Spring 2022

Applications of a Combined Approach of Kinetic Monte Carlo Simulations and Machine Learning to Model Atomic Layer Deposition (ALD) of Metal Oxides

Emily Justus
Missouri State University, emilyjustus826@gmail.com

APPLICATIONS OF A COMBINED APPROACH OF KINETIC MONTE CARLO SIMULATIONS AND MACHINE LEARNING TO MODEL ATOMIC LAYER DEPOSITION (ALD) OF METAL OXIDES

A Master's Thesis

Presented to

The Graduate College of

Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science, Material Science

By

Emily Justus

May 2022

APPLICATIONS OF A COMBINED APPROACH OF KINETIC MONTE CARLO SIMULATIONS AND MACHINE LEARNING TO MODEL ATOMIC LAYER DEPOSITION (ALD) OF METAL OXIDES

Physics, Astronomy, and Material Science

Missouri State University, May 2022

Master of Science

Emily Justus

ABSTRACT

Metal-oxides such as ZnO or Al2O3 synthesized through Atomic Layer Deposition (ALD) have been of great research interest as the candidate materials for ultra-thin tunnel barriers. In this study, I have applied a 3D on-lattice Kinetic Monte Carlo (kMC) code developed by Timo Weckman's group to simulate the growth mechanisms of the tunnel barrier layer and to evaluate the role of various experimentally relevant factors in the ALD processes. I have systematically studied the effect of parameters such as the chamber pressure temperature, pulse, and purge times. The database generated from the kMC simulations was subsequently used as descriptors in the subsequent analyses via Machine Learning algorithms. The simulated results of a combined approach of kMC and ML were then compared to the experimental results.

KEYWORDS: kinetic Monte Carlo algorithm, simulation, zinc oxide, atomic layer deposition, machine learning, linear regression, multilayer perceptron, M5P

APPLICATIONS OF A COMBINED APPROACH OF KINETIC MONTE CARLO SIMULATIONS AND MACHINE LEARNING TO MODEL ATOMIC LAYER DEPOSITION (ALD) OF METAL OXIDES

By

Emily Justus

A Master's Thesis
Submitted to the Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master of Science, Material Science

May 2022

Approved:

Ridwan Sakidja, Ph.D., Thesis Committee Chair

Tiglet Besara, Ph.D., Committee Member

Kartik Ghosh, Ph.D., Committee Member

Julie Masterson, Ph.D., Dean of the Graduate College (This line shows the format to use.)

ACKNOWLEDGEMENTS

The support from NSF (EPMD Division) (Award No. 1809284) is greatly acknowledged.

I would like to thank the following people for their support during my graduate studies: Dr. Sakidja for the continuous guidance in my research, Dr. Frodermann and Dr. Besara for introducing and teaching me the world of physics, and Melissa, my sister, for constantly encouraging and helping me grow through tough times.

I dedicate this thesis to Bean and Luna. The two of you have taught me patience and forgiveness. You are blessings in my life that I will cherish forever.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Thin-film materials are widely researched for their properties. Atomic layer deposition (ALD), a form of chemical vapor deposition, can produce thin films. When creating any thin film there are multiple parameters needed to create it, such as pressure, temperature, and precursors are just a couple of examples. Depending on what properties are needed, the best parameters are different. Although finding the best parameters of a thin film can be expensive from the equipment and materials and can take time. For researched materials, the most efficient parameters may already be known. Another approach to estimating this procedure is through simulations which can alleviate issues. An example would be to use kinetic Monte Carlo (kMC) simulations. This process can customize the different parameters like ALD and can show a thorough understanding of what is happening throughout the entire process. Although kMC is monetarily inexpensive and can be quicker than the ALD process, there is a large amount of data and calculations needed before the simulation can happen. In recent years machine learning has become prevalent in research in material science. With machine learning, it can combine everything learned from the simulations to create an algorithm to be able to estimate any input which is the goal without the need for prior calculations.

Using machine learning in this way for thin films is still new to the material science community. The best way to start this topic is through a studied material such as zinc oxide (ZnO). Zinc oxide is a semiconductor that is known to have a multitude of different applications for its electrical and optical properties, such as solar cells and sensors [1]. Since this material is well known and researched, then it will be useful to compare the experimental results with the simulated results.

Atomic Layer Deposition Process

Atomic layer deposition can produce thin films through a series of pulses and purges of a precursor and an inert gas on the surface to create layers [2, 3]. For zinc oxide, the precursors are diethylzinc (also known as, DEZ or $(C_2H_5)_2Zn$) and water. DEZ is pulsed onto the zinc oxide (100) surface for a specified length of time, ideally until fully saturated. During this time, the DEZ will bind to the surface creating a new layer on the film. An inert gas is used to purge any excess DEZ off the surface. Water is pulsed after, reacting with the DEZ surface for a specified length of time and because of its self-limiting surface reactions, the layers are easily achieved. Monoethylzinc (MEZ) can be formed when DEZ is exposed to water and if water is still present then can remove any ethyl-ligands. Inert gas is used again to purge excess water off the surface. This process is repeated until the desired number of layers or film width is achieved [3].

Kinetic Monte Carlo Simulation

A way to approximate ALD is to use kinetic Monte Carlo (kMC) simulations which have been shown with separate materials such as ZnO, HfO_2, ZrO_2, and Al_2O_3 [4-7]. These are derived from Shirazi and Elliott, who created the ALD algorithm using a stochastic parallel particle kinetic simulator (SPPARKS) for HfO_2 [8]. From this program, other computationalists have modified the program for varied materials, as stated before. For ZnO, there is a modified program made by Weckman, T., et al. [4]. This program contained an example of ZnO and HfO_2. From this, we can modify the parameters to find different outcomes.

KMC Process. The kMC algorithm runs off a random number generator which is derived from Monte Carlo simulations. Although those types of simulations have trouble with advanced ideas such as surface reactions, which the kinetic Monte Carlo algorithm can manage. The state-

to-state evolution is based on the Bortz–Kalos–Lebowitz (BKL) algorithm. The sum of the rate coefficients for all events, k_{tot}, is defined in Eqn. 1. Then to start the kMC, it picks two random numbers, ρ_1 and ρ_2, between 0 and 1. The first random number is used to pick the random event, q, which must satisfy Eqn. 2. When this event occurs, the event list is updated and any event that reacts as a by-product of the initial event will occur. Any secondary event that occurs updates the event list, as well. Although this only happens for a limited number of events. The second random number is used to develop the time evolution shown in Eqn. 3. The time step must reach a certain threshold, or this step would restart. This process is repeated until there are no reactions left to occur. The time step is independent of the event selected but dependent on the sum of the rate constants. In general, this means that when there is a high rate of constants total, this means there is a substantial number of reactions available, then the time step will slow down as shown in Eqn. 3. When a considerable number of reactions occur causing the time step to slow down, this will result in the program slowing down as well. Once the lattice is fully saturated with no more open positions available, then there will be no more mass gain. However, when running the program, we cannot assume that if the mass gain is no longer increasing that this means the lattice is fully saturated. This means that there are no more reactions to occur, which can be achieved by having a lower temperature making the reactions less likely to occur.

$$k_{tot} = \sum_{n=1}^{N} k_n \qquad \text{Eqn. 1}$$

$$\sum_{n=1}^{q-1} k_n \leq \rho_1 k_{tot} \leq \sum_{n=1}^{q} k_n \qquad \text{Eqn. 2}$$

$$t = t - \frac{\ln(\rho_2)}{k_{tot}} \qquad \text{Eqn. 3}$$

The kMC algorithm uses an on-lattice structure meaning that every point along the structure is predefined. Atoms or species can more around the lattice, but only in those defined points. Although this method is not a replica of real life, it can estimate the process well when compared to experimental values [4, 9, 10]. There are different methods of kMC which would provide an off-lattice structure, known as adaptive kMC, but it will not be discussed in this paper[11].

KMC Shortcomings. Running kMC simulations can explore the different possible outcomes with slightly different parameters to find the best optimal outcome. With ZnO, we know the best outcomes, so if we can work with common materials and understand the program then we can work with uncommon materials more easily. However, to start running the program you will need all the density functional theory (DFT) calculations to calculate the activation energies and the reaction list for every reaction [4, 8]. These items themselves can take time to acquire depending on the material. Although, being able to use these files in simulation repeatedly trumps the amount of time it takes to build. Using these files to get a better understanding of the material and being able to manipulate them to find the best outcome through kMC can provide use to experimentalists when trying to make a better material.

Another downfall of using kMC is the amount of time it takes to run. When running bigger lattice models, meaning a bigger box with more atoms, or at higher temperatures, it will take longer for the simulation to run. Having a large model takes a longer amount of time due to the mathematics used in kMC. The more rate constants that occur, then the longer it will take because of the considerable number of reactions that can occur making the time step smaller. As

the temperature rises reactions are more willing to work together overcoming the activation energy, which also produces a smaller time step.

KMC Advantage. One of the benefits is stated above, such as the ability to repeat a simulation as many times as desired at no cost. Having the ability to change the parameters minutely can optimize results. Also, from the outputs of the simulation, it is possible to learn more than what can be achieved through experimentation. For example, we can determine the number of occurrences of each reaction and each species. We have an understanding that kMC is a good approximation to experimental data by comparing the growth-per-cycle/angstrom over temperature [4], hydrogen and carbon content over temperature, and oxygen to zinc ratio versus temperature [10]. From this, we can get an approximation of what is occurring to the species.

Machine Learning

To combat the shortcomings of kMC we have begun to look toward artificial intelligence (AI). Artificial intelligence tries to mimic the human mind to learn and make decisions. A branch of AI is machine learning (ML) which is a tool for data-driven models to solve problems. Although ML is trying to mimic human reason, if given enough data, then it can be proven to be smarter than the human mind. Sometimes ML can show patterns that may not have been noticed.

ML Types. There are, in its simplistic form, four distinct types of ML, supervised, semi-supervised, unsupervised, and reinforcement learning. Supervised learning is used when one result is desired, without knowing the exact parameters required to achieve that result. This technique requires an input of data, which will split into either the training dataset or the testing dataset. Training data is used to train the model. Testing data is used to evaluate the model

predicted. Both data sets are used to make the final prediction model. There are two distinct types of supervised learning, which are classification and regression. Classification is typically used for qualitative outputs, while regression is used for quantitative outputs. An example of supervised learning would be determining the species of an Iris based on pedal length and width and the sepal length and width [12]. Unsupervised learning is used when the input is randomized and not properly labeled. This method can find patterns in the input and can find groups of similar data through clustering. Clustering looks through the data and combined them into groups of like attributes. An example of unsupervised learning is fraud detection by trying to find an outlier outside of usual purchases [13]. Semi-supervised learning is a combination of supervised and unsupervised learning. Initially, the data set starts with a majority as unlabeled, and a small amount as labeled. Then the semi-supervised method uses the few that are labeled to create a model that can predict the remaining unlabeled. An example of semi-supervised learning is speech analysis [13]. The final machine learning method is reinforcement learning. This method uses a reward/penalty system to train the data. An agent is observing their environment and will decide. If the decision is wrong, then it will be punished and if the decision is current, then it will be rewarded. Based on this, this ML method will try to avoid the penalty and try to go to the reward [13, 14]. A common example of reinforcement learning is video games. Mainly when working with a numeric set of data to predict models like ours, supervised learning is the way to go, which is commonly used in chemical research as an example [15]. There is a multitude of different algorithms that can be used for supervised learning and each one is different in its own way. There is a simplistic model such as linear regression or more complicated examples such as multilayer perceptron, which will be explored later.

COMPUTATIONAL METHODS

Kinetic Monte Carlo Inputs

In the kinetic Monte Carlo simulations, there are many different initial parameters needed, such as pulse and purge length, pressure, and temperature. However, the only parameters that changed were the temperature, pulse length, and purge length. The kinetic Monte Carlo simulation gives an output containing the mass gain and the number of occurrences for oxygen, hydrogen, zinc, MEZ, and ligands. As well as the number of occurrences of each reaction and each species.

The two input files needed to run the kMC simulation (Appendix A and Appendix B) contain all the reactions and the lattice positions. From Figure 1, a portion of appendix A can be seen. The first two columns depict what type of Event is occurring, either 1, 2, or 3. Event 1 refers to reactions that only change the species of a site. Event 2 changes the species of the site and the second nearest neighbor site. Event 3 changes the species of the site and the first nearest neighbor site. The format of the rows differs between Event 1 and Events 2 and 3. For the case of Event 1, the species before and after the reaction can be labeled in 2 columns. While in Events 2 and 3, the species before and after the reaction are labeled in 4 columns. Otherwise, the following columns are all the same. A list of all the types of sites with their description is found in Table 1 [4]. Referring to Figure 1, the third column represents before the reaction and the fourth column represents after the reaction. The fifth column represents the rate of adsorption given by Maxwell-Boltzmann statistics when in adsorption or represents the pre-factor otherwise, and the sixth column represents the exponential factor in the Arrhenius equation. The seventh column represents the activation energy and the eighth column represents the coordination number of the

species in the first site. The ninth column represents whether the reaction with be available for all reactions (0), just for DEZ pulse (1), or just for water pulse (2). From Figure 2, the third and fifth columns are the species before the reaction and the fourth and sixth columns are the species after the reaction.

```
event   1       OH      ZnX2OH  41822.40082     0       0       1       1       ZnX2(g)+OH(s)->ZnX2...OH(s)
event   1       OH      ZnX2OH  41822.40082     0       0       2       1       ZnX2(g)+OH(s)->ZnX2...OH(s)
event   1       OH2     ZnX2OH2 41822.40082     0       0       1       1       ZnX2(g)+OH(s)->ZnX2...OH2(s)
event   1       OH2     ZnX2OH2 41822.40082     0       0       2       1       ZnX2(g)+OH(s)->ZnX2...OH2(s)
event   1       O       ZnX2O   41822.40082     0       0       1       1       ZnX2(g)+O(s)->ZnX2...O(s)
event   1       O       ZnX2O   41822.40082     0       0       2       1       ZnX2(g)+O(s)->ZnX2...O(s
event   1       O       ZnX2O   41822.40082     0       0       3       1       ZnX2(g)+O(s)->ZnX2...O(s)

event   1       OH      ZnX2OH  41822.40082     0       0       -9      1       ZnX2(g)+OH(s)->ZnX2...OH(s)
event   1       OH      ZnX2OH  41822.40082     0       0       -8      1       ZnX2(g)+OH(s)->ZnX2...OH(s)
event   1       OH2     ZnX2OH2 41822.40082     0       0       -9      1       ZnX2(g)+OH(s)->ZnX2...OH2(s)
event   1       OH2     ZnX2OH2 41822.40082     0       0       -8      1       ZnX2(g)+OH(s)->ZnX2...OH2(s)
event   1       O       ZnX2O   41822.40082     0       0       -9      1       ZnX2(g)+O(s)->ZnX2...O(s)
event   1       O       ZnX2O   41822.40082     0       0       -8      1       ZnX2(g)+O(s)->ZnX2...O(s)
event   1       O       ZnX2O   41822.40082     0       0       -7      1       ZnX2(g)+O(s)->ZnX2...O(s)
```

Figure 1. A portion of the input file, Appendix A, contains the reaction list with the activation energy and coordination number.

```
event   3       O       OH      OH2ZnX  OHZnX   8.33E+12        0       0.6     -9      0       O->OH
event   3       OH      OH2     OH2ZnX  OHZnX   8.33E+12        0       0.6     -9      0       OH->OH2
event   3       OH      O       OHZnX   OH2ZnX  8.33E+12        0       0.6     -9      0       OH->O

event   3       O       OH      OH2ZnX  OHZnX   8.33E+12        0       0.6     -8      0       O->OH
event   3       OH      O       OHZnX   OH2ZnX  8.33E+12        0       0.6     -8      0       OH->O

event   3       O       OH      OH2ZnX  OHZnX   8.33E+12        0       0.7     -7      0       O->OH
event   3       OH      O       OHZnX   OH2ZnX  8.33E+12        0       0.6     -7      0       OH->O
```

Figure 2. A portion of the input file, Appendix A, contains the reaction list of Event 3.

From the initial ZnO file received, the default temperature of the simulated chamber during the ALD process is 400K (0.03447 eV), the default pulse and purge length are 0.05 seconds each for a half-cycle with a total cycle length time of 0.2 seconds, and the default pressure inside the simulated chamber is 20 Pa (0.15 Torr) which can be seen in Figure 3 [10]. There are a total of 1280 sites in the lattice which is 26.7992 Å x 20.8264 Å x 29.0119 Å.

The final row states how long the kMC simulation time will run. For the cycle time of 0.2 seconds with a full run time of 2 seconds, then this will produce 10 cycles. Under the row diag_style, that is where the list of outputs is stated. QCM represents the quartz-crystal microbalance which is used to measure the mass gain over a unit area.

The labels oxygen, zinc, hydrogen, MEZ, and ligands output the number of occurrences for each of those in total respectively. The terms with the letter 'v' followed by a numeric value represent a specific reaction. Each reaction has its own identification. For Event type 1, the beginning letter will be 's', for Event type 2, it will be 'd', and for Event type 3, it will be 'v'. The numeric value represents the reaction number in the reaction list.

The other input file (Appendix B) contains the number of sites, the size of the box where the sites are located, and the position of the sites, neighbors, and their values. This file remains the same if the initial structure or placement of the species does not wish to be changed.

From the first input file, we can change the temperature and the pulse/purge length. To change the temperature, the value of temperature shown in Figure 3 must be changed to its respective value in electron volts and the rate coefficient must be changed using Eqns. 4 and 5 depending on if it is the adsorption rate or desorption rate respectively [4, 8, 10]. P is the chamber pressure in Pascals, A_{ads} is the area of the adsorption site, and m_{ads} is the mass of the adsorped in kilograms. K_B is the Boltzmann constant, T is the temperature in Kelvin, and h is Planck's constant.

Table 1. Description of different sites.

Type	Description
VACANCY	An empty site
O	An oxygen atom
OH	A hydroxyl group
OH2	A surface water molecule
ZnX2O	An oxygen atom with an adsorbed DEZ molecule
ZnX2OH	A hydroxyl group with an adsorped DEZ molecule
ZnX2OH2	A water molecule with an adsorped DEZ molecule
ZnXO	A MEZ on an oxygen atom
ZnXOH	A MEZ on a hydroxyl group
ZnO	A zinc atom on an oxygen atom
ZnOH	A zinc atom on a hydroxyl group
Zn	A zinc atom
ZnX	A MEZ group
OH2Zn	A water molecule adsorbed to a zinc atom
OH2ZnX	A water molecule absorbed into a MEZ group
OHZn	A hydroxyl group on a zinc atom
OHZnX	A hydroxyl group on a MEZ group
OZn	An oxygen atom on a zinc atom

```
pulse_time      0.05    0.05    #T1     T3
purge_time      0.05    0.05    #T2     T4      and     cycle   =       T1+T2+T3+T4

#       temperature     in      units   of      eV

temperature     0.03447 #       400     K

#       pressure        in      units   of      torr

diag_style      ald/zno stats   yes     list    events  QCM     OXYGEN  ZINC    HYDROGEN MEZ    LIGANDS v11     v12     v13     v14     v15

stats   0.001
dump    1       text    0.005   dump.ald        id      i1      i2      x       y       z

run     2
```

Figure 3: Input parameters for the pulse and purge length, temperature, desired outputs, and length of time run in total.

$$K_{Ads} = \frac{\sigma PA}{\sqrt{2m_{ads}k_BT}} \approx \frac{PA_{ads}}{\sqrt{2m_{ads}k_BT}} \quad \text{Eqn. 4}$$

$$K_{Des} = \frac{k_BT}{h} e^{\frac{-E_{ads}}{k_BT}} \approx \frac{k_BT}{h} \quad \text{Eqn. 5}$$

From the outputs of these files, the number of total oxygen, zinc, and hydrogen molecules can be found along with the QCM. These values can be compared against each other depending on the temperature and the pulse/purge lengths over time. This can demonstrate the ALD process more thoroughly. Since we're looking into the different cycle lengths, to compare them we choose to look at the values after the third and fifth cycles. Similarly for the temperature; however, looking at the temperature after more time has passed can show a bigger difference.

Machine Learning Algorithms

Before picking out what type of supervised regression algorithm, certain parameters need to be looked at to produce the best results. When looking toward models to approximate, it is advised not to use too many inputs because it can overfit the data making it more accurate, but less useful [16]. To find the best fit is to reduce as many inputs as possible while still having a good

fit. Initially, we looked toward using a few inputs such as time, QCM, zinc, hydrogen, oxygen, MEZ, and ligands with a few different machine learning algorithms. After that, we looked at having different inputs of temperature and cycle length to compare to the kMC results.

Another method to be considered to prevent overfitting is the k-fold method [17-19], which is a cross-validation technique, meaning that it splits the data into training sets and validation sets. K-fold cross-validation is when the data gets split into k different groups equally. One group is designated as the validation group, which is used to create a model, and the remaining k-1 groups are the training sets to test that model. This is iterated over k times and the results are averaged together to get a final model. Both of these processes will be incorporated into all the machine learning methods used to produce the best outcome.

The different types of machine learning methods we focused on that are good determinants of approximation are linear regression [20], multilayer perceptron [21], and M5P [22, 23]. These three algorithms can be modeled in a machine learning software known as WEKA.

Linear Regression. When using the linear regression (LR) model, it chooses the best inputs and creates a linear expression to find the best fit to the actual data. It may choose all or one of the inputs given to form a model. This type of model can show the dependent and independent variables tied to the variable we are looking for. The equation. for a simple linear regression model is shown in Eqn. 6 and a multivariate linear regression model is shown in Eqn. 7 [20]. Since the data is over some time, the multivariate linear regression would better represent the process.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \qquad \text{Eqn. 6}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \text{ where } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \text{Eqn. 7}$$

Multilayer Perceptron. Another approach to supervised learning is artificial neural networks (ANN). ANNs are often compared to biological neural networks because of their parallel computing of large data sets of single processors with many interconnections [24]. For the multilayer perceptron (MLP) model it is a feed-forward artificial neural network, Figure 4, implying that it moves in one direction and continuously moves forward, never going back.

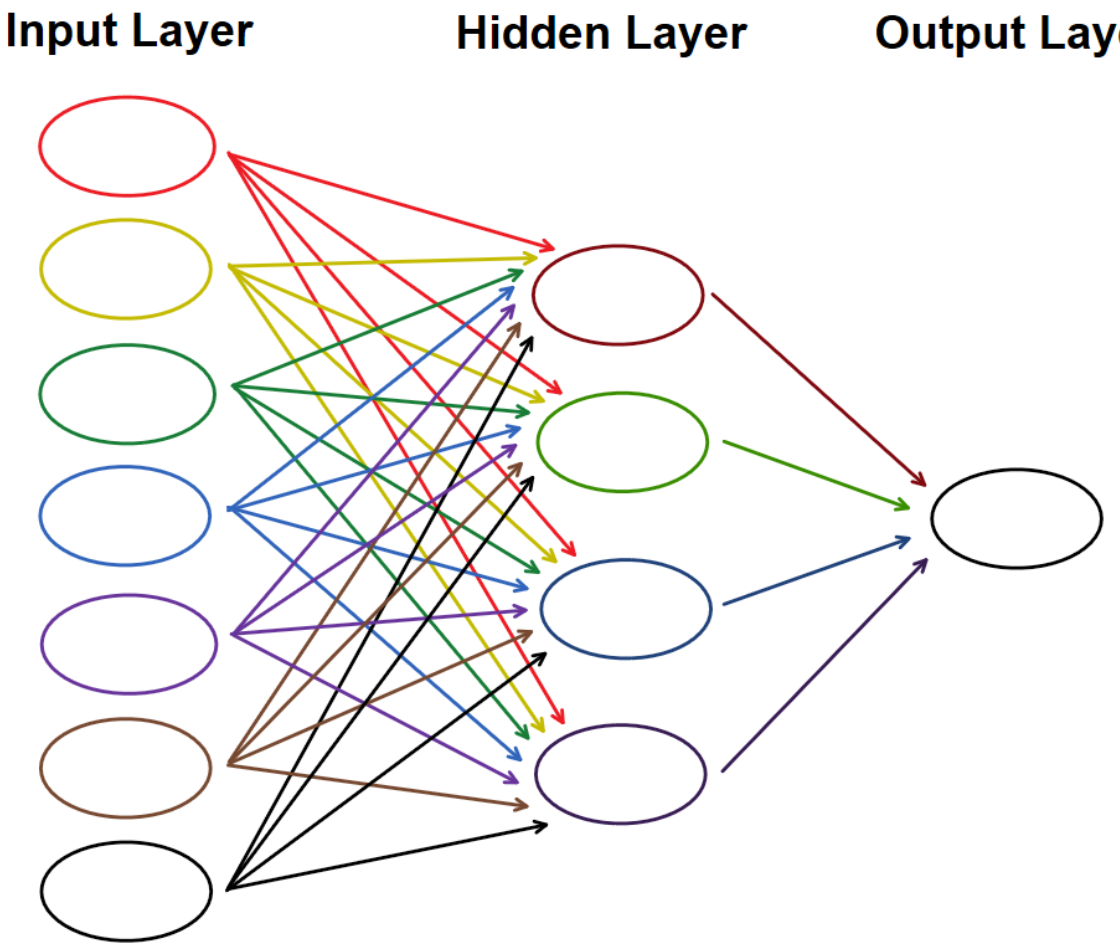

Figure 4. A general artificial neural network (ANN) diagram.

In comparison to linear regression, it uses all the inputs given. The first column contains all the input nodes which connect to the hidden layer's nodes. A hidden layer can contain as many nodes as desired and can be multiple layers thick. Every output of a node is connected to every input of the next layer with a weight. In this case, the hidden layer has one layer with four nodes as shown in Figure 4. The reason why only one layer with four nodes was shown as the hidden layer is that after testing many scenarios with a different number of layers with a combination of different amounts of nodes in each layer, the results showed that the default setting Weka sets produces the best accuracy using some of the accuracy testing stated below. The output layer connects the last row of the hidden layer's nodes to form the result [21].

M5P. Another method of machine learning we will discuss is the M5P tree algorithm, Figure 5, which is a decision tree algorithm for regression. A decision tree algorithm is based on a series of questions and answers [23]. Starting at the top of the "tree" a question will be asked with two possible answers. Depending on the answer it will dictate the path will go to find the output. When it gets to the bottom of the tree to a "leaf", then it will be able to produce an output. This algorithm grows as large as it needs to reduce the amount of error, but also tries to keep it to a minimum. Since M5P is a decision tree algorithm it will have rules or branches that will extend to create a better accuracy. There is a pruning process that trims the branches while keeping a good accuracy. As shown in Figure 5, the training data is put into varaible1, then it compares the training set value with value1. If it is higher or lower then this will determine if it will fall right or left on the tree respectively. Then it will look at the next variable, variable2, to determine if it is higher or lower than value2. This will continue until it reaches a rule. The rule will define the value for the output.

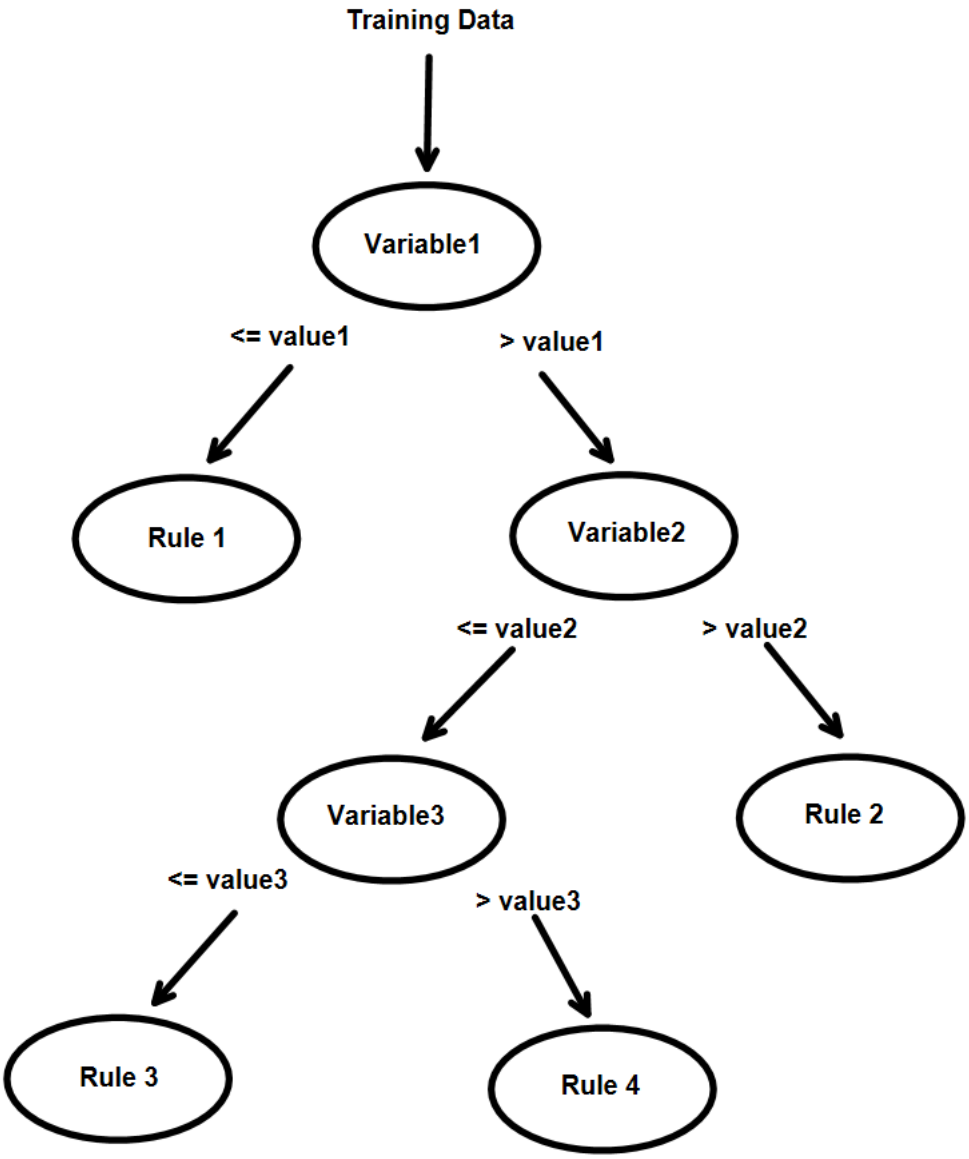

Figure 5. A general decision-tree algorithm of M5P.

Prediction Accuracy

To determine if something is a good fit, there are a few different values to look toward, such as the correlation coefficient (R) and the root mean squared error (RMSE), also known as root-mean-square deviation [25]. There are many different methods to demonstrate the strength of

the prediction, such as coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), mean absolute percent error (MAPE), and symmetric mean absolute percent error (SMAPE) [25].

R and R^2. R and R^2 are used to determine the correlation between variables. In particular, R determines the correlation between the predicted variable and the actual variable and R^2 determines the proportion of variation between the predicted and actual variables [25]. Although R and R^2 have been used to define accuracy, they are biased and only show a correlation between the variables [26]. Although these values are not advised to determine accuracy, they still can predict correlation. Thus as long as this is mind, then it can still be used. Going forward, we will only look at the R-value as it is more directly tied to the correlation. The equation for the correlation coefficient is shown in Eqn. 8 where x represents the actual data and y represents the predicted data [27].

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{Eqn. 8}$$

R values can range from -1 to 1. The closer the value is to zero, then the less correlated the predicted values are to the actual values, and the closer the value is to one or negative then the more correlated it is. If it is closer to a negative one, then this means it has a negative correlation, but it is closer to a positive one then there is a positive correlation. Depending on how close the value is to one or zero can state how well correlated the data is, shown in Table 2 [27]. For example, if in a plot of data, the points are spread out to create a sea of points then this would be close to a negligible correlation. If the points are shaped in a general positive linear-

direction then it could be considered a positive correlation with the level of correlation depending on how close the predicted versus actual points form a linear line. Similarly, if the points are shaped in a general negative linear-direction then it could be considered a negative correlation.

Table 2. Interpretation of R values.

Size of Positive Correlation	Interpretation	Size of Negative Correlation	Interpretation
0.90 to 1.00	Very High Positive Correlation	-0.90 to -1.00	Very High Negative Correlation
0.70 to 0.90	High Positive Correlation	-0.70 to -0.90	High Negative Correlation
0.50 to 0.70	Moderate Positive Correlation	-0.50 to -0.70	Moderate Negative Correlation
0.30 to 0.50	Low Positive Correlation	-0.30 to -0.50	Low Negative Correlation
0.00 to 0.30	Negligible Correlation	-0.00 to -0.30	Negligible Correlation

MAPE and SMAPE. MAPE and SMAPE are considered percentage-based measurements. Mean absolute percent error finds the average of the percent error. A disadvantage to using this method is that there is a difference between positive and negative errors that don't reflect. Therefore it is not a good representation of accuracy. Symmetric MAPE can have the reflection between positive and negative errors that the MAPE can't achieve [25].

MSE, RMSE, and MAE. MSE, RMSE, and MAE are based on a scale. The MSE finds the mean squared error of the predicted versus the actual while the RMSE find the average magnitude of error between them. Similar to RMSE, MAE finds the average magnitude of the absolute errors between the predicted and actual data [25]. Although R and R^2 are not considered a form of an accuracy measurement, RMSE, MSE, and MAE are good methods to determine accuracy [26].

Since all three of these methods have been proven to show good results, then going forward the RMSE value will be used to determine the accuracy of the predicted model in comparison to the actual data, shown in Eqn. 9. The outputs of RMSE will range from zero to positive infinity and the closer the value is to zero the more accurate the model is.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad \text{Eqn. 9}$$

From all these correlation and accuracy methods, the R and RMSE are the two values that will be looked at. These methods will be used for every Weka prediction model discussed before.

RESULTS & DISCUSSION

Looking at the initial model, there is a total of 1,280 sites available for reactions to occur where the pulse and purge time step is 0.05 seconds with a total cycle time of 0.2 seconds for a total time of 2 seconds. The temperature inside the simulated chamber is 0.03447 eV (400 K).

Initial Run

After running it with those parameters, the QCM was still inclining. Therefore the new initial total time for 400K was set to 3 seconds. In Figure 6, the QCM over time can be seen.

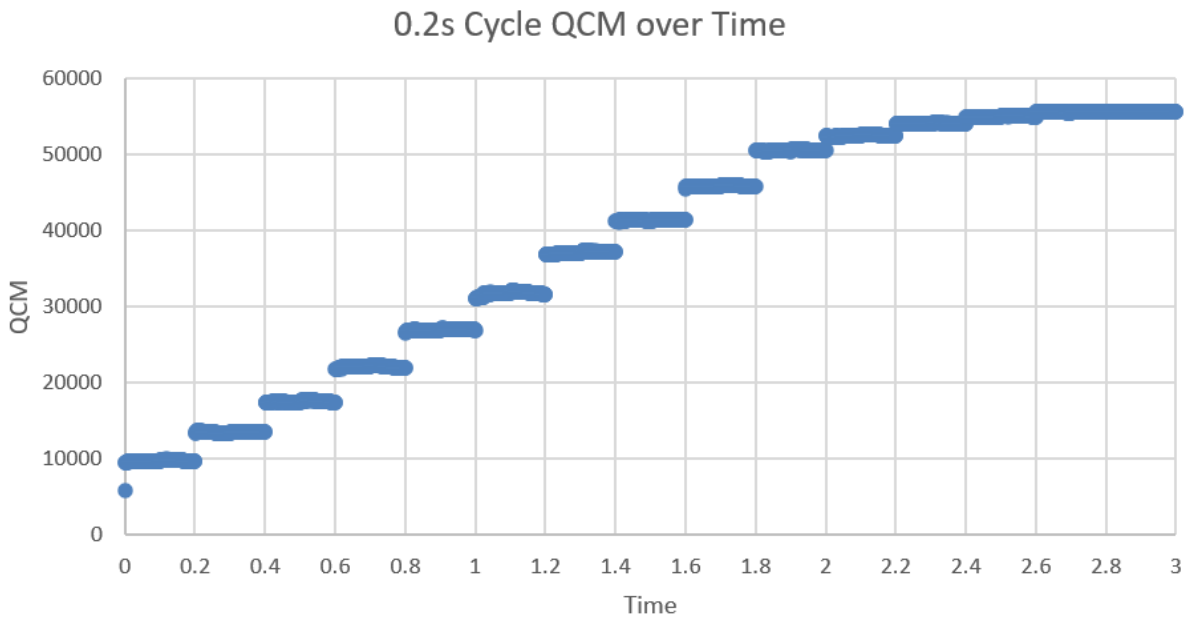

Figure 6. QCM over time for 0.2 second cycle at 400K.

This shows that there is a steady increase in mass gain per area over time until it reaches its eleventh cycle where the QCM begins to close down. At this point, the lattice is almost fully saturated not allowing for additional reactions to occur. The major jump of mass gain is at the

beginning of the DEZ pulse due to the mass influx of reactions occurring and bonding to the surface. During the purge part of the cycle, there is no change in the mass gain because the QCM only accounts for the attached species. Thus any byproducts will be blown off without changing the mass gain. During the water pulse cycle, there is no shown QCM increase because the amount of byproducts released is close to the value of mass added. Thus the QCM even outs. In Figure 7, the points on the graph represent the change in QCM over time. At the start of each cycle length, there is a spike into the thousands in the beginning, but as it nears the eleventh cycle or at the two-second mark, the change over time decreases dramatically which corroborates with Figure 6.

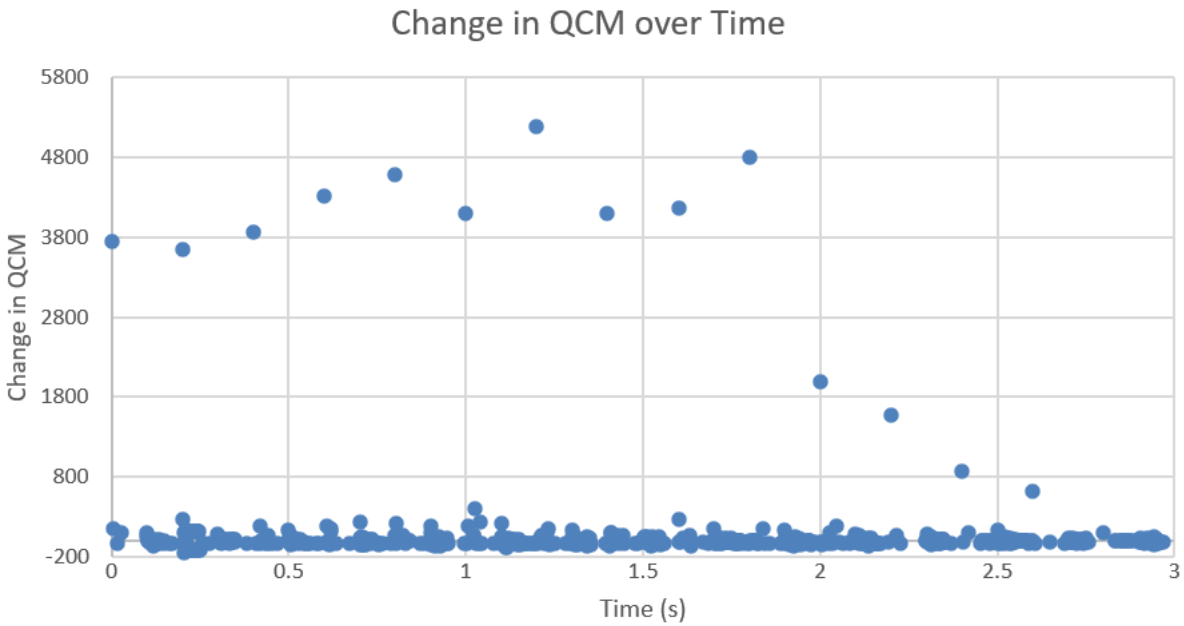

Figure 7. Graph of change in QCM over time using initial parameters.

The initial surface is a ZnO layer with a layer of water on top. The largest amount of reactions occurs at the initial moment of each cycle because when the DEZ is pulsed on top of

the surface, the DEZ reacts with the water creating MEZ and then creating a new layer of ZnO. All of those steps happen almost instantaneously. This repeats for every cycle. Depending on the pulse time will determine how many reactions will occur in that instant. The molecules ethane and methane gas are created as byproducts during this process which do not interact with the surface which isn't part of the QCM total. Thus during the purge period, there isn't much of a change in the QCM. During the remainder of the pulse length, very few reactions occur causing the QCM to bounce very slightly, which can be seen in Figure 7. This total process creates a step-like pattern, which has been shown [4, 10].

Varying the Cycle Length

Looking at the initial Figures 6 and 7, a basic understanding of what is to happen overall can be seen. Thus trying to compare the QCM for different cycle lengths can show a better understanding of the process, Figure 8.

The length of the cycle will determine how quickly the saturation is reached. However, the number of cycles it takes to reach full saturation is longer when the cycle length is shorter to a certain extent. It appears that the one through four-second cycles all have the same mass gain for each of their cycles meaning that there is a threshold of how much mass can be gained given the cycle length, also shown in Figure 9. Thus running the cycles for an extended period longer than two seconds would deem to be not useful. At three cycles the QCM is more varied than at five cycles and this may be due to it being too early to measure to have variety.

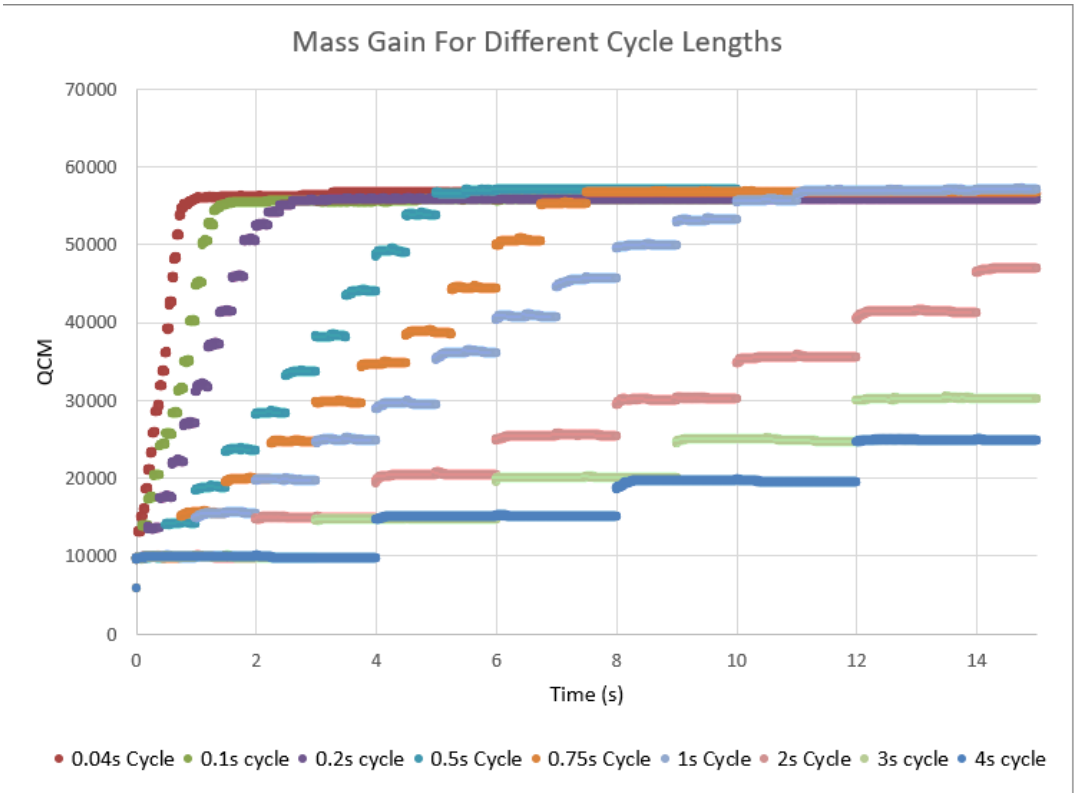

Figure 8. Graph of QCM over time compared to different cycle lengths.

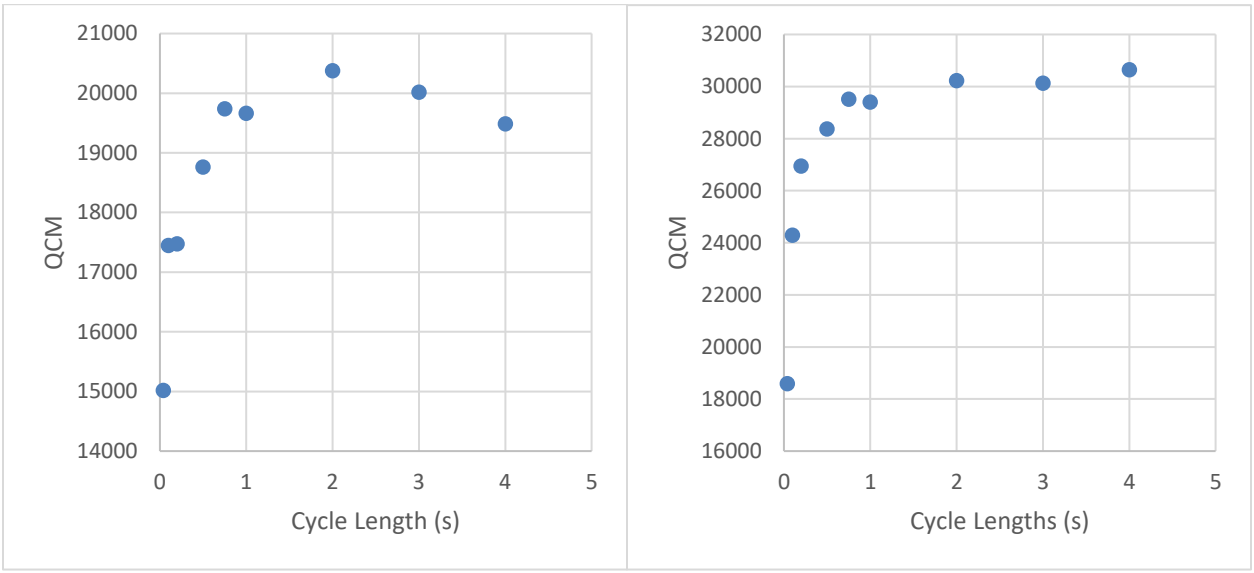

Figure 9. Graph of QCM at the end of three cycles (left) and five cycles (right) for different cycle lengths.

Figures 10-13 show the total number of hydrogen, oxygen, and zinc given the different cycle lengths for three and five cycles. The fifth cycles show that each one increased, especially the zinc. They all show the same general pattern that as the cycle length increased, so did the number of occurrences except when it reached the one to two-second mark. Once it reached that mark, then the number of occurrences for each remains similar.

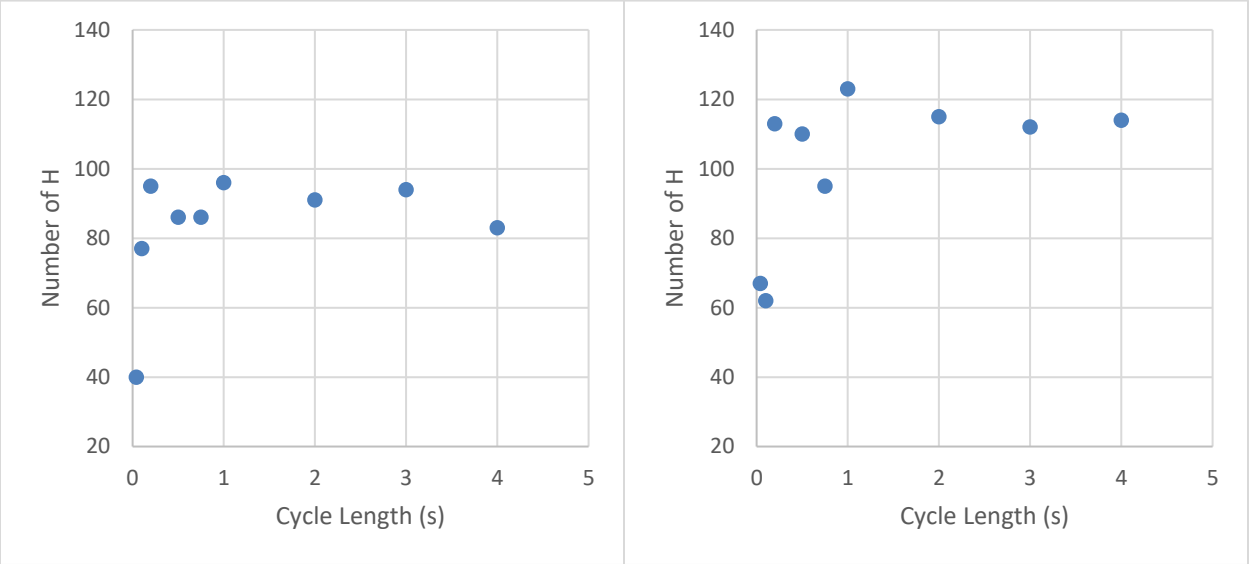

Figure 10. Graph of the number of hydrogen at the end of three cycles (left) and five cycles (right) for different cycle lengths.

In the most ideal case, the hydrogen should have remained the same or only increased a little. However, the hydrogen did increase meaning there may be some defects in the film where not all the hydrogen was released during the ALD process. From the 66.7% increase in cycles, there is only an approximate 27% increase in hydrogen. This shows that over time, more hydrogen gets taken from the layers.

Figure 11 shows the number of occurrences for oxygen. Going from three cycles to five cycles, there is an approximated 43% increase in the number of occurrences.

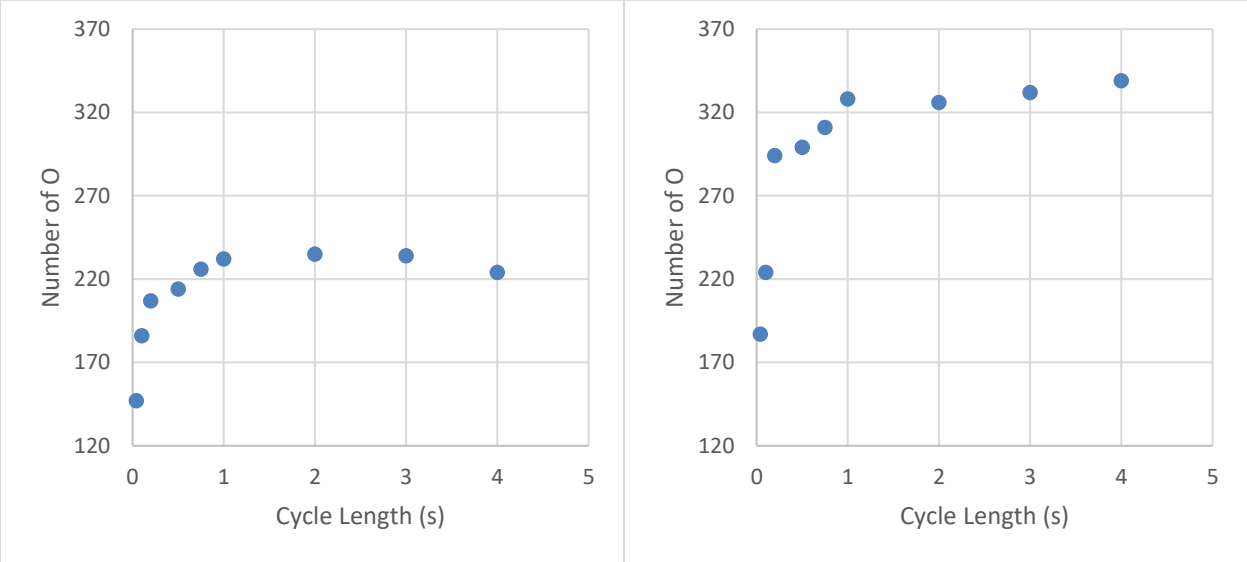

Figure 11. Graph of the number of oxygen at the end of three cycles (left) and five cycles (right) for different cycle lengths.

Figure 12 shows the number of occurrences for zinc. There is roughly a 51% increase in zinc when comparing cycles three to five. Oxygen and zinc have both increased more than hydrogen, which is what is desired. However, the difference in the number of occurrences between the hydrogen versus the oxygen and zinc is about 154% to 140% for the third cycle comparison, while for the fifth cycle the oxygen and zinc have about 186% more occurrences than hydrogen.

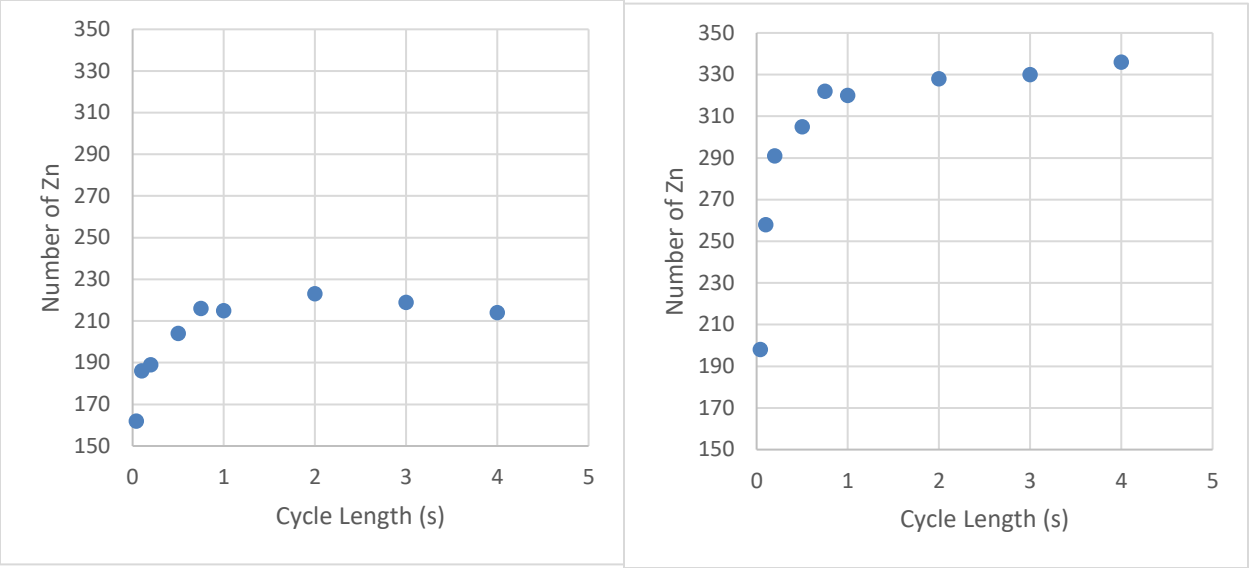

Figure 12. Graph of the number of zinc at the end of three cycles (left) and five cycles (right) for different cycle lengths.

Machine Learning Recreation. Using Weka, we can find the predicted model for the QCM and the number of hydrogen, oxygen, and zinc using the three different machine learning algorithms mentioned before, linear regression, multilayer perceptron, and M5P. The inputs used for Weka are time, QCM, number of hydrogen, oxygen, zinc, MEZ, and ligands, and the cycle length. Table 3 has all the correlation coefficients and accuracy results from the three different algorithms. Looking at the linear regression shows that the predicted values and the actual simulated values have a very high positive correlation, but the RMSE for hydrogen is a little high in comparison to the other LR RMSE. MLP can predict the actual values well since the R-value of 1 meaning that it is perfectly correlated and RMSE is less than one meaning that the predicted values are very accurate. For M5P, the R-value is very close or is one meaning that it has a very high positive correlation. The RMSE for hydrogen, oxygen, and zinc is low meaning that it is a good fit. However, the RMSE for QCM is a very high value meaning that the predicted values aren't as accurate of a representation. This can be seen in Figure 13. Another

key factor to M5P that has to be taken into account is the number of rules. A model is considered better with a smaller amount of rules, as long as the RMSE is small as well. Thus, the M5P algorithm decided that the best model for QCM has an RMSE value of 97.029, but has only 2 rules. This either means that adding a lot more rules doesn't change the RMSE much or that adding more rules makes the RMSE worse. In either case, it isn't desired. From hydrogen, it can be seen that sometimes adding more rules benefits the RMSE value. Having 204 rules is considered a lot of rules, making it a less desirable algorithm for hydrogen.

Table 3. Correlated/Accuracy results with LR, MLP, M5P for QCM, Hydrogen, Oxygen, and Zinc for varying cycle lengths.

	LR		MLP		M5P		
Predicted Variable	R	RMSE	R	RMSE	R	RMSE	No. of Rules
QCM	1	0.3057	1	0.7564	0.9999	97.029	2
Hydrogen	0.9844	6.6091	1	0.0135	0.9994	1.2999	204
Oxygen	1	0	1	0.0247	1	0.2097	2
Zinc	1	0	1	0.0091	0.9999	0.8684	9

From Figure 13, the actual results are the red dots. The predicted LR results can not be seen because they are behind the MLP results which are nearly perfectly centered on the actual results. However, the blue dot, or the M5P results, are varied. For every M5P algorithm, the smallest amount of rules it can have is two. The two rules for QCM are whether the number of zinc is greater than 235, or less than or equal to 235, Figure 14. Thus something happens that results in a change in the mass gain around 235 occurrences.

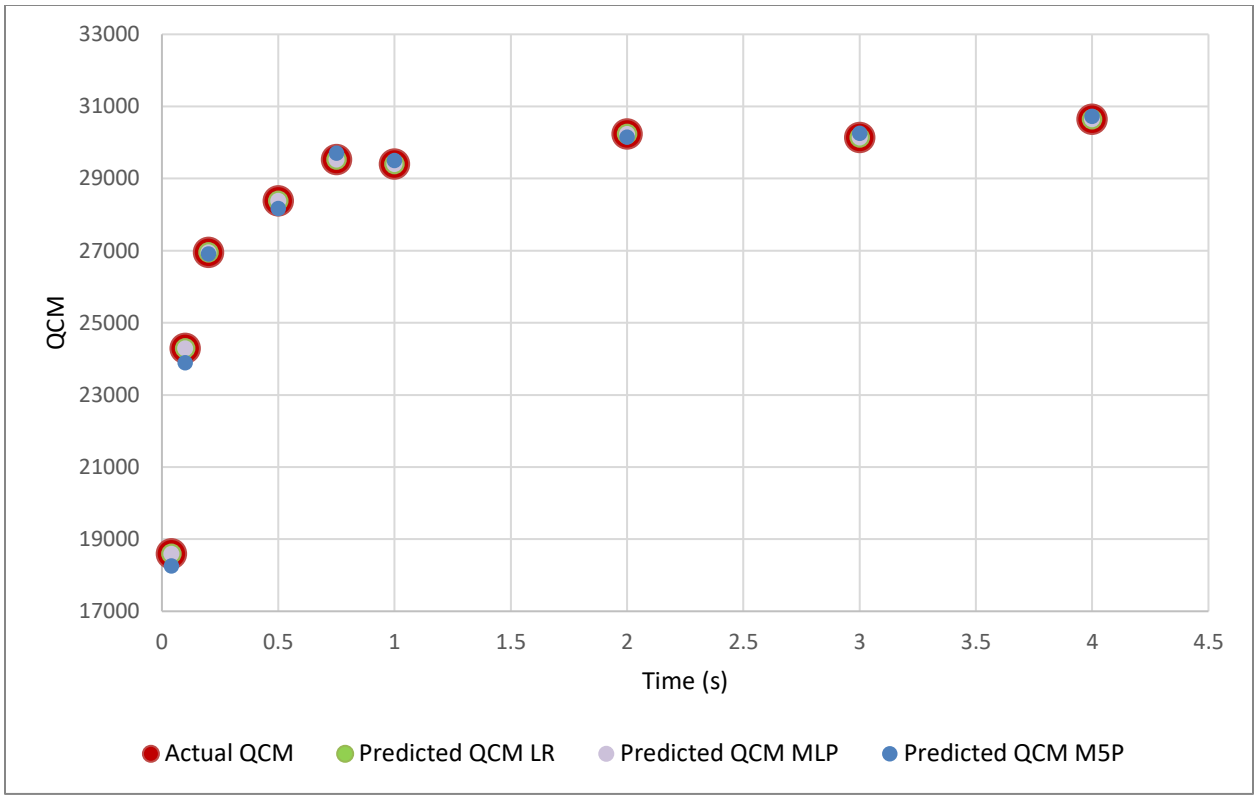

Figure 13. Machine Learning predicted results for the QCM in comparison to the actual.

With each rule, there is an equation that is produced to find the value. For this case:

LM1: QCM = -23.2319*TIME + 94.0457*ZINC – 0.0001*HYDROGEN -360.9156

LM2: QCM = -16.1753*TIME + 91.6126*ZINC – 0.119*HYDROGEN + 272.392

The equations produced, LM1 and LM2, show that the QCM is mainly dependent on time and zinc which would be expected because the purpose of this process is to build ZnO.

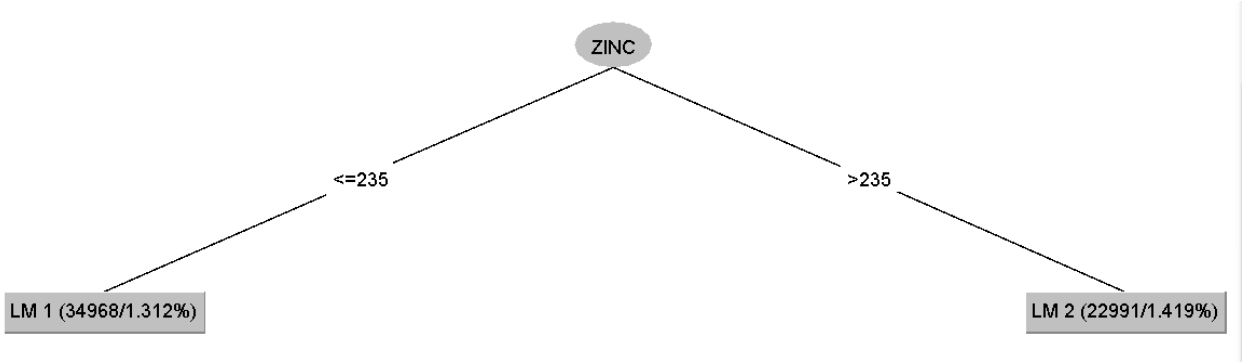

Figure 14. The M5P decision tree for QCM when varying cycle lengths.

With LR, there is only one equation to derive the entire process. The linear regression equation used:

$$QCM = 0.0205*TIME + 53.9575*OXYGEN + 27.4203*ZINC – 17.9657*HYDROGEN – 0.0111*MEZ + 48.0513*LIGANDS – 0.0615*CYCLELENGTH – 0.4835$$

$$HYDROGEN = -1.2823*TIME + 0.7767*OXYGEN – 0.931*MEZ + 0.2323*CYCLELENGTH – 18.7335$$

$$OXYGEN = 0.9997*ZINC + 0.5*HYDROGEN – 0.5001*LIGANDS$$

$$ZINC = 1*OXYGEN – 0.5*HYDROGEN + 0.5*LIGANDS$$

The oxygen and zinc equations are dependent on each other, hydrogen, and the number of ligands. This would be able to determine what the surface is doing and whether there are a lot of reactions occurring or not. The amount of hydrogen and the total QCM have to rely on the cycle length because depending on the cycle length can determine how many reactions will occur at a given time.

Looking at the hydrogen the linear regression R-value is very highly correlated, but the RMSE value is high. Figure 15 shows that correlation doesn't mean that there is a good fit. The

other two algorithms represent the value well. This shows the idea that the correlation could be good, while the RMSE is bad. However, if the correlation is bad, then the RMSE will be bad too. In either case, the output from that machine learning algorithm will not be desired.

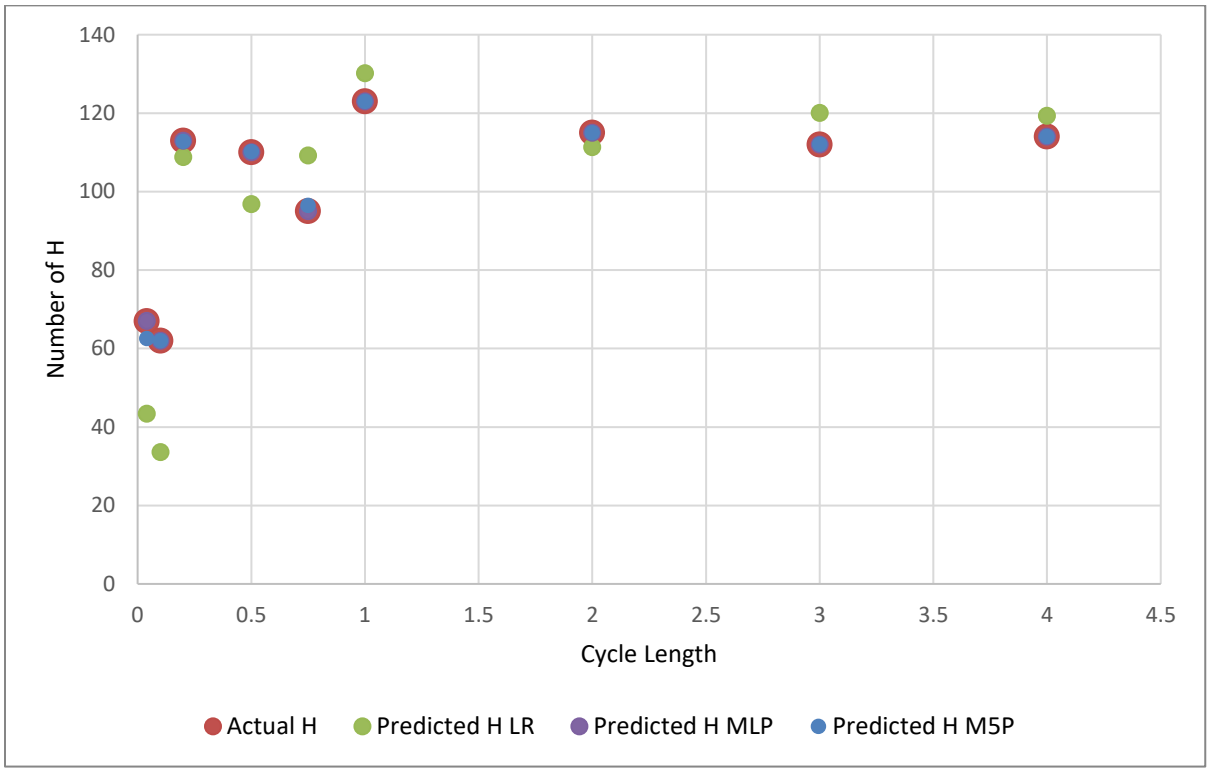

Figure 15. Number of hydrogen over cycle length for different machine learning algorithms.

Varying Temperatures

Another variable to compare is the temperature. One of the main contributing factors to ALD growth is temperature. If the temperature isn't hot enough, then reactions are less likely to occur. However, if the temperatures are too hot, then some atoms could be burned off. Some of the temperatures looked at can be seen in Figure 16.

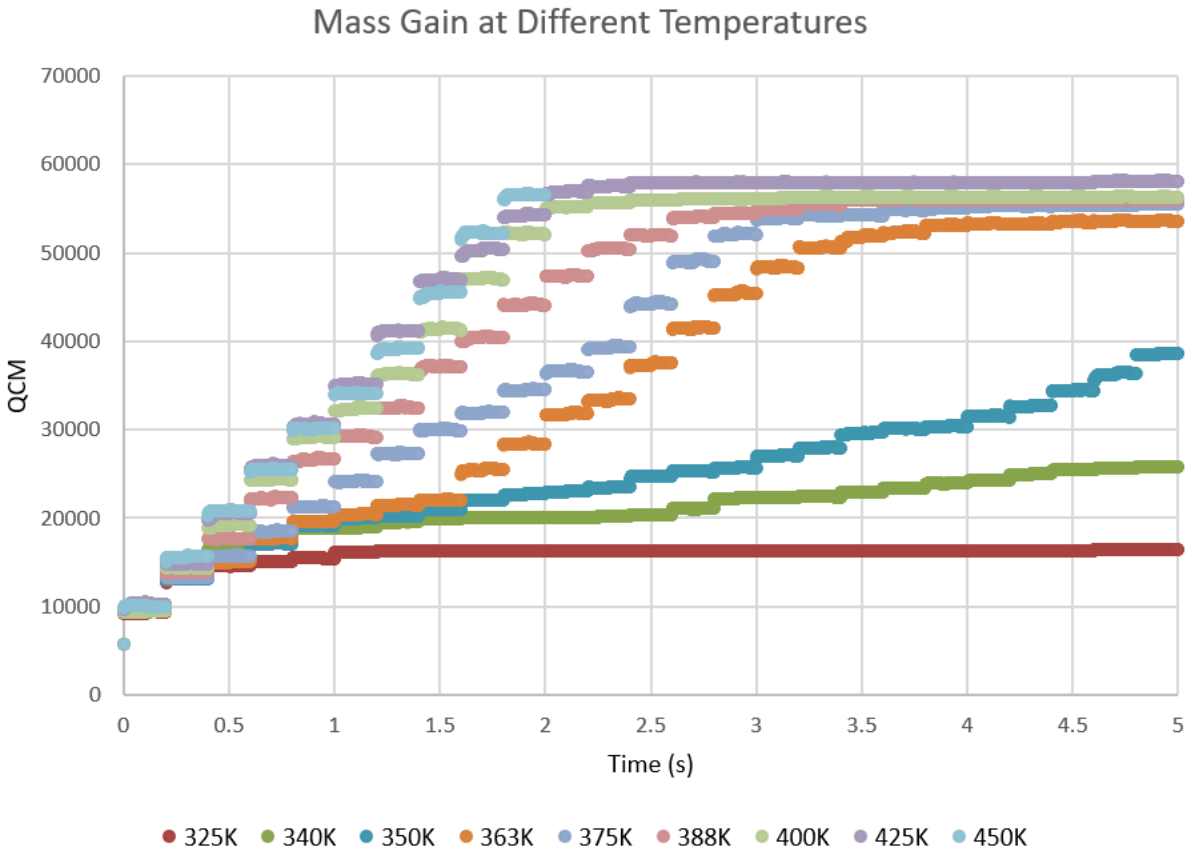

Figure 16. Graph of QCM over time with varying different temperatures.

As stated before, if the temperature isn't hot enough then the reaction will not happen which can be seen for temperatures 325-350K. Although as the temperature increases the overall QCM increases as well. Once the temperature reaches a certain threshold, the increase in QCM per cycle becomes constant. To reach saturation, more cycles are needed if the temperature is lower. The reason why they all initially start with the same amount of reactions is that the initial surface is a zinc oxide layer with water on top. This surface doesn't contain a defect, thus making it easier to react with. After the first couple of cycles, more defects or leftover carbons or hydrogens remain to make it hard for bonds to occur. From Figure 17, the graph contains the

total QCM at five cycles over temperature. This reiterates what was mentioned before. As the temperature grows, so does the QCM.

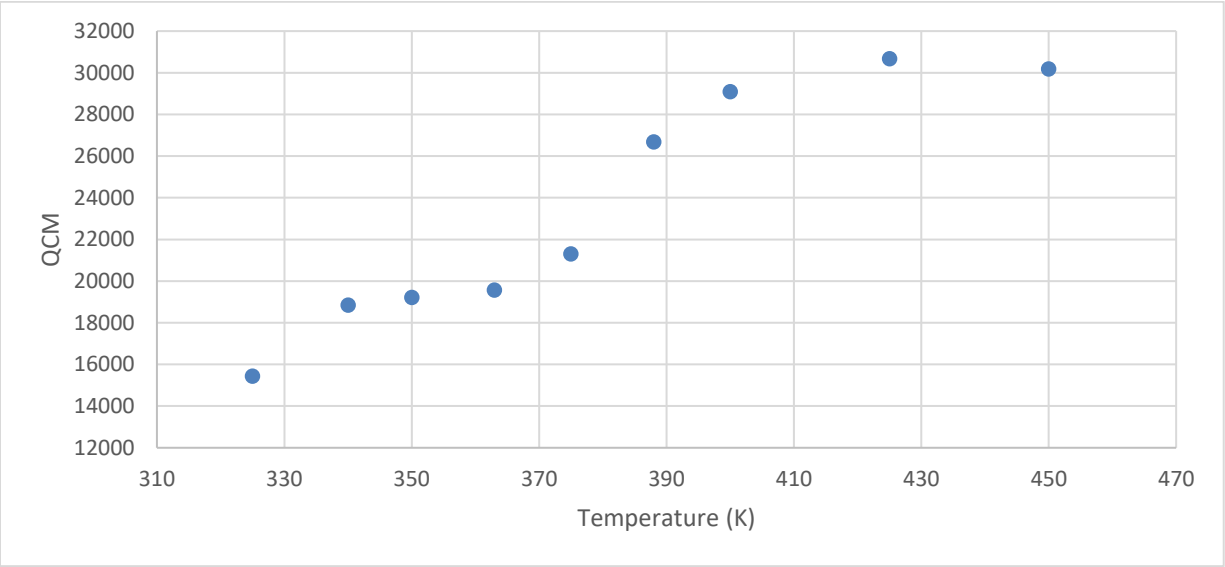

Figure 17. Graph of QCM after five cycles with varying temperatures.

Attempting to raise the temperature higher could danger the thin film in real life and is not necessary to research since results can be achieved at a smaller temperature which is what ZnO is known for [4, 10, 28]. From Figure 18, the number of hydrogen can be seen with varying temperatures. Between 340-388K there is a peak in the number of hydrogen. This could be due to not reaching high enough temperatures resulting in the hydrogen being adsorped to the surface. After 388K, the number of hydrogen in the film decreases as the temperature increases which can be due to the higher temperatures making it easier for the species to overcome the activation energy. Before 340K, the number of hydrogen is small in comparison to its surrounding temperatures and this is due to the lack of reactions happening overall on the surface which can be seen in Figures 19 and 20.

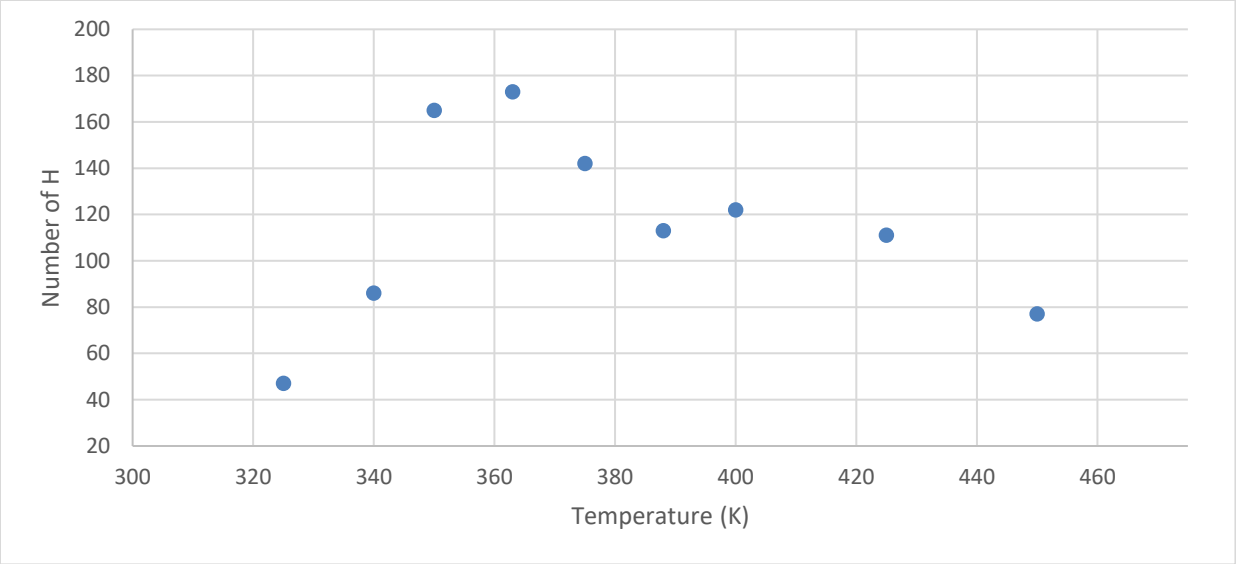

Figure 18. Graph of the number of hydrogen after five cycles with varying temperatures.

In Figure 19, the number of oxygen starts small and then begins to build and stabilize as the temperature increases. Before 363K, the temperature is not high enough for all the reactions to occur from the reactions having difficulty overcoming the activation energy. Thus as the temperature increases, it becomes easier for the activation energy to be reached. After 363K, the number of oxygen remains approximately the same and this could represent that all the reactions that could happen with oxygen, have happened.

In Figure 20, the number of zinc occurrences can be seen. The overall shape is similar to Figure 19 meaning that has similar ideas. Once the temperature reaches 363K, then it reaches the threshold of the maximum number of reactions to occur. Before that temperature, it tries to produce as many as possible, but the activation barrier is too high to fully overcome for some equations. For Figures 19 and 20, the initial values are very low, similar to the hydrogen. From this, we can say that at lower temperatures between 325K and 340K, the lower hydrogen doesn't signify the quality of the film since the zinc and oxygen are lower as well.

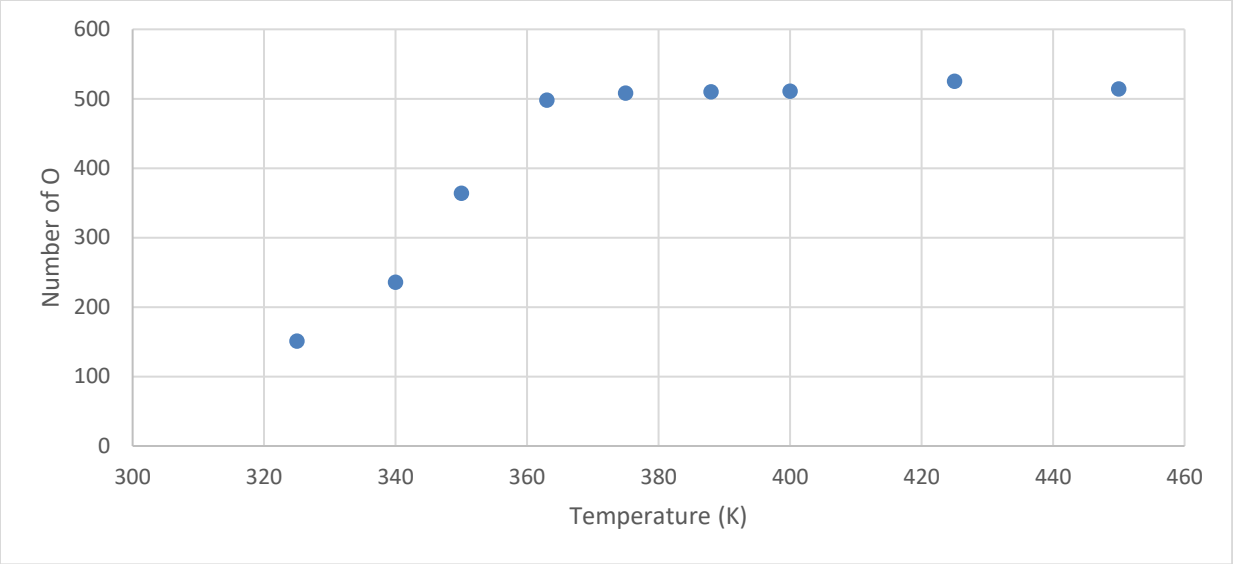

Figure 19. Graph of the number of oxygen after five cycles with varying temperatures.

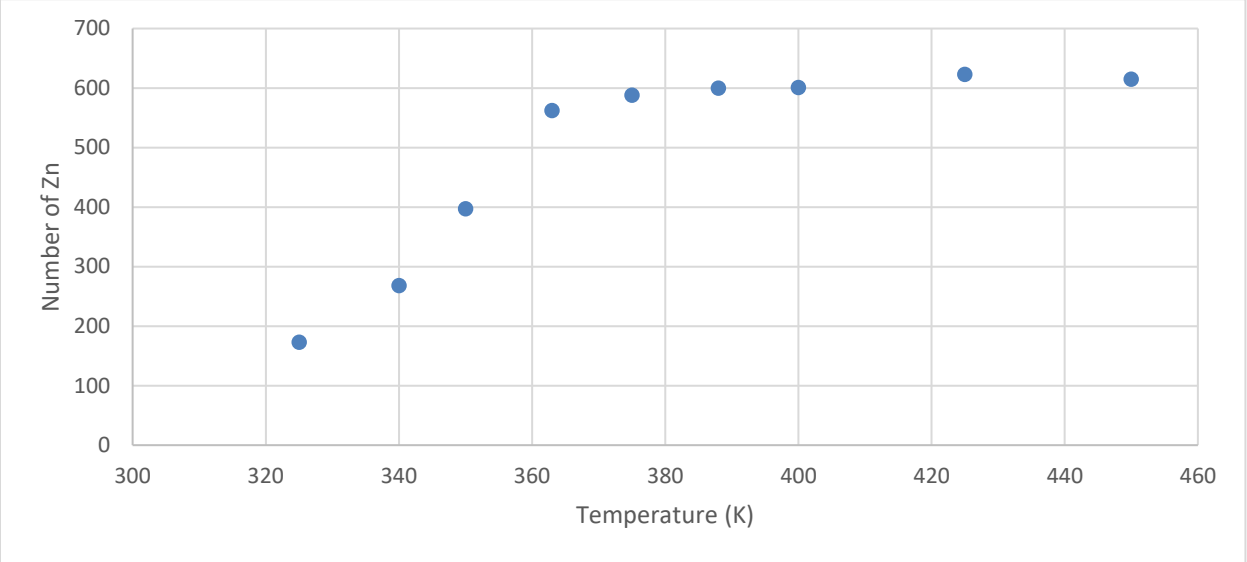

Figure 20. Graph of the number of zinc after five cycles with varying temperatures.

Machine Learning Recreation. Similar to what was done with the varying cycle length. We will use Weka to recreate the QCM, hydrogen, oxygen, and zinc with the different machine learning algorithms. These inputs used for Weka are the same used from the varying cycle lengths. Table 4 shows the correlation and accuracy results of each of the different machine

learning algorithms. This tells us that M5P has trouble predicting QCM, and LR and M5P have trouble predicting hydrogen.

Table 4. Correlation/Accuracy of QCM, Hydrogen, Oxygen, and Zinc using different machine learning algorithms, LR, MLP, and M5P for varying temperatures.

	LR		MLP		M5P		
Predicted Variable	R	RMSE	R	RMSE	R	RMSE	No. of Rules
QCM	1	0.2952	1	4.0396	0.9998	298.3433	3
Hydrogen	0.7636	26.498	1	0.0837	0.9986	2.2616	545
Oxygen	1	0	1	0.0389	1	0.0391	2
Zinc	1	0	1	0.0401	0.9998	3.3773	3

Figure 21 shows a high positive correlation and poor RMSE. The linear regression made it have a smoother line distribution, which is not reflected in the actual simulated data. The hydrogen was also hard to make accurate for the M5P by having 545 rules. With that machine learning algorithm, it tries its best to reduce the number of rules while keeping a good RMSE value, but if it is hard to estimate, then it will have a difficult time reducing the rules.

For all the graphs reflecting the QCM or number of hydrogen, oxygen, or zinc, there were not a lot of data points for the machine learning algorithm to reference. Thus if there were more data points for the varying cycle lengths and temperatures, then the algorithm could have been able to predict better.

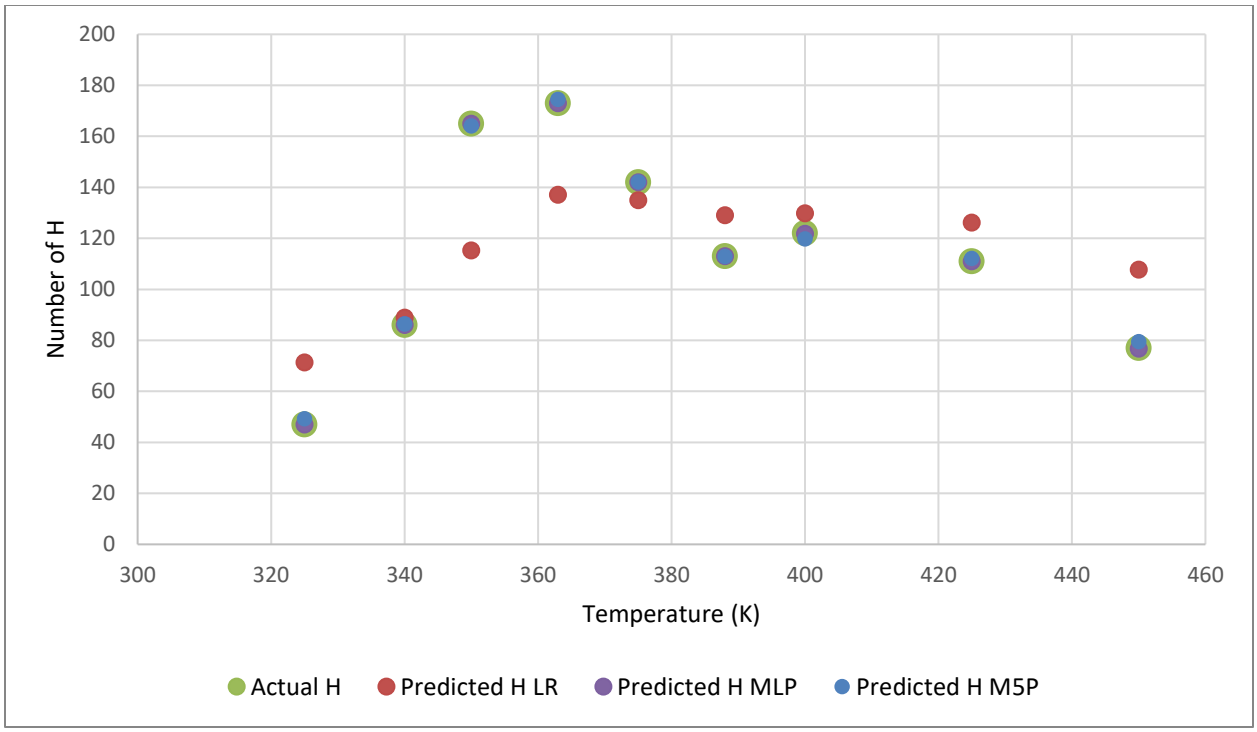

Figure 21. Machine learning prediction models of the number of hydrogen on the thin film at 5 seconds for varying temperatures.

Reactions

The kMC outputs became the machine learning inputs. The only parameters used were time, mass gain, and the number of occurrences for zinc, oxygen, hydrogen, MEZ, and ligands, as mentioned before. From there we used those inputs to find an approximation for each reaction shown in Table 5.

In the reactions above, X is defined as ethyl. Each reaction is unique. The difference between reaction V11 and V15 is their coordination number of the 1st lattice site and the only difference between V13 and V15 is their coordination number of the 1st lattice site. The difference between each of the reactions is shown in Figure 22. V15's reaction is from a harder-to-reach place in the lattice concerning the surface and has higher activation energy making it less likely to interact resulting in a smaller number of occurrences in comparison to V11.

Table 5. List of some reactions from Event type 3.

Reaction ID	Reaction	Activation Energy (eV)	Coordination Number
V11	OH2 + ZnX => OH + Zn + XH	0.72	-9
V13	OH + ZnX => O + Zn + XH	0.93	-8
V15	OH2 + ZnX => OH + Zn+ XH	0.72	-19
V17	OH + ZnX => O + Zn + XH	0.93	-18

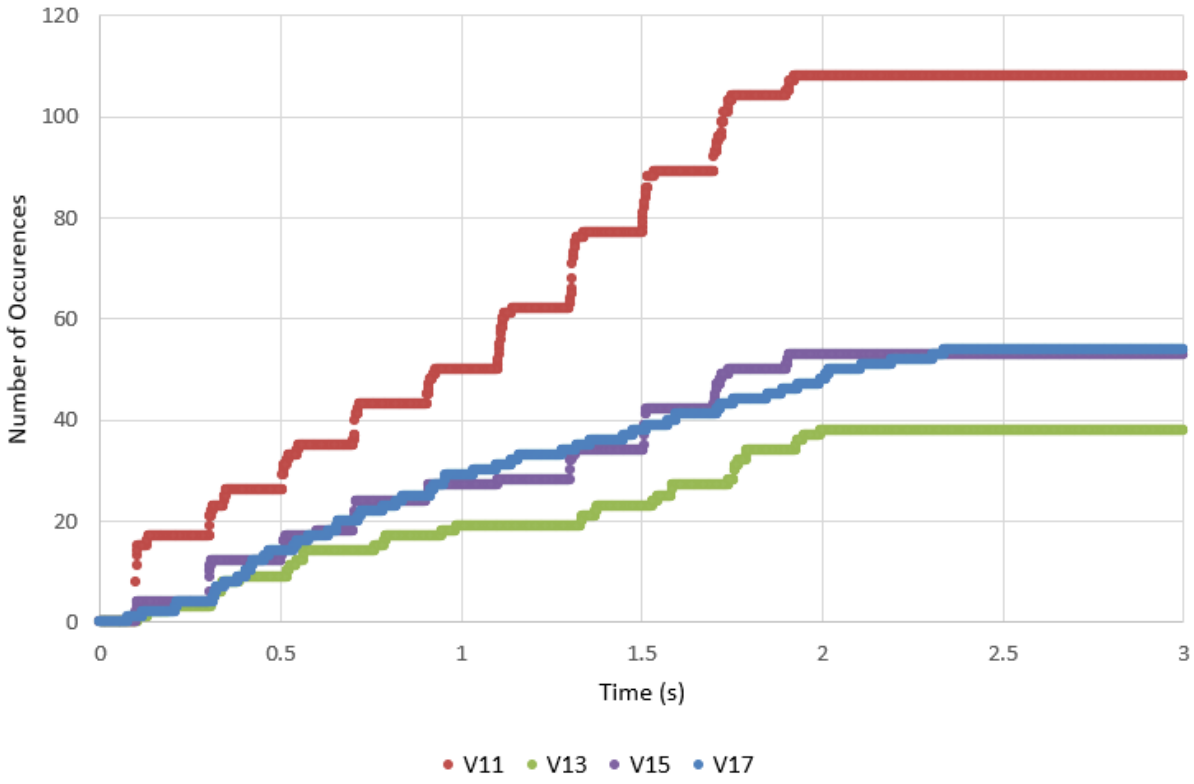

Figure 22. Graph of the number of occurrences of simulated reactions V11, V13, V15, and V17 over time at 400K with 0.2 cycle length.

Figure 22 shows the differences between each of the reactions. For comparison of each machine learning algorithm, they will be compared against each other for each reaction shown in Figures 23-26.

From each graph, the linear regression deviated the most away from the actual simulated data, Table 6. M5P has the closest comparison, but MLP is a good method of determining. In each case, the correlation is very high. Besides linear regression, the RMSE value is low for each meaning it is a good fit for the simulated results.

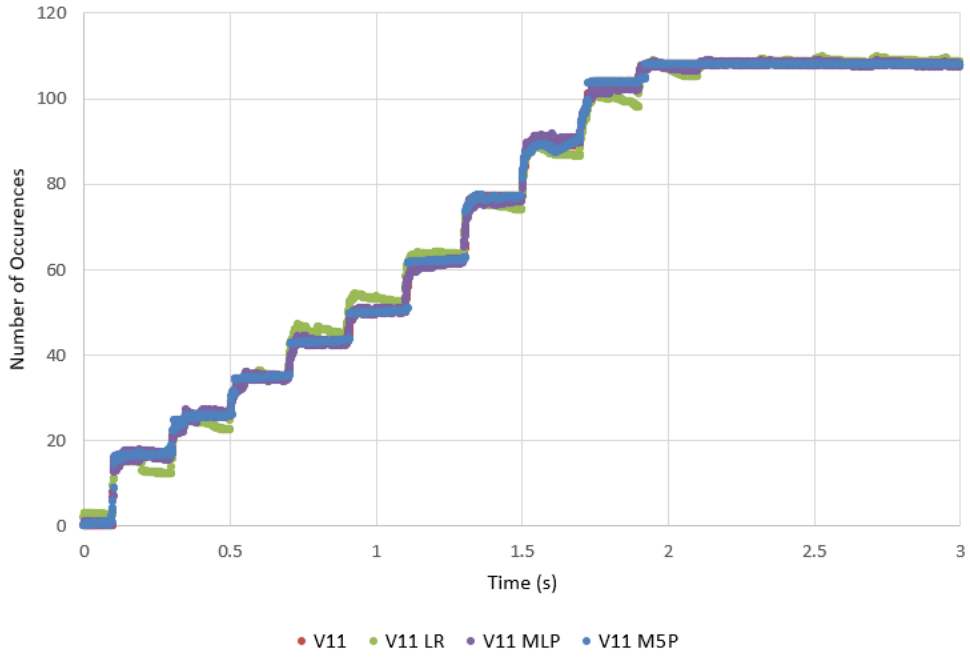

Figure 23. ML algorithm comparison of reaction V11 at 400K for three seconds.

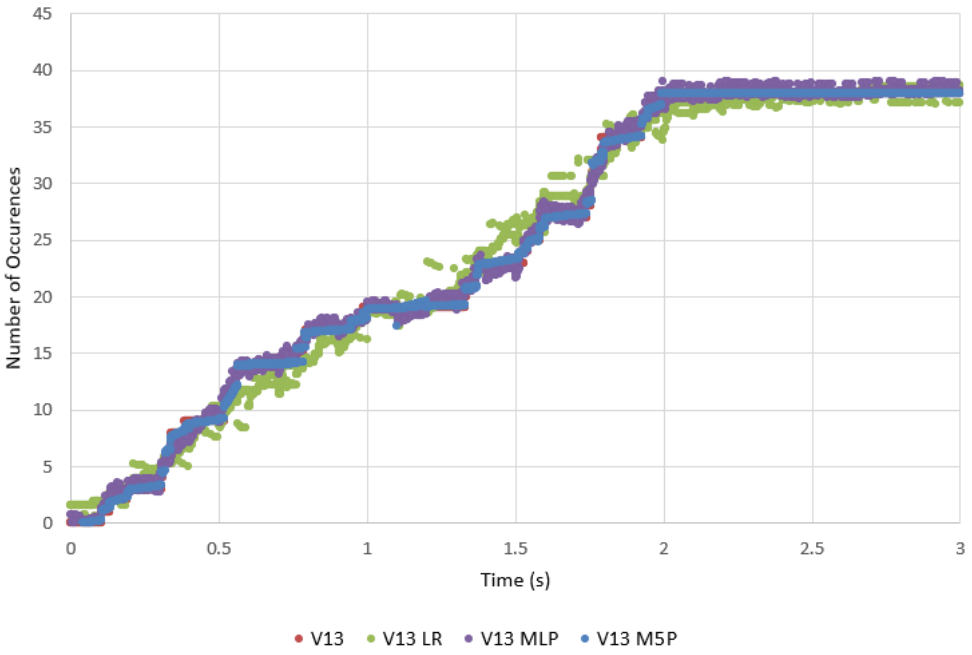

Figure 24. ML algorithm comparison of reaction V13 at 400K for three seconds.

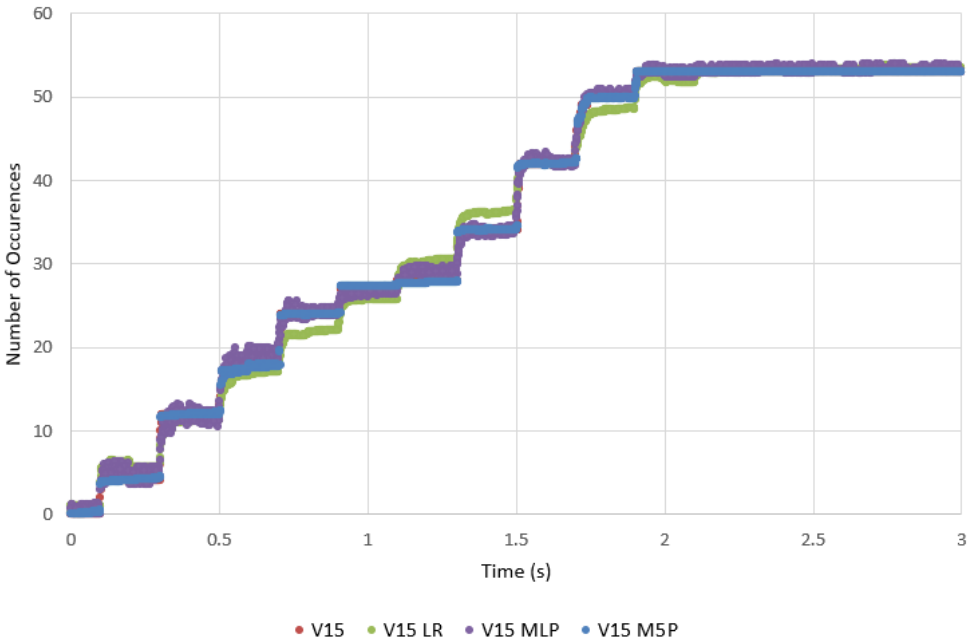

Figure 25. ML algorithm comparison of reaction V15 at 400K for three seconds.

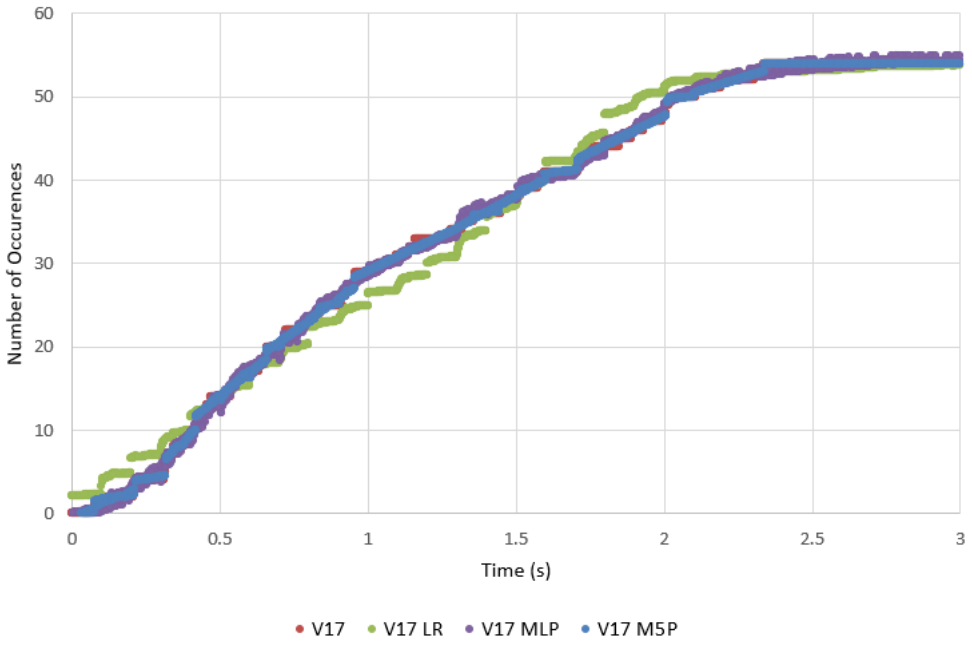

Figure 26. ML algorithm comparison of reaction V17 at 400K for three seconds.

Table 6. Correlation/Accuracy values for LR, MLP, and M5P for the reactions V11, V13, V15, and V17 for varying cycle lengths.

	LR		MLP		M5P		
Reaction	R	RMSE	R	RMSE	R	RMSE	No. of Rules
V11	0.9986	1.7474	0.9998	0.7372	0.9999	0.5096	14
V13	0.9959	1.07	0.9992	0.472	0.9999	0.1865	31
V15	0.9977	1.0914	0.9995	0.5576	0.9999	0.2576	13
V17	0.9955	1.5712	0.9996	0.4896	0.9999	0.2569	16

The equations from the linear regression are below. The stepping pattern shown in each case for the linear regression can be explained by the variables chosen, shown in Figure 27.

$$V11 = -1.7786*TIME + 0.992*OXYGEN + 0.1243*ZINC + 0.1053*HYDROGEN – 22.1644$$

$$V13 = 0.3287*TIME + 0.0576*HYDROGEN + 0.1163*MEZ – 5.2705$$

$$V15 = -0.7855*TIME + 0.0814*OXYGEN + 0.068*ZINC – 0.0466*LIGANDS – 11.901$$

$$V17 = 0.1825*TIME + 0.1151*OXYGEN + 0.0331*ZINC – 0.0563*HYDROGEN + 0.2217*MEZ – 0.2504*LIGANDS – 9.721$$

In each case, the linear regression stepped either too high or low to the actual simulated data. Reactions V11 and V15 represent water being pulsed onto zinc-ethyl which would create hydroxyl and zinc with byproducts. This step incorporates a huge portion of mass gain creating a stair-step pattern over time. Reactions V13 and V17 represent the hydroxyl reacting with zinc-ethyl creating oxygen and zinc with byproducts. This process is dependent on how many hydroxyls are made from V11, V15, and other like reactions. Thus over time the number of occurrences for V13 and V17 increases creating a smoother graph.

The reason why almost all the different methods remain to have a high correlation coefficient has to do with how the kMC formed the ALD. As shown in Figures 22-26 and Figure 27, with every cycle there is a step formation. Theoretically, all that would need to be known is how much they differ and make that into a regression model.

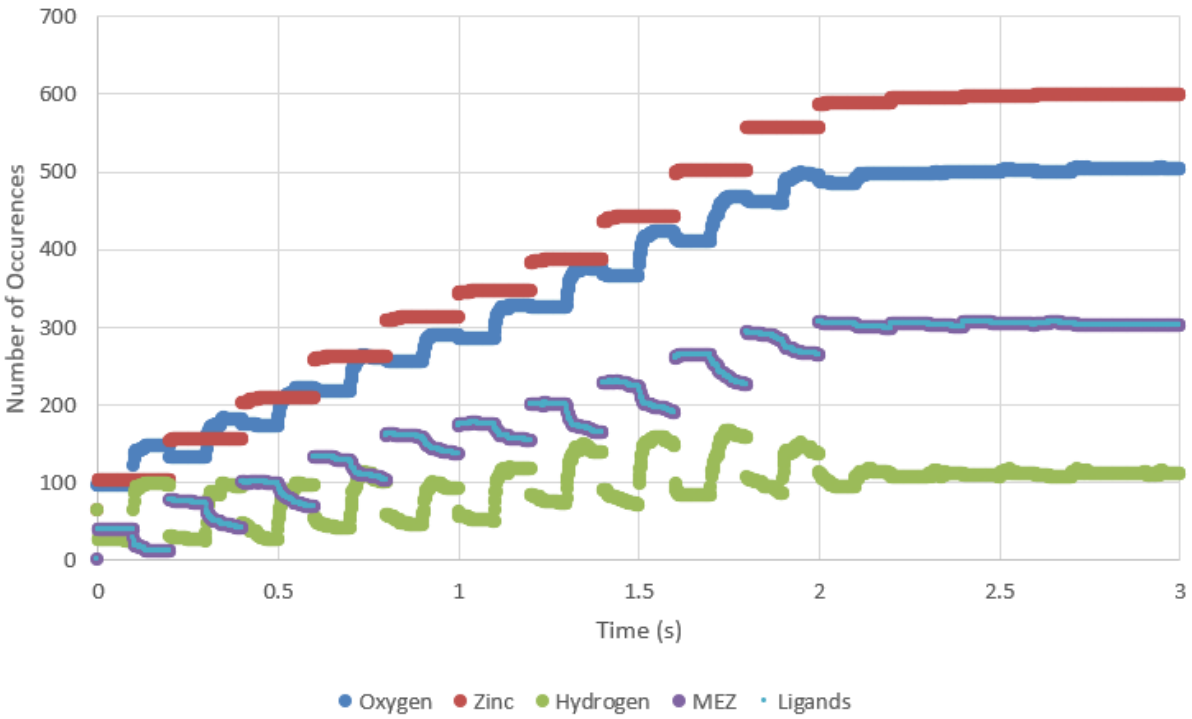

Figure 27. Species occurrences over time.

Each one of these models fits within the spectrum of being able to produce a good fit. Thus all three of these models can be used, but because of the RMSE which is an error term, it is hard to have a definitive value that defines whether it is a good fit or not. Only looking at the R values, these models correlate very highly to the actual data [27, 29]. When looking at the RMSE value, the closer it is to zero means the better it is as a fit [25, 26, 29]. Thus the M5P is a better fit than LR or MLP. Although the number of rules has to be taken into account of the complexity of the algorithm with each rule containing a linear regression model. In the simplest terms, M5P is a more complex form of LR. So it is interesting to look at the LR model and see that the correlation is high and the overall RMSE is still reasonable.

Varying Cycle Lengths Reaction Prediction. The simulated data from kMC mentioned before about the varying cycle lengths are used as an input for all three ML algorithms. The

parameters used for 400K are the same parameters used for this set. The correlation and accuracy values for the varying cycles using ML are shown in Table 7.

Table 7. Prediction values of the reactions while varying the cycle lengths using ML.

	LR		MLP		M5P		
Reaction	R	RMSE	R	RMSE	R	RMSE	No. of Rules
V11	0.9905	2.4041	0.997	1.3466	0.9996	0.5115	74
V13	0.9389	3.9168	0.9834	2.0685	0.9998	0.2493	191
V15	0.9521	3.2276	0.9939	1.1678	0.9997	0.2732	135
V17	0.9631	2.9123	0.9925	1.3264	0.9997	0.2846	222

All the correlation values for each reaction and each ML algorithm are very highly correlated. The RMSE values increased overall which can be reflected in Figures 29-32. Although the values are higher, this could be due to the number of occurrences being smaller. When predicting the reaction at 400K, the ML algorithm was predicting the entire process of the kMC at 400K for that reaction which contains roughly 5,000 data points. However, Figures 29-32 show nine distinct points that represent the number of occurrences after five cycles. The lack of data points could contribute to the higher RMSE value and lower R-value. Figure 28 shows the reactions at different cycle lengths after five cycles. V11 and V15 are the same reactions but with different coordination numbers. Thus with a lower coordination number, it is hard for more reactions to occur. It appears that after V11 has hit 0.5-second cycle lengths, then there is no impact on the number of occurrences. Although, V15 increases in the number of occurrences

overall due to having more time for those reactions to occur. V13 and V17 are similar reactions but with different coordination numbers. Both of these reactions vary depending on the cycle length which could be due to the temperature. Since V13 and V17 have higher activation energy then they happen less than V11 which has a higher coordination number and lower activation energy, but not lower than V15 because of its coordination number.

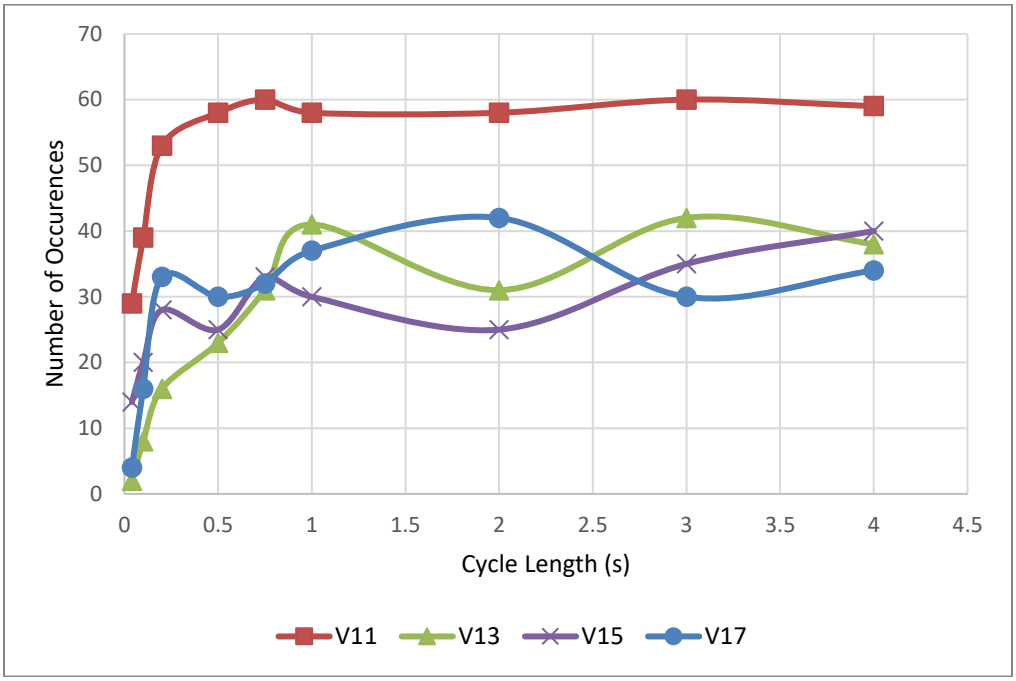

Figure 28. KMC simulated data of reactions V11, V13, V15, and V17 after five cycles for varying cycle lengths.

The red line shown in Figures 29-32 represents where the actual reaction is. For the graphs that had an irregular pattern, the LR had a harder time predicting it, such as V13-V17, which is reflected in the RMSE value in Table 7. Again, the MLP was able to predict better than the LR, but it wasn't able to outperform M5P. For M5P, the number of rules jumped to a couple of hundreds of rules, and this to due to the lack of data points.

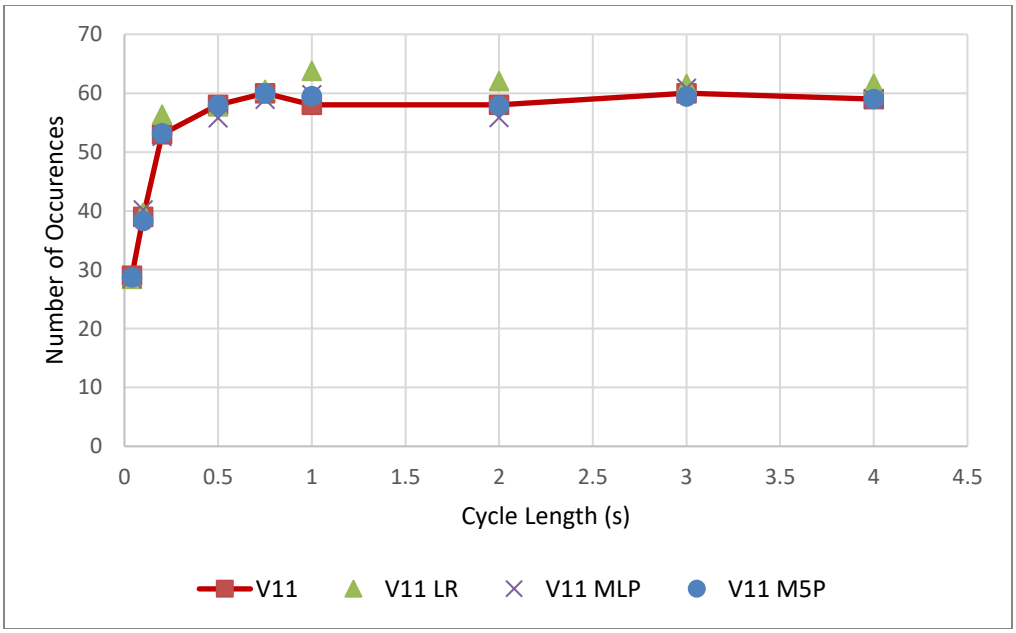

Figure 29. ML algorithm comparison of reaction V11 for varying cycles after five cycles for varying cycle lengths.

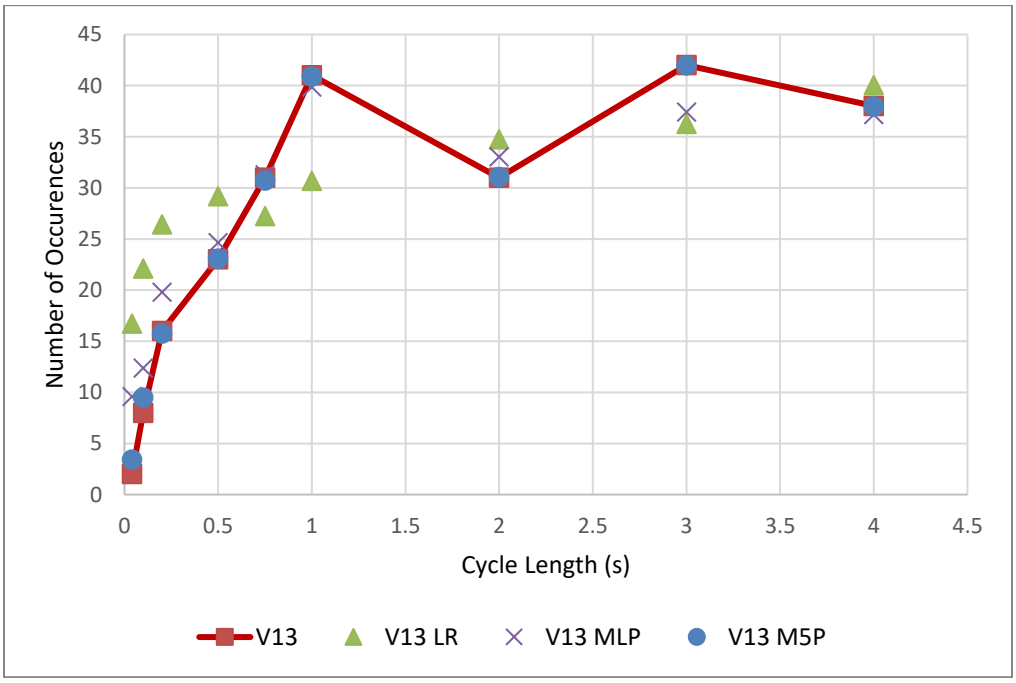

Figure 30. ML algorithm comparison of reaction V13 for varying cycles after five cycles for varying cycle lengths.

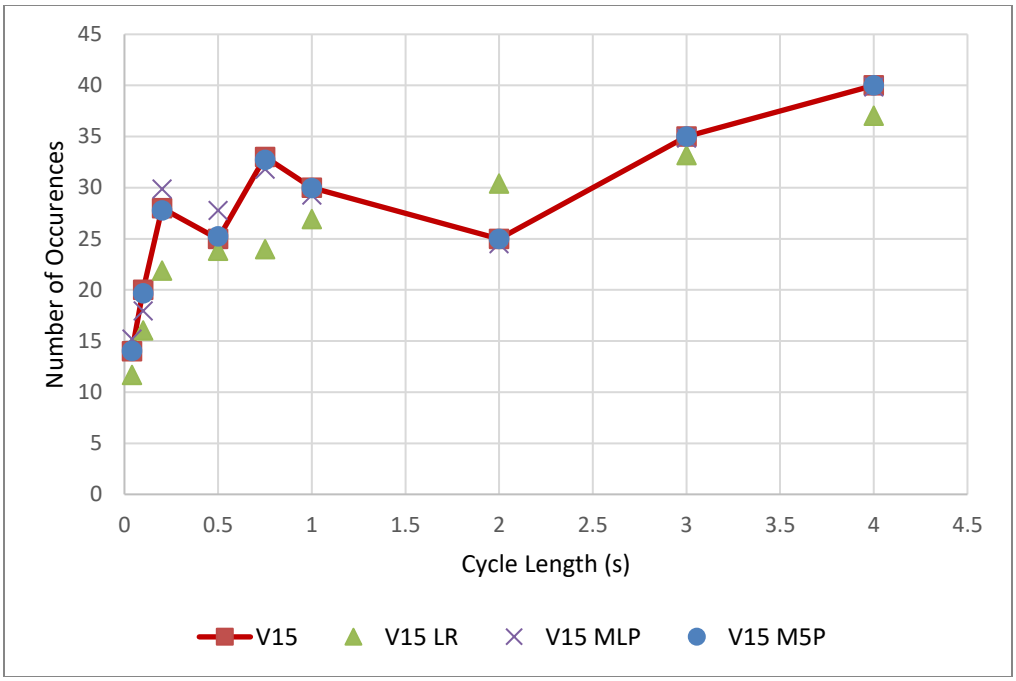

Figure 31. ML algorithm comparison of reaction V15 for varying cycles after five cycles for varying cycle lengths.

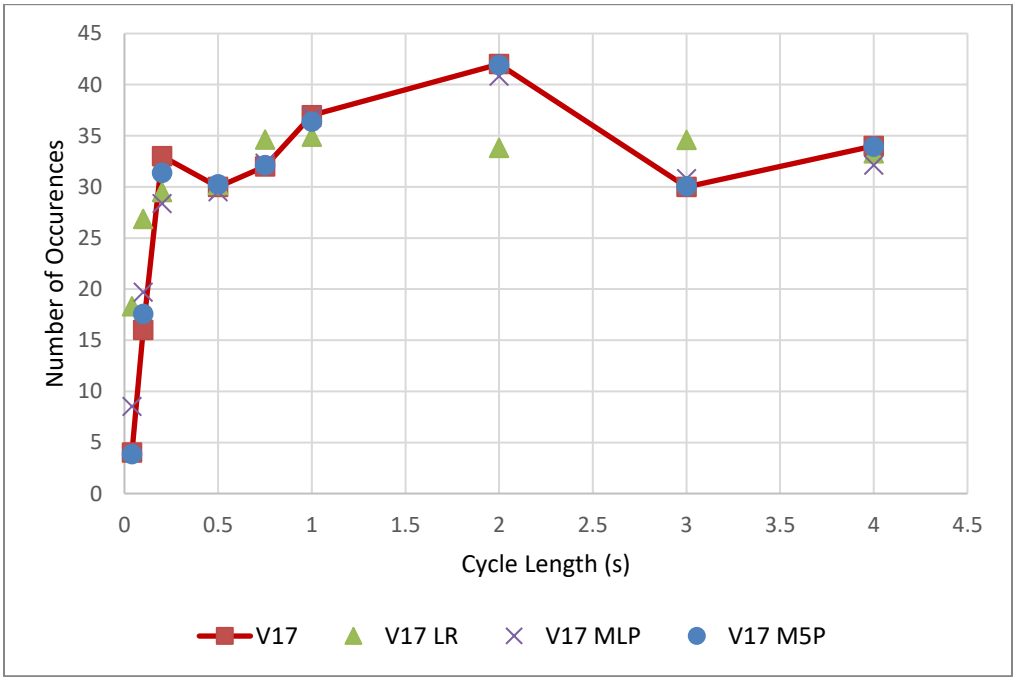

Figure 32. ML algorithm comparison of reaction V17 for varying cycles after five cycles for varying cycle lengths.

Varying Temperature Reaction Prediction. Using the same inputs as mentioned before for the machine learning algorithms, the correlation/accuracy values are shown in Table 8. For these models, they are trying to recreate an entire process, but the only values represented are nine distinct data points shown in Figures 34-37. The blue line in each of those Figures represents the kMC simulated results. Thus the R and RMSE values will be lower and higher respectively. As shown in previous models, M5P has the best accuracy and correlation to fit the simulated data. MLP is better than LR but still worse than M5P. Although in every case the R is still very highly correlated. Figure 33 shows the number of occurrences of each reaction after five seconds at a given temperature.

Table 8. Correlation/Accuracy values of the reactions while varying the temperature.

	LR		MLP		M5P		
Reaction	R	RMSE	R	RMSE	R	RMSE	No. of Rules
V11	0.9724	8.1245	0.9912	4.6565	0.9997	0.8067	121
V13	0.8906	9.6047	0.9954	2.0459	0.9998	0.4069	90
V15	0.988	3.4417	0.9951	2.217	0.9997	0.5827	135
V17	0.9517	8.7447	0.9954	2.795	0.9997	0.7397	86

V11 and V15 reduce as the temperature increases over time due to the water evaporating away, while V13 and V17 increase as the temperature increases because there is more hydroxyl available.

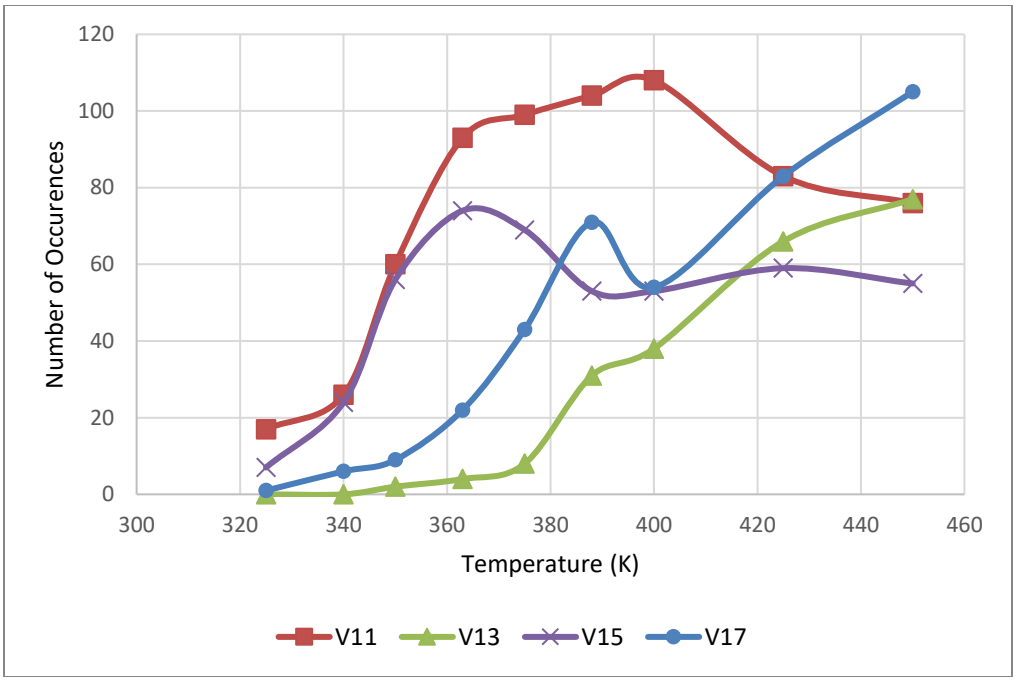

Figure 33. ML algorithm comparison of reaction V11 for varying temperatures after five seconds.

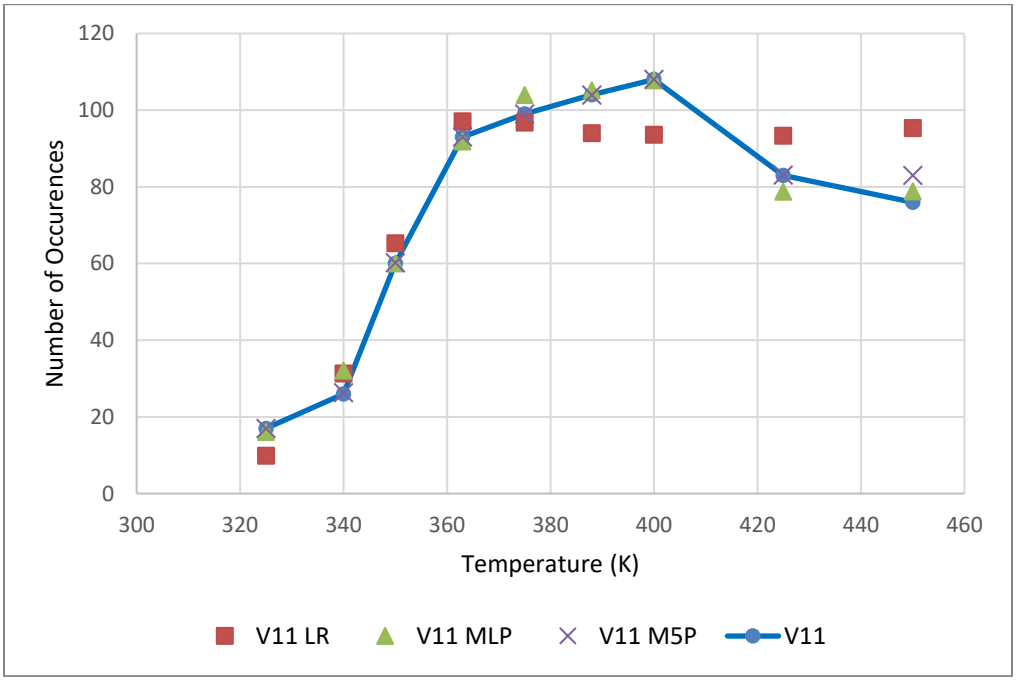

Figure 34. ML algorithm comparison of reaction V11 for varying temperatures after five seconds.

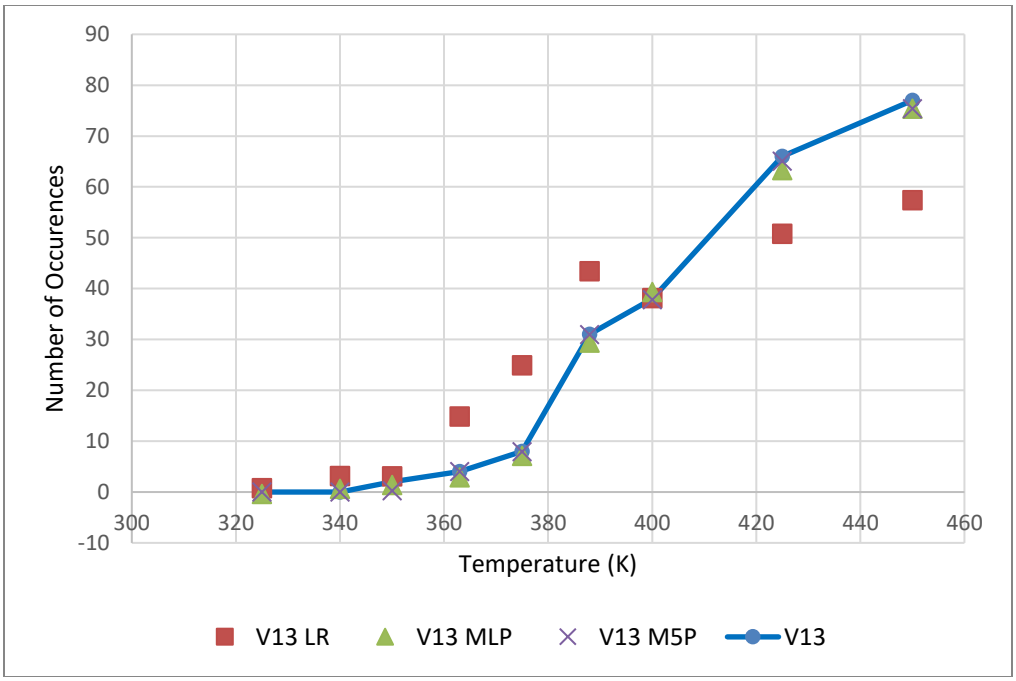

Figure 35. ML algorithm comparison of reaction V13 for varying temperatures after five seconds.

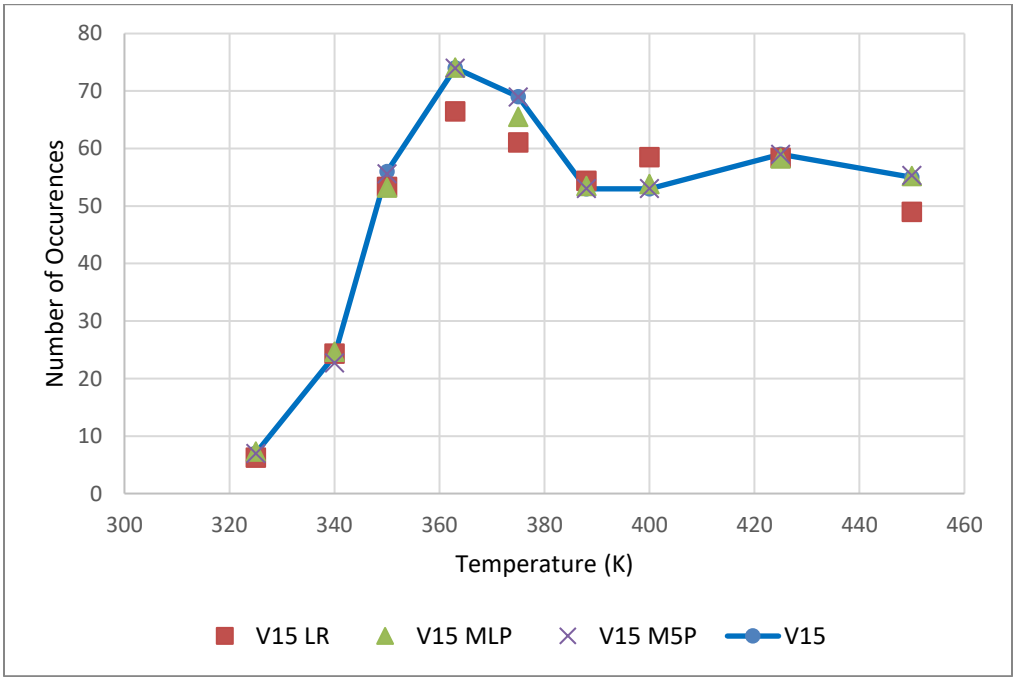

Figure 36. ML algorithm comparison of reaction V15 for varying temperatures after five seconds.

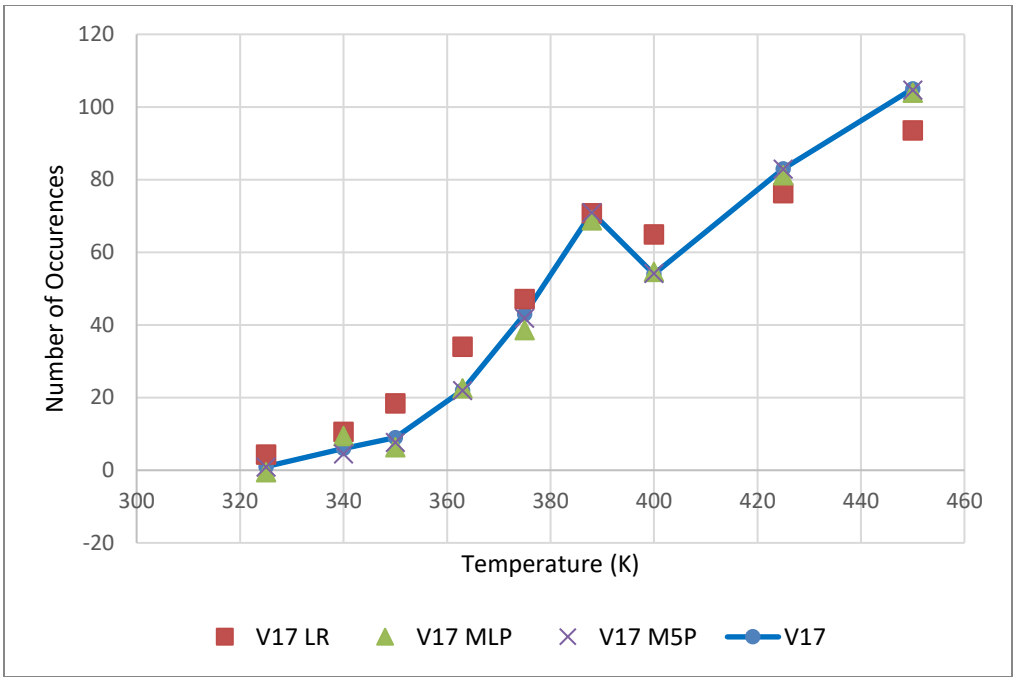

Figure 37. ML algorithm comparison of reaction V17 for varying temperatures after five seconds.

In Figure 33, for every reaction, they start with very few reactions, but after reaching a certain temperature they jump in occurrences. This is due to being able to overcome the activation energy. V11 and V15 have a lower activation energy than V13 and V17, thus being able to increase in occurrences at a lower temperature. However, for V11 after 400K and V15 after 363K, the number of occurrences goes down. For V13 and V17, after they pass the activation energy they continue to rise in the number of occurrences.

CONCLUSION

Although kMC has been shown to replicate real data, it has a lot of limitations. Through machine learning, we can bypass the complexities kMC holds and provide a more efficient and effective method to reach similar results. One of the simplest forms of machine learning, linear regression, can roughly predict a variety of different reactions that would be impossible to be known experimentally. MLP, although more complicated than LR, can predict the reactions more effectively than LR. M5P has shown to be the best predictor amongst all the simulations shown in this paper with the high R values and the low average RMSE values. To predict the QCM, hydrogen, oxygen, or zinc can be best predicted with MLP for varying cycle lengths and temperatures, while M5P can predict the reactions the best between the three algorithms tested. Although the linear regression had trouble with most scenarios, the R value was still highly correlated and the RMSE was still within a reasonable range. Since LR is the most simplistic model, it can be used for estimated values, while the MLP and M5P can perform the more accurate values. Thus using these three models together can represent the kMC simulations that have been to replicant real-life data. Using this method can help experimentalists study the ALD of zinc oxide and possibly different thin film materials if the same procedure is applied.

REFERENCES

(1) Al-Hardan, N. H.; Abdullah, M. J.; Abdul Aziz, A.; Ahmad, H.; Low, L. Y. ZnO thin films for VOC sensing applications. *Vacuum* **2010**, *85* (1), 101-106. DOI: 10.1016/j.vacuum.2010.04.009.

(2) Ding, Y.; Zhang, Y.; Ren, Y. M.; Orkoulas, G.; Christofides, P. D. Machine learning-based modeling and operation for ALD of SiO2 thin-films using data from a multiscale CFD simulation. *Chem. Engineering Research and Design* **2019**, *151*, 131-145. DOI: 10.1016/j.cherd.2019.09.005.

(3) George, S. M. Atomic layer deposition: an overview. *Chem. Rev.* **2010**, *110* (1), 111-131. DOI: 10.1021/cr900056b.

(4) Weckman, T.; Shirazi, M.; Elliott, S. D.; Laasonen, K. Kinetic Monte Carlo Study of the Atomic Layer Deposition of Zinc Oxide. *The J. of Phys. Chem. C* **2018**, *122* (47), 27044-27058. DOI: 10.1021/acs.jpcc.8b06909.

(5) Dkhissi, A.; Esteve, A.; Mastail, C.; Olivier, S.; Mazaleyrat, G.; Jeloaica, L.; Djafari Rouhani, M. Multiscale Modeling of the Atomic Layer Deposition of HfO2 Thin Film Grown on Silicon: How to Deal with a Kinetic Monte Carlo Procedure. *J. Chem. Theory Comput.* **2008**, *4* (11), 1915-1927. DOI: 10.1021/ct8001249.

(6) Deminsky, M.; Knizhnik, A.; Belov, I.; Umanskii, S.; Rykova, E.; Bagatur'yants, A.; Potapkin, B.; Stoker, M.; Korkin, A. Mechanism and kinetics of thin zirconium and hafnium oxide film growth in an ALD reactor. *Surf. Sci.* **2004**, *549* (1), 67-86. DOI: 10.1016/j.susc.2003.10.056.

(7) Mazaleyrat, G.; Estève, A.; Jeloaica, L.; Djafari-Rouhani, M. A methodology for the kinetic Monte Carlo simulation of alumina atomic layer deposition onto silicon. *Comput. Mater. Sci.* **2005**, *33* (1-3), 74-82. DOI: 10.1016/j.commatsci.2004.12.069.

(8) Shirazi, M.; Elliott, S. D. Atomistic kinetic Monte Carlo study of atomic layer deposition derived from density functional theory. *J. Comput. Chem.* **2014**, *35* (3), 244-259. DOI: 10.1002/jcc.23491.

(9) Rey, J. C.; Cheng, L. Y.; McVittie, J. P.; Saraswat, K. C. Monte Carlo low pressure deposition profile simulations. *J. of Vacuum Sci. & Technology A: Vacuum, Surfaces, and Films* **1991**, *9* (3), 1083-1087. DOI: 10.1116/1.577580.

(10) Magness, D. T. Kinetic Monte Carlo Investigations Involving Atomic Layer Deposition of Metal-Oxide ThinFilms. *MSU Graduate Theses* **2020**, *3578*.

(11) Andersen, M.; Panosetti, C.; Reuter, K. A Practical Guide to Surface Kinetic Monte Carlo Simulations. *Front Chem.* **2019**, *7*, 202. DOI: 10.3389/fchem.2019.00202.

(12) Hastie, T.; Tibshirani, R.; Friedman, J. Overview of Supervised Learning. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer New York, 2009; pp 9-41.

(13) Chinnamgari, S. K. *R Machine Learning Projects: Implement supervised, unsupervised, and reinforcement learning techniques using R 3.5*; Packt Publishing Ltd, 2019.

(14) Jo, T. *Machine Learning Foundations: Supervised, Unsupervised, and Adv. Learning*; Springer Nature, 2021.

(15) Haghighatlari, M.; Hachmann, J. Advances of machine learning in molecular modeling and simulation. *Current Opinion in Chem. Engineering* **2019**, *23*, 51-57. DOI: 10.1016/j.coche.2019.02.009.

(16) Rich Caruana, S. L., Lee Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Adv. in Neural Information Processing Systems* **2000**, *13*, 402-408.

(17) Payam Refaeilzadeh, L. T., and Huan Liu. Cross-Validation. **2009**. DOI: 10.1007/978-0-387-39940-9_565.

(18) R. Bharat Rao, G. F., Romer Rosales. On the Dangers of Cross-Validation. An Experimental Evaluation. *Proceedings of the 2008 SIAM International Conference on Data Mining (SDM)* **2008**, 588-596. DOI: https://doi.org/10.1137/1.9781611972788.54.

(19) Berrar, D. Cross-Validation. *Encyclopedia of Bioinformatics and Comput. Biology* **2018**, *1*, 542-545. DOI: 10.1016/B978-0-12-809633-8.20349-X.

(20) Maulud, D.; Abdulazeez, A. M. A Review on Linear Regression Comprehensive in Machine Learning. *J. of Appl. Sci. and Technology Trends* **2020**, *1* (4), 140-147. DOI: 10.38094/jastt1457 (acccessed 2022/04/07).

(21) Gardner, M. W.; Dorling, S. R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment* **1998**, *32* (14-15), 2627-2636. DOI: 10.1016/s1352-2310(97)00447-0.

(22) E. K. Onyari, F. M. I. Application of MLP Neural Network and M5P Model Tree in Predicting Streamflow: A Case Study of Luvuvhu Catchment, South Africa. *International J. of Innovation, Management and Technology* **2013**, *4*, 11-15.

(23) Kingsford, C.; Salzberg, S. L. What are decision trees? *Nat. Biotechnology* **2008**, *26* (9), 1011-1013. DOI: 10.1038/nbt0908-1011.

(24) Jain, A. K.; Jianchang, M.; Mohiuddin, K. M. Artificial neural networks: a tutorial. *Computer* **1996**, *29* (3), 31-44. DOI: 10.1109/2.485891.

(25) Jierula, A.; Wang, S.; OH, T.-M.; Wang, P. Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data. *Appl. Sci.* **2021**, *11* (5), 2314.

(26) Li, J. Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what? *PLoS One* **2017**, *12* (8), e0183250-e0183250. DOI: 10.1371/journal.pone.0183250 PubMed.

(27) Mukaka, M. M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **2012**, *24* (3), 69-71. PubMed.

(28) Kwon, S.; Bang, S.; Lee, S.; Jeon, S.; Jeong, W.; Kim, H.; Gong, S. C.; Chang, H. J.; Park, H.-h.; Jeon, H. Characteristics of the ZnO thin film transistor by atomic layer deposition at various temperatures. *Semiconductor Sci. and Technology* **2009**, *24* (3). DOI: 10.1088/0268-1242/24/3/035015.

(29) Chicco, D.; Warrens, M. J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623-e623. DOI: 10.7717/peerj-cs.623 PubMed.

APPENDICES

Appendix A: Input File of Reactions

The input file is too long to place here. The input file containing all the reactions and output parameters is found on the following GitHub:

https://github.com/spparks/spparks

Appendix B: Input File of Lattice Points

The input file is too long to place here. The input file of all the lattice points is found on the following GitHub:

https://github.com/spparks/spparks