MODELING OF SILVER-PLATINUM ALLOY CONFIGURATION ENERGIES

by

Lincoln B. Houghton

A senior thesis submitted to the faculty of

Brigham Young University - Idaho

in partial fulfillment of the requirements for the degree of

Bachelor of Science

Department of Physics

Brigham Young University - Idaho

December 2021

BRIGHAM YOUNG UNIVERSITY - IDAHO

DEPARTMENT APPROVAL

of a senior thesis submitted by

Lincoln B. Houghton

This thesis has been reviewed by the research advisor, research coordinator, and department chair and has been found to be satisfactory.

Date	Lance Nelson, Advisor
Date	Lord Richard Datwyler, Committee Member
Date	Summer Houghton, Committee Member
Date	Evan Hansen, Thesis Coordinator
Date	R. Todd Lines, Chair

ABSTRACT

MODELING OF SILVER-PLATINUM ALLOY CONFIGURATION ENERGIES

Lincoln B. Houghton

Department of Physics and Astronomy

Bachelor of Science

The abstract is a *summary* of the thesis, *not an introduction.* Keep in mind that abstracts are often published separately from the paper they summarize. In your abstract, give a concise synopsis of the work, emphasizing the conclusions; you need not include the supporting arguments for the conclusions. The purpose of the abstract is to help prospective readers decide whether to read your thesis, but your goal is not necessarily to persuade people to read your thesis. In fact, a successful abstract enables people to get an accurate overall view of your work without needing to read it. Usually, an abstract contains a single paragraph, but it can have more if absolutely necessary. Remember to state the subject of the paper immediately followed by a summary of the experimental or theoretical results and the methods used to obtain them. Avoid equations, graphics, and citations; if a citation is essential it must be cited fully within the abstract. Keep the abstract factual. Avoid vague statements like,

"Conclusions are drawn," or "the significance of the experiment is discussed." State the conclusions and findings outright in the abstract.

ACKNOWLEDGMENTS

This page is optional. You may acknowledge whom you will—your advisor, colleagues, family members. Please keep acknowledgments in good taste. I would like to acknowledge Dr. Kristine Hansen and Dr. Elizabeth Hedengren, whose Advanced Writing Seminar motivated this project. I also wish to thank Jean-François Van Huele, Steven Turley, and Ross Spencer for reviewing this document and for ripping it to shreds as every good advisor should do to a thesis draft.

Contents

List of Figures

0.1 Introduction

The discovery of novel, high-performing materials is the main driver for technological and industrial advances. The discovery process presents significant challenges; a material with suitable properties which is also thermodynamically stable must be identified. Rather than manufacturing every conceivable material in a laboratory, a more efficient approach involves using computer simulations and calculations to guide the metallurgist. The density functional theory (DFT) is a well-known quantum-mechanical tool for calculating formation energies, a critical quantity for determining a material's thermodynamic stability. These formation energies could then be used to construct the material's phase diagram and determine which phases are thermodynamically stable.

Although accurate, DFT calculations are computationally costly, making ground state searches (involving hundreds of thousands of calculations) and thermodynamic simulations computationally prohibitive. Another approach involves using a small set of accurate DFT data (hundreds of calculations) to construct a model which can calculate much faster. Such a model would help alleviate the stress on the main bottleneck to finding novelty alloys: computational power.

In this paper, the process of building a material model will be explained, starting from a simple toy model and later applied to a real DFT data set for Ag-Pt. Sections 0.2 and 0.3 will cover the math and modeling concepts required to understand the actual model construction in Section 0.4 and the analysis of its results in Section 0.5.

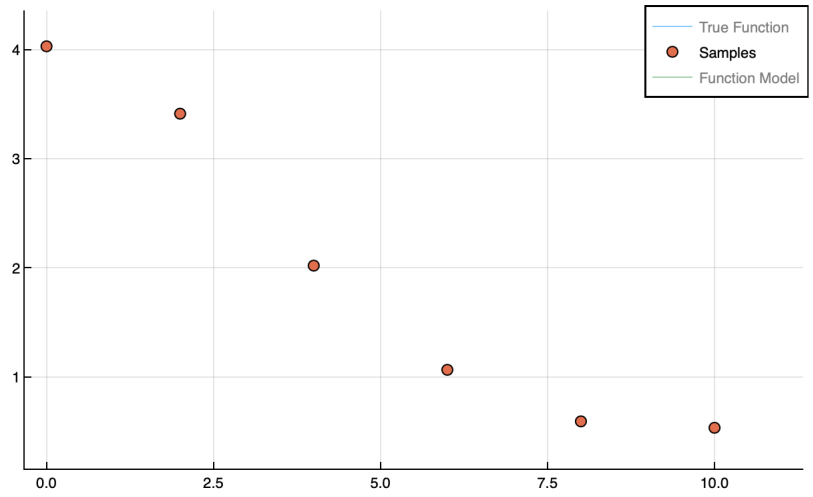


Figure 1 An unknown function sampled by 6 data points.

0.2 Background

0.2.1 Linear Algebra

A basic knowledge of linear algebra, data analysis, and signal processing is required to understand how to build simple yet useful models. Samples from an unknown function are presented in Figure 1. Building a function that matches the data will now be investigated.

It will be assumed that the function can be expanded using a basis, and in this case, a power series:

$$f(x) = \sum_{n=0}^{\infty} b_n x^n \tag{1}$$

$$= x^0 b_0 + x^1 b_1 + x^2 b_2 + \dots \ . \tag{2}$$

with the coefficients b_n to be determined via the fitting process. Evaluating equation 2 at the observed data points in figure 1 produces a system of equations:

$$
\begin{aligned}
f(0) &= b_0 \\
f(2) &= b_0 + 2b_1 + 4b_2 + \dots \\
f(4) &= b_0 + 4b_1 + 16b_2 + \dots \\
f(6) &= b_0 + 6b_1 + 36b_2 + \dots \\
f(8) &= b_0 + 8b_1 + 64b_2 + \dots \\
f(10) &= b_0 + 10b_1 + 100b_2 + \dots \qquad (3) \\
& \qquad (4)
\end{aligned}
$$

This system of equations can be expressed in the following matrix form:

$$
\begin{bmatrix}
1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\
1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\
1 & x_3 & x_3^2 & x_3^3 & x_3^4 \\
 & & \vdots & &
\end{bmatrix}
\begin{bmatrix}
b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4
\end{bmatrix}
=
\begin{bmatrix}
f(x_1) \\ f(x_2) \\ f(x_3) \\ \vdots
\end{bmatrix}. \qquad (5)
$$

And finally, $\mathbf{A}$ and $\vec{y}$ can be populated with their true values,

$$
\begin{bmatrix}
1 & 0 & 0^2 & 0^3 & 0^4 \\
1 & 2 & 2^2 & 2^3 & 2^4 \\
1 & 4 & 4^2 & 4^3 & 4^4 \\
1 & 6 & 6^2 & 6^3 & 6^4 \\
1 & 8 & 8^2 & 8^3 & 8^4 \\
1 & 10 & 10^2 & 10^3 & 10^4
\end{bmatrix}
\begin{bmatrix}
b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4
\end{bmatrix}
=
\begin{bmatrix}
f(0) \\ f(2) \\ f(4) \\ f(6) \\ f(8) \\ f(10)
\end{bmatrix}. \qquad (6)
$$

The shape of the matrix $\mathbf{A}$ determines the type of solution that can be found. If there are fewer equations (rows) than unknowns (columns), the system is underde-

termined and if it has more rows than columns it is overdetermined.

Underdetermined systems

Because the columns of an underdetermined system are linearly dependent, there are an infinite number of vectors $\vec{b}$ that can solve the system. To arrive at a single vector, the solution must be constrained using some chosen criteria. A commonly-used criteria is to minimize the ℓ_2-norm of the solution vector.

Solving an underdetermined system is done by using singular value decomposition (SVD). To solve $\mathbf{A}\vec{b} = \vec{y}$, first the eigenvalues and their corresponding orthonormal eigenvectors of $\mathbf{A}^T\mathbf{A}$ must be found. V is the matrix containing these eigenvectors and Σ is a diagonal matrix containing the singular values, the square roots of each non-zero eigenvalue, matching the shape of $\mathbf{A}$. Each column of the matrix U can be constructed by $u_k = \frac{1}{\sigma_k}\mathbf{A}\vec{v_k}$ where σ_k are the singular values and $\vec{v_k}$ are the columns of V. The final result is the complete SVD of $\mathbf{A}$ where $\mathbf{A} = U\Sigma V^T$ and $\vec{b}$ can be solved by,

$$\mathbf{A}\vec{b} = \vec{y} \tag{7}$$

$$\mathbf{A}^T\mathbf{A}\vec{b} = \mathbf{A}^T\vec{y}$$

$$\vec{b} = (U\Sigma V^T)^T\vec{y}$$

$$\vec{b} = V\Sigma^T U^T\vec{y}. \tag{8}$$

The solution vector $\vec{b}$ is the well-known least squares solution to the system. Using this solution vector, the model can be used to make predictions of the function for any x value within the range of sample points. To produce a continuous visual of the model's fit, the model can take in x values for every point in the sample range and plot its respective evaluation. The result of this process can be seen in Figure 2.

Overdetermined system

In an overdetermined system, since the column vectors don't span the space that they live in, there may be no vectors $\vec{b}$ that solve the system. The "closest" solution vector can be found by solving a slightly modified problem where the vector $\vec{y}$ is projected onto the subspace formed by the columns. This is eqivalent to the well-known least squares approach and can be written in matrix form as:

$$\mathbf{A}^T(\vec{y} - \mathbf{A}\hat{\vec{b}}) = 0 \tag{9}$$

$$\mathbf{A}^T\vec{y} = \mathbf{A}^T\mathbf{A}\hat{\vec{b}} \tag{10}$$

$$\hat{\vec{b}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\vec{y}, \tag{11}$$

where $\mathbf{A}^T$ denotes the transpose of $\mathbf{A}$.

Since DFT data points are so costly to generate, the matrices that are typically encountered when building materials models are most-often underdetermined (i.e. more basis functions than data points).

0.2.2 Basis Functions and Size of Training Set

The quality of the fit will be affected by (1)- the size of the training set (number of rows in matrix $\mathbf{A}$) and (2)- number of basis functions included in the expansion (columns in matrix $\mathbf{A}$). The convergence of our model with respect to these two parameters must be investigated. In theory there is no limit to the size of the training set or number of basis functions. However, since QM data is costly to generate, the size of the training set has a reasonable upper limit on the order of hundreds. In the example given, an increase in the traning set size or basis functions count will not dramatically affect the computational power required, but becomes a greater concern for models of systems with increasing complexity.

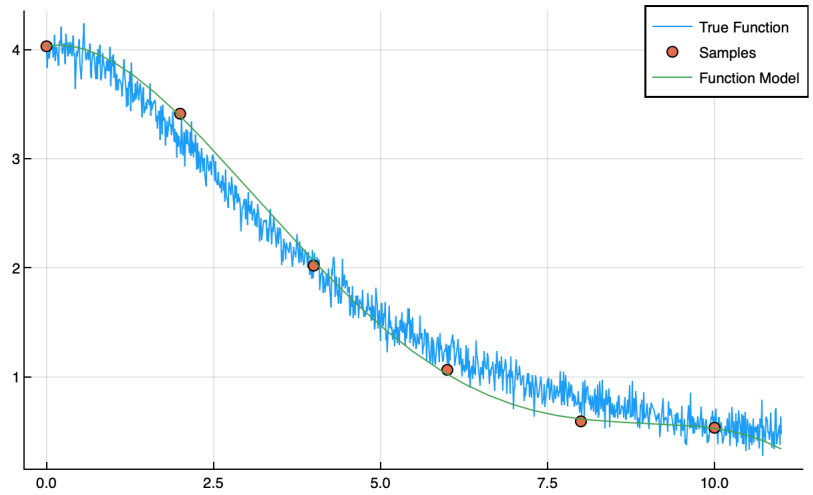


Figure 2 The "Witch of Agnesi" function with Gaussian noise is shown in blue. Samples from this function were gathered (circles) and used to construct a model using a simple polynomial basis. The fit function is depicted in green.

As discussed, the number of basis functions can make a large impact, but another important factor is quality. Though the choice of basis functions in the example above was simple, it can often be a difficult choice. Consider, for example, a Fourier basis. The equivalent of Equation 5 in this basis would be

$$\begin{bmatrix} \sin(x_1) & \sin(2x_1) & & \\ \sin(x_2) & \sin(2x_2) & \dots & \dots \\ \sin(x_3) & \sin(2x_3) & & \\ \vdots & & \ddots & \\ \sin(x_n) & \dots & & \sin(mx_n) \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_n) \end{bmatrix}, \tag{12}$$

where $\mathbf{A}$ is an $n \times m$ matrix.

When used in the proper circumstance, this Fourier basis can be an excellent choice for modeling a function, as in Figure 3.

When chosing sample points, there are again two variables to consider: number

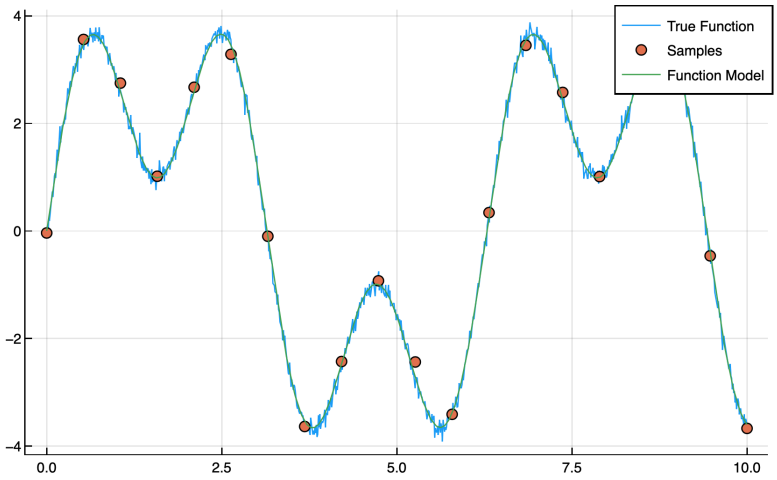


Figure 3 The sample points shown were used to build a model of the function in green, and the true function is displayed in blue. The Fourier basis resulted in an excellent fit.

and breadth. As will be seen later, the number of samples can have a significant impact on the construction time and accuracy of a model. The breadth of samples is similarly important. If all samples from Figure 3 were taken between 0 and 1, the model produced would be a poor fit for the function, as in Figure 4.

By this point it should be obvious to the reader that the decision of number and breadth of sample points as well as quantity and quality of basis functions is critical to the model's performance. With well chosen basis functions but poor breadth of samples, a model's accuracy can be greatly limited, as in Figure 4. Figure 5a shows the reverse situation, good number and breadth of samples, but poorly chosen basis functions.

0.2.3 Julia

The figures provided throughout this paper were produced using the Julia programming language. Though encouraged to use Python for the majority of my formal education, Julia is a relatively new and fast growing language in terms of popularity.

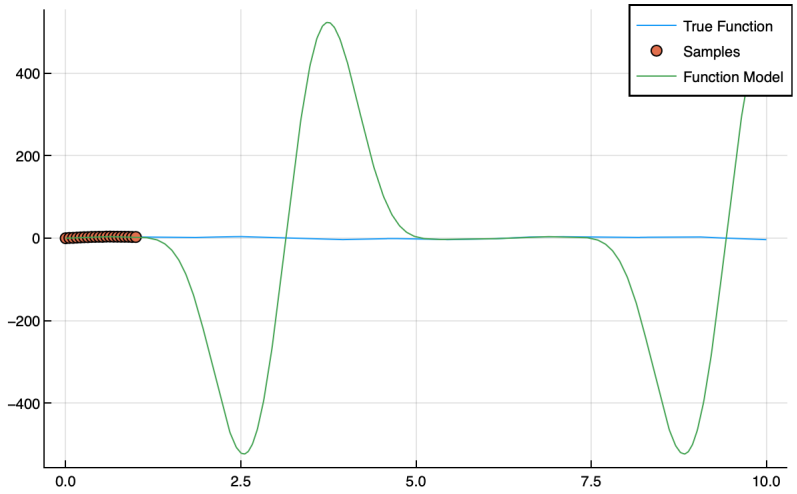
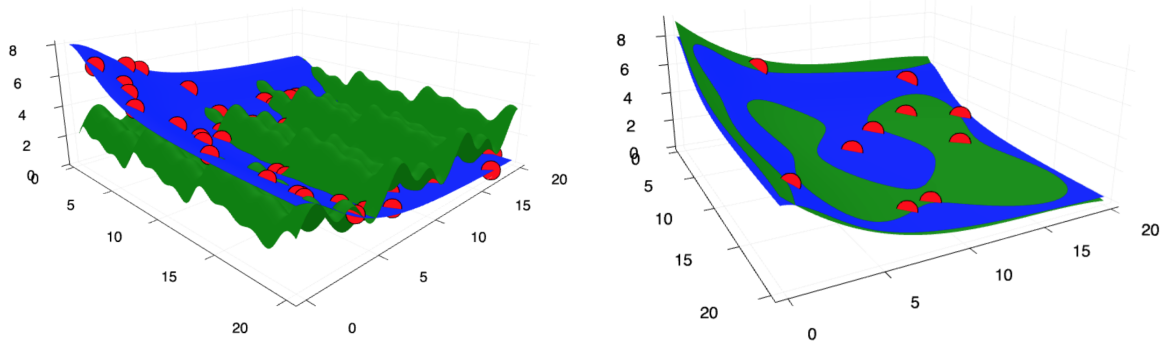


Figure 4 Gathering samples that are not uniformly distributed across the function space will produce a model that does not predict well across the space. In this case, the secret function from Figure 3 was only sampled from the interval 0 to 1. If the model were evaluated at $x = 2.5$, the fit would produce a highly inaccurate result when compared to the true function.



(a) Fourier basis. A is a 50×7 matrix.

(b) Power basis. A is a 10×7 matrix.

Figure 5 A three dimensional function in blue with a constructed model in green. The red points show the sampling of the true function. (a) In this case, a Fourier basis is a poor choice and resulted in an inaccurate model of the true function. (b) A Power series makes a convincingly better fit even with significantly fewer sample points.

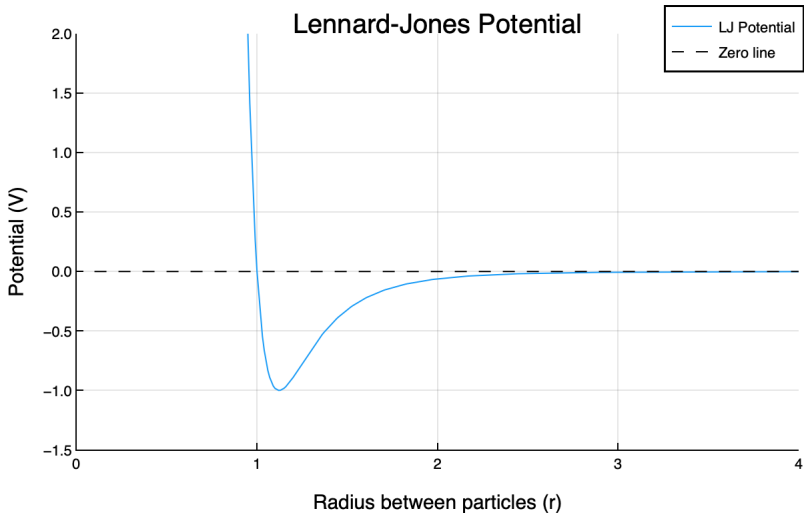


Figure 6 The Lennard-Jones potential. A simple yet realistic model of intramolecular forces.

0.3 Preliminary Modeling

0.3.1 Lennard-Jones Potential

Before jumping to the DFT data, this math and theory will be confirmed by constructing models on a simplified potential, the Lennard-Jones potential [1]. Though the Lennard-Jones potential is a simplification of reality, it does a excellent job representing real intermolecular forces of attraction and repulsion. The potential is a function of distance between two particles given by

$$V_{LJ}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right], \tag{13}$$

where ε and σ are constants for a given particle interaction. Figure 6 shows the plot of this potential.

It can be recalled that the force from a potential is given by

$$F = -\nabla U, \tag{14}$$

and thus the force between two particles is zero at the bottom of the potential well. With that location as a reference, distances any greater will produce a force that is attractive and at any lesser distances, the force is intensely repulsive.

Calculating the potential between two particles is not difficult, but finding the total potential energy of a system of several particles becomes increasingly computationaly expensive. Because of the simplicity of the Lennard-Jones potential, the computational power required to solve for the system's energy is still relatively small. In preparation for using real, quantum-mechanical data, each system's energy will be treated as expensive to compute.

0.3.2 Constructing a Model

Ensuring a sufficient number and breadth of samples as well as reasonable basis functions becomes difficult as the number of dimensions goes beyond 2 or 3. Ways to determine quality of samples will be discussed later and from studying a variety of potential basis functions, bessel functions of the second kind, $Y_\alpha(x)$, have potential to be useful.

Rewriting Equation 7 in component form yields

$$\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1m} \\ A_{21} & & & \\ \vdots & & & \vdots \\ A_{n1} & \dots & & A_{nm} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ \vdots \\ V_n \end{bmatrix}. \tag{15}$$

Recognizing each row of A as a unique sample, Equation 15 can be written as a

system of linear equations

$$\begin{aligned} A_{11}b_1 + A_{12}b_2 + A_{13}b_3 + ... + A_{1m}b_m &= V_1 \\ A_{21}b_1 + A_{22}b_2 + A_{23}b_3 + ... + A_{2m}b_m &= V_2 \\ A_{n1}b_1 + A_{n2}b_2 + A_{n3}b_3 + ... + A_{nm}b_m &= V_n. \end{aligned} \tag{16}$$

where V_n is the total potential energy of configuration n. Each element of matrix A will be populated as follows

$$\begin{aligned} A_{11} &= Y_0(\alpha_{01}r_{12}) + Y_0(\alpha_{01}r_{13}) + Y_0(\alpha_{01}r_{23}) + \dots \\ A_{12} &= Y_0(\alpha_{02}r_{12}) + Y_0(\alpha_{02}r_{13}) + Y_0(\alpha_{02}r_{23}) + \dots \\ A_{1m} &= Y_0(\alpha_{0m}r_{12}) + Y_0(\alpha_{0m}r_{13}) + Y_0(\alpha_{0m}r_{23}) + \dots \end{aligned} \tag{17}$$

where α_{0m} is the m-th zero of Y_0, the zeroth bessel function of the first kind. Each r_{xy} represents the separation vector norm (ℓ^2 norm) for a single particle, x, and its pair, y, in the configuration. Because this model only accepts the ℓ^2 norm of each separation vector, it is only a simple distance-dependent model using pair-interactions rather than a more complicated model using three body interactions (see Section 0.6.1).

An example of a random assortment of particles in a box can be seen in Figure 7. If two particles are generated too close together, it will cause a large spike in the system's potential. Thus a minimum separation distance must be enforced for each configuration generated. This minimum separation distance will ensure some degree of uniformity in the configurations and their total potential, effectively reducing the range of possible energies and requiring fewer training systems.

Through experimentation, the detailed relationship between size of the training set, number of basis functions, and the minimum separation distance can be outlined. In general, the accuracy of the model increases as the number of training sets and basis functions increases. The accuracy also increases when the minimum separation

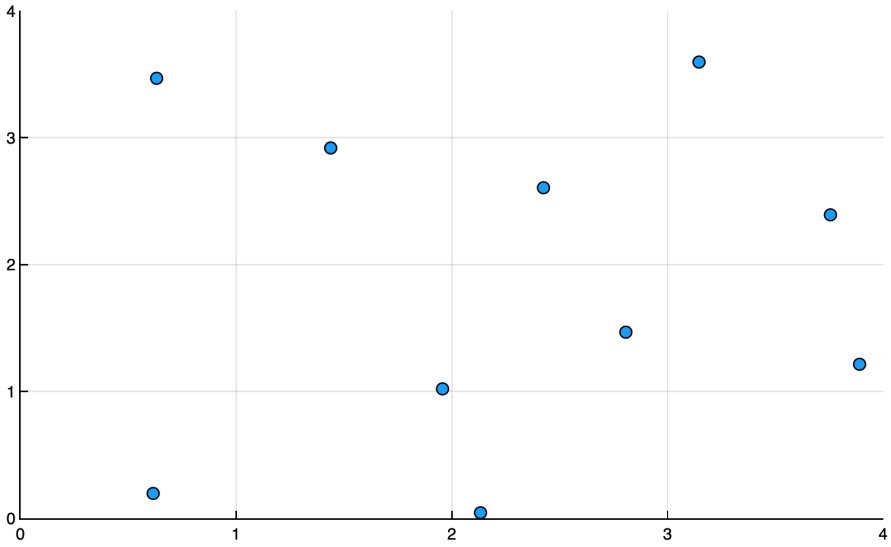


Figure 7 A random assortment of 10 particles in a 4x4 square. No two particles are allowed to be within a pre-specified distance of each other.

distance does not correspond to a particularly large potential energy. In the case of the Lennard-Jones potential in Figure 6, the minimum separation distance has a preferred minimum of 0.8 or 0.9 to avoid large errors in the constructed model.

0.3.3 Multi-Type Particle Systems

Now that a simple monotomic model has been constructed, it can be adjusted to handle two different types of particles. The different particles can be labeled type-A and type-B. This diatomic system will now require *three* different potential equations, one describing each type of interaction. The three unique interactions are type-A interacting with another type-A, a type-A interacting with a type-B, and a type-B with another type-B. Because these are arbitrary interactions (due to unspecified atomic structure), they can be defined by choosing reasonable values of ε and σ from

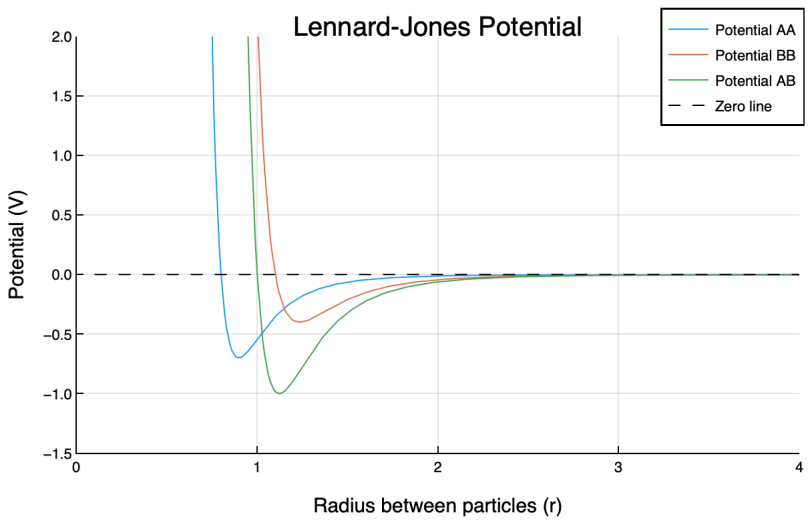


Figure 8 A Lennard-Jones potential for each particle interaction type (AA, AB, and BB). Each interaction potential has a slightly different equilibrium position and energy.

Equation 13.

$$V_{AB}(r) = 4\left[\left(\frac{1}{r}\right)^{12} - \left(\frac{1}{r}\right)^{6}\right] \tag{18}$$

$$V_{AA}(r) = 4(0.7)\left[\left(\frac{0.8}{r}\right)^{12} - \left(\frac{0.8}{r}\right)^{6}\right] \tag{19}$$

$$V_{BB}(r) = 4(0.4)\left[\left(\frac{1.1}{r}\right)^{12} - \left(\frac{1.1}{r}\right)^{6}\right] \tag{20}$$

The graph of each potential can be seen in Figure 8. Each of the equilibrium positions and energies is slightly different, but all in the same neighborhood.

To handle this increase in complexity, matrix A will need to contain *three* columns where the previous model had only one. This is also because of the three different interaction types, one column for each. Another obstacle arises in the decision for particle type ratio. If the particle number (10) and box size (4×4) remain constant, how many type-A versus type-B particles should there be to help train an effective model?

This question can be generalized to ask: what is a sufficient breadth of samples

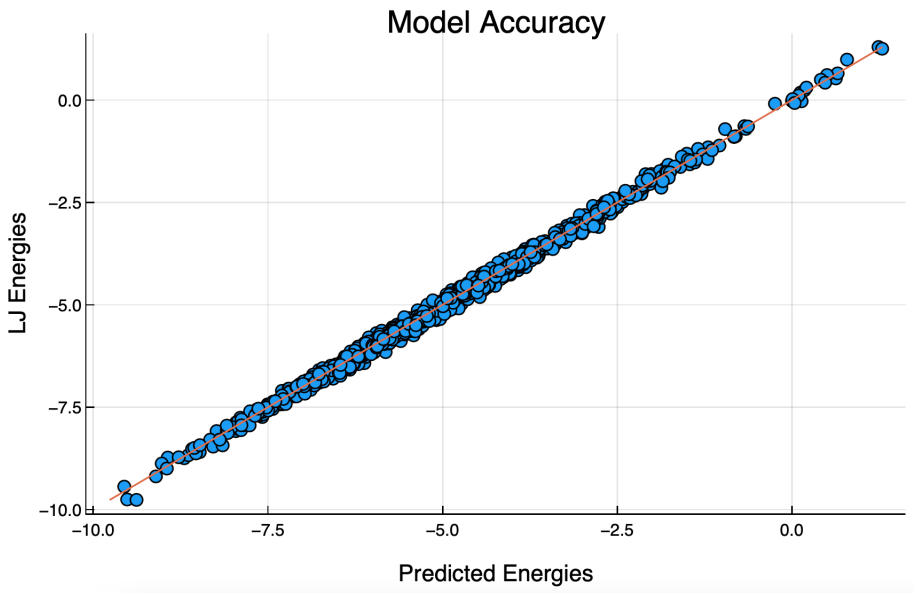


Figure 9 1000 tests of a model with a training size of 300 systems, 900 total basis functions, and a spread of particle type ratios each for a 10 particle system. The fit is very good, denoting a highly accurate model.

for this model? The answer depends greatly upon the desired range of predictions. As will be seen in Section 0.4.1, the actual configurations vary in particle number and particle type ratio. Therefore, the model constructed here should be trained and tested using a spread of values for those two variables. Each training and testing set will thus be populated by a random number and ratio of particles.

Once a model has been properly constructed and trained, its accuracy can be visually determined by plotting each system's predicted energy versus its actual energy. An example of this type of plot can be seen in Figure 9 and will be used again later on.

Now that a successful model has been created for a simplified potential, it can be adapted to fit real data in the hopes of producing useful data faster than current models.

0.4 Procedure

0.4.1 Quantum Mechanical Data

First-principles calculations were performed within the framework of AFLOW [] which employs the VASP software for computing energies. [] Projector-augmented-wave (PAW) potentials were used and exchange-correlation functionals parametrized by Perdew, Burke, and Ernzerhof under the generalized gradient approximation (GGA) []. A dense k-mesh scheme was used to perform the numeric integration over the Brillioun zone []. Optimal choices of the unit cells, by standardization of the reciprocal lattice, were adopted to accelerate the convergence of the calculations.

As previously discussed, the process of obtaining this data was computationally expensive, creating a bottleneck in the discovery of novelty alloys. The culmination of this research is to produce a simpler model that can be trained on the given data to reproduce expected configuration energies within a range of reasonable error. Figure 10 shows a histogram of all the configuration energies in the data set.

Each configuration in the data represents a unique primitive unit cell. A unit cell is the building block of any crystal structure. Each unit cell is an identical copy of every other, with the same shape, size, and contents. A *primitive* unit cell is the smallest possible unit cell which contains only one of each uniquely positioned atoms in the crystal [2]. An example of a primitive unit cell configuration can be seen in Figure 11.

To use the data in the *.txt* file, it needs to be parsed into vectors that can be easily manipulated. An example of the data being parsed can be seen in Figure 12. The number on the second line is the lattice parameter, followed by three lattice vectors in (i, j, k) coordinates. The line following contains two numbers, the first is the number of silver (Ag) atoms in the unit cell and the second is the number of platinum (Pt)

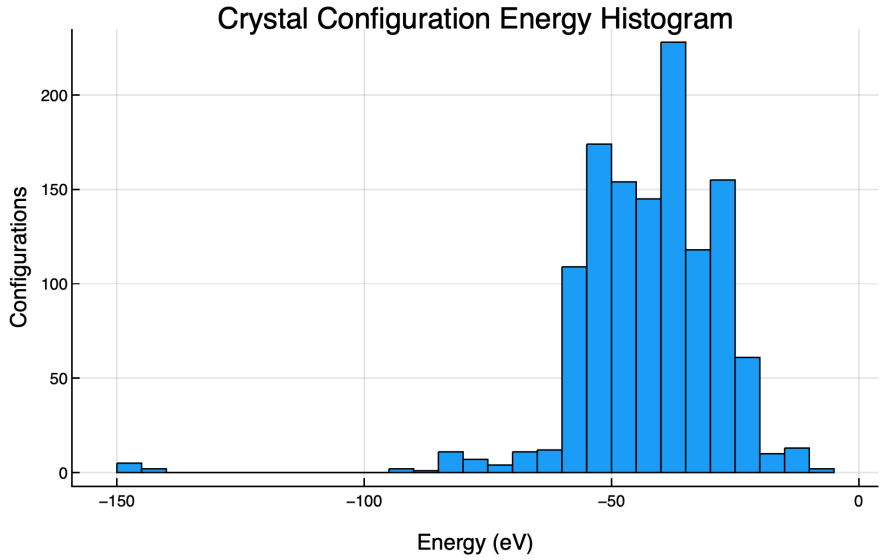


Figure 10 A histogram showing all the configuration energies in the original data file. The majority lie within the −20eV to −60eV range, with a few outliers around −150eV.

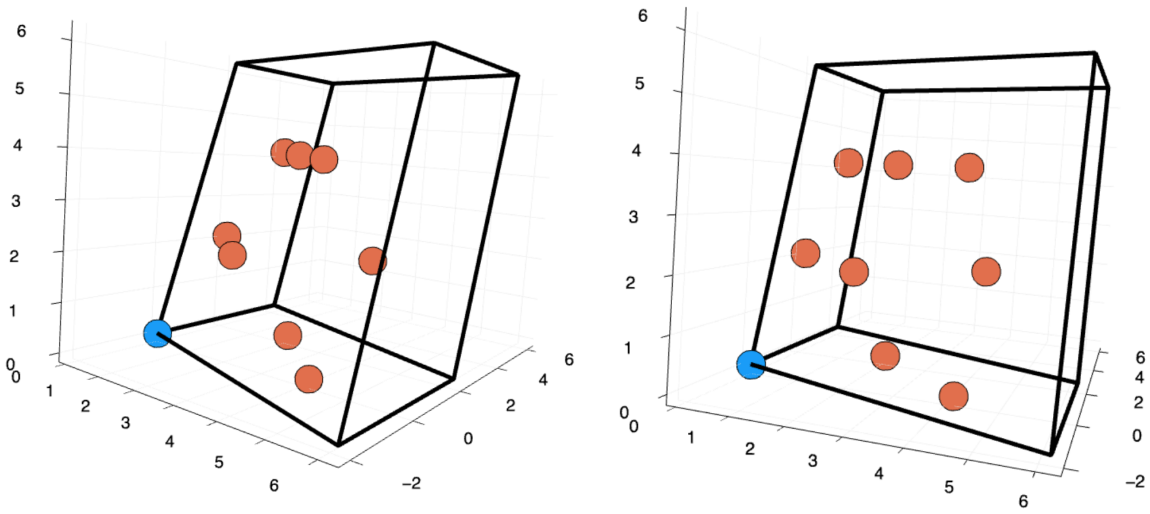


Figure 11 The primitive unit cell for configuration 65 from two different perspectives. The blue point is the single Ag atom while each red point shows each Pt atom in said configuration.

```
Ag-Pt str #: 100
   4.06600000000000
     0.4897375027131054    0.4897375027131054    0.0000000000000000
    -0.9857506444128054    0.9857506444128054    0.0001480742989182
     0.4911874202245589   -0.0014499175114533    1.5317387854017932
   3   3
Direct
  0.9962945910437568  0.0016543423460078  0.0074108179124863
  0.3399413516957068  0.8293793395082005  0.3201172866085926
  0.9989260447531194  0.5020634528287573  0.0021479104937541
  0.6671639334432413  0.6696967385982830  0.6656721431135040
  0.3328857305662609  0.3335677384713236  0.3342285288674844
  0.6647883484979076  0.1636383882474277  0.6704233130041786
 Energy:-26.60832
```

Figure 12 An example of the data given in AgPtdata.txt that needs to be used to build the model.

atoms. In direct coordinates (in terms of the lattice vectors) the positions of each silver and platinum atom are then given. The last line in this system tells the total potential energy of the unit cell configuration.

0.4.2 Model Construction

With the file parsed and the important data retrieved, the process of building a model can commence. The potential energy of a single particle is due to its interactions with all its surrounding particles, thus to account for each interaction with nearby particles the unit cell must be propagated outwards in all three dimensions. All atomic pairs can be enumerated by adding multiples of the lattice vectors. Then the relative position of each affecting particle can be determined and the vector separating the particle pair can be calculated. These separation vectors are the information that will be passed into the basis function to construct the model.

Because the affect two particles have on each other drops off as a function of distance (similar to the Lennard-Jones potential), the unit cell does not need to be propogated infinitely in each direction, only out to a radius of reasonable influence, creating an imaginary sphere beyond which all interactions are negligible. It should be clear that the choice of this radius will make a significant impact on the quality

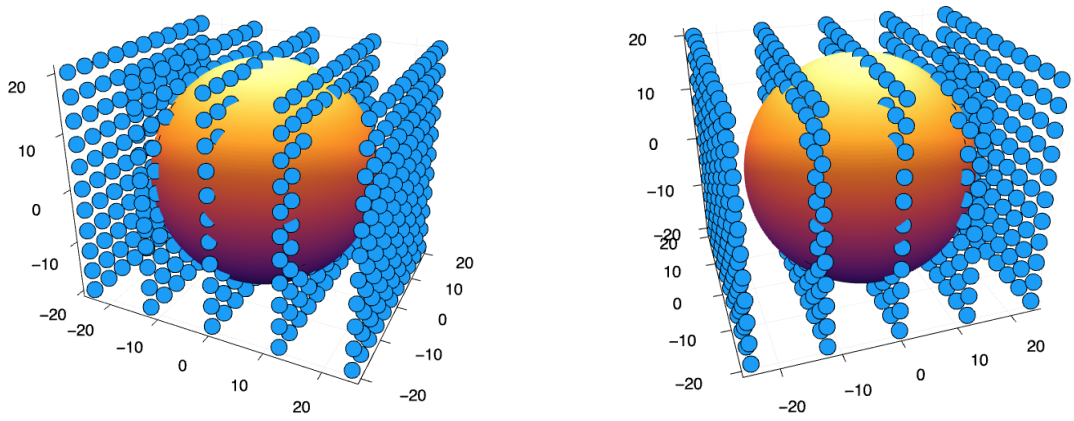


Figure 13 The sphere of influence for one silver atom is completely encapsulated by the unit cell's iterations from two different perspectives. Only one particle from each unit cell is shown.

of the model. It should also be noted that one arbitrarily chosen radius cannot be applied effectively to each unique system. A simple solution is to make the radius of this sphere of influence equal to the magnitude of the largest lattice vector multiplied by a constant. The effect of this constant on the model's precision can be tested later, but for now will be chosen to be 1.2.

The simplification of reality due to this "sphere of influence" gives a clear upper bound to our unit cell propagation. It is expected that the unit cell will be propagated out further than the radius in each direction, thus encapsulating said sphere. An example of this iterated unit cell can be seen in Figure 13.

Each unit cell in Figure 13 can now be populated with all the contained atoms. Every atom inside the sphere can then be extracted. When these atoms are stored into a new vector, they can be plotted as in Figure 14. When this process is repeated for every system in the DFT data, the result is a vector containing these "important positions" for each unique silver and platinum atom in each unit cell. The separation

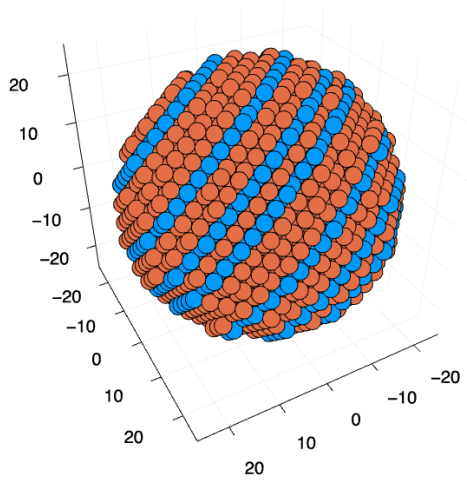


Figure 14 Every silver and platinum atom inside the sphere of interest. According to the model, these are the only atoms interacting with the central atom in question. This data is taken from the first silver atom in the third sample system.

vector norms can be quickly calculated and stored.

With all the pertinent information from the data organized into vectors, construction of the **A** matrix can begin. As in the Lennard-Jones example, each row represents a crystal structure and each column corresponds to a unique basis function. And as explained in Section 0.3.3, a single basis function will result in three columns of **A** due to each type of interaction. The method of populating **A** will be the same as was done earlier in Equation 17. With 300 basis functions, the first 300 columns of **A** will be calculated using the separation vector norms from Ag-Ag interactions. The following 300 columns from Pt-Pt interactions and the final 300 from Ag-Pt.

The size of the training set is limited to 1224, the number of systems given in the data set. Whatever systems are not included in the training set will comprise the holdout set, used for testing the model's accuracy.

Once the model is fully constructed, tests can be run to determine the effect of radius, sample number, and number of basis functions on the precision and accuracy

of the model. An increase of radius, sample number, or number of basis functions will make significant impacts on the program runtime. As the radius is increased, the unit cell must be propagated out further to encapsulate the sphere of influence and more atom interactions will be considered. For every basis function added, there will be three columns added to the **A** matrix, drawing out its construction times.

Generally, as the number of basis functions is increased and the radius of influence extended, the model's predictions will become increasingly reliable. On the other hand, as those factors increase accuracy and precision, they also increase the computational costs. It would be ideal to find a manageable trade off between the program's runtime and reliability. To investigate the quality of the model as a function of cutoff radius, sample number, and number of basis functions, a sweep can be conducted over a range of values for each and the details of the fit for each combination can be recorded. Rather than run this script on an ordinary desktop computer, it was completed on Mary Lou, BYU's supercomputer.

0.5 Results and Analysis

The quality of fit for a single model is given in Figure 15. For some models, there are predictions that differ significantly, more than $100eV$, from their true values, these are configurations that are difficult for the model to predict. In those cases, the number of outlying crystals has been counted then excluded from the average error calculation. In this way, the average error can become mostly independent of the model's outliers. When a difficult-to-predict configuration is poorly predicted and thus excluded from the average error, it can reduce the overall average error for a given model. A better trained model may then predict that configuration energy slightly better, within $100eV$ of its true value, resulting in a larger average error. Therefore

```
Number of samples:                    800
Number of basis functions:            500
Radius r:                             1.5
Number of energy predictions:         424
Number of errors greater than 100.0: 3
Average error:                        3.365961170977803
```

Figure 15 An example of the data in the output file.

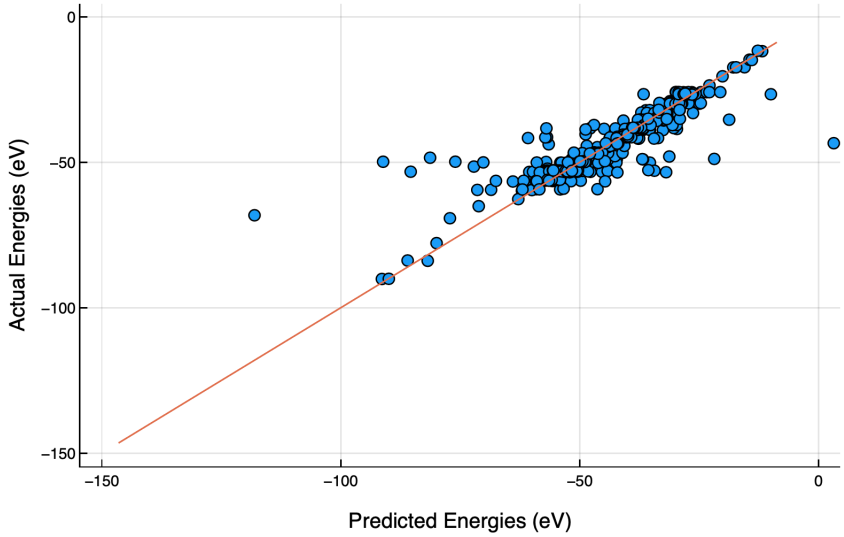


Figure 16 Predicted energies versus the actual energies. The model tested is the same as the one shown in Figure 15. Each energy prediction comes from the holdout set.

the average error and number of large errors must both be considered when assessing the value of the model.

To understand a given model's accuracy and precision visually, each prediction energy can be plotted against the actual energy for each configuration. The model referenced in Figure 15 can be seen in this format in Figure 16.

This method of visual analysis in Figure 16 works very well for observing one model at a time but interpreting each of the 270 unique models this way would difficult and time consuming. It would be preferable if several data sets could be interpreted at one time.

One of many possible solutions is to use a series of heatmaps showing the average

error of each model with it's respective parameters. But as discussed above, the average error does not tell the whole story, the number of large errors must also be accounted. Figure 17 shows five heatmaps, one for each size of training set. The color in each square represents the average error for a specific model. The scale for each subplot is set to be identical to make its analysis easier. A similar cluster of plots can be seen in Figure 18, showing the number of large errors for each model.

The true accuracy and precision of each model can only be realized when comparing its results from Figures 17 *and* 18. One without the other does not show the full picture. For example, the lower left corner of Figure 17a appears to be surprisingly accurate, but when compared with the corresponding squares in Figure 18a, it can be seen that these models actually produce a considerable number of large errors. These particular models are not of interest. The models that *are* of interest will be the squares from corresponding subplots in Figures 17 and 18 that are both blue or blue-ish.

The most successful model was trained on 1000 configurations, used 40 basis functions, and had an effective radius of 1.0. This model produced 0 large errors and had an average error of only 1.77eV. The second and third best models had very similar input parameters, a one-step change in the size of the training set or the number of basis functions.

0.6 Conclusion

The models produced were satisfactorily successful. Each provided great insight into which parameters build useful models. Though some models are useful, none are without errors. Using the "best" model, as mentioned in Section 0.5, new data on silver-platinum crystal configurations can be produced.

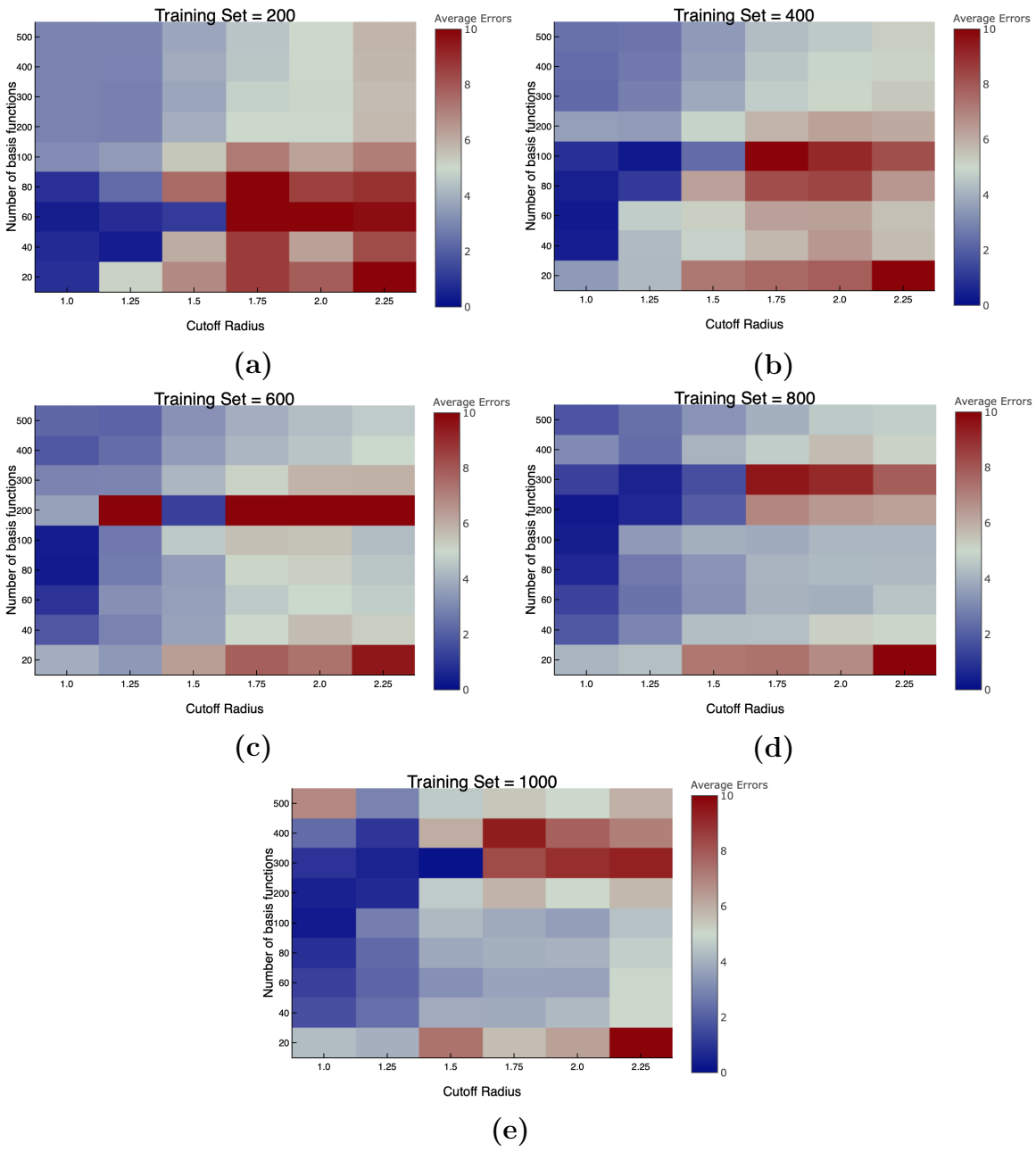


Figure 17 Heatmaps showing the average error of each model produced. The scale for each is fixed from 0eV to 10eV. The effect from removing the large errors from the average can be seen by comparing the more red areas from Figure 18 with the dark blue areas here.

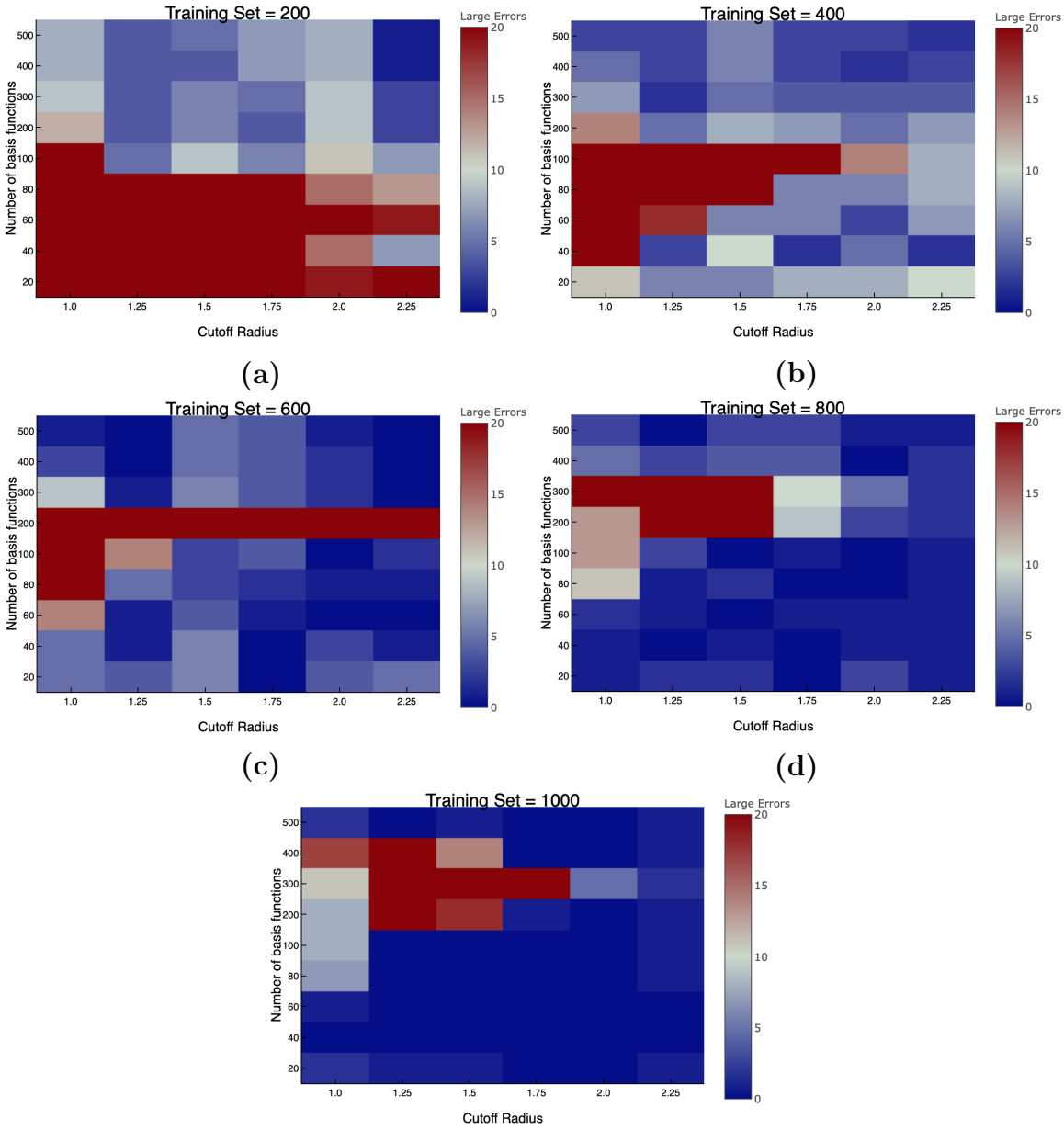


Figure 18 Heatmaps showing the number of large errors of each model produced. The scale for each is fixed from 0 to 20 errors. Because the scale is fixed, any square with a deep red color has a minimum of 20 errors.

0.6.1 Future Work

Using a small set of accurate DFT data, the code written to produce the model in Section 0.4.2 can be easily adapted to construct models for compounds other than Ag-Pt. Such research could be used to determine which compounds can be accurately predicted using a simple pair-interaction model. A successful model can quickly predict formation energies for a variety of crystal configurations. These formation energies could then be used to construct the material's phase diagram and determine which phases are thermodynamically stable.

Because the models produced in this paper used only pair-interactions, their effectiveness is limited by their simplicity. Implementation of three-body interactions has the potential to greatly increase the accuracy of these models.

Citations yet to be included: [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16]

Bibliography

[1] J. E. Jones, "On the determination of molecular fields.—I. From the variation of the viscosity of a gas with temperature," Proceedings of the Royal Society A **106,** 441–462 (1924).

[2] H. T. Stokes, *Solid State Physics for Advanced Undergraduates*, 4th ed. (Brigham Young University, 2007).

[3] S. Curtarolo, G. L. W. Hart, W. Setyawan, R. Chepulskii, O. Levi, and D. Morgan, "AFLOW:software for high throughput calculation of materials properties,".

[4] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects," Phys. Rev. **140,** A1133–A1138 (1965).

[5] D. Morgan, G. Ceder, and S. Curtarolo, "High-throughput and data mining with ab initio methods," Measurement Science and Technology **16,** 296 (2005).

[6] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, "Predicting crystal structures with data mining of quantum calculations," Phys. Rev. Lett. **91,** 135503 (2003).

[7] G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," Phys. Rev. B **59,** 1758 (1999).

[8] G. Kresse and J. Hafner, "Ab initio molecular dynamics for liquid metals," Phys. Rev. B **47,** 558 (1993).

[9] P. E. Blöchl, "Projector augmented-wave method," Phys. Rev. B **50,** 17953 (1994).

[10] G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," **6,** 15–50 (1996).

[11] H. Monkhorst and J. Pack, "Special points for Brillouin-zone integrations," Phys. Rev. B **13,** 5188–5192 (1976).

[12] J. Sanchez, F. Ducastelle, and D. Gratias, "Generalized cluster description of multicomponent systems," Physica A: Statistical and Theoretical Physics **128,** 334–350 (1984).

[13] D. Laks, L. Ferreira, S. Froyen, and A. Zunger, "Efficient cluster expansion for substitutional systems," Phys. Rev. B **46,** 12587 (1992).

[14] D. Lerch, O. Wieckhorst, G. L. W. Hart, R. W. Forcade, and S. Müller, "UNCLE: a code for constructing cluster expansions for arbitrary lattices with minimal user-input," Modelling and Simulation in Materials Science and Engineering **17,** 055003 (2009).

[15] E. Cockayne and A. Van De Walle, "Building effective models from sparse but precise data: Application to an alloy cluster expansion model," Phys. Rev. B **81,** 012104 (2010).

[16] L. J. Nelson, G. L. Hart, F. Zhou, and V. Ozoliņš, "Compressive sensing as a paradigm for building physics models," Physical Review B **87,** 035125 (2013).