MODELING VARIANCE: A VARIANCE-MOTIVATED APPROACH TO MOLECULAR PREDICTION

Na'il Mitchell

Chemathon Inc, Berkeley, CA 94705

(Dated: November 15, 2020)

ABSTRACT

Using machine learning to predict molecular properties is an exciting research area at the interface of computer science, statistics, chemistry, and physics. Thus far, a great deal of work has been done on training various ML models to predict ground state energy, using the so-called 'Coulomb matrix', a global molecular descriptor as inputs. In this variance-motivated study, the variance of a multi-thousand molecular dataset of Coulomb matrices is analyzed; a variance analysis. This paper presents novel statistical methods and models that can aide in prediction and analysis of molecular properties and molecules using the Coulomb matrix. Analysis is performed after a detailed literature review of molecular prediction and Normality of data assessment. It is also hoped that the models introduced in this variance analysis can be generalized to other areas of interest.

INTRODUCTION

Quantum mechanics has been combined with machine learning to predict molecular properties.[1] Various ML models for this purpose have been introduced[1,2,3]; they generally share one thing in common: models are trained to predict ground state energy. One common input for such an ML model is the Coulomb Matrix (CM), made popular by Dr. Matthias Rupp. CM is a global 3D molecular descriptor. 1D molecular descriptors capture composition, 2D captures the molecular graph, and 3D captures shape. Global descriptors characterize the entire molecule.

CM is an alternative to using the Schrodinger equation to determine molecular properties, which is rather cumbersome. Using the Hamiltonian, a component of the Schrodinger equation, and the Schrodinger equation to determine molecular properties is the classical quantum mechanical approach. The Coulomb matrix is a square and symmetric matrix with size equal to the number of atoms in the molecule, squared. It represents electronic interactions of the atoms of a molecule with themselves and each other. Diagonal elements represent a polynomial fit of atomic energies to nuclear charge, hence, they represent the fitted atomic energies of the atoms. Off-diagonal elements represent Coulomb repulsion between atoms.

Figure 1 displays a CM representation of carbon dioxide, a greenhouse gas.

O=C=O

C:C	C:O1	C:O2
O1:C	O1:O1	O1:O2
O2:C	O2:O1	O2:O2

Figure 1. CM representation of carbon dioxide below the structure of carbon dioxide.

As can be seen in figure 1, the number of elements is the number of atoms, squared, hence there are 9 elements. The diagonal elements represent an atom interacting with itself, which is represented in the CM as a polynomial fit: $.5*Z_{atom}^{2.4}$. Where Z is the atomic number of the atom. Hence element 1,1 (row 1, column 1) has a fitted atomic energy of 36.86. All elements that are not diagonal represent Coulomb repulsion, as stated earlier; the numeric representation is: $Z_IZ_J/|R_I-R_J|$, where R_I-R_J is the distance between atoms I and J.

It has been shown computationally[1] that a plot of an ML model that used the Coulomb matrix as inputs, based on thousands of molecules, showed a 1:1 relationship between predicted atomization energies and reference atomization energies. The model was said to be a better predictor of atomization energy than semi-empirical quantum chemistry. The energy estimates used in the ML model are a weighted Gaussian sum. Molecules used in the study were from a Molecular Generated Database (MGD) of nearly 1 billion stable organic molecules.

Regression tree algorithms have also been used to predict ground state energies[2], using CM as input-"features". Using a dataset that contained over 16,000 molecules, ground state energies were computed on trained ML models. The model, boosted regression tree, was shown to have increased accuracy and reduced computational cost. The computational based study has potential applications in molecular discovery and informatics. The actual prediction is pseudo-atomization energy, which is a function of ground state energy and pseudo atomic energy. The absolute value of pseudo-atomization energy for the sample size of 16,242 molecules was shown to have a Normal distribution.

Not all seek alternatives to solving the Schrodinger equation for molecular prediction. AFLOW[4], a high throughput (HT) framework uses ab initio calculations to solve the Schrodinger equation to yield information such as energies and electron densities. AB initio calculations use only restrains and coulomb interactions as inputs. AFLOW is also used to calculate crystal structure properties of alloys, inorganic compounds, and intermetallic compounds. AFLOW is designed to run atop structure energy software (DFT is common). After selection of starting structures from a database, AFLOW adjust lattice parameters, and creates an input file with all parameters necessary for relaxations, and static and bond structure runs. AFLOW computes structure total energies and electronic bond structures. It also can perform Monte Carlo calculations, generate nanoparticle structure files, and identify interstitial sites inside any crystallographic structure (from an input of atomic positions).

Whether AFLOW, which uses classical quantum mechanics, or the CM is used, both can perform powerful molecular characterizations. Both however require a detailed and descriptive list of inputs to function. In addition to using the CM in ML to predict atomization energies, it has also been stated that the CM can be used to for rotational spectra interpretation[3]. In the computationally driven study revolving around molecular isomerism, the author used an ETKDG algorithm to generate 1000 random conformations of 309 isomers. The prediction task was to distinguish between a given isomer from all

other isomers. The misclassification rate was defined as proportion of incorrect assigned labels (incorrectly identifying an isomer). Various ML models were used: logistic regression, decision tree, random forest, gradient boosted trees, support vector machine, and k-nearest neighbor. Models were trained and cross-validated. The decision tree ML model had the highest misclassification rate and support vector machine had the lowest. The mean largest eigenvalue of CM (averaged over conformers) was said to be inefficient at distinguishing isomers.

Now it's time to discuss a blended computational-laboratory study on the using ML to label the degree of peptide reactivity with chemicals that contain allergens[5]. Chemical allergens react with proteins, inducing skin sensitization[5]. The majority of allergens are electrophilic and react with nucleophilic amino acids. The purpose of the study was to determine whether and to what extent reactivity correlates with skin sensitization potential. They evaluated 82 allergen containing chemicals (of different potencies) and non-allergen containing chemicals for their ability to react with amino-acid containing molecules. After a set reaction time, UV detection was used to quantify the depleted amount (reacted amount). The reactivity data and existing data was used to build a classification tree that allowed ranking of reactivity: minimal, low, moderate, high. The classification tree ML model had 89% prediction accuracy (based on cysteine and lysine amino acids). The splitting rule for the tree was based on average peptide depletion exceeding a threshold.

Thus far various ML models have been presented that predicted molecular properties. In addition, a non-ML (quantum mechanic) model was presented (AFLOW) that had a much broader spectrum of molecular characterization, though it relied heavily on input files. Predictions for the ML models were usually limited to atomization energy, though several other molecular predictions were presented as well. Isomerism was predicted via comparison of various ML models and the reactivity of chemicals containing allergens was predicted using classifications trees. Except for the last study and AFLOW, all others presented used the CM as input for the ML model.

One may wonder how this relates to This study of CM variance. The purpose of This study is to model the variance of the fitted atomic energies for a random sample of molecules from a large molecular database, in an effort to make inference to broader populations of molecules, using current and novel models. The goal also is to use smaller and simpler inputs to predict molecular properties. To analyze and model variance, parametrically, it is necessary to assign a distribution to variances of fitted atomic energies. Of note is the fact that fitted atomic energy variance is not the same as sampling variance. To begin, let's briefly review some literature on characterizing and assessing probability distributions and data.

The Normal distribution is a common distribution for modeling data, especially of a large size. Hence, it can model and statistically summarize Normal data. Common methods to assess normality are graphical, usually the data quantiles are plotted against the standard Normal quantiles. Another common test, that is non-graphical is the Shapiro-Wilk test, which is generally only recommended for sample sizes 50 and under. Box-plots can also be used as a simple quantile graphical tool to assess normality. A histogram plot of the data is yet another common graphical method for assessing normality. It is often that analysts will use these graphical tools to deem data that 'roughly' fits as normal or 'approximately Normal'. Let us look at a study that presents an alternative graphical method to assess Normality that was compared with various tests using Monte Carlo testing[6]. The results of the study suggest a potential more evident way to reject un-Normal data.

According to the study, in finance literature, a plot of empirical and fitted normal densities on the log scale is regularly preferred as a graphical means to assess normality. The study argues that interpretation of quantile-quantile plots can be compromised; assessing degree of curvature in the plot to designate data as not Normal is largely subjective. They then discuss a graphical alternative, the log-density (empirical density) plot. The graphical procedure involves plotting empirical density alongside fitted Normal density. The log-scale is used to clearly display the tails of non-Normal data. One weakness of this technique is that it was said to be possibly misleading for small sample sizes (which shouldn't be a surprise). In the study, a 1,000 simulation Monte Carlo test was performed to simulate a p-value for the log density method, which was compared to other Normality assessing tests (Shapiro-Wilk, Anderson-Darling, and Cremer von-Mises). Power was used to compare the various tests. Data from three distributions were used: Cauchy, t(with 6 degrees of freedom), and Gumbel (extreme value). In all but the extreme value distribution, the Monte Carlo log density test had the highest power. The author argues this is due to the sensitivity of the Monte Carlo Log Density to fat-tailed departures from normality, but being weaker for skewed (extreme value) distributions. Another test comparison between quantile-quantile plots and log density plots showed a Normal appearing quantile-quantile plot, but an obvious non-Normal log density plot. One of the biggest drawbacks of the log density method presented in the study is that it is cumbersome to compute empirical density, hence most would likely prefer the more easily implemented quantile-quantile plots for approximating normality.

A brief literature review of skewness[7] will conclude the data characterization review. The paper is essentially a review of skewness and argues in favor of incorporating it into statistical education. There are various ways to assess skewness, graphical and quantitative. One quantitative method is in the form of a skewness statistic, which compares the mean to the median. Visual tools to assess skewness include beam-and-fulcrum plots, boxplots, and dotplots; dotplots being deemed the best visual tool to assess skewness. The Fisher-Pearson coefficient of skewness, measures skewness and is the ratio of average distance cubed from the mean to average distance squared from the mean (the denominator is raised to the 3/2 power). Of note is that the statistic presented in the paper consists of sample distance from the mean; the mean being the sample average. The brief review of skewness was necessary, because it is often the case that data exhibits non-Normal deviation in the quantile-quantile plots at the extreme ends (tails), and it is necessary to have a means to measure tail deviation as significant skewness, or not. Now that the literature review has been completed, it is time to discuss the experiments and data analysis.

COMPUTATIONAL METHODS

The QM7 database[8] was used to obtain the CMs. The dataset contains 7165 molecules and is a subset of the GDB-13 database (contains nearly 1 billion stable organic molecules). QM7 contain molecules no larger than 23 atoms, and contains the CMs for each molecule and atomization energies. ML models commonly aim to predict the atomization energies using CMs as input.

CMs were stored as row elements in the QM7 database, so it was necessary to regenerate the CMs in matrix form. A random sample of size 500 was taken from the 7165 molecule database, each sample corresponding to a molecule (stored as a CM and atomization energy). Atomic energies (diagonals of CM) were extracted for each CM in the sample. Mean atomic energy and variances were computed for each molecule and stored in vectors. Only molecules with variance of no more than 1000 (units: protons$^{5.76}$) were considered, reducing the sample size to 477. Unless noted otherwise, all calculations

are rounded to the nearest whole number. Unless stated otherwise, variance will refer to sample variance. Unless stated otherwise, atomic energy refers to CM model fitted atomic energy ($.5Z^{2.4}$)

RESULTS AND DISCUSSION

The density histogram of atomic energy (fitted) variance of the random sample of size 477 molecules is shown in Figure 2. The sample mean for this distribution is 562 protons$^{5.76}$ and the corresponding standard error is 118 protons$^{5.76}$. Note that the units here are the same, because the measured quantity is variance. The sample median for this distribution (557 protons$^{5.76}$) is very near in value to the sample mean, indicating the symmetry of this distribution and providing good justification to further assess normality of data.

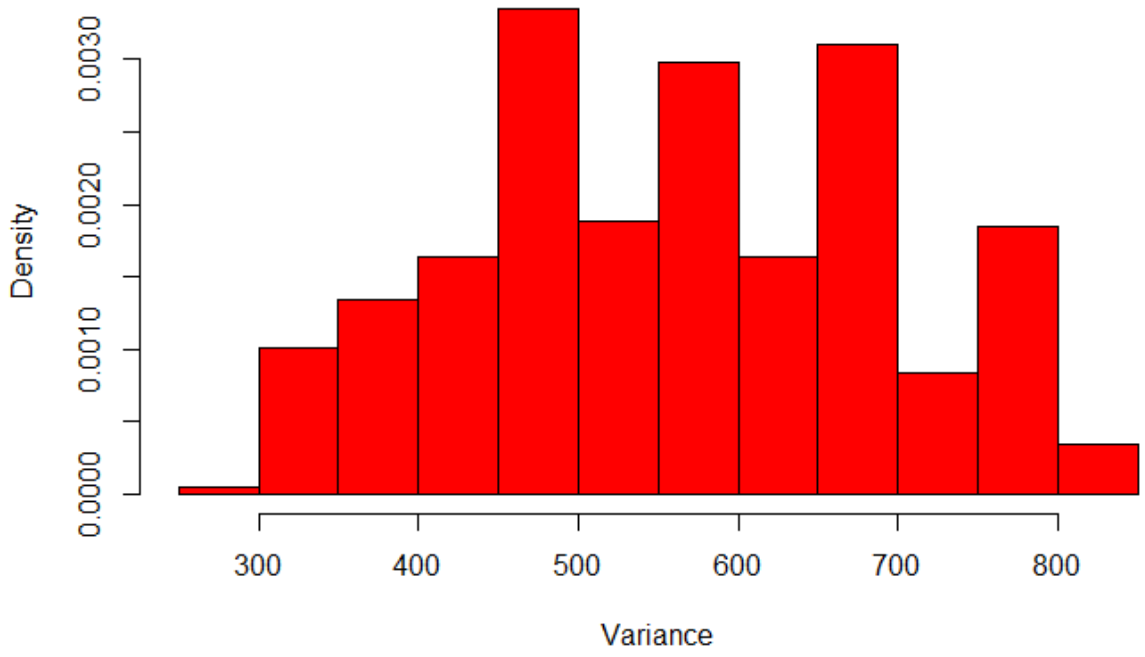

Figure 2. Density histogram of atomic energy variance for the 477 molecule sample

Figure 3 shows the density histogram of the mean fitted atomic energies for the random sample of size 477. For this distribution, the sample mean, median, and variance are 20 protons$^{2.4}$, 20 protons$^{2.4}$, and 19 protons$^{5.76}$. Though the mean and median are equal, the distribution is visually not symmetric with

high variance relative to the sample mean; they are approximately equal. Both distributions are visibly continuous. Figure 4 shows the density histogram of atomic energy variance with fitted normal density superimposed. From the histogram, the Normal distribution is a good approximation of the sample data.

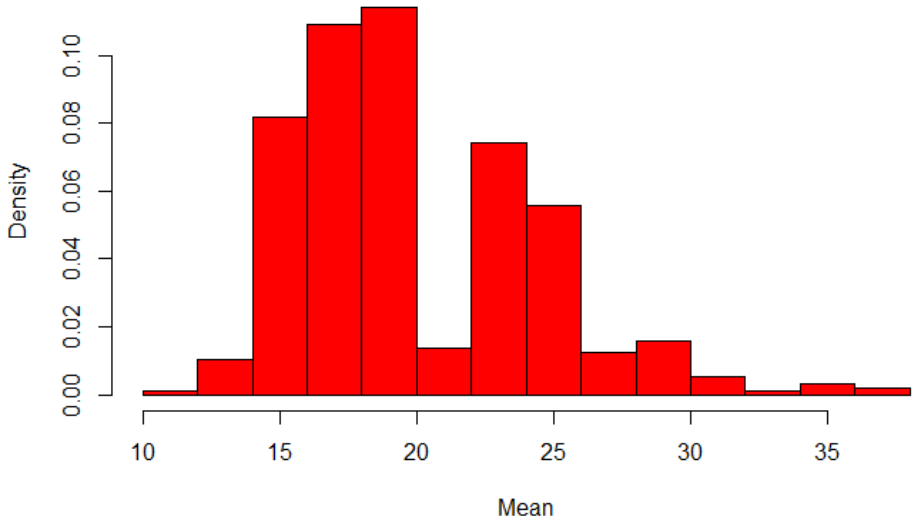

Figure 3. Density histogram of mean atomic energy for the 477 molecule sized sample

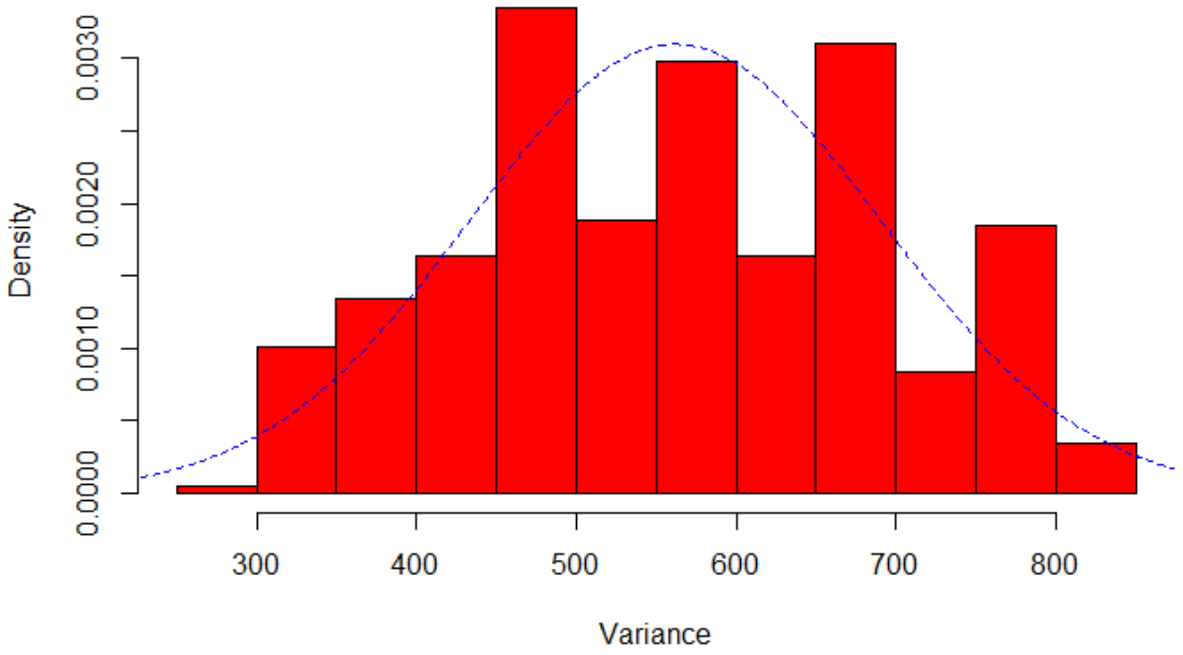

Figure 4. Density histogram of atomic energy variance for the 477 molecule sized sample, with fitted Normal density superimposed (blue)

To further assess normality of atomic energy variance of the sample, a quantile-quantile plot was graphed (Figure 5), which showed bulk normality, but normality deviation for lower and higher quantiles.

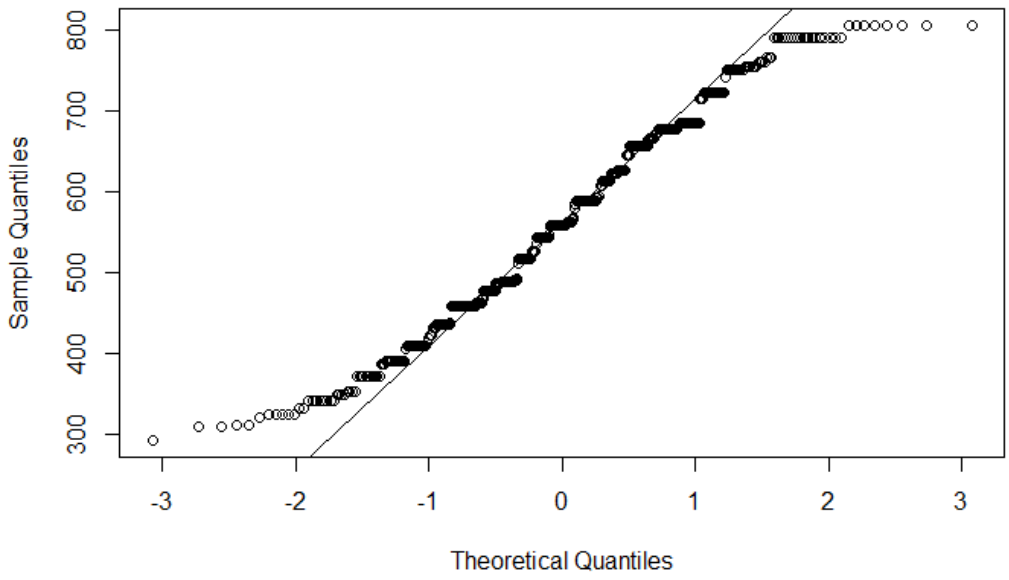

Figure 5. quantile-quantile plot of atomic energy variance for the 477 molecule sized sample.

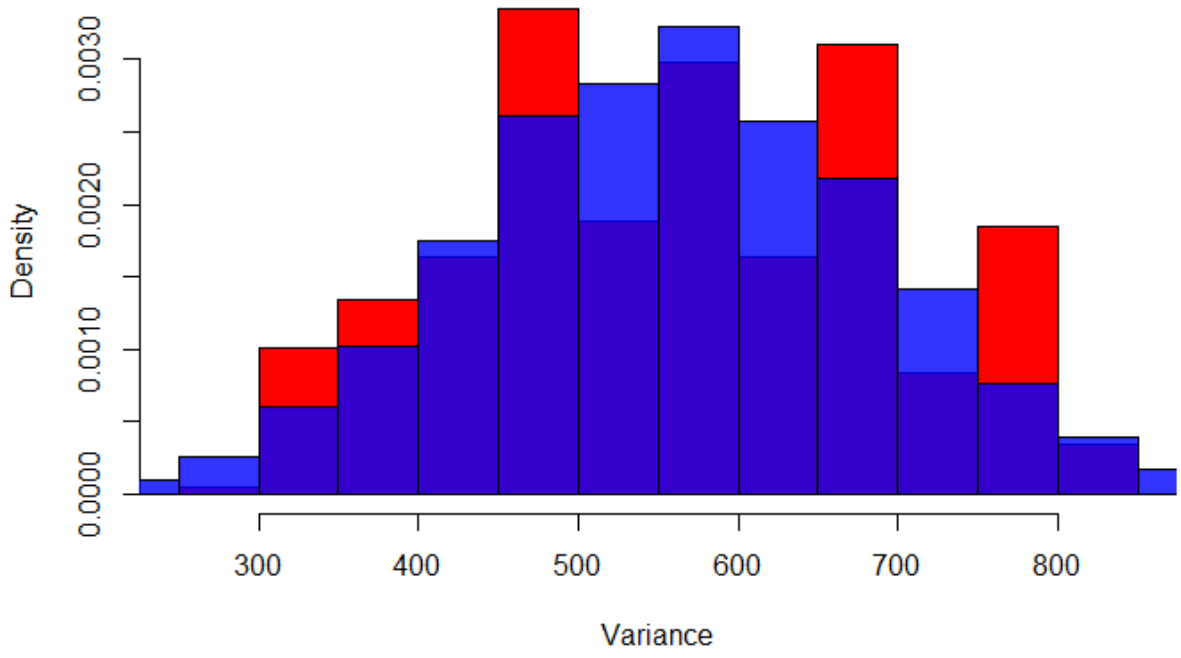

Figure 6. Density histogram of atomic energy variance for the 477 molecule sample, with fitted Normal density histogram superimposed (blue).

To assess normality further, an algorithm was created that superimposed Normal density over the density of atomic energy variance. The Normal density represented a population of atomic energy variance with mean(atomic energy variance) equal to the estimated mean (562 protons$^{5.76}$) and standard deviation equal to estimated standard deviation (128 protons$^{5.76}$); note that the units are the same due to the measured quantity being variance. The size of the population was set equal to the original population size of the QM7 dataset, 7,165 molecules. Each simulation (algorithm run) superimposed randomly generated Normal density with mean variance and standard deviation given by the estimates. The density histogram of a simulation is shown in Figure 6. The results further support that the Normal probability distribution is a good model to summarize and approximate the atomic energy variance sample data.

Next, an alternative model to classify data as approximately Normal will be introduced. The model is a combination of a classification tree and statistical hypothesis testing, and will be referred to as TDT (Test-performing Decision Tree). Figure 7 displays the decision tree. The advantage of the test over visual (subjective) methods to assess and 'measure' normality is that TDT performs a hypothesis test for Normality. The test ends in a composite test statistic, which is used to make a decision that the data is Normal or not Normal. The test is also fully automated; user interpretation is not a component of the test. Note that Normal refers to being Normal from an input population (or parent data). The decision tree splits are based on the data proportions of a Normal distribution: 68% of the data falls within 1 standard deviation units of the mean, 95% of the data falls within 2 standard deviation units of the mean, and 99.7% of the data falls within 3 standard deviation units of the mean. The actual acceptance criteria for proportions is relaxed, since the TDT tests for approximate Normality: 58% to 83% of the data must fall within 1 standard deviation unit of the mean, 85% or more of the data must fall within 2 standard deviation units of the mean, and 95% or more of the data must fall within 3 standard deviation units of the mean. The test tallies the counts in each class and computes proportions. The proportions form the composite test statistic. Data is deemed Normal only if the proportions fall within the percentage ranges discussed prior. TDT performs a hypothesis test, the null hypothesis is that the Data is Normal (with mean equal to the population mean and standard deviation equal to the population standard deviation). The alternate (research) hypothesis is that the data is not Normal (with mean and standard deviation equal to that of the population). The model inputs are the sample and the parent sample (or population).

To assess the strength of the model at correctly identifying Normal data from a given population as Normal, 10,000 simulations were performed to measure type 1 error (rejecting the null hypothesis when the null hypothesis is true). TDT inputs were a population (size 7,165) of randomly generated realizations from the standard normal distribution and a random sample (size 477) from the population. The sizes of the population and sample correspond to the size of the QM7 dataset and variance random sample used in This study. Three iterations of 10,000 simulations were performed, each having a type 1 error rate of 0. TDT was then used on the atomic variance sample of interest, which classified the data as Normal. Figure 8 shows the R output of the TDT test.

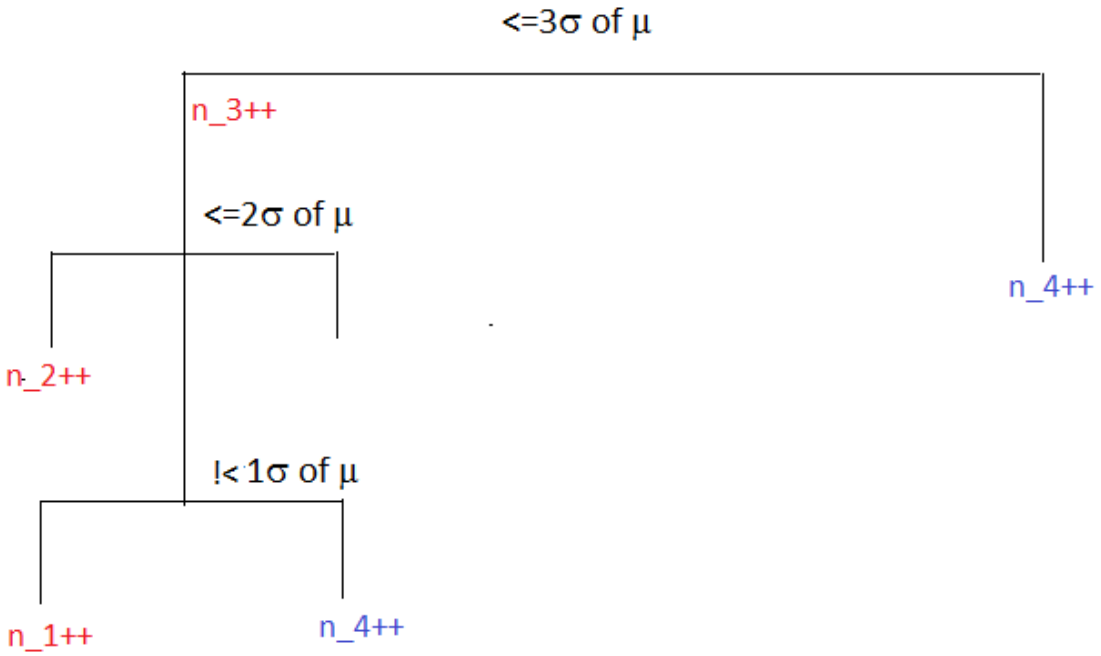

Figure 7. Classification tree for TDT model. The proportion of n_1 to N represents data within 1 standard deviation of mean.

```
test statisticA, p_1sd, is:  0.6771488  .test statisticB, p_2sd, is:
0.9979036 . test statisticC, p_3sd, is:  1 . Accept null. X ~
Normal(mean_population,sd_population)
```

Figure 8. Output of Normality classification algorithm, classifying 477 size atomic energy sample as Normal

The next part of the analysis is determining if the atomic energy variance can be modeled as a linear function of mean atomic energy. Figure 9 is a scatter plot of the 477 molecule sized sample.

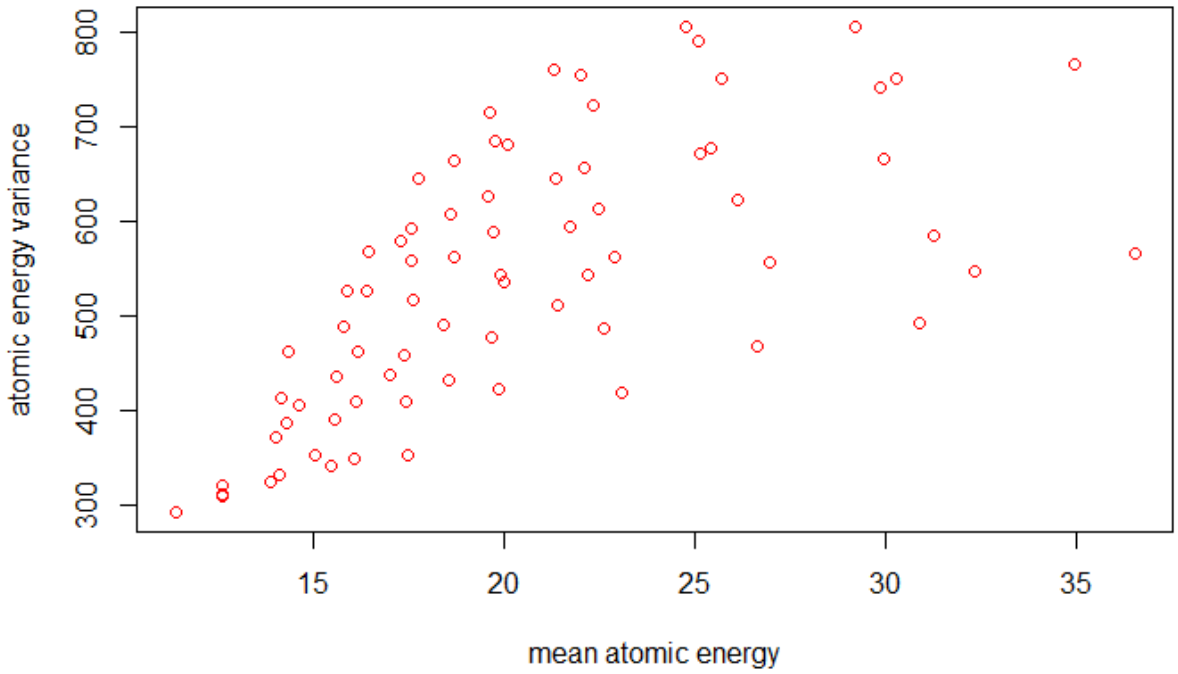

FIGURE 9. Scatter plot of atomic energy variance against mean atomic energy for the 477 sized random sample

The plot shows a clear positive correlation between atomic energy variance and mean atomic energy, however it also shows a fanning pattern and it is not hard to see that a line fit through the data would have residuals of nonconstant variance, a violation of linear regression. To reduce the fanning so that a linear model could be fit to the data, only molecules with a maximum variance of 350 protons$^{5.76}$ were considered, shrinking the sample size to 25. Examining the new condensed sample showed that numerous molecules shared the same mean atomic energy and/or atomic energy variance, which weakens the ability to use linear regression as a model for the statistics of interest, however the data will still be shown and briefly discussed. Figure 10 shows the scatter plot, showing a linear relationship between atomic energy variance and mean atomic energy. Figure 11 is the same plot with the best fit regression line shown.

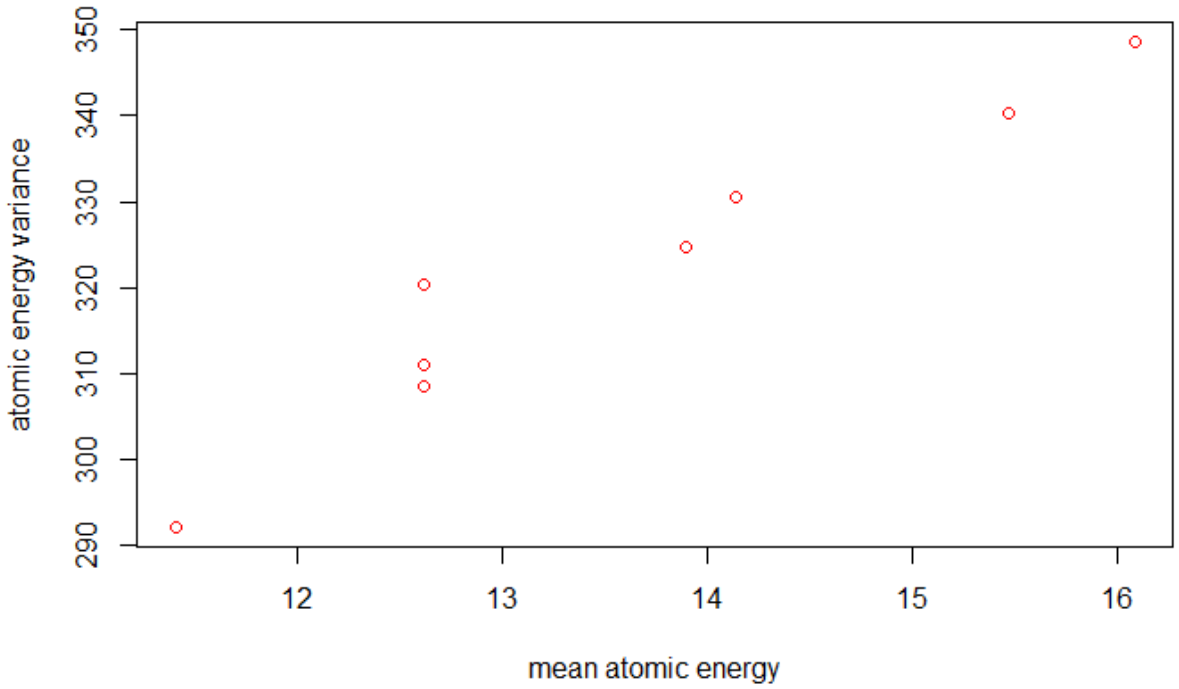

FIGURE 10. Scatter plot of atomic energy variance against mean atomic energy, maximum variance restricted to 350 (protons$^{5.76}$)

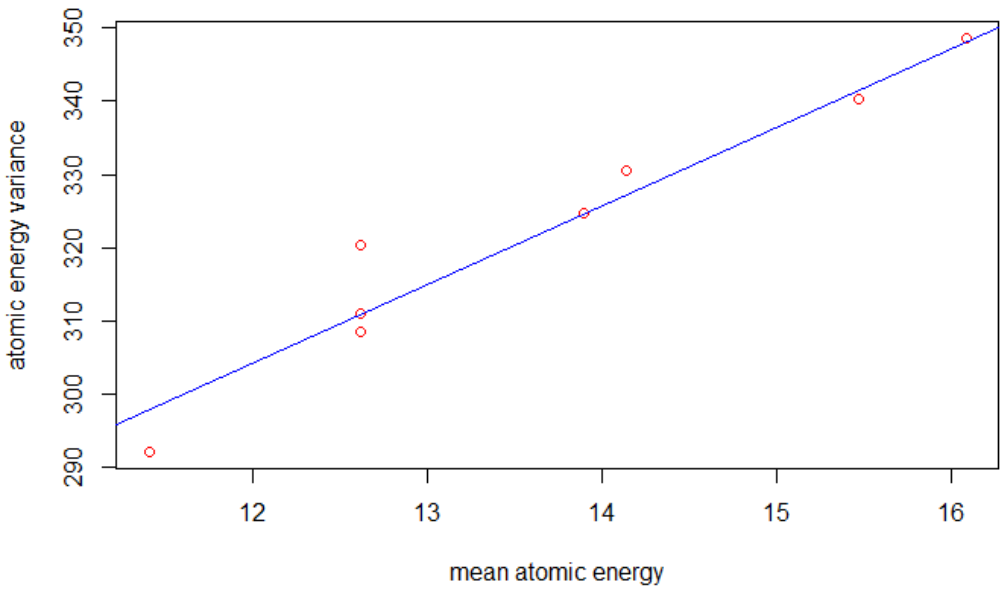

Figure 11. Scatter plot of atomic energy variance against mean atomic energy, maximum variance restricted to 350 (protons$^{5.76}$). Best fit line shown.

Figure 12 shows the R output of the linear model, and figure 13 is a plot of the residuals against mean atomic energy. The plot shows approximate constant variance, but there are some outlier residuals.

```
Call:
lm(formula = atomic_energy_variance ~ atomic_energy_mean)

Coefficients:
       (Intercept)  atomic_energy_mean
            175.72               10.71
```

FIGURE 12. R output of variance restricted (350 protons$^{5.76}$ maximum) linear model of fitted atomic energy variance as a function of mean atomic energy

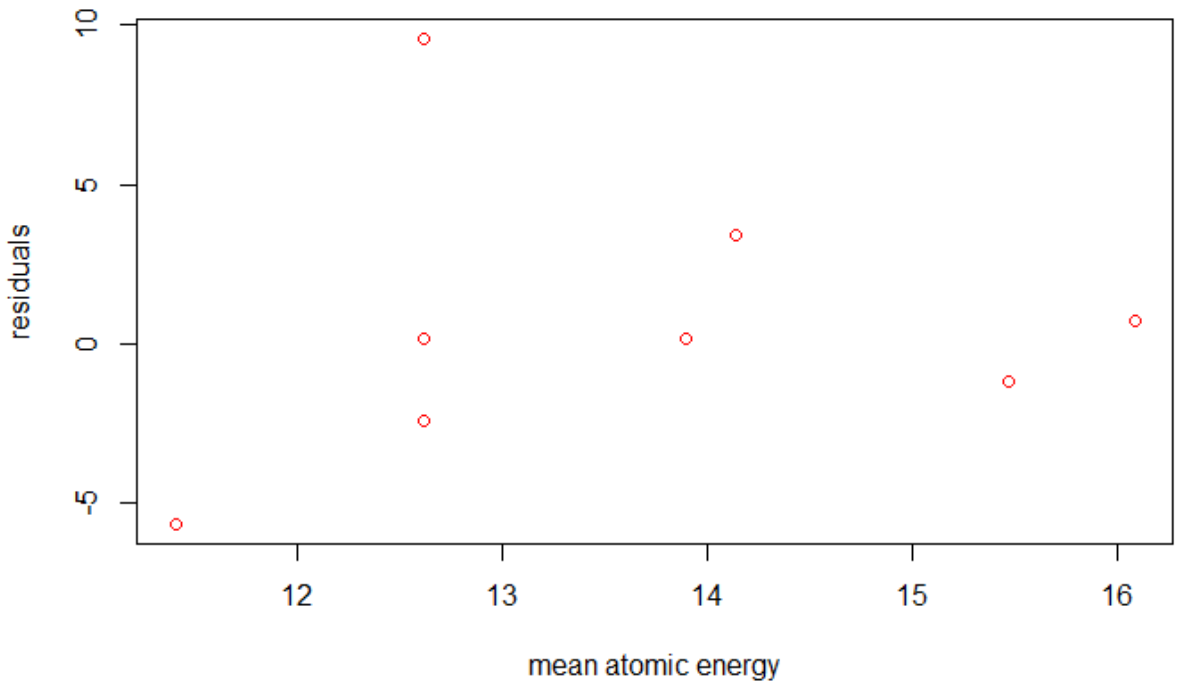

Figure 13. Scatter plot of residuals vs mean atomic energy for variance restricted linear model.

The final portion of the data analysis concludes with variance visualization. Though commonly used to characterize data in statistical analysis, visualization of variance is not regularly used in statistical analysis and data visualization. This portion of the analysis presents a model to visual variance. Atomic energy sample variance is known for each of the 477 molecules in the random sample of interest. Sample variance is the average squared distance from the sample mean; emphasis here is put on the

fact that it is a distance. The distance between two points on a line can be calculated as the hypotenuse of a triangle: $c=(a^2+b^2)^{1/2}$, where a is the horizontal segment length between the two points and b is the vertical segment length between the two points. The variance visualization technique involves computing the vertical component of sample variance for a fixed horizontal component (this set represents point B), then plotting a line containing 0,0 (point A) and point B. The length of this line is the sample variance. The variance, represented as a vector, is translated to the origin (all variance vectors use the origin as point A). Figure 14 is a plot of a random sample of size 30 from the atomic energy variance sample (size 477). The mean, median, and standard deviation of this sample is 584 protons$^{5.76}$, 583 protons$^{5.76}$, and 121 protons$^{5.76}$, which are near in value to the values of the parent sample. The horizontal component of the variance was fixed at 1. Figure 14 shows the distribution of variance, visualized as vectors representing atomic energy sample variances. Figure 15 is a larger sized sample (100) with half the points assigned a horizontal variance component of 1, the other half assigned -1. Since the horizontal component is fixed (magnitude 1), the vertical component in the plots, at the scale of variance in This study, capture almost all sample variance. Visualization of variance will be the subject of further research in the near future. Note that this visualization of variance (distribution of variance) model is nonparametric.

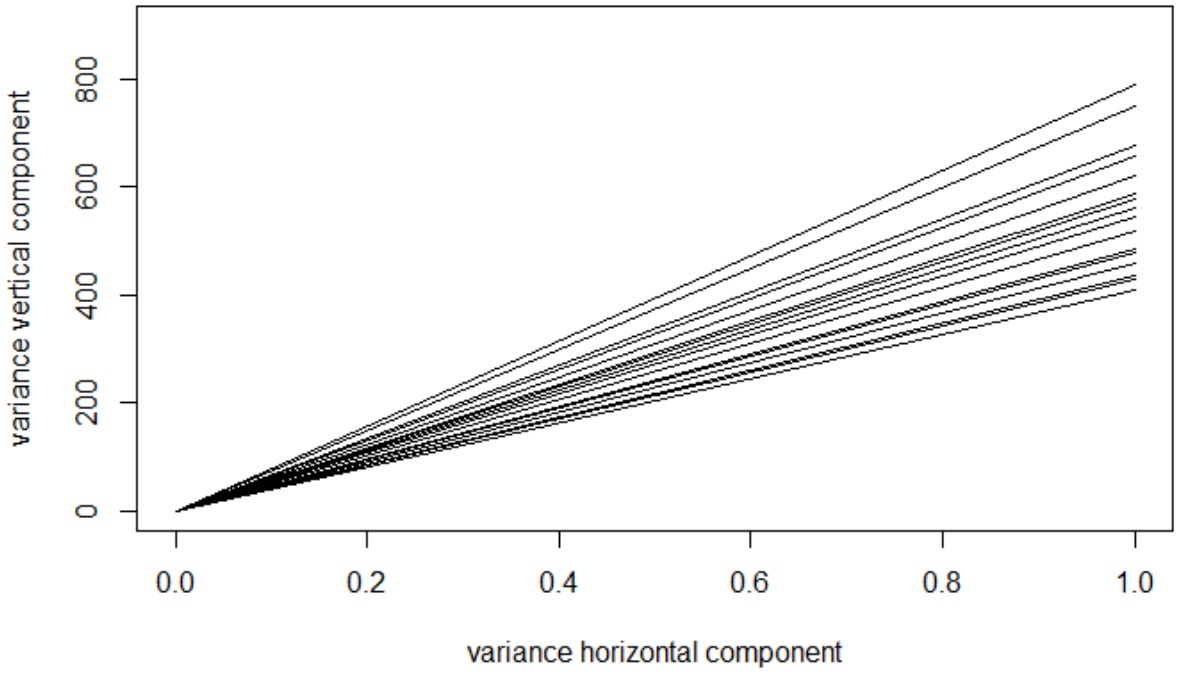

Figure 14. Distribution of variance translated about the origin for a random sample of size 20.

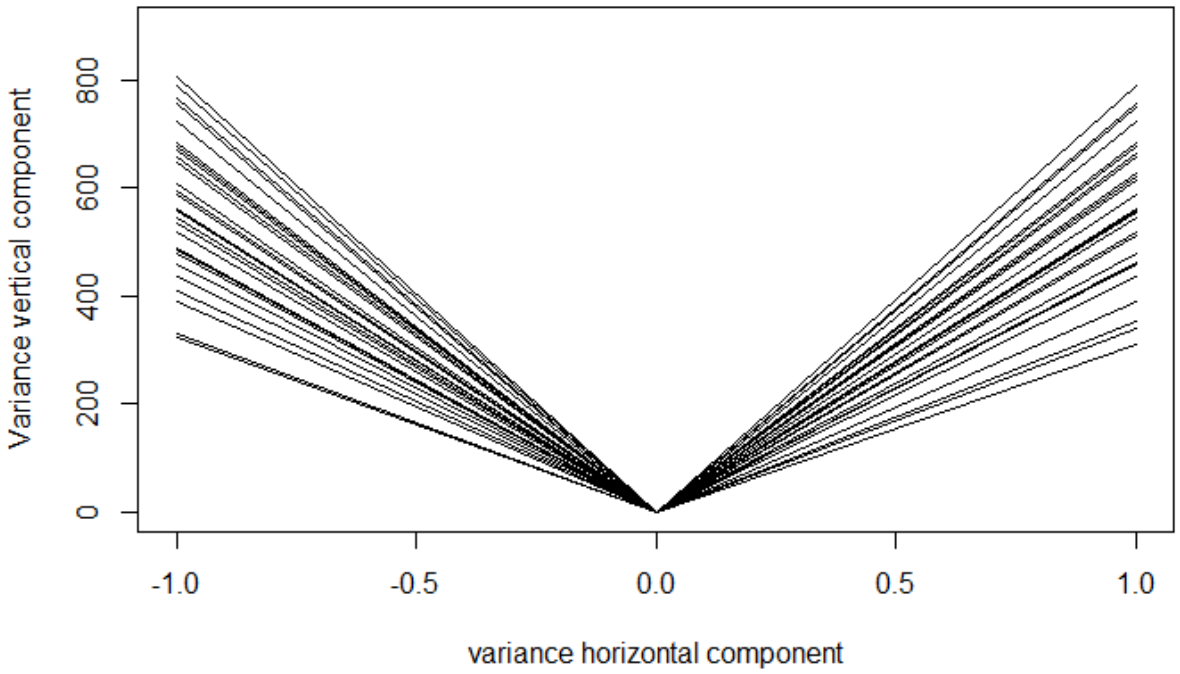

Figure 15. Distribution of variance translated about the origin for a random sample of size 100. X is fixed at 1 for half the sample, -1 for other half.

Conclusion

Variance was modeled for a random sample of molecules from the QM7 database. Modeled variance represented fitted atomic energy variance. A rigorous assessment of normality was performed using simulation and multiple visualization methods, including quantile-quantile plots. An alternate model to assess normality was introduced, the TDT (Test-performing Decision Tree), which computes a test statistic for accepting or rejecting data as Normal based on proportion amounts of data that are within 1,2, and 3 standard deviations of the mean. TDT is a combination of statistical hypothesis testing and machine learning/statistical learning classification trees that uses sample and population data (or parent sample data) as inputs. Three iterations of 10,000 simulations were performed using TDT, each resulting in a type 1 error rate of 0. TDT classified the fitted atomic energy variance sample as Normal. TDT is a fully automated model for assessing normality that is independent of user interpretation. Linear regression was attempted to determine if the fitted atomic energy variance is a linear function of mean fitted atomic energy. A scatter plot showed a strong positive correlation, but the fanning pattern in the plot was evidence linear regression cannot be used on the sample. Maximum variance was restricted (lowered) and linear regression was attempted. Though the best fit line fit the data well, numerous points shared mean fitted atomic energy and/or fitted atomic energy variance, providing evidence linear regression should not be used. For future analysis, similar groups will be considered from QM7 (example atmospheric or greenhouse gases, greenhouse related gases) and linear regression will be reattempted. The variance-driven analysis concluded with the introduction of a variance visualization nonparametric

model, which provided visualization of variance about a fixed location (the origin). It is hoped that the models in this paper will be considered and used to augment statistical analysis, in numerous areas of application.

REFERENCES

(1) Rupp, M.; Tkatchenko, A.; Muller, K.; Llienfled, O. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *PRL* 2012, *108*, 058301-1-058301-5.
(2) Himmetoglu, B. Tree based machine learning framework for predicting ground state energies of molecules. *J. Chem. Phys.* 2016, *145*.
(3) Schrier, J. Can One Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?. *J. Chem. Inf. Model* 2020, *60*, 3804-3811.
(4) . AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* 2012, *58*, 218-226.
(5) Gerberick, G.F.; Vassallo, J.D.; Foertsch, L.M.; Price, B.B.; Chaney, J.G.; Lepoittevin, J. Quantification of Chemical Peptide Reactivity for Screening Contact Allergens: A Classification Tree Model Approach. *Toxicol. Sci. 2007, 97*, 417-427.
(6) Hazelton, M.L. A Graphical Tool for Assessing Normality. *AM STAT* 2003, *57*,285-288.
(7) Doane, P.D.; Seward, L.E. Measuring Skewness: A Forgotten Statistic? *J. Educ. Stat.* 2011, *19*,1-18.
(8) QUANTUM-MACHINE.ORG. http://www.quantum-machine.org (accessed November 08, 2020)
(9) Sheather, S.J. A Modern Approach to Regression with R, Springer: New York, 2009.
(10) James, G., Witten, D.; Hastie, T.; Tibshirani, R. An Introduction to Statistical Learning, Springer: New York, 2013