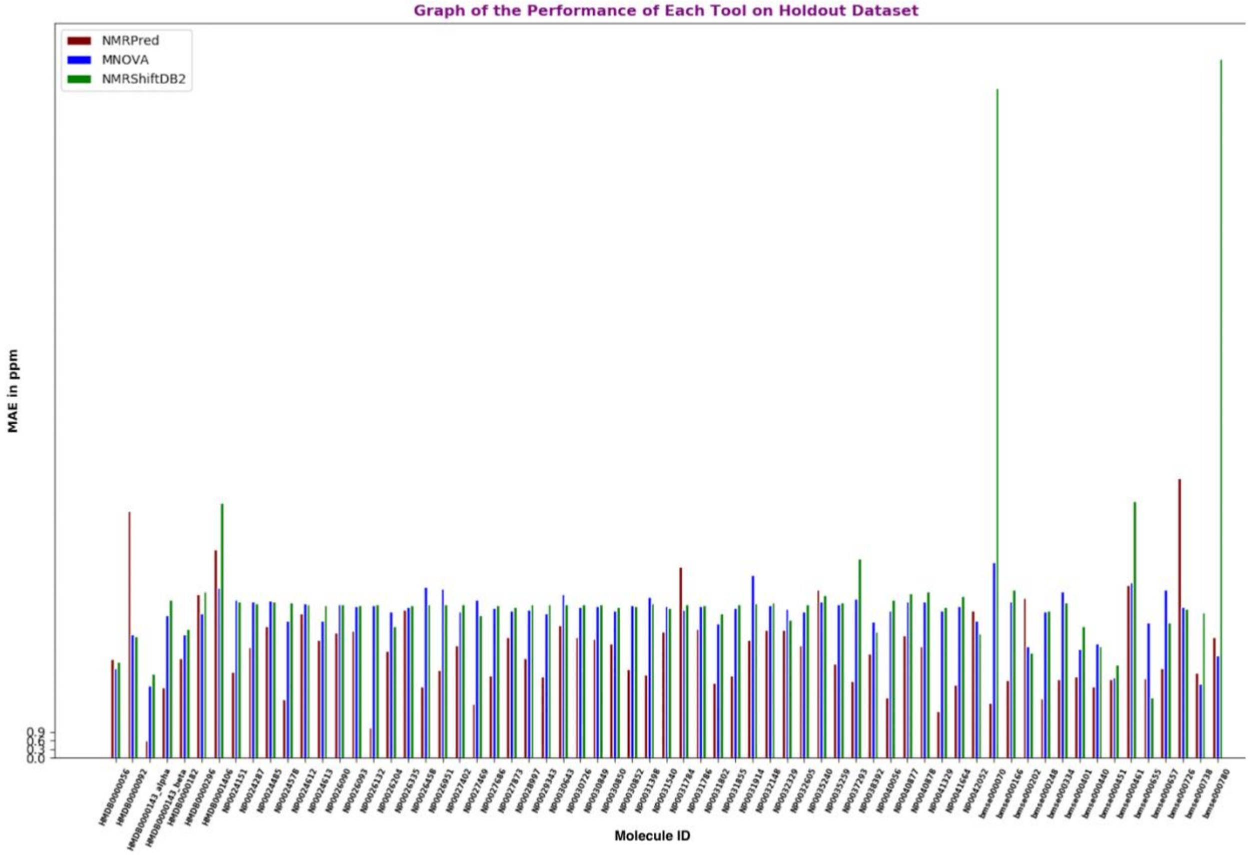
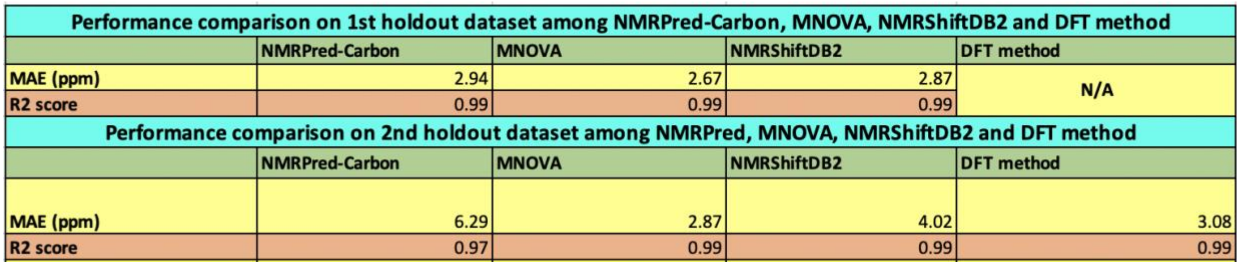


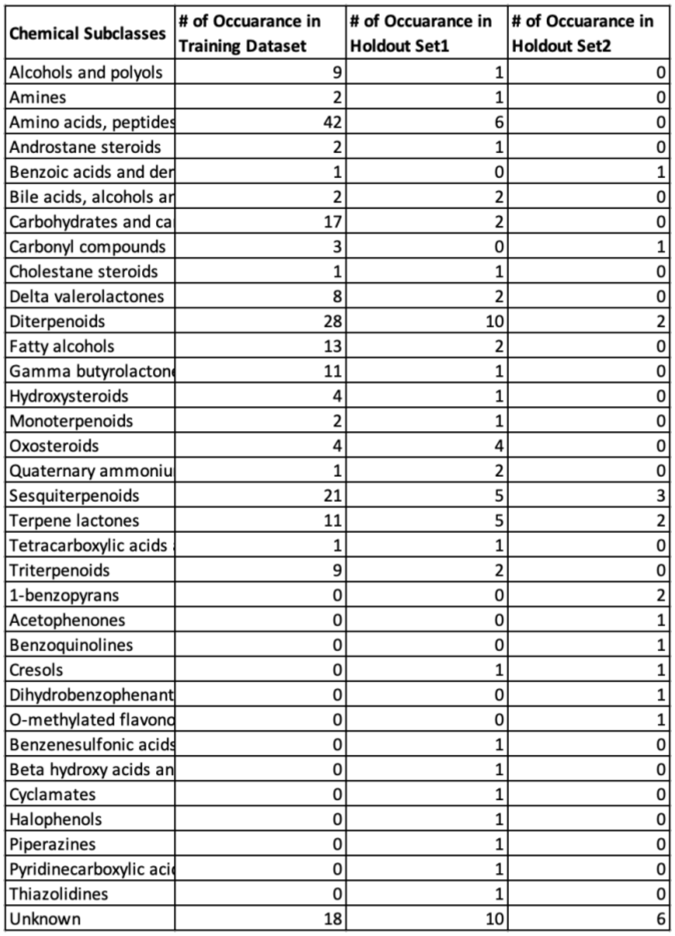
Figure 3.19: A comparison of the mean absolute errors (MAEs) for each of the 65 molecules in the first holdout dataset for the different predictors (NMRPred-Carbon, MNOVA, NMRShiftDB2).



| Performance comparison on 1st holdout dataset among NMRPred-Carbon, MNOVA, NMRShiftDB2 and DFT method | | | | |
|---|---|---|---|---|
| | NMRPred-Carbon | MNOVA | NMRShiftDB2 | DFT method |
| MAE (ppm) | 2.94 | 2.67 | 2.87 | N/A |
| R2 score | 0.99 | 0.99 | 0.99 | |
| **Performance comparison on 2nd holdout dataset among NMRPred, MNOVA, NMRShiftDB2 and DFT method** | | | | |
| | NMRPred-Carbon | MNOVA | NMRShiftDB2 | DFT method |
| MAE (ppm) | 6.29 | 2.87 | 4.02 | 3.08 |
| R2 score | 0.97 | 0.99 | 0.99 | 0.99 |

Figure 3.20: Performance comparison between the 1st and 2nd holdout datasets among the four predictors: NMRPred-Carbon, MNOVA, NMRShiftDB2 and the DFT predictor.

between the first and second training dataset, the smaller overlap in the second dataset may have been enough to tip the balance. Likewise, the fact that ClassyFire failed to classify so many structures in all three datasets may have also led to an underestimate in the true structure class overlap.



| Chemical Subclasses | # of Occuarance in Training Dataset | # of Occuarance in Holdout Set1 | # of Occuarance in Holdout Set2 |
|---|---|---|---|
| Alcohols and polyols | 9 | 1 | 0 |
| Amines | 2 | 1 | 0 |
| Amino acids, peptides | 42 | 6 | 0 |
| Androstane steroids | 2 | 1 | 0 |
| Benzoic acids and der | 1 | 0 | 1 |
| Bile acids, alcohols ar | 2 | 2 | 0 |
| Carbohydrates and ca | 17 | 2 | 0 |
| Carbonyl compounds | 3 | 0 | 1 |
| Cholestane steroids | 1 | 1 | 0 |
| Delta valerolactones | 8 | 2 | 0 |
| Diterpenoids | 28 | 10 | 2 |
| Fatty alcohols | 13 | 2 | 0 |
| Gamma butyrolacton | 11 | 1 | 0 |
| Hydroxysteroids | 4 | 1 | 0 |
| Monoterpenoids | 2 | 1 | 0 |
| Oxosteroids | 4 | 4 | 0 |
| Quaternary ammoniu | 1 | 2 | 0 |
| Sesquiterpenoids | 21 | 5 | 3 |
| Terpene lactones | 11 | 5 | 2 |
| Tetracarboxylic acids | 1 | 1 | 0 |
| Triterpenoids | 9 | 2 | 0 |
| 1-benzopyrans | 0 | 0 | 2 |
| Acetophenones | 0 | 0 | 1 |
| Benzoquinolines | 0 | 0 | 1 |
| Cresols | 0 | 1 | 1 |
| Dihydrobenzophenant | 0 | 0 | 1 |
| O-methylated flavonc | 0 | 0 | 1 |
| Benzenesulfonic acids | 0 | 1 | 0 |
| Beta hydroxy acids an | 0 | 1 | 0 |
| Cyclamates | 0 | 1 | 0 |
| Halophenols | 0 | 1 | 0 |
| Piperazines | 0 | 1 | 0 |
| Pyridinecarboxylic aci | 0 | 1 | 0 |
| Thiazolidines | 0 | 1 | 0 |
| Unknown | 18 | 10 | 6 |

Figure 3.21: The chemical subclass distribution for the NMRPred-Carbon training dataset as well as the 1st holdout dataset and the 2nd holdout dataset (from NP-MRD) as indicated by membership in chemical subclasses.

Another reason for the poor or inconsistent performance in NMRPred-Carbon may have had to do with the lack of distance information or bond geometry information in our feature set.

Recall that in the feature identification stage, similar to the $^{1}$H chemical shift feature selection stage, we calculated the frequency of certain chemical functional groups up to a certain bond distance from the target atom. As noted previously, $^{13}$C chemical shifts are heavily influenced by their α, β and γ substituents. As a result, only considering the frequency of occurrence of these substituents, rather than their orientation, distance or bond connectivity, did not provide enough information to properly estimate $^{13}$C shifts – especially in molecules with never-before-seen geometry. This oversight in feature encoding likely contributed to NMRPred-Carbon’s poor overall performance.

Overall, we can conclude that NMRPred-Carbon was both under-trained in terms of diverse molecular structures or molecular classes (in the training dataset) and that it lacked sufficiently informative geometrical and distance features, all of which led to its poor performance, especially on the second holdout dataset. Undertraining arose because we did provide a sufficient number of molecules belonging to the chemical subclasses represented in the 2nd holdout dataset. This was not totally unexpected given the NMRPred-Carbon's training set was relatively modest in both size and chemical diversity.

## Conclusion

While the overall performance for NMRPred-Carbon were less than satisfactory, I believe I accomplished a number of important goals. I successfully put together, curated, and cleaned a sizeable dataset of experimentally measured $^{13}$C chemical shifts from several well-known NMR libraries, including NP-MRD, BMRB, and HMDB. This dataset, which was used to train and test my machine learning models, required a significant amount of manual remediation. In the end, I created a dataset consisting of 318 structurally diverse molecules with 4983 correctly assigned $^{13}$C chemical shifts. This dataset is an important contribution to the field and can certainly be used by others wishing to further develop $^{13}$C chemical shift predictors.

As was described in the methods, this dataset was then divided into a training set (containing of 253 molecules and 3896 $^{13}$C chemical shifts) and a test set (holdout set #1 consisting of 65 molecules and 1087 $^{13}$C chemical shifts). To prevent data leaking, the split was performed using an 80% to 20% ratio on molecules rather than on $^{13}$C chemical shifts. I also constructed a second dataset, designated holdout set #2 that had 22 compounds and a total of 653 $^{13}$C chemical shifts. I used literature-derived data, RDKit and CDKit, and the $^{13}$C chemical shift dataset (which consisted of 3D structures and experimentally assigned $^{13}$C chemical shifts for each molecule) to construct a set of molecular and atomic features that were expected to be useful for $^{13}$C chemical shift prediction. These features included essential geometrical and atomic characteristics known to contribute to $^{13}$C chemical shift effects. Intelligent feature selection was able to reduce the initial size of the feature set from thousands to just 212 characteristics.

I tested four different machine learning (regressor) algorithms to assess their performance in predicting $^{13}$C chemical shifts. 5-fold cross validation was used to optimize the hyperparameters for each of the regressors on the training set. The first holdout set was used to evaluate their performance via MAE and $R^2$ calculations and to select the best performing model. Unlike the $^1$H chemical shift prediction experiment, we decided not to use nested cross validation to measure the model performance because the ratio of our holdout dataset to the size of the training dataset was relatively conventional. To ascertain whether any of the regressor algorithms were over-trained, a training error assessment was also performed. In the end, the Gradient Boost Regressor (GBR) was found to be the best-performing algorithm and was implemented into the NMRPred-Carbon chemical shift prediction tool.

NMRPred-Carbon is able to accept an organic molecule as a SMILES string, produce a 3D chemical structure (in SDF format), add hydrogens to the input molecule, calculate all the necessary atomic and geometric features of the input molecule, and then pass these features to the GBR model to predict the $^{13}$C chemical shifts using DSS as a chemical shift reference. Using chemical shift reference corrections noted by Wishart at el. (30, 31), NMRPred-Carbon can adjust chemical shift references for TMS and TSP.

MNOVA, NMRShiftDB, and a DFT approach (NWChem) were some of the commercial or well-known $^{13}$C chemical shift predictors against which NMRPred-Carbon was compared. Using holdout set #1 we found that NMRPred-Carbon had an MAE of 2.94 ppm. This result was reasonably close to the performance of MNOVA (MAE = 2.67 ppm) and NMRShiftDB2 (MAE = 2.87 ppm). Further comparisons revealed that NMRPred-Carbon showed the best prediction results for 83.08% of the molecules in holdout dataset #1, whereas MNOVA showed the best results for 4.61% of the molecules in holdout dataset #1 and NMRShiftDB2 showed the best results for just 1.54% of the molecules in the holdout dataset.

On the other hand, when NMRPred-Carbon was tested on holdout set #2, the MAE was 6.29 ppm. This was somewhat worse than the other predictors, including the DFT predictor (3.08 ppm), MNOVA (2.87 ppm), and NMRShiftDB2 (MAE = 4.02 ppm). These results were rather unsatisfactory and indicate that NMRPred-Carbon (and its GBR model) was insufficiently trained. We determined that undertraining was the likely culprit since the testing error for the GBR model was only 1.36 times greater than the training error. The modest ratio between the test to training error indicates that no overfitting occurred. Therefore, undertraining seems to be the main problem with NMRPred-Carbon's performance. Too few examples of key functional groups, key chemical properties or key molecular geometries were available in the training set to allow the model(s) to properly learn or handle the novel structures seen in holdout dataset #2. Additionally, the lack of feature information on bond geometry, bond distances and α, β and γ substituent positions also led to less than adequate (or informative) feature set. To overcome the problem of undertraining, a larger (2-3X), more chemically diverse training dataset would need to be generated. Likewise an expanded set of features that included both bond geometry and path information would also need to be created.

Furthermore, as was noted in Chapter 2, insufficient numbers of training examples and too many features in the training data can lead to other challenges (such as overtraining). When this occurs, the model will often perform extremely well on the training dataset but significantly worse on the holdout dataset.

Overall, there were at least three issues with this NMRPred-Carbon predictor: 1) insufficient training (caused by an inadequately sized training set); 2) an excessive number of uninformative features or high dimensionality in features and 3) too many missing features relating to substituent geometry and substituent distances from the target atom. Unfortunately, I was unable to enhance the atomic feature space or enlarge the $^{13}C$ NMR chemical shift dataset before the MSc program's mandated conclusion date.

# Chapter 4: Summary and Future Directions

NMR spectroscopy continues to be the gold standard for characterizing small organic molecules – both naturally occurring and laboratory synthesized varieties. In particular, NMR allows chemists to assign specific ($^1$H and $^{13}$C) chemical shifts to specific atoms within a molecule. Once a molecule has been "assigned" (i.e., all relevant chemical shifts are assigned to specific atoms), it is possible to use this chemical shift information along with experimentally measured J-coupling constant data and NOE (i.e., distance) data to determine bond connectivity, stereochemistry, atomic geometry, interatomic atomic distances and ultimately the structure of the molecule. Key to the entire structure elucidation process (via NMR) is the chemical shift assignment step. However, the chemical shift assignment process is often very manually intensive and can be slow, tedious and prone to error. If $^1$H and $^{13}$C chemical shifts could be accurately predicted from a known, suspected or hypothesized structure, then many of the major issues (time, high error rates, misinterpretation) associated with chemical shift assignment could be reduced.

In this thesis, I provided a brief review of the theory of NMR chemical shifts and the methods used to predict NMR $^1$H and $^{13}$C chemical shifts, ranging from quantum mechanical *ab initi*o techniques to rule based or classical physics techniques, to database or look-up methods; and, most recently, machine learning (ML) based methods. The most promising of these methods are the ML approaches as they are very fast and generally more accurate than quantum or rule-based methods. While several ML-based methods for $^1$H and $^{13}$C chemical shift prediction have been developed over the past few years, most of them do not achieve the level of accuracy expected of them nor do they fully account for all the known solvent, chemical shift reference and diastereotopic effects that are commonly seen in real samples. It is because of these known deficits with ML-based chemical shift prediction that I chose to try to develop a better, faster and more accurate chemical shift predictor. More specifically, for this thesis I hypothesized that it would be possible to develop ML-based methods that can accurately predict $^{1}H$ and $^{13}C$ chemical shifts of small molecules with an MAE of <0.20 ppm for $^{1}H$ shifts and an MAE of <2 ppm for $^{13}C$ shifts.

To test this hypothesis, I first created a large and very "clean" database of carefully curated, correctly assigned, and carefully referenced $^{1}H$ and $^{13}C$ chemical shift assignments for hundreds of organic molecules where both the solvent and chemical shift references were known. I then implemented a set of programs (using RDKit) that could convert a chemical SMILES (text) string into a robust 3D chemical structure. I then used known information about the atomic, molecular and geometric effects on $^{1}H$ and $^{13}C$ chemical shifts to create a "smart" feature set that could be rapidly calculated from any 3D structure or any organic molecules (using CDKit and RDkit). Using these derived features (along with various feature selection methods) and the associated experimental $^{1}H$ and $^{13}C$ chemical shifts in my database, I then tested several ML algorithms for $^{1}H$ and $^{13}C$ chemical shift prediction. These included support vector machine algorithms, random forest algorithms along with a Gradient Boost regressor, an XGBoost regressor and a CatBoost regressor. Training and model optimization were done using standard K-fold, cross-validation methods and various assessments of training completeness (over and undertraining) were performed. The best performing algorithms for each of the $^{1}H$ and $^{13}C$ chemical shift predictors were then tested on two different sets of holdout datasets as a validation step. The resulting $^{1}H$ chemical shift predictor was called NMRPred and the resulting $^{13}C$ chemical shift predictor was NMRPred-Carbon.

As I noted several times throughout this document, the most difficult and time-consuming part of this study was the data collection and curation process. This was because the quality of most chemical shift databases was questionable. For this study, I selected and cleaned data from several well-known chemical shift databases, including HMDB, NP-MRD, BMRB, and the GISSMO library. All chemical shifts selected for this training dataset were referenced (or re-referenced) to a consistent solvent ($D_2O$), a standard chemical shift reference (DSS), and a standard pH value (7.0 – 7.4). To create a clean $^1$H chemical shift dataset, we selected 577 molecules from 2693 molecules that satisfied these criteria. To create a clean $^{13}$C chemical shift dataset we selected 253 molecules from 4802 molecules that satisfied these criteria. As part of the curation process, the correct molecular structure, correctly matched spectra, the correct 3D structures for the molecules, the correct atom index mapping, and the examination and correction of incorrect chemical shift assignments through literature all had to be performed by hand. Chemical shift re-assignments or corrections were based on information provided in the Reich database, comparison with predicted values from NMRShiftDB2 and MNOVA, and consultation with NMR experts.

After "informed" feature preparation and selection I constructed several ML models and identified two optimal models for chemical shift prediction. One was for the prediction of $^1$H chemical shifts that used a Random Forest Regressor (RFR). This model (and the finalized program called NMRPred) had an MAE of 0.11 ppm with a standard deviation of 0.18 ppm for the first holdout dataset and 0.36 ppm with a standard deviation of 0.56 ppm for the second holdout dataset. The exact size, chemical composition and other details of each of the holdout sets was described in Chapters 2 and 3. Comparisons of NMRPred to other commercial or freeware programs showed that this program performed well. Unlike other programs, NMRPred also has the capability of accurately predicting the chemical shifts of diastereotopic protons and accurately adjusting chemical shifts to different NMR solvents. The other ML model that was developed was for the prediction of $^{13}$C chemical shifts. This model used a Gradient Boost Regressor (GBR). This model (and the finalized program called NMRPred-Carbon) had an MAE of 2.94 ppm with a standard deviation of 4.20 ppm for the first holdout dataset and 6.29 ppm with a standard deviation of 8.65 ppm for the second holdout set. Comparisons of NMRPred-Carbon to other commercial or freeware programs showed that this program did not perform particularly well.

While the performance of NMRPred (the $^{1}$H chemical shift predictor) met the initial performance criteria (<0.2 ppm MAE) on the first holdout dataset, it failed to meet the performance criteria on the second holdout data set (>0.2 ppm). On the other hand, the performance of NMRPred-Carbon (the $^{13}$C chemical shift predictor) did not meet the performance objective (<2.0 ppm MAE) on any of the holdout datasets. My analysis indicated that both the $^{1}$H and $^{13}$C predictors suffered from undertraining (too few examples with too few structure classes), with the $^{13}$C shift predictor suffering most severely. Other problems that contributed to the poorer-than-expected performance included issues with high dimensional features (both the $^{1}$H and $^{13}$C predictors) and incomplete or improper feature sets (the $^{13}$C predictor). Despite not achieving the broad performance objective, I believe this work led to some useful advances and some important new resources. These are summarized below.

## Contributions

Following is a summary of what I believe were the main contributions of this thesis:

- The development of two large, correctly assigned, correctly referenced, correctly structured and richly annotated (with solvent, pH and temperature data) $^1$H and $^{13}$C chemical shift/structure databases. These databases will be of significant value to anyone wishing to develop ML chemical shift predictors and to understand the impact of solvent, pH, temperature and chemical shift references on measured chemical shifts.
- The development of methods that can automatically and accurately adjust $^1$H chemical shift values in various solvents ($D_2O$, $CDCl_3$, DMSO).
- The development of an effective strategy for predicting $^1$H chemical shifts for diastereotopic protons.
- The development of an effective strategy for predicting or re-scaling predicted $^{13}$C chemical shifts using different chemical shift references (TMS, DSS, TSP).
- The development and implementation of a generic pipeline of molecular/atomic feature calculation tools and ML models for chemical shift prediction that could be easily improved by providing a larger, more chemically diverse training dataset.
- NMRPred has already been used in NP-MRD to predict $^1$H chemical shifts for ~60,425 compounds. Using those NMRPred predicted chemical shift values as an input for an in-house software package, called JPred, at total of ~604,250 $^1$H NMR spectra have been generated over 10 different NMR spectrometer frequencies. All of these predicted NMR spectra are available in the NP-MRD website.

## Future Work

As noted earlier, both the $^1$H and $^{13}$C predictors suffered from undertraining (too few examples with too few structure classes), with the $^{13}$C shift predictor suffering most severely. Other problems that contributed to the poorer-than-expected performance included issues with high dimensional feature sets (too many features) and incomplete or improper feature sets (which did not include bond connectivity data, route learning and bond geometry – especially for the $^{13}$C predictor).

To address the problem of undertraining, I believe it will be important to significantly expand the size of our training datasets. A dataset of 5K molecules was recently assembled (21) using data from NMRShiftDB2. We would like to incorporate that dataset into our model(s). However, the experimental chemical shift values in NMRDhiftDB2 are not consistently referenced. Therefore, some effort will be required to adjust the chemical shift values appropriately so that they would correspond to those in the same solvent, with the same pH, and with the same chemical shift reference. This would take 2-3 months of manual effort, but I believe it is possible.

In terms of capturing more relevant features, I would propose to modify the feature calculation to take into account all atoms that are within 4-5 Angstroms of the target atom, as opposed to just capturing effects of the three closest (by distance) atoms to the target atom. The current feature calculation determines the existence of several chemical functional groups at various distances using one-hot encoding. This led to a very high dimensional feature set. A better approach would be to assign numerical labels to various chemical functional groups (rather than using one-hot encoding) in order to reduce the dimensionality. This would require using only one column in the feature set (to indicate the presence of the chemical functional groups). Furthermore, by taking into account all functional groups that are within 6-7 Angstroms of the target atom and including their geometric orientation relative to the target atom, I believe more useful functional group information could be integrated into the model. Additionally, the implementation of a more sophisticated deep machine learning model, such as a Graph Neural Network would likely improve the overall performance of my original ML-based predictors. I believe that if these modifications were made, it would be possible to meet the original objectives of my thesis and to exceed the top performing chemical shift predictors now on the market.

# References

1. Andrew,E.R. (2009) Nuclear Magnetic Resonance. *Cambridge, UK: Cambridge University Press*. ISBN: 0521114330, 9780521114332

2. Mountford, C. E., Stanwell, P., Lin, A., Ramadan, S., & Ross, B. (2010). Neurospectroscopy: the Past, Present and Future. *Chemical Reviews*, *110*(5), 3060-3086.

3. Marion, D. (2013). An Introduction to Biological NMR Spectroscopy. *Molecular & Cellular Proteomics*, *12*(11), 3006-3025.

4. Wishart, D. S., Cheng, L. L., Copié, V., Edison, A. S., Eghbalnia, H. R., Hoch, J. C., Gouveia, G. J., Pathmasiri, W., Powers, R., Schock, T. B., Sumner, L. W., & Uchimiya, M. (2022). NMR and Metabolomics-A Roadmap for the Future. *Metabolites*, 12(8), 678.

5. Kamal, G. M., Uddin, J., Tahir, M. S., Khalid, M., Ahmad, S., & Hussain, A. I. (2021). Nuclear Magnetic Resonance Spectroscopy in Food Analysis. *Techniques to Measure Food Safety and Quality: Microbial, Chemical, and Sensory*, 137-168.

6. Zloh, M. (2019). NMR Spectroscopy in Drug Discovery and Development: Evaluation of Physico-Chemical Properties. *ADMET and DMPK*, *7*(4), 242-251.

7. Rakhmatullin, I. Z., Efimov, S. V., Klochkov, V. V., & Varfolomeev, M. A. (2006). Nuclear Magnetic Resonance Characterization of Petroleum. *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation*, 1-9.

8. Lecoq,L., Schledorn,M., Wang,S., Smith-Penzel,S., Malär,A.A., Callon,M., Nassal,M., Meier,B.H. and Böckmann,A. (2019).100 kHz MAS Proton-Detected NMR Spectroscopy of Hepatitis B Virus Capsids. *Frontiers in Molecular Biosciences*, 6, 58.

9. Wider, G. (1998). Technical Aspects of NMR Spectroscopy with Biological Macromolecules and Studies of Hydration in Solution. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *32*(3), 193-275.

10. Popov, M. (Ed.). (1990). Modern NMR Techniques and their Application in Chemistry. *CRC Press.*

11. Gáspári, Z. (Ed.). (2020). Structural Bioinformatics: Methods and Protocol*s*. *Humana Press*.

12. Steinbeck, C., Krause, S., & Kuhn, S. (2003). NMRShiftDB Constructing a Free Chemical Information System with Open-Source Components. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1733-1739.

13. SAITO, T., & KINUGASA, S. (2011). Development and Release of a Spectral Database for Organic Compounds-Key to the Continual Services and Success of a Large-Scale Database. *Synthesiology English edition*, *4*(1), 35-44.

14. Wishart, D. S., Sayeeda, Z., Budinski, Z., Guo, A., Lee, B. L., Berjanskii, M., ... & Cort, J. R. (2022). NP-MRD: the Natural Products Magnetic Resonance Database. *Nucleic Acids Research*, *50*(D1), D665-D677.

15. Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., ... & Querengesser, L. (2007). HMDB: the Human Metabolome Database. *Nucleic Acids Research*, 35(suppl_1), D521-D526.

16. Smurnyy,Y.D., Blinov,K.A., Churanova,T.S., Elyashberg,M.E. and Williams,A.J. (2008) Toward More Reliable $^{13}$C and $^{1}$H Chemical Shift Prediction: A Systematic Comparison of Neural-Network and Least-Squares Regression Based Approaches. *J Chem Inf Model*, 48, 12.

17. Blinov, K. A., Smurnyy, Y. D., Churanova, T. S., Elyashberg, M. E., & Williams, A. J. (2009). Development of a Fast and Accurate Method of $^{13}$C NMR Chemical Shift Prediction. *Chemometrics and Intelligent Laboratory Systems*, 97(1), 91-97.

18. Jonas, E., Kuhn, S., & Schlörer, N. (2022). Prediction of Chemical Shift in NMR: A review. *Magnetic Resonance in Chemistry*, 60(11), 1021-1031.

19. Bühl, M., Kaupp, M., Malkina, O. L., & Malkin, V. G. (1999). The DFT Route to NMR Chemical Shifts. *Journal of Computational Chemistry*, 20(1), 91-105

20. Lodewyk, M. W., Siebert, M. R., & Tantillo, D. J. (2012). Computational Prediction of $^{1}$H and $^{13}$C Chemical Shifts: A Useful Tool for Natural Product, Mechanistic, and Synthetic Organic Chemistry. *Chemical Reviews*, 112(3), 1839-1862.

21. Guan, Y., Sowndarya, S. S., Gallegos, L. C., John, P. C. S., & Paton, R. S. (2021). Real-time Prediction of $^{1}$H and $^{13}$C Chemical Shifts with DFT Accuracy Using a 3D Graph Neural Network. *Chemical Science*, 12(36), 12012-12026.

22. Grant, D. M., & Paul, E. G. (1964). Carbon-13 Magnetic Resonance. II. Chemical Shift Data for the Alkanes. *Journal of the American Chemical Society*, 86(15), 2984-2990.

23. Clerc, J. T., & Sommerauer, H. (1977). A Minicomputer Program Based on Additivity Rules for the Estimation of $^{13}$C-NMR Chemical Shifts. *Analytica Chimica Acta*, 95(1), 33-40.

24. Fürst, A., & Pretsch, E. (1990). A Computer Program for the Prediction of 13-C-NMR Chemical Shifts of Organic Compounds. *Analytica Chimica Acta*, 229, 17-25.

25. Pretsch, E., Furst, A., Badertscher, M., Buergin, R., & Munk, M. E. (1992). C13Shift: A Computer Program for the Prediction of Carbon-13 NMR Spectra Based on an Open Set of Additivity Rules. *Journal of Chemical Information and Computer Sciences*, 32(4), 291-295.

26. Schaller, R. B., & Pretsch, E. (1994). A Computer Program for the Automatic Estimation of $^{1}$H NMR Chemical Shifts. *Analytica Chimica Acta*, 290(3), 295-302.

27. Schaller, R. B., Arnold, C., & Pretsch, E. (1995). New Parameters for Predicting $^{1}$H NMR Chemical Shifts of Protons Attached to Carbon Atoms. *Analytica Chimica Acta*, 312(1), 95-105.

28. Bremser, W. (1978). HOSE—A Novel Substructure Code. *Analytica Chimica Acta*, 103(4), 355-365.

29. Kuhn, S., & Johnson, S. R. (2019). Stereo-Aware Extension of HOSE Codes. *ACS Omega*, 4(4), 7323-7329.

30. Wishart, D. S., & Sykes, B. D. (1994). [12] Chemical Shifts as a Tool for Structure Determination. *Methods in Enzymology*, 239, 363-392.

31. Wishart, D. S., Bigam, C. G., Yao, J., Abildgaard, F., Dyson, H. J., Oldfield, E., ... & Sykes, B. D. (1995). $^{1}$H, $^{13}$C and $^{15}$N Chemical Shift Referencing in Biomolecular NMR. *Journal of Biomolecular NMR*, 6, 135-140.

32. Kanade, V. (2022, August 30). What is Machine Learning? Definition, Types, Applications, and Trends for 2022. From https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/

33. Machine Learning. Coursera. (n.d.). From https://www.coursera.org/specializations/machine-learning-introduction

34. Wishart, D. S., Tian, S., Allen, D., Oler, E., Peters, H., Lui, V. W., ... & Metz, T. O. (2022). BioTransformer 3.0—A Web Server for Accurately Predicting Metabolic Transformation Products. *Nucleic Acids Research*, 50(W1), W115-W123.

35. Ivanov, J., Polshakov, D., Kato-Weinstein, J., Zhou, Q., Li, Y., Granet, R., ... & Aultman, C. (2020). Quantitative Structure–Activity Relationship Machine Learning Models and Their Applications for Identifying Viral 3CLpro-and RdRp-Targeting Compounds as Potential Therapeutics for COVID-19 and Related Viral Infections. *ACS Omega*, 5(42), 27344-27358.

36. Allen, F., Pon, A., Wilson, M., Greiner, R., & Wishart, D. (2014). CFM-ID: A Web Server for Annotation, Spectrum Prediction and Metabolite Identification from Tandem Mass Spectra. *Nucleic Acids Research*, 42(W1), W94-W99.

37. Djoumbou-Feunang, Y., Pon, A., Karu, N., Zheng, J., Li, C., Arndt, D., ... & Wishart, D. S. (2019). CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. *Metabolites*, 9(4), 72.

38. Wang, F., Liigand, J., Tian, S., Arndt, D., Greiner, R., & Wishart, D. S. (2021). CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Analytical Chemistry*, 93(34), 11692-11700.

39. Aires-de-Sousa, J., Hemmer, M. C., & Gasteiger, J. (2002). Prediction of $^{1}$H NMR Chemical Shifts Using Neural Networks. *Analytical Chemistry*, 74(1), 80-90.

40. Umakant_Shinde. (2021, January 6). Machine Learning With Flowchart. From https://medium.com/analytics-vidhya/machine-learning-with-flowchart-696ff42f8aff

41. Zhang, A. (2021, November 12). Data Types From a Machine Learning Perspective With Examples. From https://towardsdatascience.com/data-types-from-a-machine-learning-perspective-with-examples-111ac679e8bc

42. Kvasnicka, V., Sklenak, S., & Pospichal, J. (1992). Application of Recurrent Neural Networks in Chemistry. Prediction and Classification of Carbon-13 NMR Chemical Shifts in A Series of Monosubstituted Benzenes. *Journal of Chemical Information and Computer Sciences*, 32(6), 742-747.

43. Kvasnička, V., Sklenák, Š., & Pospichal, J. (1992). Application of Neural Networks With Feedback Connections in Chemistry: Prediction of $^{13}$C NMR Chemical Shifts in A Series of Monosubstituted Benzenes. Journal of Molecular Structure: *THEOCHEM*, 277, 87-107.

44. Sklenak, S., Kvasnicka, V., & Pospichal, J. (1994). Prediction of $^{13}$C NMR Chemical Shifts by Neural Networks in A Series of Monosubstituted Benzenes. *Chem. Papers*, 48, 135-140.

45. Meiler, J., Meusinger, R., & Will, M. (2000). Fast Determination of $^{13}$C NMR Chemical Shifts Using Artificial Neural Networks. *Journal of Chemical Information and Computer Sciences*, 40(5), 1169-1176.

46. Meiler, J., Maier, W., Will, M., & Meusinger, R. (2002). Using Neural Networks for $^{13}$C NMR Chemical Shift Prediction–Comparison With Traditional Methods. *Journal of Magnetic Resonance*, 157(2), 242-252.

47. Binev, Y., & Aires-de-Sousa, J. (2004). Structure-Based Predictions of $^{1}$H NMR Chemical Shifts Using Feed-Forward Neural Networks. *Journal of Chemical Information and Computer Sciences*, 44(3), 940-945.

48. Binev, Y., Corvo, M., & Aires-de-Sousa, J. (2004). The Impact of Available Experimental Data on the Prediction of $^{1}$H NMR Chemical Shifts by Neural Networks. *Journal of Chemical Information and Computer Sciences*, 44(3), 946-949.

49. Kuhn, S., Egert, B., Neumann, S., & Steinbeck, C. (2008). Building Blocks for Automated Elucidation of Metabolites: Machine Learning Methods for NMR Prediction. *BMC Bioinformatics*, 9, 1-19.

50. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo-and Bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2), 493-500.

51. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo-and Bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2), 493-500.

52. Jonas, E., & Kuhn, S. (2019). Rapid prediction of NMR Spectral Properties With Quantified Uncertainty. *Journal of Cheminformatics*, 11(1), 1-7.

53. Kwon, Y., Lee, D., Choi, Y. S., Kang, M., & Kang, S. (2020). Neural Message Passing for NMR Chemical Shift Prediction. *Journal of Chemical Information and Modeling*, 60(4), 2024-2030.

54. Jonas, E., Kuhn, S., & Schlörer, N. (2022). Prediction of chemical shift in NMR: A Review. *Magnetic Resonance in Chemistry*, 60(11), 1021-1031.

55. Weininger, D. (1988). SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1), 31-36.

56. Landrum, G. (2013). RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling. *Greg Landrum*, 8.

57. Landrum, G., Lewis, R., Palmer, A., Stiefl, N., & Vulpetti, A. (2011). Making Sure there's a" Give" Associated With the" Take": Producing and Using Open-Source Software in Big Pharma. *Journal of Cheminformatics*, 3, 1-1.

58. Dashti, H., Wedell, J. R., Westler, W. M., Tonelli, M., Aceti, D., Amarasinghe, G. K., ... & Eghbalnia, H. R. (2018). Applications of Parametrized NMR Spin Systems of Small Molecules. *Analytical Chemistry*, 90(18), 10646-10649.

59. Kaushik,M. (2014) A Review of Innovative Chemical Drawing and Spectra Prediction Computer Software. *Mediterranean Journal of Chemistry*, 3, 759–766.

60. Csizmadia,F. (2000) JChem: Java Applets and Modules Supporting Chemical Database Handling from Web Browsers. *Journal of Chemical Information and Computer Sciences* , 40, 323–324.

61. Dashti,H., Westler,W.M., Markley,J.L. and Eghbalnia,H.R. (2017) Unique Identifiers for Small Molecules Enable Rigorous Labeling of Their Atoms. *Scientific Data 2017 4:1*, 4, 1–9.

62. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: A Public Information System for Analyzing Bioactivities of Small Molecules. *Nucleic Acids Research*, 37, W623–W633.

63. Willcott,M.R. (2009) MestRe Nova. *Journal of the American Chemical Society*, 131, 13180–13180.

64. Meragelman,T.L. (2005) Basic One- and Two-Dimensional NMR Spectroscopy. *J Nat Prod*, 68, 1578–1579. ISBN 3-527-31233-1

65. Kan, R. O. (1964). A Correlation of Chemical Shifts With Inductive Effect Parameters. *Journal of the American Chemical Society*, 86(23), 5180-5183.

66. Reusch, W. (n.d.). Bonding & Molecular Structure. From https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/intro2.htm#strc3b .

67. Libretexts. (2022, October 3). 5.3: Factors That Influence NMR Chemical Shift. *Chemistry LibreTexts*. From https://chem.libretexts.org/Courses/Providence_College/Organic_Chemistry_I/05%3A_Analytical_Methods_for_Structure_Elucidation/5.03%3A_Factors_That_Influence_NMR_Chemical_Shift

68. Schaefer, T., Reynolds, W. F., & Yonemoto, T. (1963). Possible Intramolecular Van der Waals Contributions to Proton and Carbon-13 Shifts in Aliphatic and Aromatic Halogen Compounds. *Canadian Journal of Chemistry*, 41(12), 2969-2976.

69. Sutter, K., Aucar, G. A., & Autschbach, J. (2015). Analysis of Proton NMR in Hydrogen Bonds in Terms of Lone-Pair and Bond Orbital Contributions. *Chemistry–A European Journal*, 21(50), 18138-18155.

70. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo-and Bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2), 493-500.

71. Bonaccorso, G. (2018). Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning. *Packt Publishing Ltd*. ISBN: 9781789347999

72. Suykens, J. A., & Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, 9, 293-300.

73. Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32.

74. Sharma, N. (2018). XGBoost. The Extreme Gradient Boosting for Mining Applications. Munich: *GRIN Verlag*. ISBN: 9783668660618.

75. Valiev, M., Bylaska, E. J., Govind, N., Kowalski, K., Straatsma, T. P., Van Dam, H. J. J., ... & De Jong, W. A. (2010). NWChem: A Comprehensive and Scalable Open-Source Solution for Large Scale Molecular Simulations. *Computer Physics Communications*, 181(9), 1477-1489.

76. Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., ... & Wishart, D. S. (2016). ClassyFire: Automated Chemical Classification With a Comprehensive, Computable Taxonomy. *Journal of Cheminformatics*, 8, 1-20.

77. Prakash, A. (n.d.). Working With Sparse Features in Machine Learning Models. *KDnuggets*. From https://www.kdnuggets.com/2021/01/sparse-features-machine-learning-models.html

78. Kuss, O. (2002). Global Goodness-of-Fit Tests in Logistic Regression With Sparse Data. *Statistics in Medicine*, 21(24), 3789-3801.

79. Mandal, P. K., & Majumdar, A. (2004). A Comprehensive Discussion of HSQC and HMQC Pulse Sequences. Concepts in Magnetic Resonance Part A: *An Educational Journal*, 20(1), 1-23.

80. Facey, G. (n.d.). HMQC vs HSQC. From http://u-of-o-nmr facility.blogspot.com/2009/01/hmqc-vs-hsqc.html

81. Libretexts. (2022, October 4). 4.4: Factors in Chemical Shift- Carbon Geometry. *Chemistry LibreTexts*. From

https://chem.libretexts.org/Bookshelves/General_Chemistry/Book%3A_Structure_and_Reactivity_in_Organic_Biological_and_Inorganic_Chemistry_(Schaller)/Structure_and_Reactivity_in_Organic_Biological_and_Inorganic_Chemistry_II%3A_Practical_Aspects_of_Structure_-_Purification_and_Spectroscopy/04%3A_Nuclear_Magnetic_Resonance_Spectroscopy/4.04%3A_Factors_in_Chemical_Shift-_Carbon_Geometry

82. Admin. (2022, May 4). Hybridization of Carbon - Molecular Geometry and Bond Angles. From https://byjus.com/jee/hybridization-of-carbon/

83. Organic Chemistry Data. (n.d.). 6-CMR-2 Origin of Chemical Shifts. *NMR spectroscopy*. From https://organicchemistrydata.org/hansreich/resources/nmr/?index=nmr_index%2Finfo&amp;page=06-cmr-02-shifts%2F

84. Schneider, H. J., & Hoppen, V. (1978). Carbon-13 Nuclear Magnetic Resonance Substituent-Induced Shieldings and Conformational Equilibriums in Cyclohexanes. *The Journal of Organic Chemistry*, 43(20), 3866-3873.

85. Organic Chemistry Data. (n.d.). 6-CMR-4 13C Chemical Shift Effects on sp2 and sp Carbons. *NMR Spectroscopy*. From https://organicchemistrydata.org/hansreich/resources/nmr/?index=nmr_index%2Finfo&amp;page=06-cmr-04-shifts-vinyl%2F