


5-2021

Molecular Cluster Fragment Machine Learning Training Techniques to Predict Energetics of Brown Carbon Aerosol Clusters

Emily E. Chappie
William & Mary

Follow this and additional works at: <https://scholarworks.wm.edu/honorsthesis>

 Part of the [Data Science Commons](#), [Environmental Chemistry Commons](#), [Other Chemistry Commons](#),
and the [Other Computer Sciences Commons](#)

Recommended Citation

Chappie, Emily E., "Molecular Cluster Fragment Machine Learning Training Techniques to Predict Energetics of Brown Carbon Aerosol Clusters" (2021). *Undergraduate Honors Theses*. William & Mary. Paper 1622.

<https://scholarworks.wm.edu/honorsthesis/1622>

This Honors Thesis -- Open Access is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.

Molecular Cluster Fragment Machine Learning Training Techniques to Predict Energetics of
Brown Carbon Aerosol Clusters

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Science in Chemistry from
William & Mary

by

Emily Elisabeth Chappie

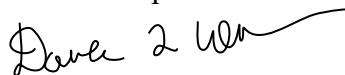
Accepted for Honors in Chemistry



Dr. Nathanael M. Kidwell, Director



Dr. Randolph A. Coleman



Dr. Dana L. Willner



Dr. Daniel P. Tabor (Texas A&M University)

Williamsburg, VA

May 3, 2021

ABSTRACT

Density functional theory (DFT) has become a popular method for computational work involving larger molecular systems as it provides accuracy that rivals *ab initio* methods while lowering computational cost. Nevertheless, computational cost is still high for systems greater than ten atoms in size, preventing their application in modeling realistic atmospheric systems at the molecular level. Machine learning techniques, however, show promise as cost-effective tools in predicting chemical properties when properly trained. In the interest of furthering chemical machine learning in the field of atmospheric science, I have developed a training method for predicting cluster energetics of newly characterized nitrogen-based brown carbon aerosols that can undergo tautomerization. By creating a training dataset of cluster fragment and functional group DFT calculations, I can effectively train machine learning models to predict overall energetics of previously unknown brown carbon clusters while improving computational efficiency.

TABLE OF CONTENTS

Acknowledgements	iii
List of Tables	iv
List of Figures	v
Chapter 1. INTRODUCTION	1
1A. Brown Carbon	1
1B. Density Functional Theory and Machine Learning	1
1C. Fragment Background	4
1D. Cluster Background	5
1E. Fragments for Training	6
Chapter 2. ABOUT THE DATA	8
2A. Train-Test Splits	8
2B. Molecular Structure Overview	8
2C. Data Curation	9
Chapter 3. METHODS	11
3A. DFT Methods	11
3B. Machine Learning Methods	11
3C. Data Collected	13
Chapter 4. RESULTS AND CONCLUSION	17
4A. Metrics	17
4B. Model Performances	17
4C. Feature Importance Ranking	20
4D. Conclusion	22
Chapter 5. FUTURE WORK	24

Supporting Information	25
Bibliography	32

ACKNOWLEDGEMENTS

Professor Kidwell, thank you for taking a chance on me by adding me to your group. You have never ceased to believe that I can turn my scientific dreams into reality and have opened an immeasurable number of doors for me because of it. I will forever be grateful to you for putting me on the path I am on today. Professor Tabor, thank you for taking me on as a mentee. I greatly appreciate your support for my ideas and you sharing your knowledge, time, and enthusiasm throughout this project. Professor Willner and Professor Coleman, thank you for agreeing to be on my committee. Your encouragement and kindness towards me and my project have been very appreciated.

Tyler and Gabriel, you were a wonderful addition to my team and I am sincerely thankful for your help with running calculations. Jalen and Joseph, thank you for letting me be your mentor for your own senior capstone this year - I enjoyed every minute of it. Charles, thank you for carrying on with the work I have started here. All five of you have been a joy to work with and I look forward to hearing about your future endeavors. Brianna, David, and Chris - thank you for all of the laughs, advice, and support. And to the Kidwell Group as a whole, I cannot imagine a better group to start doing research with. Thank you all for a wonderful year and a half.

To my friends and grandparents, your willingness to hear about my project has not gone unnoticed. I am thankful for all of you and your encouragement. I could not have done this without such a wonderful group of people around me. To my parents, thank you for unconditionally being my biggest supporters. I could not have asked for a better family.

Lastly, thank you to William & Mary Research Computing and the Texas A&M University Department of Chemistry for providing me with computational resources and technical support for this work as a student and Visiting Scholar respectively.

LIST OF TABLES

1. Average R^2 and scaled RMSE values across all algorithm and training method combinations	20
2. Parent molecule structure information	25-26
3. Fragment structure information	27-29
4. Testing dataset extrapolated cluster configurations	30
5. Components of extrapolated clusters	31

LIST OF FIGURES

1. Deriving a modified sub-fragment from a parent molecule	5
2. Depiction of a cluster	6
3. Stereochemistry variations of a parent molecule	9
4. Diagram of a hypothetical random forest	12
5. Training data sources and flow	13
6. How to create a Coulomb matrix	16
7. Graph of percent error values across all algorithm and training method combinations for predicted energetics of extrapolated clusters	19
8. Fragment trained random forest feature rank	21
9. Cluster trained random forest feature rank	21
10. Fragment trained stochastic gradient descent feature rank	22
11. Cluster trained stochastic gradient descent feature rank	22

CHAPTER 1. INTRODUCTION

A. Brown Carbon

Brown carbon molecules are atmospheric chromophores that span the ultraviolet and visible spectra and are found naturally in aerosol clusters (1). While it is known that brown carbon comes from fuel and biomass burning, such as forest fires, as well as general atmospheric reactions, the composition of these clusters largely remains a mystery. Of the characterized molecules, however, most have been found to be polar, somewhat water-soluble, and containing reactive oxygen-based functional groups. A large number of brown carbon molecules are partially made of nitrogen (1).

The chromophoric nature of brown carbon aerosols is what makes them particularly important to understand. These small clusters are found throughout the troposphere where they have the ability to change light absorption patterns of other aerosols, causing large atmospheric implications. Specifically, brown carbon light absorption can cause water evaporation and inhibit photolysis based atmospheric reactions which respectively lead to atmospheric warming and a decrease in ozone (1). Considering the large scale effects these unknown molecules have on the atmosphere, more tools, such as this machine learning training technique, are needed to increase knowledge on brown carbon characteristics.

B. Density Functional Theory and Machine Learning

Density functional theory (DFT) is a powerful quantum mechanical computational technique that has been utilized here for effective model training. The accuracy of calculated results using DFT rivals that of traditional *ab initio* calculations (e.g. post-Hartree-Fock methods such as MP2) because of its quantum mechanical roots but it is able to significantly decrease

computational time in comparison. DFT also has the ability to calculate spectroscopic values due to its quantum mechanical approach, which produces more computationally derived unique identifiers and provides insight into chemical properties beyond energetics, both of which are valuable in model training (2). While only ground state calculations were performed for this research, it is worth noting that DFT is not as reliable a method for excited state calculations as the method does not accurately treat bond breaking, so further work modeling brown carbon excitation due to sunlight should consider another computational technique (2).

Machine learning is fundamentally based on math and has no physics foundation on which to base its predictions on. It may seem counterintuitive to utilize techniques not dependent on quantum mechanics to improve chemical predictive processes, but the customizable nature of machine learning makes it a powerful tool to be able to reduce the computational resources required. Also, relatively simple models with clear variable relationships can be used to elucidate which molecular features best predict energetics. Keeping in mind that correlation is not causation, finding molecular characteristic relationships is valuable for the machine learning community even though it is known how to solve for the values computationally.

Two main types of machine learning are supervised and unsupervised learning. Supervised learning models use background information on the data, such as what should be predicted, to analyze data while unsupervised learning decides its own path and gives descriptive results of data. Within the supervised learning realm there are two subtypes, regression and classification. Regression refers to predicting numeric values while classification aims to use known groupings of data to predict characteristics of similar but previously unseen data (3). *As the goal of this work is to develop models that can predict energy values of unknown brown carbon molecular*

clusters, supervised learning regression models have been developed. In this case, the types of regression models used are random forests and stochastic gradient descent.

A significant contributor to good predictive power and a robust machine learning model is a dataset made of accurate and appropriately varied data. If the machine learning algorithms are fitted to training data that does not correctly represent the sample space trying to be predicted, it is traditionally likely to perform poorly due to overfitting (4). Part of this challenge is ensuring that instances of certain phenomena are representative of their natural occurrence, which in a computational chemistry application means curating a dataset that has a representative sample of optimized geometries found with highly probable energetics. I have chosen to accomplish this by looking at optimized molecular and cluster structures at their energetic minima as they indicate the most stable and therefore most probable atomic arrangements of targeted brown carbon chromophores.

The challenge of needing to train models on accurate, representative datasets is where the combination of machine learning and DFT computational methods become strong. As ground state DFT methods rival the accuracy of *ab initio* calculations, often seen as the gold standard of computational chemistry, high quality data is being used to train and will likely lead to more accurate predictions. Furthermore, due to the speed of DFT relative to *ab initio* methods, it is possible to make a much larger and more representative dataset in the same amount of time.

While it may appear unorthodox to use computational resources with higher resource costs to create a ‘low cost’ machine learning tool, creating a high-quality dataset is an investment. Empirical computational chemistry modeling methods require representative, experimental datasets to estimate chemical values more accurately, and the same is necessary here (although a computational dataset will be used instead). In fact, dataset limitations are why different empirical

computational methods do not work effectively on the same types of molecules (5). Model developers often try to get around these limitations by taking on the cost to create large datasets as they typically explain a larger sample space of the problem while avoiding overfitting (6). However, a smaller yet representative dataset should accomplish the same goal with fewer resources used. *My fragment-based training method has utilized this, making it more feasible to predict energetics of future characterized brown carbon clusters that will be considered extrapolated data relative to a current brown carbon dataset.*

C. Fragment Background

To be able to describe why fragments are a viable method to create low cost, accurate, representative training datasets, it first is necessary to describe what is being deemed a fragment in this research. As the goal is to predict energetics of newly characterized brown carbon clusters, a cluster is considered a full chemical unit and a fragment is any subsection of a cluster. The fragment dataset that has been created for this work includes all brown carbon molecules included in the clusters, water, subsections of brown carbon molecules, and functional groups. Moving forward, I will refer to whole brown carbon molecules in the fragment dataset as ‘parent molecules’ and subsections of parent molecules as ‘sub-fragments’. Some functional groups and sub-fragments are partially modified to keep the calculations more consistent and generalizable. An example of a brown carbon parent molecule and its modified sub-fragment can be seen in Figure 1.

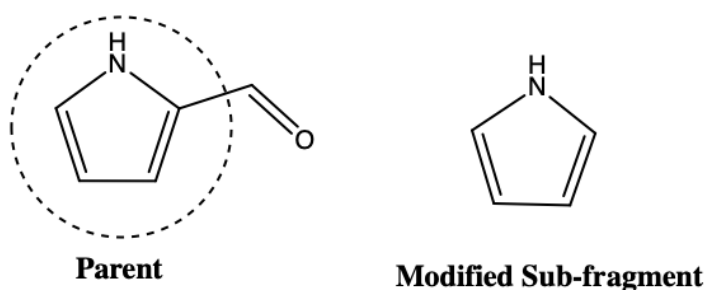


Figure 1. The ‘parent’ molecule on the left is a pyrrole based brown carbon molecule found in the fragment dataset and the ‘modified subfragment’ is a modified subsection of the parent also seen in the fragment dataset. The modified subfragment is modified as a true subsection of the parent would include a radical on the aromatic carbon attached to the aldehyde while this structure has a hydrogen. The circled region of the parent molecule shows the origin of the modified subfragment.

D. Cluster Background

Clusters are the other type of molecular structure utilized in this research, and as mentioned previously, are the structures of interest when predicting energetics. A cluster in the current dataset is made of one or two brown carbon parent molecules surrounded by one to three water molecules held together by non-covalent interactions. As there are many possible ways to arrange the molecules relative to each other to make a cluster, the program ABCluster was used to aid in finding various arrangements. ABCluster utilizes the ABC clustering method and a rigid force field to position molecules in ways it predicts will be the lowest energy, meaning that it uses a swarm technique to find the best arrangements of geometrically pre-optimized molecules without changing their structure (7). While one could use ABCluster to find one arrangement of each cluster, it is likely that there are multiple viable isomer forms, so the two-hundred predicted lowest energy arrangements of the molecules in each cluster have been created for the current brown carbon cluster training dataset. For clusters made from such a small number of molecules, it is

likely that there will be virtually identical clusters produced, but it is also probable that all or almost all viable molecular arrangements have been identified. These brown carbon cluster variations have also been supplemented with thirty variations each of a water dimer and trimer so there is training data on how water molecules interact independently of brown carbon parent molecules. After all clusters have been formed, they are re-optimized using DFT and new energetics and cluster characteristics are computed.

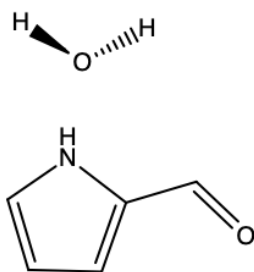


Figure 2. One of the two-hundred versions of this cluster made using ABCluster and geometry pre-optimized molecular structures.

E. Fragments for Training

Now that the fragment and cluster classification systems have been detailed, you may be wondering why I would suggest training models solely on fragment calculations in order to predict brown carbon cluster energetics, especially after emphasizing the importance of having a representative training dataset. However, when looking to predict highly varied unseen data, fragments are actually the more representative sample space. It is easy to see this when thinking in terms of vehicles. If you go to a golf course and see dozens of different models of golf carts, it is relatively easy to imagine what the next golf cart you see will look and function like. However, if you have only learned about golf carts and are asked to describe a car, your prediction will be

severely flawed. If you had instead learned about the mechanical parts and properties of a golf cart, you likely would have had a much better idea of how a car would act considering you would have fundamental, generalizable knowledge on motors, wheels, steering, and braking. This predictively powerful generalizable training is what I have brought to machine learning through the use of a fragment dataset. By training on small pieces of brown carbon clusters and common chemical functional groups, models are provided with fundamental chemical information that can be applied to highly varied problems, such as future characterized brown carbon clusters.

Generalizable training data is far from the only benefit of using fragments, however. Compared to training with clusters, fragments provide a computationally inexpensive process overall. Considering there is just one molecule being worked with at a time, ABCcluster never needs to be used and DFT calculations are only performed once. This is especially helpful for reducing computational time as it takes a large amount of resources to use DFT to optimize cluster geometries and compute cluster characteristics due to the intermolecular interactions. Also, since there is only one version of each fragment instead of hundreds, the size of the training dataset is significantly reduced, shrinking the time required to fit a model.

CHAPTER 2. ABOUT THE DATA

A. Train-Test Splits

The data is split into two sets of training and testing datasets for this work. The first train-test split effectively serves as the null hypothesis as it uses all non-extrapolated brown carbon clusters as training data and extrapolated clusters as testing data. The train-test split used to explore the efficacy of fragment training predictive power includes all fragment data for training and all extrapolated clusters for testing. It is worth noting that all testing datasets are deemed extrapolated data in machine learning as the models have not previously seen that subset of data. In this context, however, I describe data as being extrapolated with a more generalized scientific definition, referring to the clusters in the testing dataset being dissimilar to the clusters and fragments used for training.

B. Molecular Structure Overview

Although characterized brown carbon molecules are not all nitrogen based, this cluster dataset includes only clusters made with parent molecules containing nitrogen (excluding water dimer and trimer clusters). All non-extrapolated clusters are also restricted to being made with aromatic brown carbon parent molecules, but this is not a limitation in the fragment or extrapolated cluster datasets. Additionally, I have chosen to include solely parent molecules that can undergo keto-enol tautomerization in order to ensure there is at least one reactive oxygen based functional group in the molecule.

Stereochemistry is a focal point of the molecules in this dataset. With the entire parent molecule collection having the ability to tautomerize, it is possible to give an E-Z alkene stereochemistry assignment to each parent molecule when in its enol state. In an effort to encourage

the computer to “link” the keto and enol structures of a single parent molecule, I have carried over the atomic arrangement of each enol structure to the keto structure and assigned identical stereochemistry designations even though the keto arrangement cannot have an alkene stereochemistry value by IUPAC standards. Fragment stereochemistry assignment may also break IUPAC rules as fragments directly descending from a parent molecule are given the same stereochemistry designation as the parent. It is important to consider stereochemistry when optimizing parent structures with DFT methods as the starting molecular structure will determine which potential energy surface local or global minima that structure optimization will finish at. With this in mind, all parent molecules in the training dataset have been optimized in their keto and enol states with both E and Z stereochemistry and clusters have been made with all four variations of the parent.

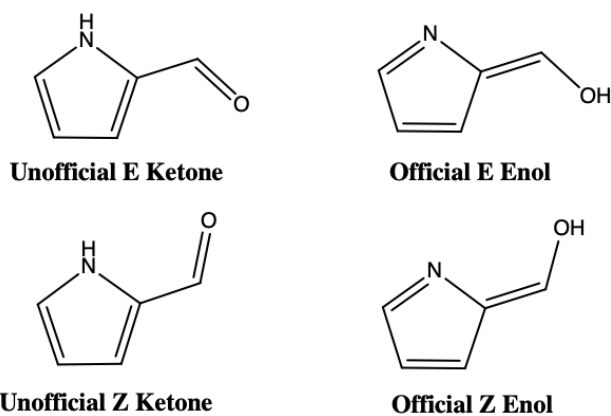


Figure 3. Official and unofficial alkene stereochemistry assignments are shown for the keto and enol structures of a parent molecule.

C. Data Curation

Making two-hundred versions of each brown carbon cluster means that not every cluster iteration will be a viable molecular arrangement. Some cluster calculations do not reach an optimal

geometry convergence and therefore will not have calculated energetics or vibrational frequency values. As molecular energetics is what is being predicted and a set of harmonic vibrational frequencies is a unique identifier, I have chosen to remove any instances of data that are missing one or both of these. I have also used vibrational frequency data to remove instances of transition states as they are not part of a representative sample space in a non-reaction-based problem. Finally, all duplicate clusters have been removed from the dataset as they may inadvertently cause bias in the dataset. While it would seem logical that a more frequent occurrence of a cluster would correspond to a higher probability of that cluster existing in nature, that is not necessarily the case as cluster optimization calculations could become trapped in an energy local minimum that misrepresents natural occurrence rates. By removing duplicates, the likelihood of a cluster existing remains dependent on energetics.

CHAPTER 3. METHODS

A. DFT Methods

All calculations included in this work have been optimized using density functional theory (DFT). A geometry optimization and frequency calculation with ω B97X-D functional and 6-311++G(d,p) basis set was run on each fragment and cluster. This combination of functional and basis set was chosen as the ω B97X-D functional has been found to be effective in modeling atmospheric cluster systems and the 6-311++G(d,p) basis set has been shown to be robust with a large number of functionals, likely increasing the accuracy of calculations (8). Fragment calculations only needed to be run once. Clusters had parent and water molecules optimized individually before creating a cluster and the resulting cluster arrangement was also optimized.

B. Machine Learning Methods

Different types of machine learning algorithms are used here to predict the absolute energy of a cluster (reported in kJ mol^{-1}). Energetics will be of particular interest when studying newly characterized molecules as the Boltzmann distribution explains that overall cluster energy is directly related to the probability of finding the cluster in nature and lower energy arrangements are more likely due to their increased stability (9). While I do not have a newly characterized brown carbon cluster to test models on, the extrapolated dataset of hypothetical brown carbon clusters takes its place in testing the efficacy of fragment training to predict overall cluster energetics.

Besides removing data based on the criteria discussed previously, I have scaled all data to a value between 0 and 1 inclusive (10). As this simply changes the range of the data proportionally, there is no change in the underlying distribution. Keeping the underlying distribution is preferred

in this scenario as cluster energetics should already be following the Boltzmann distribution and there is no reason to disrupt a data spread based on scientific principle. This type of scaling is also useful for working with molecular energetics applications as it is not robust to outliers (10), meaning that it will not alter the data in a way that will make it insensitive to how large geometric shifts change cluster energy.

The types of machine learning models used are regression-based random forests and stochastic gradient descent. The random forests were used with the default parameters while stochastic gradient descent models were run with a hybrid L1-L2 penalty in order to introduce a bit of feature selection. The random forest algorithm works by running multiple iterations of trees that identify the best ways to separate data so that each subsequent split in the tree leads to more specific similarities in the subgroup of data. A visualization of a collection of trees can be found in Figure 4. A regression-based stochastic gradient descent algorithm works to find the best fit of a linear model by randomly choosing data from the training dataset to estimate the slope of the cost function at a certain point and move toward the cost function minimum (11, 12).

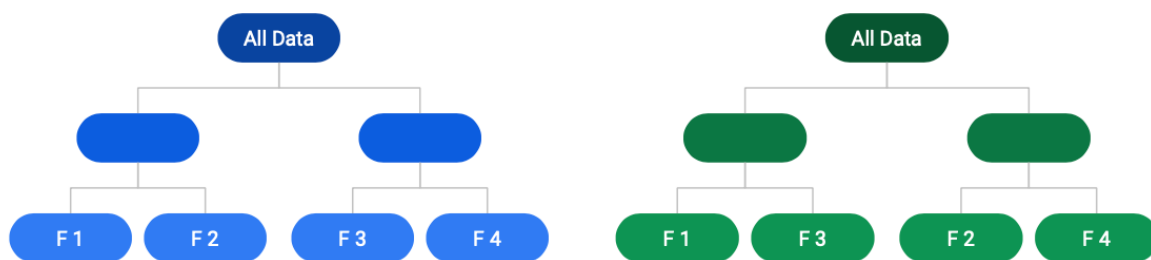


Figure 4. Two possible decision trees run with unlimited depth are shown. The combination of trees could function as a hypothetical random forest. 'F' with a number shows an instance of a single fragment.

C. Data Collected

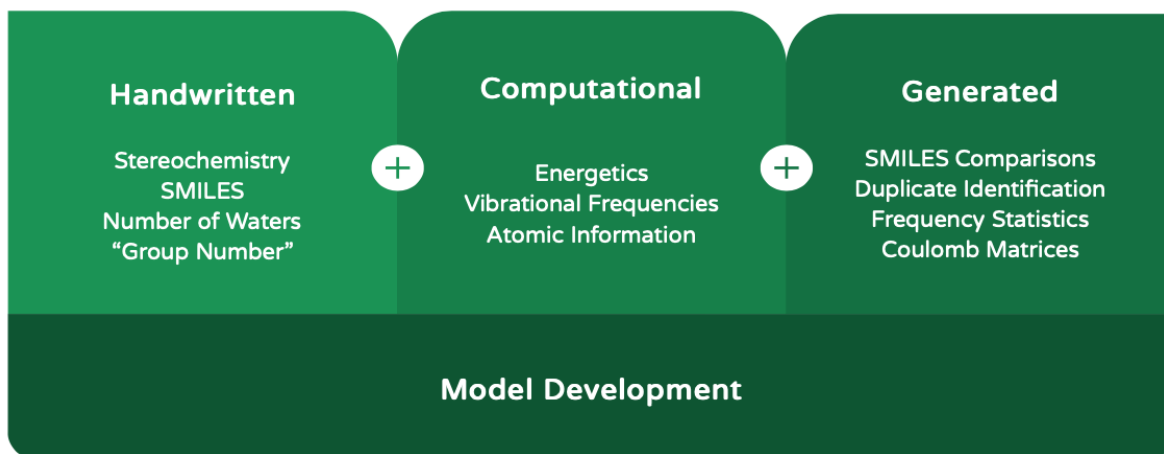


Figure 5. The flow of data from three sources to machine learning model development. Types of data listed include features used to pre-process data as well as features included in modeling.

Data for this work comes from three different sources: handwritten spreadsheets, computational calculation results, and data generated from analyses. While some data feeds directly into model development, other data, such as handwritten SMILES strings are used to create data down the pipeline in the ‘generated’ category that will be used for model fitting.

Handwritten data includes descriptive variables that are impossible or nearly impossible to automatically generate from a calculation. Stereochemistry and SMILES strings of clusters, fragments, and parent molecules, number of waters in a cluster, and cluster ‘group number’ are all values supplied in a handwritten spreadsheet. SMILES strings have been generated from structures that include all implicit hydrogens. These strings are an optimal cluster or fragment descriptor to use with machine learning as SMILES were designed for efficient chemical analysis with computers (13). Cluster ‘group number’ data is an arbitrary numerical categorical value where clusters with the same type of parent molecule (keto-enol tautomers of each other with varying stereochemistry) are assigned a value to encourage the models to associate the clusters with each other. As these characteristics are generalizable to all two-hundred clusters generated for each

combination of parent molecule and number of water molecules, only one entry needs to be written for each type of cluster which keeps data entry time low. One entry must be made for each fragment in addition.

The other raw data source is computational data output from DFT methods. Energetics, vibrational frequencies, and atomic information are all of interest. For energy values, electronic plus zero-point energy is recorded in Hartrees and converted to kJ mol^{-1} . Vibrational frequencies are rounded to the nearest whole wavenumber and recorded at each vibrational mode. These rounded vibrational frequency values are used both as raw data for model development and are crucial in developing generated data down the data pipeline. As a raw data input, frequencies are an important set of values as they act as a molecular fingerprint and can therefore be influential in aiding models in differentiating between similar molecules and their characteristics. Finally atomic number of atoms, number of atoms in the cluster, and atomic mass can all be collected from computational output and are crucial in making energetics predictions as they become directly representative of the optimized geometry of the structures. Knowing about the atomic positioning and characteristics of optimized structures gives fundamental information on atomic attraction and repulsion, leading to a better understanding of how high or low molecular energetics will be.

The final data source is data generated from analyses on both handwritten and computationally derived data. SMILES string comparisons are the one type of generated data originating from handwritten information. These comparisons arise from use of the ‘fuzzywuzzy’ Python library and built in string comparisons. The package uses the Levenshtein string distance method which calculates the number of letters that would have to be changed or added to get matching sequences (14). Fuzzywuzzy functions ‘ratio’ and ‘partial_ratio’ have the capability to compare two strings in their entirety as well as compute a similarity score for the entirety of the

shorter string and a substring in the larger respectively (15). I have utilized this library by running a 'ratio' and 'partial_ratio' string comparison for each cluster or fragment SMILES string with every other SMILES string in the database, including itself. By utilizing the ratio function on SMILES strings, I am able to look at overall similarity of clusters. The 'partial_ratio' looks at molecular or cluster details and allows for the identification and subsequent similarity estimate of a substructure or subcluster in the compared entities. Beyond these tools, I have utilized the Python 'in' string comparison method which looks for identical strings or identical substrings in a pair of strings. Unlike the fuzzywuzzy 'partial_ratio' function, there is no flexibility in identifying a subcluster and the entire smaller string must be found in the larger for a substring, and therefore a representative subcluster, to be acknowledged. Effectively the 'partial_ratio' and 'in' comparison methods serve the same purpose but the built in string comparison provides a more exact subcluster identification.

A large portion of generated data comes from computationally derived vibrational frequencies. Frequency statistics comprises a large amount of frequency based generated data, and this includes saddle point identification and median and standard deviation calculations for each vibrational mode. Median and standard deviations at each vibrational mode are calculated across the set of two hundred variations of each cluster type before data removal occurs. This is because median and standard deviation are used in this case as population based statistics that can illuminate a general frequency baseline and range respectively. Saddle point identification involves looking for an imaginary frequency value in a molecule or cluster, indicating that the structure is a transition state. In other settings a transition state structure could be useful information, but as this is not reaction or molecular dynamics based work, a transition state is not representative of the dataset and will need to be identified and removed.

The other portion of frequency-based data in this category is duplicate cluster identification. To identify duplicate clusters in the dataset, vibrational frequencies from two clusters that are rounded to the nearest whole number are compared. If eighty five percent of the rounded vibrational frequencies at each mode are identical, the clusters are considered duplicates. While the cutoff of eighty five percent is arbitrary, the utilization of rounded values and eighty five percent similarity yield results that rival comparisons done by hand. If two or more clusters are deemed duplicates of each other, one is included in the dataset and what is left is removed for reasons related to energy distributions discussed previously.

The last category of analysis generated data is Coulomb matrices. A Coulomb matrix is a way to represent atomic interactions based on distance and atomic number in a condensed form. The formula for computing elements of a Coulomb matrix can be found below in Figure 6. Although there is a lot of atomic information packed into one matrix, clusters or fragments that are large in size will have massive matrices, leading to a set of high dimensional data. In order to reduce the dimensionality, the eigenvalues of Coulomb matrices have been used as model features instead of the matrix elements (16). Resulting eigenvalues do not have any physical meaning and are simply used to represent the matrices efficiently.

$$M_{ij} = \begin{cases} 0.5Z_i^{2.4} & \text{if } i = j \text{ (diagonal)} \\ \frac{Z_i Z_j}{R_{ij}} & \text{if } i \neq j \text{ (off diagonal)} \end{cases}$$

Figure 6. The formula to create a Coulomb matrix is shown above where 'Z' represents atomic number and 'R' is the distance between atoms.

CHAPTER 4. RESULTS AND CONCLUSION

A. Metrics

Both random forest and stochastic gradient descent models are judged on their predictive performance using R^2 , root mean square error, and percent error metrics. The R^2 metric measures the amount of variance in the data that is explained by the model and therefore the model fit. It has a maximum value of one which represents one hundred percent of variance being explained by the model. Negative R^2 values mean that the model fits the data poorly enough that it performs less well than random guessing (17). Root mean square error (RMSE) is a measurement of the distance between the model and each predicted data point on a graph, representing the amount of error there is for each instance of data based on the fit of the model. Due to the naturally large predicted cluster energetics values, and therefore errors in predictions, the RMSE value has been scaled to become a relative factor of the average of all actual extrapolated cluster energy values. The percent error of each predicted value will also be used to compare model efficacy. All models have been run twenty five times and the reported R^2 and RMSE values are the average of all twenty five R^2 and RMSE values calculated for one model. Percent error values are calculated using the average predicted energy value for each cluster across all twenty five runs of a model. The exact orientations of each cluster used in the extrapolated cluster testing dataset can be found in the supporting information section in Table 4.

B. Model Performances

The four machine learning algorithm and training method combinations tested here are fragment trained random forests, cluster trained random forests, fragment trained stochastic gradient descent, and cluster trained stochastic gradient descent, and each has been run twenty five

times as mentioned above. The resulting data shows that model performance is highly dependent on the algorithm type and cluster being predicted, meaning that there are pockets of good predictive performance in all four model-training method combinations tested.

When looking at Figure 7, which displays the percent error of each cluster for each model and training combination, it is immediately clear that random forest and stochastic gradient descent algorithms have their strengths with different subsets of data. Random forest models, whether fragment or cluster trained, perform very well with the aldehyde and enol single brown carbon parent molecule clusters. However, stochastic gradient descent performs relatively poorly with this subset of data and instead better predicts energetics of the larger two brown carbon parent molecule clusters, shown on the far right side of the graph. Just as with the best performing subset of data with the random forest models, the training method is fairly irrelevant to the predictive power of the stochastic gradient descent models when looking at the two parent molecule clusters. That being said, the average stochastic gradient descent fragment trained model performs significantly better than its cluster trained counterpart for single parent molecule clusters. When discussing the stark differences in performance based on cluster type, it is worth noting that the two parent molecule clusters, labeled as '2Mol h2o' in Figure 7, have these known brown carbon parent molecules included in the training datasets, whether individually in the fragment dataset or as a parent molecule in a one, two, or three water molecule cluster. The aldehyde and enol single parent molecule clusters, seen in the same figure labeled as 'Ald h2o' and 'Enol h2o' respectively, are farther away from the training data sample space, however, as the parent molecule is a hypothetical linear brown carbon molecule and there are no linear parent molecules included in the cluster or fragment training datasets and linear fragments only show up in the fragment training dataset as functional groups or small pieces of parent molecules. Considering this, the random forest models

seem to perform consistently better on data farther from the sample space, regardless of training method.

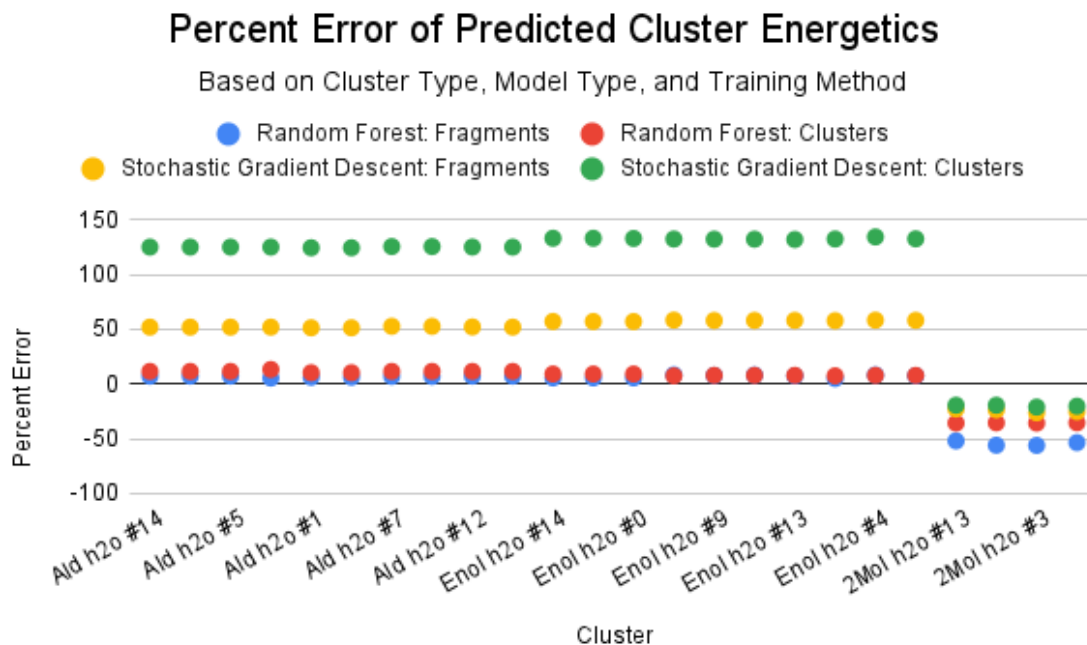


Figure 7. Percent error values for each cluster based on training method and algorithm type are shown. ‘Ald’ and ‘Enol’ clusters listed are keto-enol tautomers of each other and consist of a linear potential brown carbon parent molecule and a water molecule. ‘2Mol’ clusters are clusters made of two known brown carbon parent molecules and one water molecule. Numbers listed along with cluster names indicate which version of the cluster is being described.

As for overall performance of the models, Table 1 shows the large spread of average R^2 and scaled RMSE values for each algorithm and training method combination. When balancing the errors and fits across both types of extrapolated clusters, the cluster trained random forest performs the best based on both average R^2 and scaled RMSE. The fragment trained stochastic gradient descent model is not far behind, however. As seen in Figure 7 and discussed above, the two models have vastly different strengths, but overall they perform quite well and similarly.

Algorithm and Training Types	Average R ²	Average Scaled RMSE
Random Forest: Fragments	0.377196458	0.603320239
Random Forest: Clusters	0.728528922	0.398345220
Stochastic Gradient Descent: Fragments	0.685853648	0.428204842
Stochastic Gradient Descent: Clusters	-0.110674184	0.805791349

Table 1. R² and RMSE values were averaged across twenty five runs for each model.

C. Feature Importance Ranking

Figures 8-11 shown below display the most important features for each algorithm and training method combination graphically. While stochastic gradient descent models did have negatively weighted features, the random forest models did not, so the ten most positively weighted and ten most negatively weighted features are shown for stochastic gradient descent models while just the ten most positively weighted features were found for the random forests.

As for feature patterns among the models, eigenvalue features comprise the large majority of positively weighted features with fragment training methods, regardless of the algorithm type. Additionally, frequency and frequency statistic features dominate with cluster training, but are positively weighted for random forest models and negatively weighted when used with stochastic gradient descent. When looking at trends across algorithms, random forests uniquely include some indicator of cluster or fragment size in their most important feature list. There appear to be no algorithm specific trends for stochastic gradient descent, however. Instead, fragment trained stochastic gradient descent has atomic numbers rounding out most of its most negative features and cluster trained models see a uniquely large influence coming from SMILES comparisons.

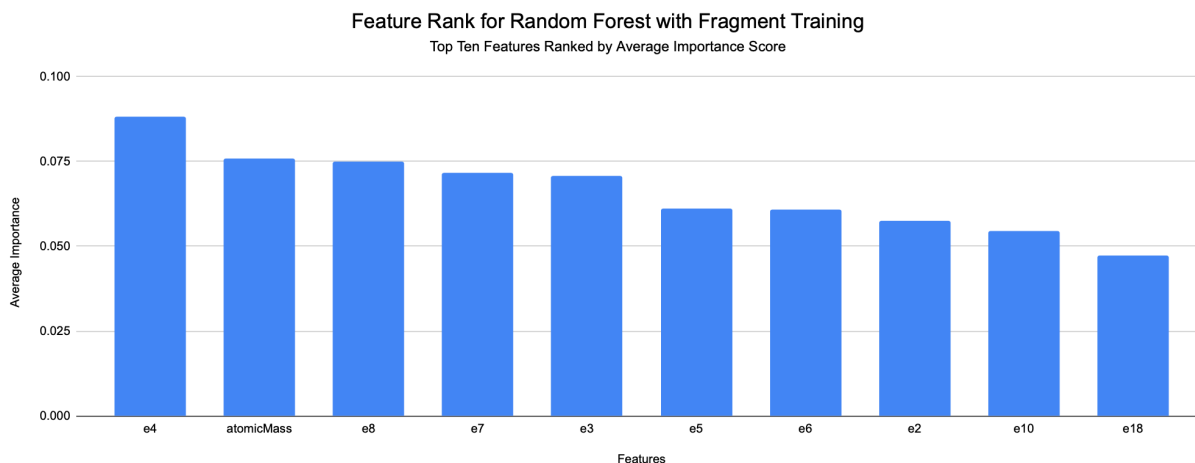


Figure 8. Average feature ranking for random forests fitted using fragment training. All ‘e’ based features represent the resulting eigenvalues from Coulomb matrices. Eigenvalues are not organized in order of increasing value but are instead left in the order calculated based on Coulomb matrix row arrangement being determined by increasing atomic number. This means that ‘e4’, for example, is the fourth resulting eigenvalue and the fourth resulting eigenvalue has consistently been found to be the most influential feature when predicting cluster energetics with fragment trained random forests. Nine of the top ten most influential features are Coulomb matrix eigenvalue features.

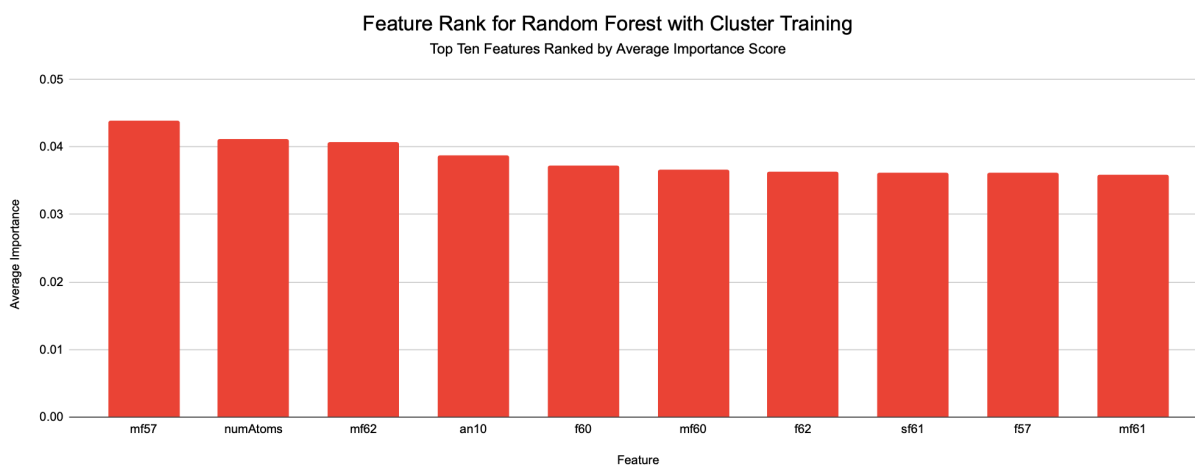


Figure 9. Average feature ranking for random forests fitted using cluster training. All ‘f’ based features are vibrational frequencies at that numerical position in the list of frequencies where vibrational frequencies are arranged from least to greatest. Features with an ‘mf’ description represent the median value of frequencies at that numerical position within one type of cluster. ‘Sf’ is representative of the standard deviation but otherwise works identically to ‘mf’ features. The ‘an’ feature refers to the atomic number at that position in the list of atomic numbers as given in the DFT output. Eight of the top ten most influential features are frequency or frequency statistic features.

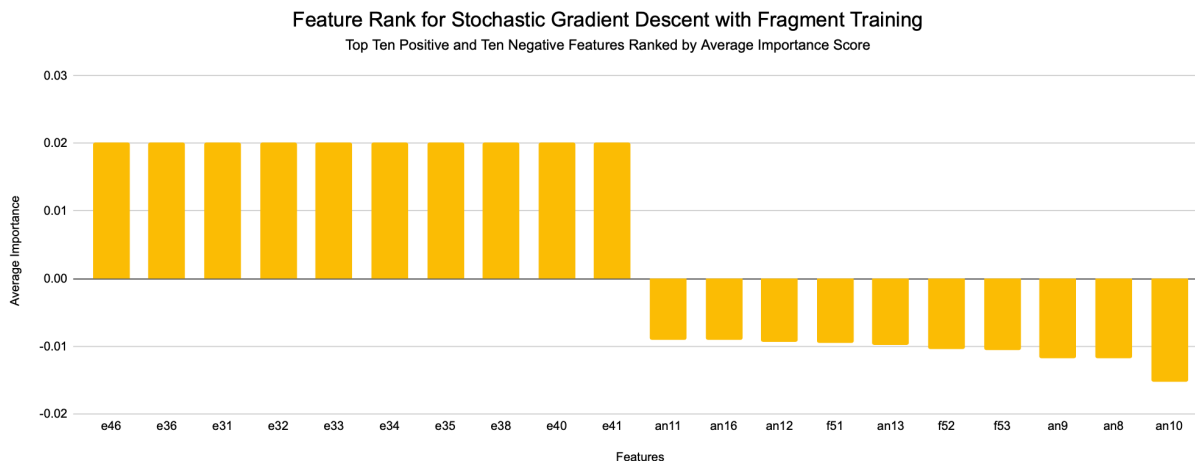


Figure 10. Average feature ranking for fragment trained stochastic gradient descent models. Eigenvalue, vibrational frequency, and atomic number based features comprise the ten most influential positive and ten most influential negative features.

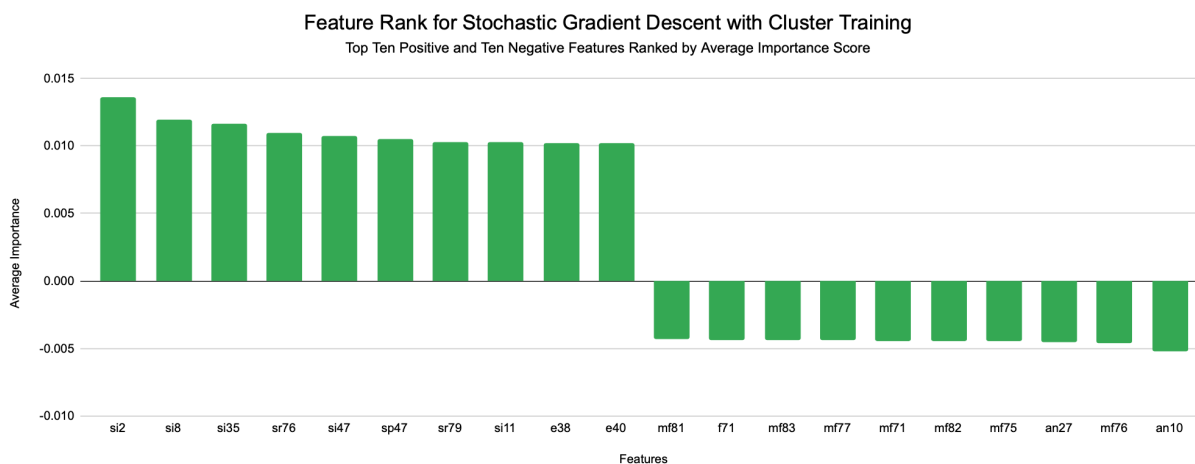


Figure 11. Average feature ranking for cluster trained stochastic gradient descent models. ‘Si’, ‘sr’, and ‘sp’ are features based on the smiles ‘in’ string comparison, smiles fuzzywuzzy comparison ratio, and smiles fuzzywuzzy comparison partial ratio respectively. The features displayed here show the most wide variety of features being included across the four algorithm-training method combinations.

D. Conclusion

It is clear to see from Figure 7 that when algorithm types are performing at their best, the training methods produce nearly identical results. Additionally, Table 1 shows that cluster trained random forests perform only slightly better than fragment trained stochastic gradient descent

overall. And considering that fragment training takes a minute fraction of the amount of computational resources to create a representative training dataset, this comparable performance shows that fragment training is a viable training option for predicting brown carbon cluster energetics when Coulomb matrix eigenvalues are used as the driving force.

CHAPTER 5. FUTURE WORK

While I have managed to show that fragments can be used to efficiently and effectively train predictive models for atmospheric applications, this work should be expanded upon to get a more complete understanding of the process of fragment training. Firstly, the testing dataset should be split so each dataset is made of similar iterations of each cluster type. In this case, the testing dataset would be split into two where one dataset includes all keto-enol tautomer clusters with the linear parent molecule and the other dataset has just clusters with two aromatic brown carbon parent molecules. I believe this is an important next step as the models show strong performance with one type of cluster but not both, so separating them should give a more clear perspective on how the training method influences predictive power. Additionally, feature selection methods should be implemented on all models to see if the computational time required to make the training dataset can be reduced across all techniques. Finally, fragment structures should be run with more than one configuration in order to get a variety of viable structures that may appear in different clusters.

SUPPORTING INFORMATION

Parent Molecule Name	SMILES	Stereochemistry	Source
1H-pyrrole-2-carbaldehyde	<chem>[H]C(C1=C([H])C([H])=C([H])N1)=O</chem>	e	(1)
1H-pyrrole-2-carbaldehyde	<chem>O=C([H])C1=C([H])C([H])=C([H])N1</chem>	z	(1)
(E)-(2H-pyrrol-2-ylidene)methanol	<chem>O/C([H])=C1C([H])=C([H])C([H])=N1</chem>	e	(1)
(Z)-(2H-pyrrol-2-ylidene)methanol	<chem>[H]/C(O)=C1C([H])=C([H])C([H])=N1</chem>	z	(1)
2-(1H,1'H-[2,2'-biimidazol]-1-yl)-2-hydroxyacetaldehyde	<chem>OC(C([H])=O)([H])N1C([H])=C([H])N=C1C2=NC([H])=C([H])N2</chem>	e	(1)
2-(1H,1'H-[2,2'-biimidazol]-1-yl)-2-hydroxyacetaldehyde	<chem>OC(C([H])=O)([H])N1C([H])=C([H])N=C1C2=NC([H])=C([H])N2</chem>	z	(1)
(E)-1-(1H,1'H-[2,2'-biimidazol]-1-yl)ethene-1,2-diol	<chem>O/C(N1C([H])=C([H])N=C1C2=NC([H])=C([H])N2)=C([H])/O</chem>	e	(1)
(Z)-1-(1H,1'H-[2,2'-biimidazol]-1-yl)ethene-1,2-diol	<chem>O/C(N1C([H])=C([H])N=C1C2=NC([H])=C([H])N2)=C([H])\O</chem>	z	(1)
2-(2-(dihydroxymethyl)-1H-imidazol-1-yl)-2-hydroxyacetaldehyde	<chem>OC(O)([H])C1=NC([H])=C([H])N1C(O)([H])C([H])=O</chem>	e	(1)
2-(2-(dihydroxymethyl)-1H-imidazol-1-yl)-2-hydroxyacetaldehyde	<chem>OC(O)([H])C1=NC([H])=C([H])N1C(O)([H])C([H])=O</chem>	z	(1)
(E)-1-(2-(dihydroxymethyl)-1H-imidazol-1-yl)ethene-1,2-diol	<chem>OC(O)([H])C1=NC([H])=C([H])N1/C(O)=C([H])\O</chem>	e	(1)
(Z)-1-(2-(dihydroxymethyl)-1H-	<chem>OC(O)([H])C1=NC([H])=C([H])N1/C(O)=C(O)\[H]</chem>	z	(1)

imidazol-1-yl)ethene-1,2-diol			
(E)-4-(2-oxoethylidene)-1,2,3,4-tetrahydropyridine-2,6-dicarboxylic acid	[H]/C(C([H])=O)=C1C([H])=C(C(O)=O)NC(C(O)=O)([H])C1([H])[H]	e	(18)
(Z)-4-(2-oxoethylidene)-1,2,3,4-tetrahydropyridine-2,6-dicarboxylic acid	O=C(O)C(NC(C(O)=O)=C/1[H])([H])C([H])([H])C1=C([H])\C([H])=O	z	(18)
4-(2-hydroxyvinylidene)-1,2,3,4-tetrahydropyridine-2,6-dicarboxylic acid	OC([H])=C=C1C([H])=C(C(O)=O)NC(C(O)=O)([H])C1([H])[H]	e	(18)
4-(2-hydroxyvinylidene)-1,2,3,4-tetrahydropyridine-2,6-dicarboxylic acid	O=C(O)C(NC(C(O)=O)=C1[H])([H])C([H])([H])C1=C=C([H])O	z	(18)
(E)-6-(dihydroxymethylene)-4-(2-oxoethylidene)-1,4,5,6-tetrahydropyridine-2-carboxylic acid	[H]/C(C([H])=O)=C1C([H])=C(C(O)=O)N/C(C1([H])[H])=C(O)\O	e	(18)
(Z)-6-(dihydroxymethylene)-4-(2-oxoethylidene)-1,4,5,6-tetrahydropyridine-2-carboxylic acid	O/C(O)=C(NC(C(O)=O)=C/1[H])/C([H])([H])C1=C([H])\C([H])=O	z	(18)
Water	[H]O[H]		

Table 2. Training data parent molecule list and structure information. Parent molecules have been used individually as listed as well as with the addition of water molecules to create clusters.

Fragment Name	SMILES	Stereochemistry
formaldehyde	<chem>[H]C([H])=O</chem>	
ethenol	<chem>[H]/C([H])=C([H])/O</chem>	
1H-pyrrole	<chem>[H]C1=C([H])C([H])=C([H])N1</chem>	
(E)-ethene-1,2-diol	<chem>[H]/C(O)=C(O)/[H]</chem>	e
(Z)-ethene-1,2-diol	<chem>[H]/C(O)=C([H])/O</chem>	z
2-hydroxyacetaldehyde	<chem>[H]C(C(O)([H])[H])=O</chem>	e
2-(dimethylamino)-2-hydroxyacetaldehyde	<chem>[H]C(C(O)([H])N(C)C)=O</chem>	e
(E)-1-(dimethylamino)prop-1-ene-1,2-diol	<chem>C/C(O)=C(O)/N(C)C</chem>	e
1-methyl-1H,1'H-2,2'-biimidazole	<chem>CN1C(C2=NC([H])=C([H])N2)=NC([H])=C1[H]</chem>	
1-(dimethylamino)-1-hydroxypropan-2-one	<chem>O=C(C(N(C)C)([H])O)C</chem>	z
1-(dimethylamino)propane-1,2-diol	<chem>OC(C(N(C)C)([H])O)([H])C</chem>	z
3-methylbut-2-enal	<chem>C/C(C)=C([H])\C([H])=O</chem>	
acrylaldehyde	<chem>[H]/C(C([H])=O)=C([H])\[H]</chem>	
4-methylene-1,2,3,4-tetrahydropyridine	<chem>[H]/C([H])=C1C([H])=C([H])NC([H])([H])C/1([H])[H]</chem>	
4-(propan-2-ylidene)-1,2,3,4-tetrahydropyridine	<chem>C/C(C)=C1C([H])=C([H])NC([H])([H])C/1([H])[H]</chem>	
1,2,3,4-tetrahydropyridine	<chem>[H]C1([H])C([H])=C([H])NC([H])([H])C1([H])[H]</chem>	
(E)-4-(2-oxoethylidene)-1,2,3,4-tetrahydropyridine-2-carboxylic acid	<chem>O=C([H])/C([H])=C1C([H])=C([H])NC(C(O)=O)([H])C/1([H])[H]</chem>	e
(E)-2-(2,3-dihydropyridin-4(1H)-ylidene)acetaldehyde	<chem>[H]/C(C([H])=O)=C1C([H])=C([H])NC([H])([H])C/1([H])[H]</chem>	e

(E)-6-formyl-4-(2-oxoethylidene)-1,4,5,6-tetrahydropyridine-2-carboxylic acid	<chem>O=C([H])/C([H])=C1C([H])=C(C(O)=O)NC(C([H])=O)([H])C/1([H])[H]</chem>	e
(E)-6-formyl-4-(2-oxoethylidene)-1,2,3,4-tetrahydropyridine-2-carboxylic acid	<chem>O=C([H])/C([H])=C1C([H])=C(C([H])=O)NC(C(O)=O)([H])C/1([H])[H]</chem>	e
(E)-4-(2-oxoethylidene)-1,2,3,4-tetrahydropyridine-2,6-dicarbaldehyde	<chem>O=C([H])/C([H])=C1C([H])=C(C([H])=O)NC(C([H])=O)([H])C/1([H])[H]</chem>	e
(E)-4-(2-oxoethylidene)-1,4,5,6-tetrahydropyridine-2-carbaldehyde	<chem>O=C([H])/C([H])=C1C([H])=C(C([H])=O)NC([H])([H])C/1([H])[H]</chem>	e
(Z)-4-(2-oxoethylidene)-1,2,3,4-tetrahydropyridine-2-carboxylic acid	<chem>[H]C1([H])C([H])(C(O)=O)NC([H])=C([H])/C1=C(C([H])=O)/[H]</chem>	z
(Z)-2-(2,3-dihydropyridin-4(1H)-ylidene)acetaldehyde	<chem>[H]/C(C([H])=O)=C1C([H])=C([H])NC([H])([H])C/1([H])[H]</chem>	z
(Z)-6-formyl-4-(2-oxoethylidene)-1,4,5,6-tetrahydropyridine-2-carboxylic acid	<chem>[H]/C(C([H])=O)=C1C([H])=C(C(O)=O)NC(C([H])=O)([H])C/1([H])[H]</chem>	z
(Z)-6-formyl-4-(2-oxoethylidene)-1,2,3,4-tetrahydropyridine-2-carboxylic acid	<chem>[H]/C(C([H])=O)=C1C([H])=C(C([H])=O)NC(C(O)=O)([H])C/1([H])[H]</chem>	z
(Z)-4-(2-oxoethylidene)-1,4,5,6-tetrahydropyridine-2-carbaldehyde	<chem>[H]/C(C([H])=O)=C1C([H])=C(C([H])=O)NC([H])([H])C/1([H])[H]</chem>	z
(Z)-4-(2-oxoethylidene)-1,2,3,4-tetrahydropyridine-2,6-dicarbaldehyde	<chem>[H]/C(C([H])=O)=C1C([H])=C(C([H])=O)NC(C([H])=O)([H])C/1([H])[H]</chem>	z
1H-imidazole	<chem>[H]C1=C([H])NC([H])=N1</chem>	
(1H-imidazol-2-yl)methanediol	<chem>[H]C1=C([H])NC(C(O)(O)[H])=N1</chem>	
(1-(1-hydroxyethyl)-1H-	<chem>[H]C1=C([H])N(C(C)([H])O)C(C(O)(O)</chem>	

imidazol-2-yl)methanediol	[H])=N1	
2-hydroxy-2-(1H-imidazol-1-yl)acetaldehyde	[H]C1=C([H])N(C(C([H])=O)([H])O)C([H])=N1	e
2-hydroxy-2-(1H-imidazol-1-yl)acetaldehyde	[H]C1=C([H])N(C(C([H])=O)([H])O)C([H])=N1	z
(E)-1-(1H-imidazol-1-yl)ethene-1,2-diol	[H]C1=C([H])N(/C(O)=C([H])\O)C([H])=N1	e
(Z)-1-(1H-imidazol-1-yl)ethene-1,2-diol	[H]C1=C([H])N(/C(O)=C(O)\[H])C([H])=N1	z
(E)-1-(dimethylamino)ethene-1,2-diol	[H]/C(O)=C(N(C)C)\O	e
(Z)-1-(dimethylamino)ethene-1,2-diol	O/C([H])=C(N(C)C)\O	z
2-(dimethylamino)-2-hydroxyacetaldehyde	[H]C(C(N(C)C)([H])O)=O	e
2-(dimethylamino)-2-hydroxyacetaldehyde	O=C([H])C(N(C)C)([H])O	z

Table 3. A list of fragment structure characteristics listed as used.

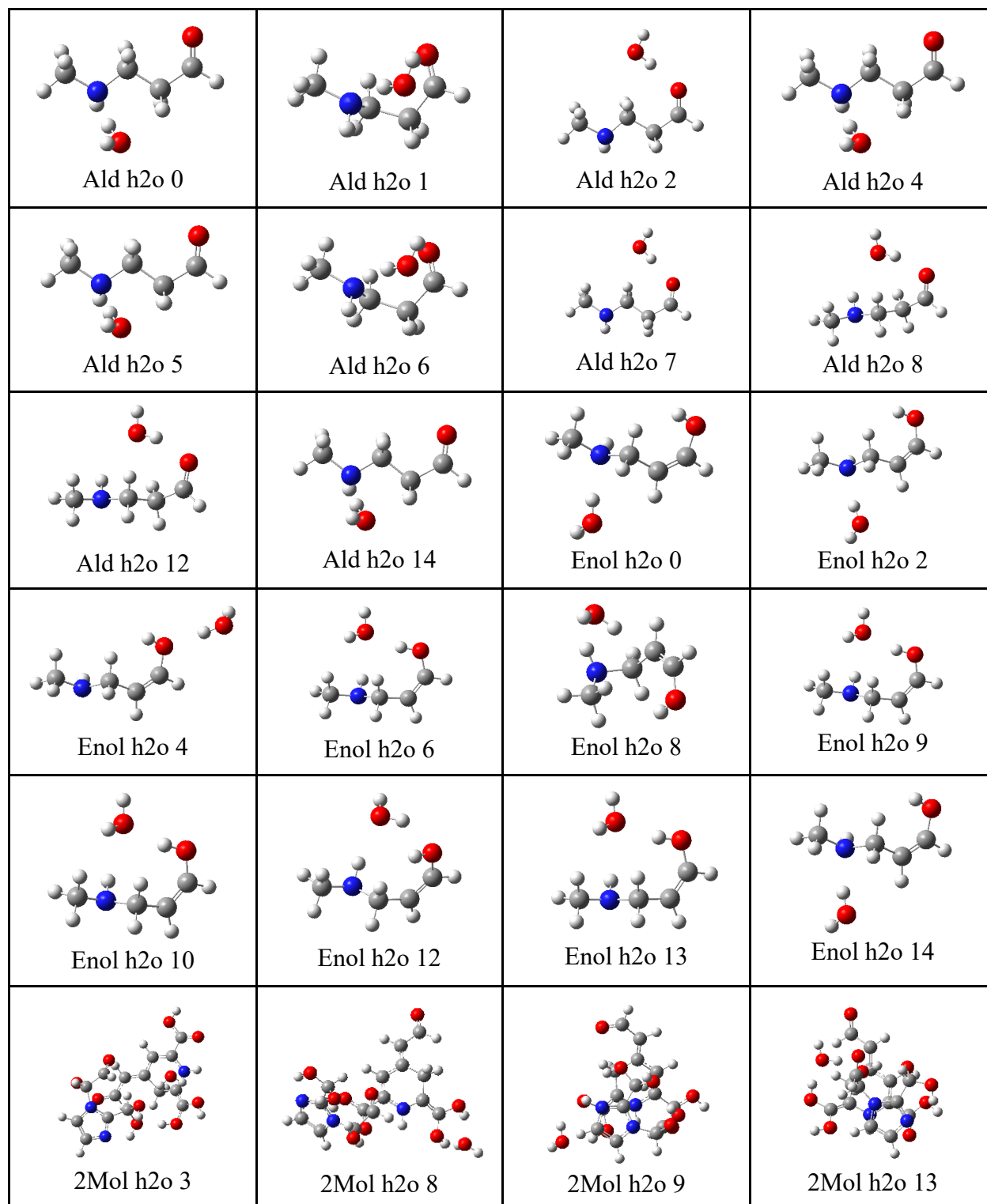


Table 4. Testing dataset extrapolated cluster configurations are shown above. Numbers included indicate the version of the cluster shown. Clusters deemed appropriate to include in the test dataset follow the rules previously listed for data removal. All clusters include one water.

Cluster Nickname	Parent Molecule Name(s)	SMILES	Sources
Ald h2o	3-(methylamino)propanal	CNCCC=O	
Enol h2o	(Z)-3-(methylamino)prop-1-en-1-ol	CNC/C=C\O	
2Mol h2o	2-(2-(dihydroxymethyl)-1H-imidazol-1-yl)-2-hydroxyacetaldehyde AND (E)-6-(dihydroxymethylene)-4-(2-oxoethylidene)-1,4,5,6-tetrahydropyridine-2-carboxylic acid	OC(O)([H])C1=NC([H])=C([H])N1C(O)([H])C([H])=O AND [H]/C(C([H])=O)=C1C([H])=C(C(O)=O)N/C(C\1([H])[H])=C(O)O	(1, 18)

Table 5. Components of extrapolated clusters shown in Table 4.

BIBLIOGRAPHY

- (1) Laskin, Alexander, Julia Laskin, and Sergey A. Nizkorodov. "Chemistry of atmospheric brown carbon." *Chemical reviews* 115.10 (2015): 4335-4382.
- (2) van Mourik, Tanja, Michael Bühl, and Marie-Pierre Gaigeot. "Density functional theory across chemistry, physics and biology." (2014): 20120488.
- (3) Delua, Julianna. "Supervised vs. Unsupervised Learning: What's the Difference?" *IBM*, IBM, 12 Mar. 2021, www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.
- (4) Schmidt, Jonathan, et al. "Recent advances and applications of machine learning in solid-state materials science." *npj Computational Materials* 5.1 (2019): 1-36
- (5) Genheden, Samuel, et al. "Computational Chemistry and Molecular Modelling Basics." (2017): 1-38.
- (6) Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." *Journal of Big Data* 6.1 (2019): 1-48.
- (7) Zhang, Jun, and Michael Dolg. "Global optimization of clusters of rigid molecules using the artificial bee colony algorithm." *Physical Chemistry Chemical Physics* 18.4 (2016): 3003-3010.
- (8) Elm, Jonas. "Toward a Holistic Understanding of the Formation and Growth of Atmospheric Molecular Clusters: A Quantum Machine Learning Perspective." *The Journal of Physical Chemistry A* (2020).
- (9) Libretexts. "17.2: The Thermal Boltzman Distribution." *Chemistry LibreTexts*, Libretexts, 30 Aug. 2020, [chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Map%3A_A_Physical_Chemistry_\(McQuarrie_and_Simon\)/17%3A_Boltzmann_Factor_and_Partition_Functions/17.02%3A_The_Thermal_Boltzman_Distribution](https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Map%3A_A_Physical_Chemistry_(McQuarrie_and_Simon)/17%3A_Boltzmann_Factor_and_Partition_Functions/17.02%3A_The_Thermal_Boltzman_Distribution).
- (10) "Compare the Effect of Different Scalers on Data with Outliers." *Scikit-Learn*, Scikit-Learn Developers, scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html.
- (11) Srinivasan, Aishwarya V. "Stochastic Gradient Descent-Clearly Explained!!" *Towards Data Science*, Medium, 6 Sept. 2019, towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31.
- (12) "Sklearn.linear_model.SGDRegressor." *Scikit-Learn*, Scikit-Learn Developers, scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html.
- (13) Weininger, David. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." *Journal of chemical information and computer sciences* 28.1 (1988): 31-36.
- (14) Nam, Ethan. "Understanding the Levenshtein Distance Equation for Beginners." *Medium*, Medium, 26 Feb. 2019, medium.com/@ethannam/understanding-the-levenshtein-distance-equation-for-beginners-c4285a5604f0.
- (15) Gitau, Catherine. "Fuzzy String Matching in Python." *Towards Data Science*, Medium, 5 Mar. 2018, towardsdatascience.com/fuzzy-string-matching-in-python-68f240d910fe.
- (16) Schrier, Joshua. "Can one hear the shape of a molecule (from its Coulomb matrix eigenvalues)?" *Journal of Chemical Information and Modeling* 60.8 (2020): 3804-3811.
- (17) "Sklearn.metrics.r2_score." *Scikit-Learn*, Scikit-Learn Developers, scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html.
- (18) Lin, Peng, et al. "Molecular chemistry of atmospheric brown carbon inferred from a nationwide biomass burning event." *Environmental science & technology* 51.20 (2017): 11561-11570