Prediction of 1H NMR Spectra in Water using machine learning models

Tanvir Sajed **Russ Greiner** **David Wishart**
Department of Computer Science, University of Alberta
{tsajed,rgreiner,dwishart}@ualberta.ca

Abstract

In the field of nuclear magnetic resonance (NMR) spectroscopy, there have been limited attempts to predict an 1D proton NMR spectra from the structure of a small molecule. Steinbeck's team and Aires De Sousa's team have developed algorithms using machine learning techniques to solve the problem while only De Sousa's algorithm is freely available online. In this paper, we propose an algorithm called NmrPred that can calculate atomic features from a structure, predict its 1H NMR chemical shifts in water and perform NMR simulation to generate a 1D NMR spectra. Using a training dataset of 1,022 protons, our best classifier Random Forest achieved a mean absolute error of 0.21 ppm. With a hold out set of 88 protons for 10 molecules, our algorithm achieved a mean absolute error of 0.13 ppm while De Sousa's algorithm managed 0.17 ppm on the same hold out set. Our algorithm beat De Sousa's by a small margin. There are commercial algorithms that have also made attempts to solve this problem but their software and dataset are not open source.

1 Introduction

Nuclear Magnetic Resonance [1] spectroscopy is applied to metabolomics for identification of compounds in a sample and calculation of its concentration. The sample is prepared in a liquid solvent, like water or chloroform, and rotated inside an external magnetic field. Under the influence of an external magnetic field, two spin states of nuclei exists, +1/2 and -1/2. Radio frequency signals are transmitted to the sample and are absorbed by the nuclei. When the RF signals are shut down, some nuclei will undergo a spin relaxation process where they emit radio frequency signals which are then captured by a receiver and analyzed. By changing the external magnetic field, signals in the time domain are acquired by observing the emitted radio frequencies of the nuclei. The radio frequency signal decays and that leads to the creation of a free induction decay (fid). In order to convert the signal from the time domain to the frequency domain, fast Fourier transformation [2] is applied to the signal to create a graph of intensities against resonant spin frequencies of different nuclei, measured in ppm. The graph can be 1 dimensional if we are analyzing 1D NMR or 2 dimensional [3] if we are analyzing 2D NMR. 2D NMR, which is a graph of frequency against frequency, is generally analyzed for complex spectra. NMR spectra will vary if we are only looking at spins of Hydrogen compared to Carbon [4]. There will also be variance in the spectra if we use a solvent like water as opposed to chloroform [5]. In this paper, we are only looking at 1 Dimensional NMR spectra [6] involving spins of Hydrogen and the solvent is water. In existing literature, there have been limited attempts on predicting NMR spectra from a structure. Kuhn et. al. [7] and De Sousa et. al. [8] have built machine learning algorithms to predict 1D Hydrogen NMR spectra from a given structure. Chloroform is the main solvent that they use. A prediction algorithm will help build a library of predicted NMR spectra that can be later used for identification and quantification of compounds in a sample.

2 Problem Statement

Given a chemical structure, predict the chemical shifts of all the hydrogen atoms. The chemical shifts are real number measured in ppm. A chemical structure can be represented by molfile or SDF. A SDF file is an XML file that represents the atom positions and bonding patterns of a molecule. Not all the hydrogen atoms register a chemical shift in NMR. Hydrogen atoms connected to Nitrogen and Oxygen get dissolved in water and do not emit any radio frequency pulses. We have proposed two different ways to predict the chemical shifts, by classification and regression.

2.1 Classification

Chemical shifts generally range between 0 and 10 ppm. We divide the chemical shift range by 1000 to get 1000 different classes of spin frequencies. Therefore, 0.01 ppm belongs to the 1st class out of 1000 assigned classes. The next class is 0.02 ppm and each subsequent class is added by 0.01 ppm. The real chemical shifts are converted to the nearest 0.01 ppm and assigned to its class. Although it introduces a 0.005 error, we allow since the error is negligible. We have tried to solve the classification problem using Random Forests, J48 Decision trees, SVM, Artificial Neural Networks and Naive Bayes.

2.2 Regression

Defining the problem of regression is much easier. Since all chemical shifts are real numbers, the problem intuitively falls into a regression. We have tried Linear Regression and Artificial Neural Networks to predict a chemical shift value in ppm from the chemical features of a structure.

3 Literature Survey

Kuhn et. al. [7] have published a paper that predicts chemical shift of hydrogen atoms in 1D NMR in chloroform. They have reported a mean absolute error of 0.18 ppm with Decision Tree J48 over 10 fold cross validation on their 18,692 instances from 1,829 unique molecules. However, they did not test it on a hold out set or compare their models with another state of the art NMR prediction algorithm.

In 2002, De Sousa et. al. [8] published a prediction algorithm that uses neural network to predict hydrogen chemical shifts in 1D NMR using chloroform as a solvent. A server is published with the algorithm to predict NMR, along with a spin simulation that renders the NMR spectra graph on the server [9]. They sampled a training set of 744 protons and an independent test set of 259 protons over which they achieved a mean absolute error of 0.25 ppm.

In 2004, Binev et al. [10] published another feed forward artificial neural network algorithm to predict 1H chemical shifts in chloroform. Using a training set of 744 hydrogen atoms and an independent test set of 952, they achieved a mean average error of 0.29 ppm for the independent test set.

Binev et. al. [11] have also published another prediction algorithm to predict J-coupling constants [12] required for NMR simulation. They have used neural networks similar to the ones used to predict chemical shifts of 1H NMR.

ACD labs [13] built a commercial chemical shift prediction software in which they have achieved a mean standard error of 0.22 ppm over approximately 50,000 training instances of hydrogen atoms. CambridgeSoft ChemDraw 8.0 [7] achieved a mean standard error of 0.45 ppm over a similar training set. Both are commercial software and their dataset is not publicly available.

In this paper, we try to predict hydrogen chemical shifts and then simulate NMR, rendering a 1D NMR spectra. Although we have created a pipeline for that, we have not yet created a server that supports NMR prediction and simulation. We also take J-coupling constants to be 4 Hz always, and do not offer any algorithm to predict or calculate J-coupling constants.

4 The Dataset

We have a total of 1,022 hydrogen chemical shifts generated from 169 molecules by mining HMDB [14]. We have also collected an independent hold out set of 88 hydrogen atoms over 10 molecules. We also downloaded 3d SDF files from HMDB for all the molecules in both training and hold out set. We used these structures for calculating numerical features of instances.

4.1 Manual Labeling

The 1H NMR chemical shifts were downloaded as text files from the HMDB server. The chemical shift of a hydrogen was be mapped to a particular carbon atom it is bonded to. However, the ordering of atoms changes from structure to structure. For example, using Marvin Sketch, a software for drawing molecules, one might start from a hydrogen atom. So the first hydrogen atom he/she has drawn will be numbered 1. If someone else started drawing from the first carbon atom, that carbon atom will be numbered 1. Therefore, the numbering of atoms in a molecule varies depending on how it was done in the first place. In HMDB, the molecular structures do not have the same pattern of numbering. Although the numbering system initially used was saved in an image, the structures were not. The structures are independent, downloaded from HMDB, with a totally different numbering system. Figure 4 shows the problem in detail. Figure 1 shows a table of assignments of carbon atom numbers to hydrogen chemical shifts in ppm. These carbon numbers correspond to those in Figure 2. However, the structures were never saved in HMDB. The same structure with different atom numbering is found in Figure 3, the SDF file downloaded form HMDB.

In order to rectify the problem, we compared the original picture of structure to the SDF structure downloaded from HMDB. Using Marvin Sketch from ChemAxon [15], we rotated the structure around different axes in order to align it with the original structure. For each molecule, we mapped the two structures looking at their images side by side and then changed the atom numbers the chemical shifts were assigned to. After the change, the chemical shifts pointed to atomic numbers represented by the SDF file that we use to calculate atomic features for that molecule. The table of assignments will have actual hydrogen atom position rather than carbon atom position as depicted in Figure 1. The attempt was time consuming since each molecule took about 10 minutes for remapping of atomic numbers.

4.2 Training Set

Some molecules in the training set were selected by prior knowledge. Some amino acids and nucleotides were selected because they represent rings and have a complex structure. This was done in order to increase variance in the dataset. Others were selected sequentially from HMDB00001 to HMDB00412, but we had no prior idea of what these molecules might be. We also faced some molecules without any proper 1H NMR data. Sometimes there were missing chemical shifts. We discarded them to avoid outliers and learning incorrect data.

4.3 Hold out Set

As mentioned before, we selected 10 molecules and 88 hydrogen atoms for hold out set. These were selected sequentially from HMDB00413 to HMDB00439. HMDB000969, Tryptophan, was selected since we wanted to have the hold out set represent an aromatic amino acid. We had no prior knowledge of the other molecules in the hold set except for HMDB00969. There are a total of 5 aromatic molecules, and 5 aliphatic molecules in the hold out set. As done for the training set, the atoms in the text files downloaded from HMDB needed to be manually labeled and renumbered.

5 Evaluation

We used mean absolute error (MAE) to define our performance measure. This means the higher the mean absolute error, the less accurate our algorithm is. Equation (1) shows the equation where y_j is

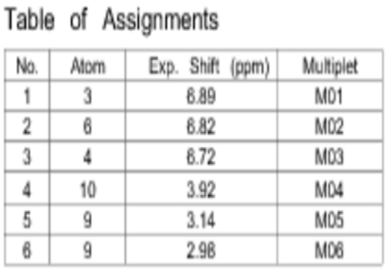
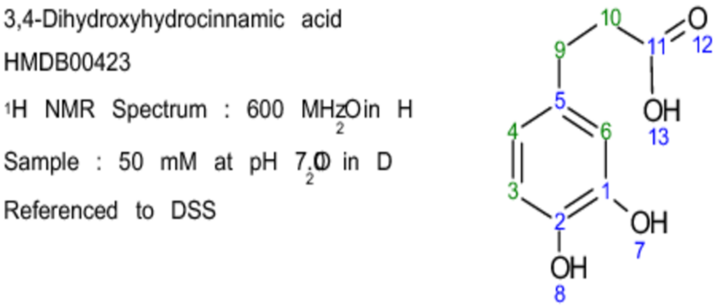


Table of Assignments

No.	Atom	Exp. Shift (ppm)	Multiplet
1	3	6.89	M01
2	6	6.82	M02
3	4	6.72	M03
4	10	3.92	M04
5	9	3.14	M05
6	9	2.98	M06

Figure 1: Table of assignments of chemical shifts of hydrogen connected to carbon atoms. The numbers in the Atom column refer to carbon atom positions drawn in Figure 2



3,4-Dihydroxyhydrocinnamic acid

HMDB00423

^{1}H NMR Spectrum : 600 MHz in H_2O

Sample : 50 mM at pH 7.0 in D_2O

Referenced to DSS

Figure 2: Drawn structure of HMDB00423 with carbon and oxygen atoms numbered in figure. These numbers are mapped to chemical shifts of hydrogen in Figure 1

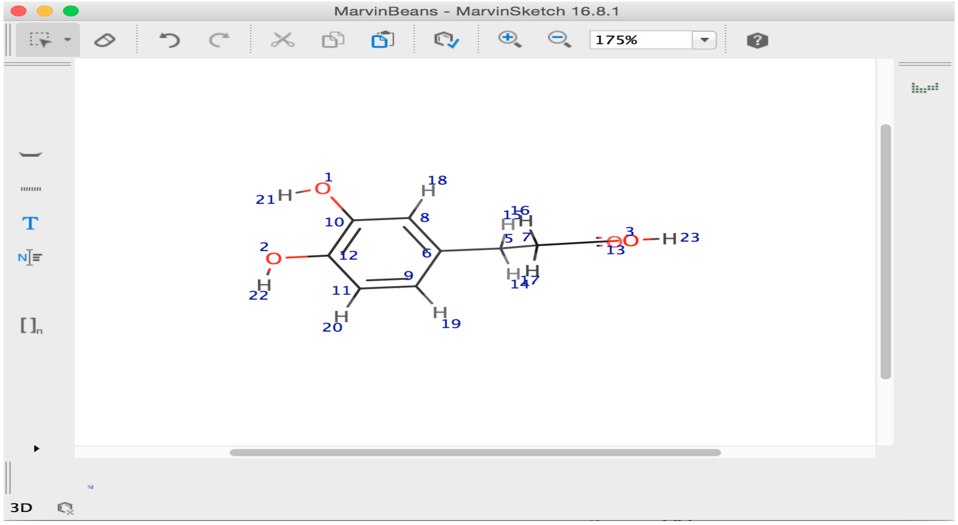


Figure 3: Structure of HMDB00423 downloaded from HMDB with atom numbering in Marvin Sketch

Figure 4: The process of manual labeling for HMDB00423. Map the atoms of Figure 2 and Figure 3 and replace the Atom column in Figure 1 with atomic positions from Figure 3. Atomic features are calculated from structure in Figure 3 since the structure from Figure 2 is not found in HMDB, it only exists as a figure

the chemical shift of jth instance, $\hat{y_j}$ is the predicted chemical shift of the jth instance in the dataset and N is the total number of instances or hydrogen atoms in the dataset.

$$\text{MAE} = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y_j}| \tag{1}$$

6 Feature Selection

For each molecule, the downloaded SDF files from HMDB were used to generate atomic features using the Chemical Development Kit developed by Steinbeck et. al. [16]. In this kit, there is a library called QSAR for calculation of atomic and molecular properties. We used this library to generate 29 atomic features for each atom present in the molecule. Figure 5 shows the first 14 atomic features and Figure 6 shows the rest. There was an additional unnamed feature that made the total features for each atom, 29. In order to represent the surrounding of the hydrogen atom, we used all 29 features of that hydrogen atom and the 29 features of the 3 nearest atoms. Therefore, each data instance will have a total of 29 X 4 = 116 atomic features. In the feature space the hydrogen atom's first 29 features come first, followed by closest atom's 29 features, followed by 2nd closest atoms 29 features and finally followed by the 3rd closets atom's 29 features. Table 1 shows how the feature space is aligned for each instance. All features are numeric.

Table 1: The sequence of all 116 features that are calculated from CDK package

Dataset Instance	Feature Space			
	29 features of H atom	29 features of 1st nearest atom	29 features of 2nd nearest atom	29 features of 3rd nearest atom
1	...	...	...	...
2	...	...	...	...
1022	...	...	...	...

6.1 Automated Feature Selection

Using the 116 calculated atomic features for each data instance, we also implement two different algorithms of feature dimensionality reduction namely Principal Component Analysis [17] and Correlation Feature Selection Subset Evaluation [18]. Both algorithms were implemented from the Weka Package [19].

6.1.1 Principal Component Analysis

Principal Component analysis is an orthogonal linear transformation on variables to reduce its dimension so that the principal component has the greatest variance. We used a variance covered of 1% to find all the components and selected them as features. Thus, our feature space underwent dimensionality reduction from 116 to 74.

6.1.2 Correlation Feature Selection

Correlation Feature Selection or CFS is another feature selection technique that selects the best features that are highly correlated to the classes. These features are less correlated to each other. After implementing CFS subset Evaluation, the number of features were reduced from 116 to 33.

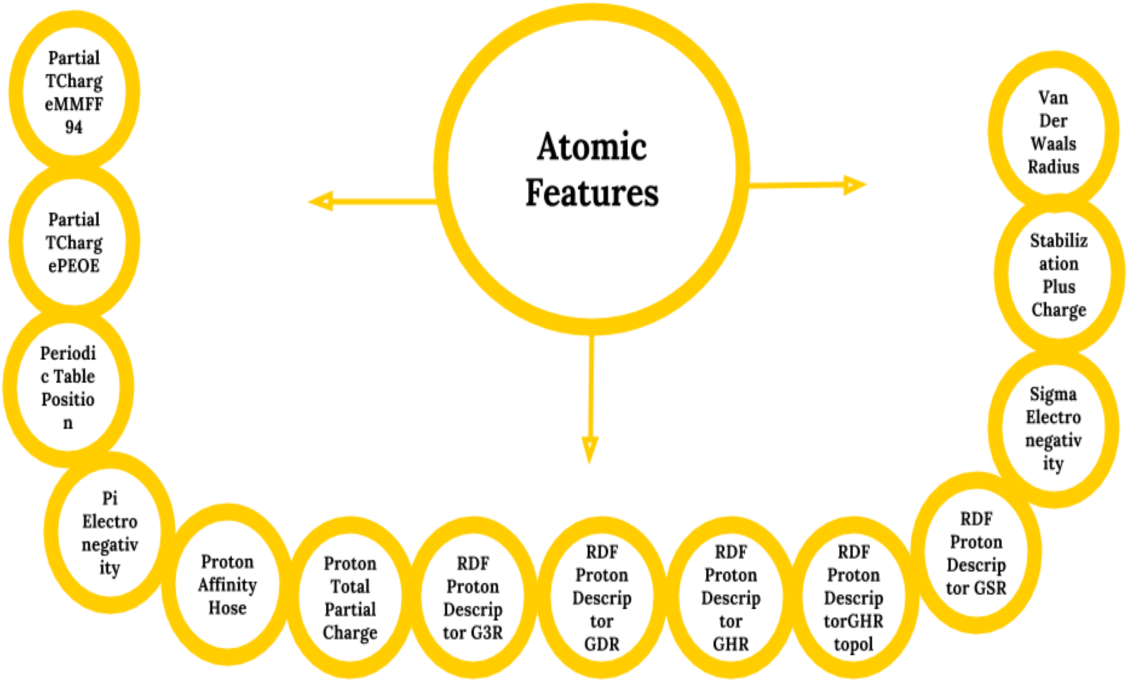

Atomic Features
Partial TChargeMMFF94
Partial TChargePEOE
Periodic Table Position
Pi Electronegativity
Proton Affinity Hose
Proton Total Partial Charge
RDF Proton Descriptor G3R
RDF Proton Descriptor GDR
RDF Proton Descriptor GHR
RDF Proton DescriptorGHR topol
RDF Proton Descriptor GSR
Sigma Electronegativity
Stabilization Plus Charge
Van Der Waals Radius

Figure 5: One subset of atomic features that are calculated from a structure using CDK package. All are numeric.

7 Methodology

We tried 1 trial, 10 fold cross validation of the entire training set for selecting the best performing algorithm. The training set was randomly shuffled by a particular seed and this was used to evaluate all the different algorithms. After choosing the best algorithm, it was tested on a hold out set and compared against De sousa's algorithm [8] using mean absolute error defined in evaluation section.

7.1 Classification

7.1.1 Random Forest

Random Forests [20] are bagged decision trees that reduce the problem of overfitting by using multiple decision trees [21] on random subset of features. The class that occurs the most in all the trees is the predicted class. Using random forest, we achieved a mean absolute error of 0.21 ppm on the 10 fold cross validation. Bagged and boosted [22] random forests did not change the error. The performance did not improve with features that are automatically selected by PCA or CFS subset evaluation.

7.1.2 Decision Tree J48

J48 [23] is a decision tree algorithm based on C4.5 [24]. It chooses the best feature at each node based on the information gain. The feature with the highest information gain is chosen to split the data at that node. At the leaf of the tree, it outputs the predicted class for the instance. J48 achieved a mean absolute error of 0.27 ppm. Bagged and boosted J48 got lower error of 0.24 ppm. This is done using 116 features selected initially. Using automated feature selection with PCA and CFS subset eval did not decrease the mean absolute error on 10 fold cross validation.

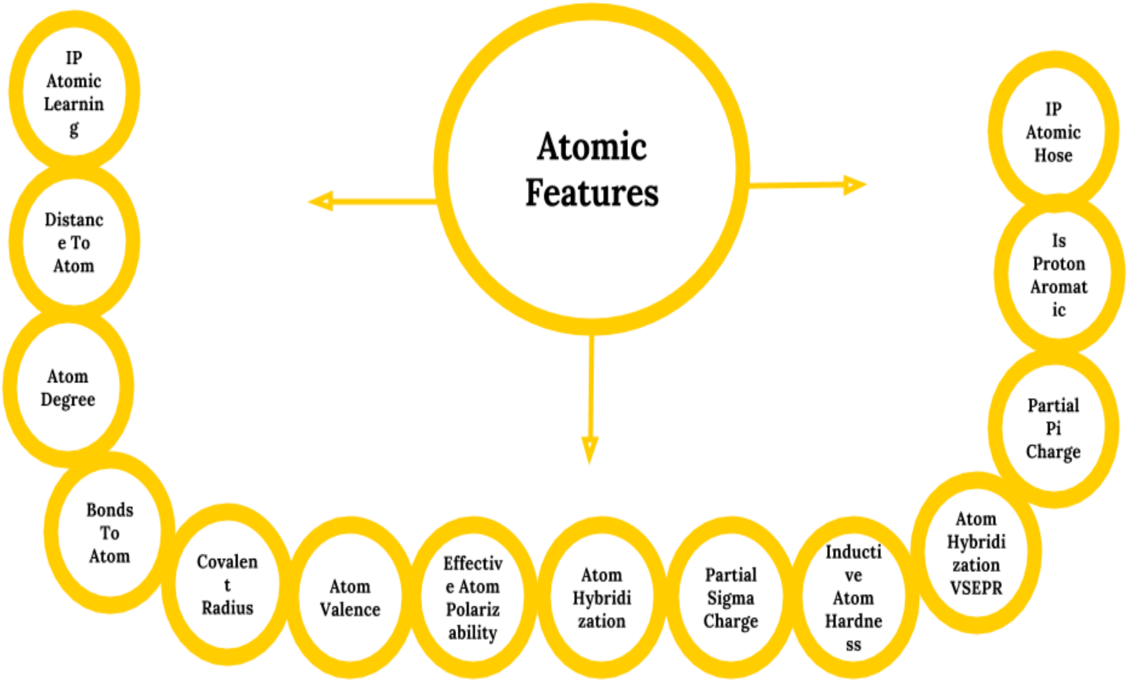


Figure 6: The other subset of atomic features that are calculated from a structure using CDK package. All are numeric

7.1.3 Artificial Neural Networks

Artificial Neural Networks [25] are non-linear models that maps features by transforming it in hidden layers with weights and finally producing an output that gives the probabilities of all the classes involved in the classification task. We tried MultiLayer Perceptron model from WEKA framework to define an artificial neural network with 4 hidden layers, starting from 116 nodes in its first 2 layers to 50 nodes in its last 2. The learning rate was 0.1, momentum was 0.2 and the total number of Epochs was 500. 10 fold Cross validation took longer time than other algorithms with around 2 hours to complete 1 trial. The CV accuracy was 0.31 ppm with 116 features. Using CFS subset eval or PCA algorithm did not improve the performance of the ANN.

7.1.4 Support Vector Machines

A Support Vector Machine [26] 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, it optimizes the hyperplane, separating the classes as much as possible. The best results with SVM were obtained by using a radial kernel. Cost and gamma were tuned by the algorithm in WEKA package. The 10 fold cross validation error was 0.3 ppm.

7.1.5 Naive Bayes

Naive Bayes [27] is a probabilistic classifier that makes the assumption that the features are independent of each other. Based on this assumption, it applies Bayes theorem to calculate the probability of a class given the features. The class with the highest probability is chosen as the predicted class. With Naive Bayes, the prediction on our problem was not accurate enough. It produced a 10 fold cross validation mean absolute error of 0.45 ppm which is the highest among all other classifiers. With CFS subset eval features, the results did not improve.

7.2 Regression

7.2.1 Artificial Neural Networks

Just like in classification, 4 layers were stacked to create an artificial neural network for regression. The first two layers had 116 nodes, the number of features, and they were reduced to 50 nodes each for the last two layers. The final layer maps 50 nodes to 1 node which gives the real valued chemical shift of the hydrogen atom it is predicting. The learning rate was 0.1, the momentum 0.2 and the number of epochs 500. The 10 fold cross validation mean absolute error of 0.33 ppm for regression using ANN was higher than for classification with the same parameters, meaning classification outdid regression in predicting NMR chemical shifts of hydrogen. The time taken for 10 fold cross validation for regression was not as long as classification with ANN.

7.2.2 Linear Regression

Linear regression [28] is a mapping of linear combination of feature variables into an outcome variable. A best fit line is drawn to map the relationship between feature variables and outcome variable. Our outcome variable is the chemical shift we are predicting. 10 fold cross validation on linear regression registered a mean absolute error of 0.39 ppm, higher than the classification algorithms that performed better for our task.

7.3 NMR Simulation

Using Spinach matlab package [29], we developed an NMR simulation module that takes a list of chemical shifts for hydrogen atoms along with J-coupling constants to create a liquid 1D NMR system for protons. Using parameters zerofill of 2,048, offset of 1,200, sweep of 12,000, npoints of 30,000 and a magnetic field of 12 Hz, It creates the fid for the molecule. Using the fid, we perform crisp-1d apodization and finally, a fast fourier transform to create the spectrum for the NMR system. This spectrum is then displayed on a graph of intensities vs ppm as any normal NMR 1d graph.

We decided to give every couplings between hydrogen atoms a value of 4 Hz. It should be noted that the graphs will look more original if original J-coupling values were learned and used. De Sousa's algorithm [8] uses NMR simulation software developed by Castillo et. al. [30] and is much faster in computing spectra compared to our simulation module.

8 Results

Random Forest performed the best with a mean absolute error of 0.21 ppm on 10 fold cross validation and was therefore, selected for testing hold out set. In the hold out set, our algorithm, NmrPred, beat De Sousa's on mean absolute error by only 0.04 ppm. Testing the two algorithms on a larger hold out set with more variety in molecules will result in more meaningful results.

8.1 Cross Validation

Figure 7 shows the comparison of 10-fold cross validation errors between different algorithms for NmrPred. Random Forest's mean absolute error of 0.21 ppm beats all other algorithms' performance measure. Bagged Decision Tree J48 came in closest with mean absolute error of 0.24 ppm. Other algorithms may have overfit the training data. This comparison is made for training with 116 features, the initial feature space.

Figure 8 shows a scatter plot of predicted values against true values when using Random Forest during a fold of the 10 fold cross validation trial. Figure 10 shows a similar scatter plot for Decision Tree J48 for the same test fold. Comparing the two graphs, points in random forests are less scattered than points in Decision Tree J48. Figure 9 and Figure 11 shows training set scatter plot for random forests and decision tree j48 respectively. Training error for random forest is 0.03 ppm and cross validation error is around 0.21 ppm which suggests to some degree of overfitting. Training mean absolute error for decision tree j48 is 0.13 ppm, more than that of random forest.

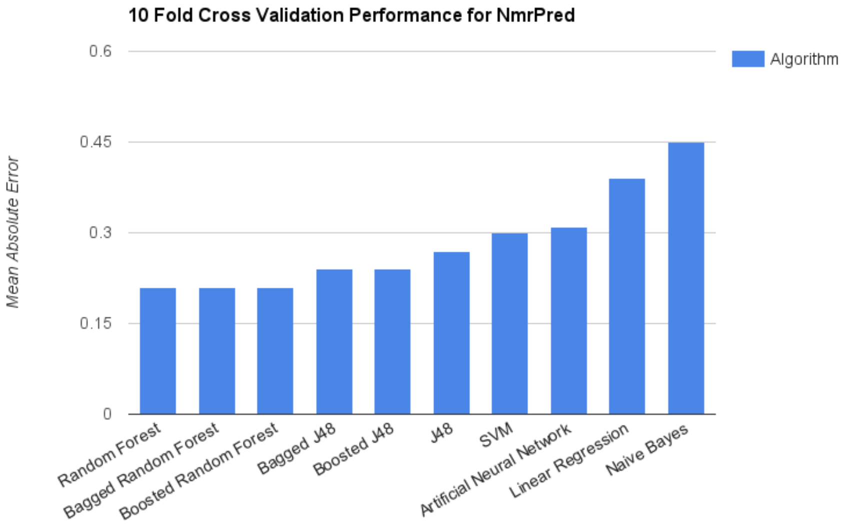


Figure 7: Comparison of all algorithms' 10 fold cross validation performance. Random Forest does the best with the least mean absolute error of 0.21 ppm. Naive Bayes performs the worst with a mean absolute error of 0.45 ppm. The 10 fold cross validation is carried on the same training set shuffled randomly with the same seed.

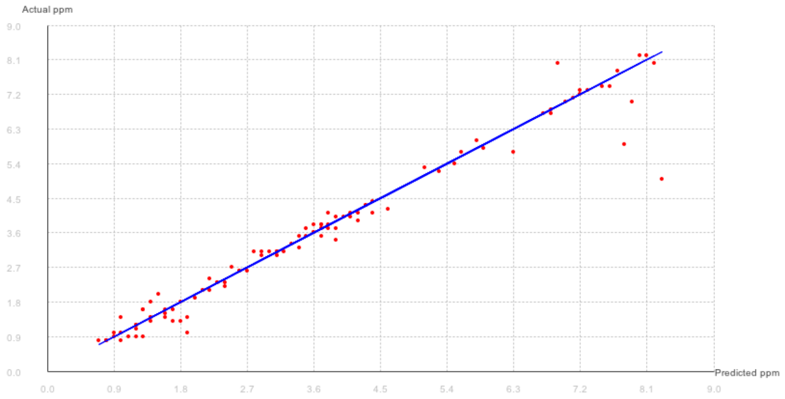


Figure 8: A graph of predicted chemical shifts vs true chemical shifts using Random Forest on the final fold of 10 fold cross validation. This fold gave a mean absolute error of 0.196 ppm. Some test instances around 9.0 ppm have larger error since there are more scattered points around the optimal line.

8.1.1 CFS Subset Evaluation

Feature dimensions were reduced by applying Cfs Subset Eval algorithm described in Feature Selection section. The best features that represent the data were selected and using these features, all the algorithms were run on the 10 fold cross validation. Figure 12 shows how the mean absolute error changed between algorithms. Again, random forests proved to outperform with the reduced feature space. Figure 13 shows a comparison between the performance of two different feature selection methods. Using original 116 features outperforms using CFS reduced features for all algorithms.

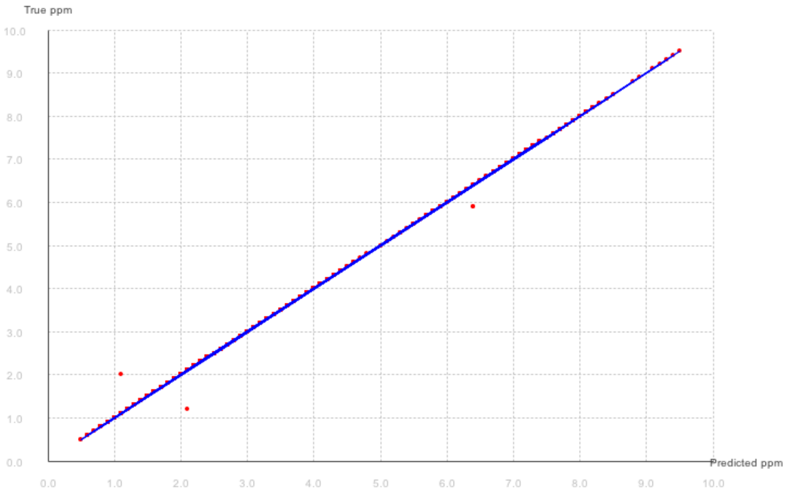


Figure 9: A graph of predicted chemical shifts vs true chemical shifts using Random Forest on the training set. Testing on the training set gave a mean absolute error of 0.03 ppm. It may suggest overfitting due to the very low absolute error.

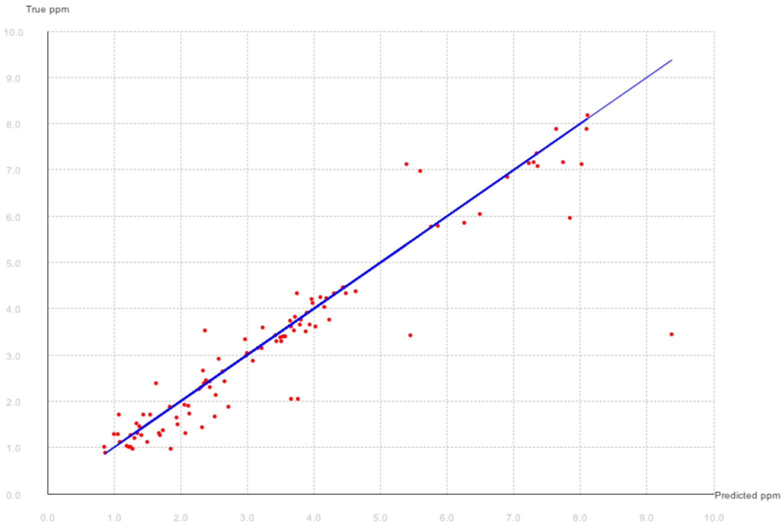


Figure 10: A graph of predicted chemical shifts vs true chemical shifts using J48 on the final fold of 10 fold cross validation. This fold gave a mean absolute error of 0.28 ppm while Random Forest gave a mean absolute error of 0.196 ppm on the same fold.

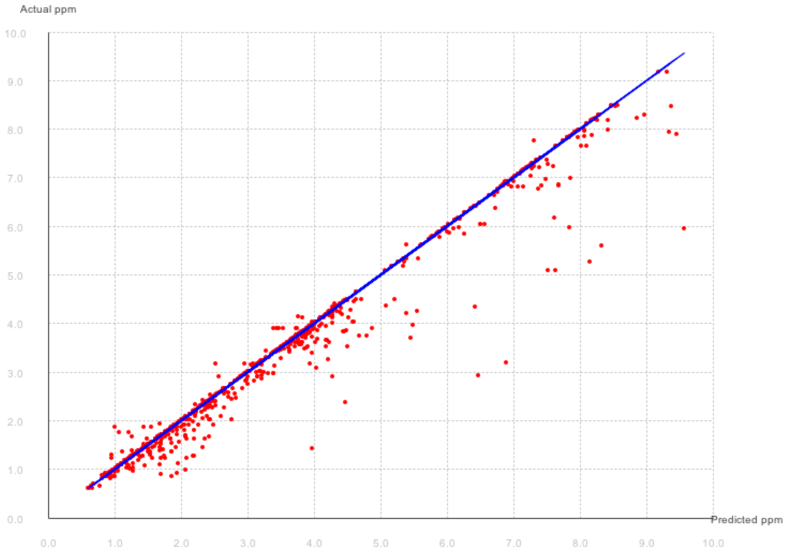


Figure 11: A graph of predicted chemical shifts vs true chemical shifts using J48 on the training set. Testing on the training set gave a mean absolute error of 0.13 ppm.

8.1.2 Principal Component Analysis

Figure 14 shows the resuls with PCA derived features on a subset of algorithms. Again, as observed with using CFS eval subset, the principal components did not decrease the mean absolute error for any of the algorithms. In fact, PCA components selected as features did the worst. Random Forest with PCA's 74 features was found to have the highest mean absolute error of 0.31 ppm. J48's mean absolute error shot to 0.69 ppm. Due to the increase in mean absolute error with these two algorithms, we decided not to use PCA features with other algorithms.

Looking at the results with automated feature selection, it seems like reducing the number of features may not be the best idea. We may need to try other feature selection methods like MrMr to see if selecting a subset of features improves performance for the algorithms. Playing around with different parameters of CFS and PCA may help improve the results.

8.2 Hold Out Set

Figure 15 shows the comparison of mean absolute errors for each of the 10 molecules in the hold out set. It also shows the average mean absolute error for all 88 hydrogen atoms in the 10 molecules in hold out set. NmrPred perform better on average with a mean absolute error of 0.13 ppm compared to De Sousa's algorithm's error of 0.17 ppm. However, there are 2 molecules out of the 10 molecules over which Sousa's algorithm performed better. HMDB00421 and HMDB00439 are two molecules that NmrPred does poorly over. Figure 17 and Figure 18 shows the structures of these two molecules. However, with test molecule HMDB00423, both NmrPred and De Sousa's algorithm performed very poorly with high mean absolute errors of 0.48 ppm and 0.57 ppm respectively. This is higher than the average mean absolute error over hold out set.

8.2.1 Comparing Aliphatic and Aromatic

Both HMDB00439 and HMDB00423 contain aromatic rings. Predicting may become more difficult with aromatic rings in them. Figure 20 and Figure 21 Creffig:HMDB00440 shows molecules that

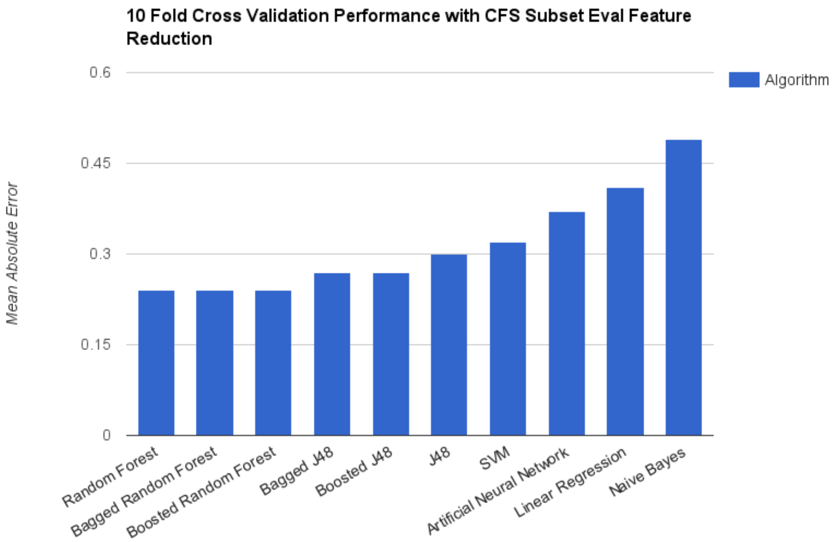


Figure 12: Comparison of 10 fold Cross Validation performance between algorithms with features extracted by CFS Subset Evaluation. Again, Random Forest performs the best with the lowest mean absolute error of 0.24 ppm.

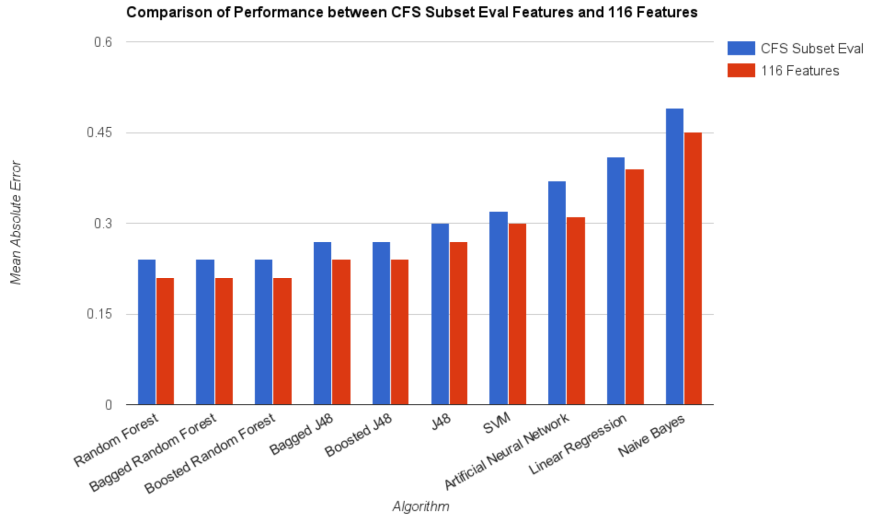


Figure 13: Comparison of performance of all algorithms with 116 features and CFS extracted features. For every algorithm, using CFS extracted features performs worse compared to using 116 features.

are very well predicted by NmrPred with mean absolute errors of 0.04 ppm, 0.08 ppm and 0.11 ppm respectively. HMDB00413 is aliphatic while HMDB00434 and HMDB00440 are aromatic molecules. HMDB00421 is an aliphatic molecule with larger mean absolute error than both HMDB00434 and HMDB00440. Figure 16 shows the mean absolute errors by NmrPred on aliphatic and aromatic molecules. Although the average mean absolute error on aliphatic molecules is less than aromatic ones, which suggests to difficulty in predicting chemical shifts in aromatic molecules, we are only dealing with a 10 molecule hold out set. We will need to have a larger hold out set with more aliphatic and aromatic molecules to see if there is a major difference in mean absolute errors.

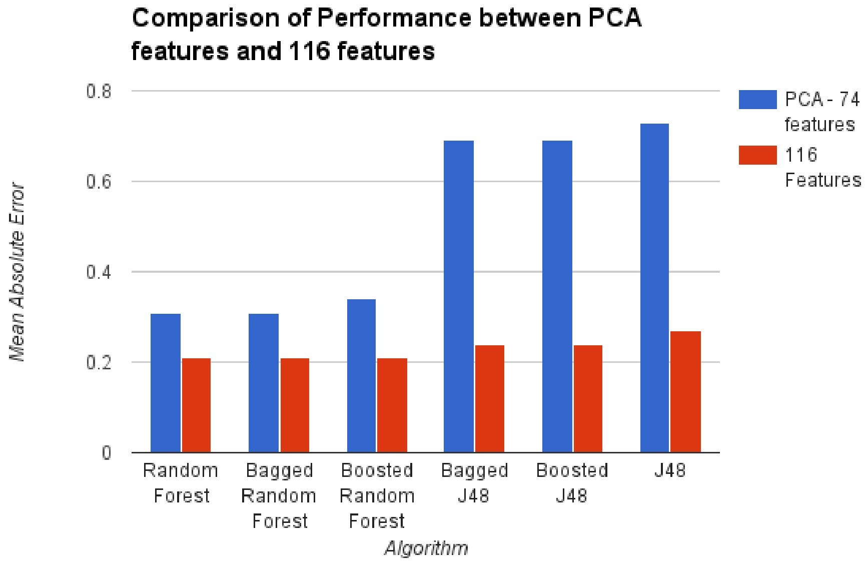


Figure 14: Comparison of performance of all algorithms with 116 features and PCA extracted features. For every algorithm, using PCA extracted features performs worse compared to using 116 features. PCA extracted features are even worse than CFS extracted features.

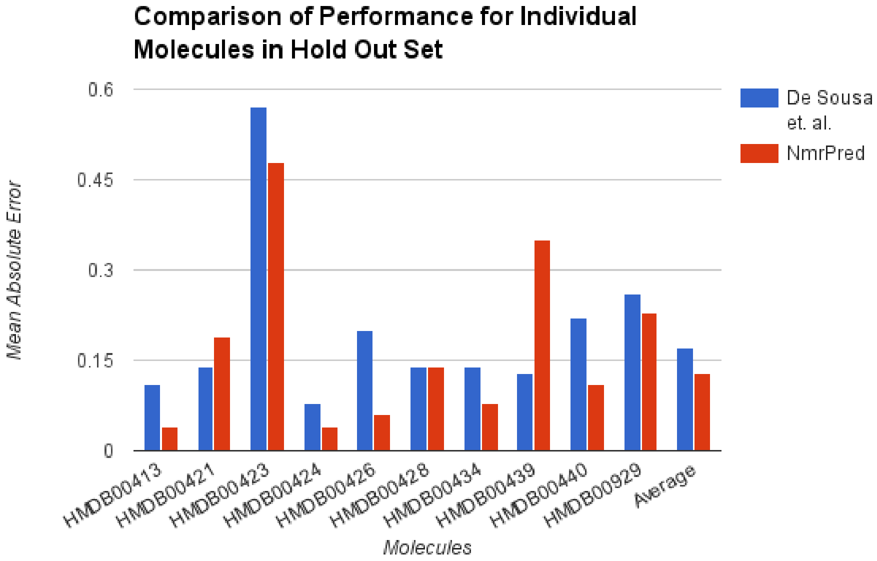


Figure 15: Comparison between NmrPred and De Sousa's algorithm on the hold out set. NmrPred was at least as good as De Sousa's algorithm on all molecules except HMDB00421 and HMDB00438. On average, NmrPred beats De Sousa's algorithm by 0.03 ppm.

8.2.2 Comparing Nmr Simulation

Figure 26 shows a comparison of NMR spectra created by NmrPred, De Sousa's algorithm and true NMR experiment with HMDB000421. The large spike near 0 ppm is similar to all three graphs but after that things vary for different images. Figure 30 a similar comparison done on HMDB00423. HMDB00423 is not a properly predicted molecule by either algorithms and therefore, the variation between spectra is quite large.

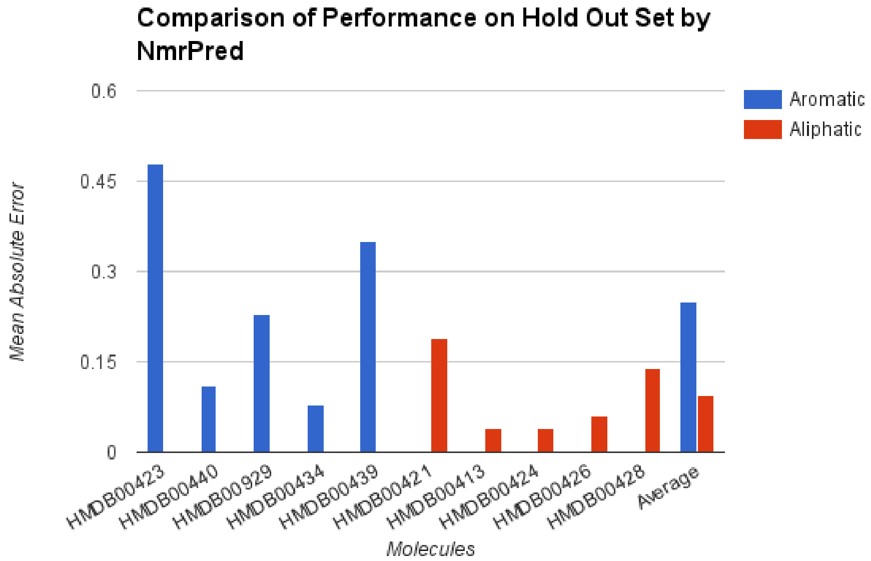


Figure 16: Comparison of aliphatic and aromatic molecules' mean absolute error. On average, chemical shift prediction with aliphatic molecules is more accurate than aromatic molecules looking at the result of 10 hold out set molecules.

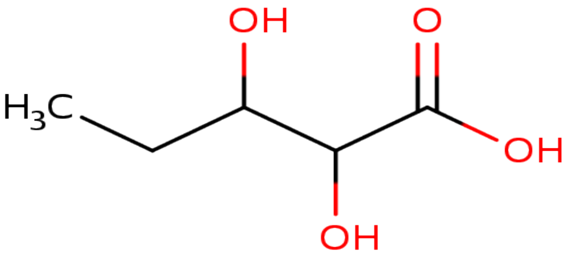


Figure 17: The downloaded structure image of HMDB00421 from HMDB.

There are several factors when comparing NMR spectra between experiments. The magnetic field is important in identifying peak intensity. The magnetic field may vary between experiments. NmrPred works with a magnetic field of 12 Hz and zerofill of 2048. It may be that De Sousa's algorithm is creating spectra for parameters like magnetic field and zerofill. Similarly, the real spectra may have been simulated by a software with different parameters. The fact that NmrPred has constant J-coupling constants of 4 Hz produces a different shape for the peaks and thereby, a different spectra altogether.

9 Conclusion

We developed an application that takes a structure of a molecule, predicts the chemical shifts of hydrogen in water in 1 dimensional NMR. With constant J-couplings of 4 Hz, the chemical shifts

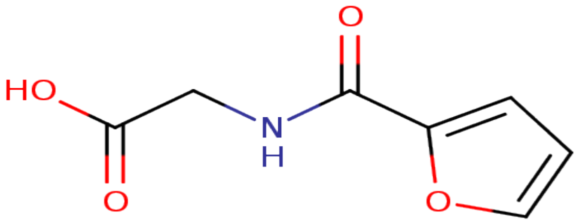


Figure 18: The downloaded structure image of HMDB00439 from HMDB.

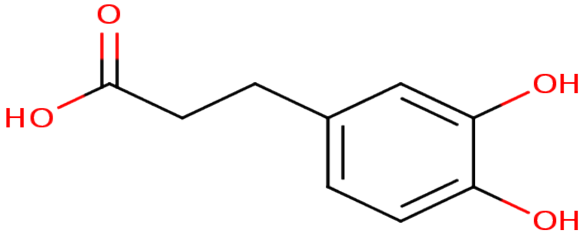


Figure 19: The downloaded structure image of HMDB00423 from HMDB. NmrPred performed poorly on this molecule.

and J-couplings are piped into a liquid nmr system that creates a NMR spectra of intensities against chemical shifts. Although our algorithm performed better than a popular algorithm for a hold out set of 10 molecules, it is worth noting that with a larger hold out test set, results may vary.

Our dataset set was also considerably small due to the fact that we had to perform manual labeling of atomic positions in order to map chemical shift values to the correct atoms. This may have introduced error since manual labeling is not entirely perfect and reliable. With multiple people doing manual labeling, the true chemical shift values may be cross checked and made more reliable. With a larger training set and hold out set, the results may also be more reliable.

Random Forest beats all other classifiers and regressors that we tried to model for NmrPred. It is less prone to overfitting that other decision trees like J48. Future work may include working with more algorithms and an in-fold cross validation to tune parameters for different algorithms. Fine tuning may result in less error.

It is also worth noting that for the classification problem, 1000 classes were used and not all classes had training instances. If we could manage a more balanced training set with multiple instances for

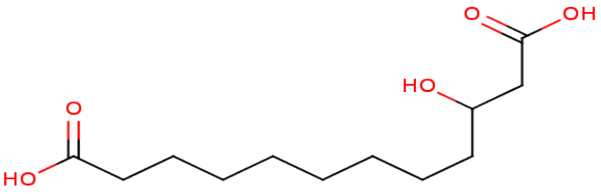


Figure 20: The downloaded structure image of HMDB00413 from HMDB. NmrPred performed really well on this molecule.

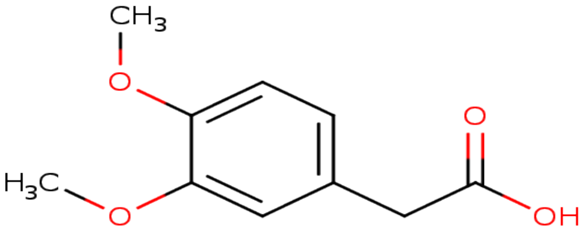


Figure 21: The downloaded structure image of HMDB00434 from HMDB.

all classes, classification can be made more accurate. It is difficult to find instances for all numbers between 0.10 to 9.99 ppm with each 0.01 ppm being a class.

In order to create a better NMR graph, J-coupling constants may need to learned either by machine learning or database knowledge. We assigned all constants to be 4 Hz. NMR simulation is also considerably slower compared to De Sousa's software. It might be worth looking into optimizing the NMR spectra creation process in matlab to reduce time taken to render a 1D NMR image. Optimizing the process will be helpful for future release of software as an embedded server.

With more features that affect chemical shifts of hydrogen, results might even be better. For example, there is no atomic feature that represents chiral centre in molecules. Chiral centres affect NMR spins. More research needs to be conducted to find out all important features that influence NMR spectra in water. With the application of proposed changes to the algorithm, it is possible to make NMR prediction more accurate.

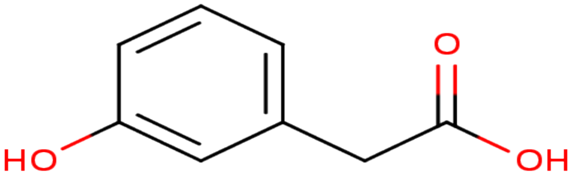


Figure 22: The downloaded structure image of HMDB00440 from HMDB.

References

[1] Edward Raymond Andrew. Nuclear magnetic resonance. *Nuclear Magnetic Resonance, by ER Andrew, Cambridge, UK: Cambridge University Press, 2009*, 1, 2009.

[2] K Hallenga. Fourier transform nmr. In *Modern NMR techniques and their application in chemistry*. 1991.

[3] Horst Friebolin and Jack K Becconsall. *Basic one-and two-dimensional NMR spectroscopy*. VCH Weinheim, 1993.

[4] Hans-Otto Kalinowski, Stefan Berger, and Siegmar Braun. Carbon-13 nmr spectroscopy. 1988.

[5] Hugo E Gottlieb, Vadim Kotlyar, and Abraham Nudelman. Nmr chemical shifts of common laboratory solvents as trace impurities. *The Journal of organic chemistry*, 62(21):7512–7515, 1997.

[6] Ernoe Pretsch, Philippe Bühlmann, and Christian Affolter. 1h nmr spectroscopy. In *Structure Determination of Organic Compounds*, pages 161–243. Springer, 2000.

[7] Stefan Kuhn, Björn Egert, Steffen Neumann, and Christoph Steinbeck. Building blocks for automated elucidation of metabolites: Machine learning methods for nmr prediction. *BMC bioinformatics*, 9(1):1, 2008.

[8] João Aires-de Sousa, Markus C Hemmer, and Johann Gasteiger. Prediction of 1h nmr chemical shifts using neural networks. *Analytical chemistry*, 74(1):80–90, 2002.

[9] Damiano Banfi and Luc Patiny. www. nmrdb. org: Resurrecting and processing nmr spectra on-line. *CHIMIA International Journal for Chemistry*, 62(4):280–281, 2008.

[10] Yuri Binev and João Aires-de Sousa. Structure-based predictions of 1h nmr chemical shifts using feed-forward neural networks. *Journal of chemical information and computer sciences*, 44(3):940–945, 2004.

[11] Yuri Binev, Maria MB Marques, and João Aires-de Sousa. Prediction of 1h nmr coupling constants with associative neural networks trained for chemical shifts. *Journal of chemical information and modeling*, 47(6):2089–2097, 2007.

[12] Horst Kessler, Christian Griesinger, Joerg Lautz, Arndt Mueller, Wilfred F Van Gunsteren, and Herman JC Berendsen. Conformational dynamics detected by nuclear magnetic resonance noe values and j coupling constants. *Journal of the American Chemical Society*, 110(11):3393–3396, 1988.

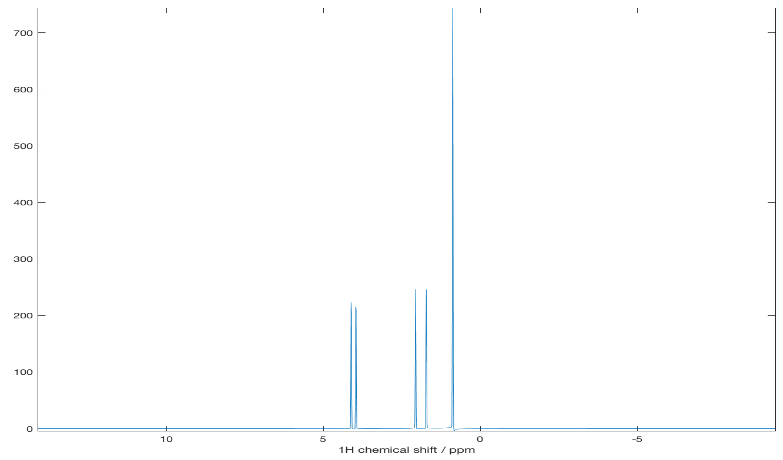


Figure 23: Simulated NMR Spectra for HMDB00421 by NmrPred

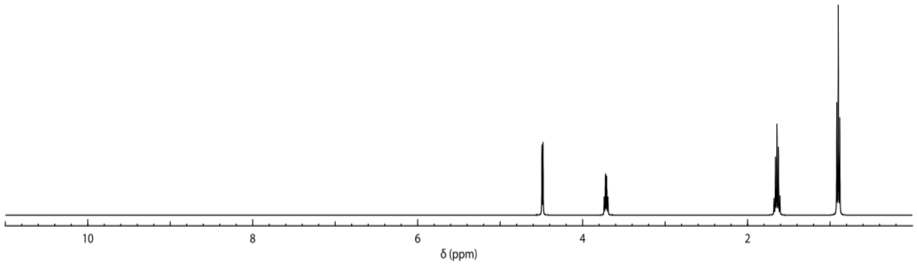


Figure 24: Simulated NMR Spectra for HMDB00421 by De Sousa's algorithm. The intensity labeling on the vertical axis is ignored by their algorithm

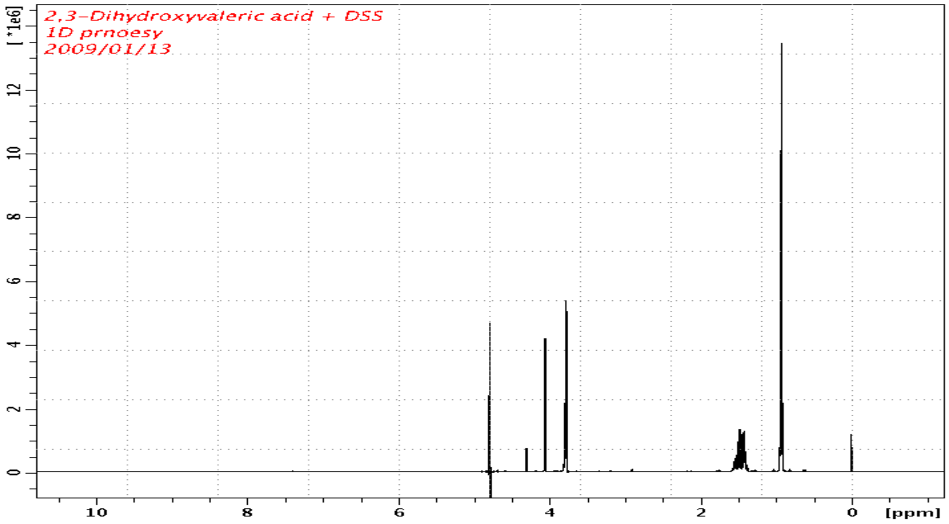


Figure 25: True NMR Spectra for HMDB00421 from HMDB

Figure 26: Comparison of NMR spectra images generated by NmrPred, Figure 23, and De Sousa's algorithm, Figure 24, against the true NMR Spectra image in Figure 25. Some peak intensities match while others do not.

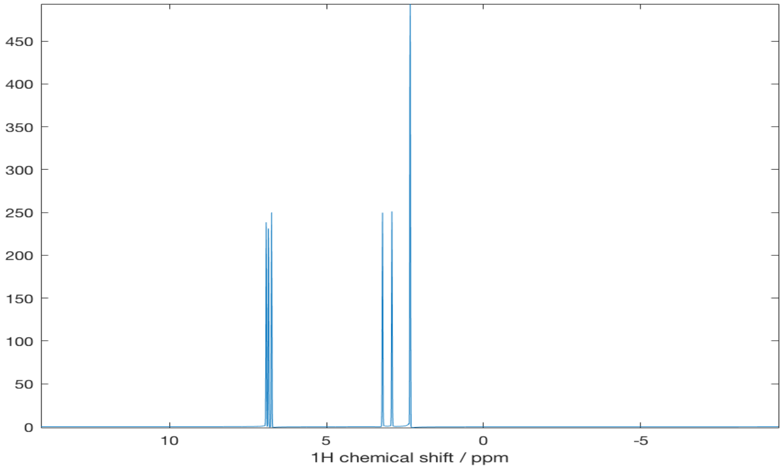


Figure 27: Simulated NMR Spectra for HMDB00423 by NmrPred

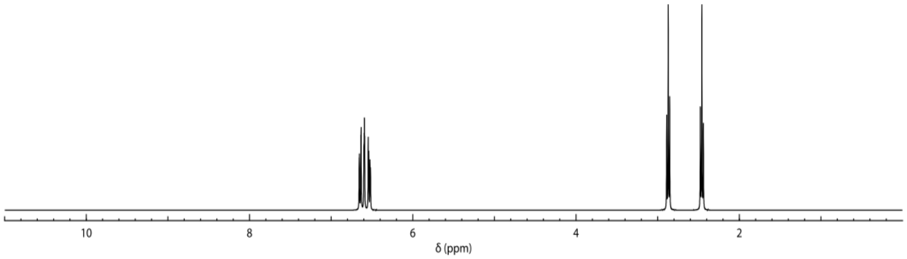


Figure 28: Simulated NMR Spectra for HMDB00423 by De Sousa's algorithm. The intensity labeling on the vertical axis is ignored by their algorithm.

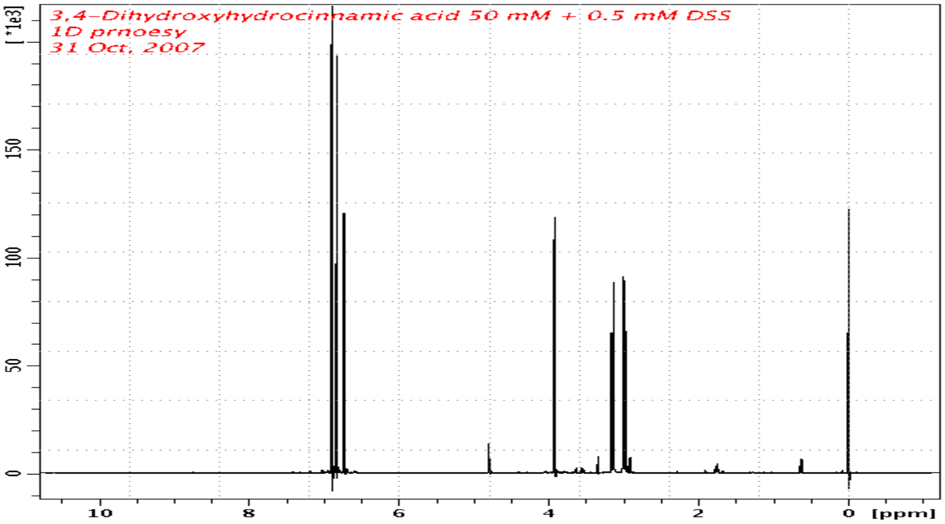


Figure 29: True NMR Spectra for HMDB00423 from HMDB

Figure 30: Comparison of NMR spectra images generated by NmrPred, Figure 27, and De Sousa's algorithm, Figure 28, against the true NMR Spectra image in Figure 29. Both algorithms performed poorly on HMDB00423 and therefore, the spectra will not match.

[13] Antony Williams, Brent Lefebvre, and Ryan Sasaki. Putting acd/nmr predictors to the test, 2006.

[14] David S Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, et al. Hmdb 3.0—the human metabolome database in 2013. *Nucleic acids research*, page gks1065, 2012.

[15] Ferenc Csizmadia. Jchem: Java applets and modules supporting chemical database handling from web browsers. *Journal of Chemical Information and Computer Sciences*, 40(2):323–324, 2000.

[16] Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. 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, 2003.

[17] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[18] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[19] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[20] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[21] J. Ross Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[22] Ludmila I Kuncheva. Bagging and boosting. *Combining Pattern Classifiers: Methods and Algorithms*, pages 203–235.

[23] Neeraj Bhargava, Girja Sharma, Ritu Bhargava, and Manish Mathuria. Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, 3(6), 2013.

[24] J Ross Quinlan. Bagging, boosting, and c4. 5. In *AAAI/IAAI, Vol. 1*, pages 725–730, 1996.

[25] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[26] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[27] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[28] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 936. John Wiley & Sons, 2012.

[29] HJ Hogben, M Krzystyniak, GTP Charnock, PJ Hore, and Ilya Kuprov. Spinach–a software library for simulation of spin dynamics in large spin systems. *Journal of Magnetic Resonance*, 208(2):179–194, 2011.

[30] Andrés M Castillo, Luc Patiny, and Julien Wist. Fast and accurate algorithm for the simulation of nmr spectra of large spin systems. *Journal of Magnetic Resonance*, 209(2):123–130, 2011.