Statistical learning for alloy design from electronic structure calculations

Scott Broderick
Iowa State University
Statistical learning for alloy design from electronic structure calculations

by

Scott R. Broderick

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Materials Science and Engineering

Program of Study Committee:
Krishna Rajan, Major Professor
Vasant Honavar
Michael Kessler
Richard LeSar
Sriram Sundrarajan

Iowa State University
Ames, Iowa
2009

Copyright © Scott R. Broderick, 2009. All rights reserved.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS v
ABSTRACT vi

CHAPTER 1: INTRODUCTION 1
  1.1 Objectives 1
  1.2 Data Mining 3
  1.3 Novelty of Work 4
  1.4 Description of Density of States 7
  1.5 Thesis Outline 10

CHAPTER 2: ELECTRONIC STRUCTURE PREDICTION VIA DATA MINING 13
  THE EMPIRICAL PSEUDOPOTENTIAL METHOD
    2.1 Introduction 13
    2.2 Classification of Form Factors 15
    2.3 Prediction of Electronic Structure 20
    2.4 Summary 26

CHAPTER 3: PARAMETER REDUCTION THROUGH STATISTICAL 27
  LEARNING BASED SCREENING
    3.1 Introduction 27
    3.2 Data Mining of Semi-Empirical Calculations 29
    3.3 Results and Parameter Reduction 31
    3.4 Summary 35

CHAPTER 4: CLASSIFICATION OF THE DENSITY OF STATES SPECTRA 36
  4.1 Introduction 36
  4.2 Inverse Problem Approach 38
  4.3 Analysis of Intermetallic Systems 41
  4.4 Analysis of Oxide Systems 46
  4.5 Summary 52

CHAPTER 5: APPROACHES FOR MODELING DENSITY OF STATES SPECTRA 53
  5.1 Introduction 53
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1 Mathematics of PCA</td>
<td>99</td>
</tr>
<tr>
<td>A.2 Mathematics of PLS</td>
<td>102</td>
</tr>
<tr>
<td>A.3 Schematic of Modeling the Density of States</td>
<td>104</td>
</tr>
<tr>
<td>A.4 Mathematics of Haydock’s Recursion Method</td>
<td>108</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>110</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my thesis advisor, Professor Krishna Rajan. He has provided me with numerous opportunities beyond what is normal for a graduate student, and his guidance has made this experience very productive. I would also like to thank my thesis committee members – Professors Vasant Honavar, Michael Kessler, Richard LeSar and Sriram Sundrarajan – for their insights.

I have been fortunate to have collaborated with numerous excellent scientists. I would like to specifically mention Professor Shuichi Iwata and his laboratory members at the University of Tokyo for their hospitality during my visits and for their insights and ideas. I would also like to thank Professor Hafid Aourag at the University Abou Bekr Belkaid in Algeria for providing me numerous ideas and suggestions during the course of this work. He has been instrumental in the development of the work discussed in this thesis.

I would also like to acknowledge my fellow group members for the numerous discussions we have had over the years. In particular, I would like to thank Dr. Changwon Suh for his advice and teachings during my beginning years in graduate school.
ABSTRACT

The objective of this thesis is to explore how statistical learning methods can contribute to the interpretation and efficacy of electronic structure calculations. This study develops new applications of statistical learning and data mining methods to both semi-empirical and density functional theory (DFT) calculations. Each of these classes of electronic structure calculations serves as templates for different data driven discovery strategies for materials science applications. In our study of semi-empirical methods, we take advantage of the ability of data mining methods to quantitatively assess high dimensional parameterization schemes. The impact of this work includes the development of accelerated computational schemes for developing reduced order models. Another application is the use of these informatics based techniques to serve as a means for estimating parameters when data for such calculations are not available.

Using density of states (DOS) spectra derived from DFT calculations we have demonstrated the classification power of singular value decomposition methods to accurately develop structural and stoichiometric classifications of compounds. Building on this work we have extended this analytical strategy to apply the predictive capacity of informatics methods to develop a new and far more robust modeling approach for DOS spectra, addressing an issue that has gone relatively unchallenged over two decades. By exploring a diverse array of materials systems (metals, ceramics, different crystal structures) this work has laid the foundations for expanding the linkages between statistical learning and statistical thermodynamics. The results of this work provide exciting new opportunities in computational based design of materials that have not been explored before.
CHAPTER 1: INTRODUCTION

The design of new materials has traditionally been based on fortuitous discovery. While electronic structure calculations have begun to address this problem by providing a physical basis on which to propose new materials for exploration, the complexity of the calculations and the amount of data created has prevented the full power of these calculations from being utilized. This thesis explores the feasibility of using data mining tools for the “soft modeling” of the density of states and identifying the relationship between electronic structure, crystal structure, and material properties. Additionally, the use of data mining for the prediction and minimization of semi-empirical parameters is explored.

1.1 Objectives

The computational branch of materials science has become increasingly important with the improved power of computers, making the calculations of complex quantum properties possible. The work discussed in this thesis combines two important areas within computational materials science: first principles calculations and materials informatics. Materials informatics is an emerging field analogous to bio-informatics and drug discovery, addressing issues in data driven discovery in materials science [1, 2]. The work discussed here as well as other work combining first principles calculations with materials informatics are laying the foundation for the informatics toolbox to become integral within computational materials science.

Even with advances in high speed computing, there are limitations to the number of new alloy chemistries that can be calculated. It is for this reason that combining data mining, an area focused on the extraction of knowledge and insight from massive databases, with first principles calculations is important for recognizing attributes of an alloy that contribute to interesting properties and guiding the selection of future calculations. Combining data mining techniques with electronic structure calculations presents numerous challenges. Foremost among these challenges is how to integrate the large amounts of data and to extract important information. This thesis addresses this challenge by developing a methodology to
allow for the improved analysis of complex and large data sets, utilizing advanced computational materials science and data mining tools.

The work discussed in this thesis had two parts. First, the application of a data mining approach to semi-empirical calculations is discussed. The statistical learning component of this step is the quantitative assessment of parameterization, while the physics impact is the development of reduced order models and the estimation of parameters that are presently unavailable. The second part applies data mining to density functional theory (DFT) calculations, and specifically the density of states (DOS). Data mining is used to classify DOS spectra to demonstrate the ability of statistical learning classification methods to detect structural and stoichiometric classifications of compounds. I then examine the role of statistical learning prediction methods on DFT calculations, with the physics impact being a new and far more accurate “soft modeling” method for the DOS than currently exists. This modeling is demonstrated for different bonding and crystal structure problems, while addressing a long standing problem that has effectively been idle for the last twenty years.

Modeling the DOS without direct consideration of Schrödinger’s or Dirac’s equation has been considered over the past half century, and thus numerous other approaches to model the DOS have been suggested. Of these, the most remarkable is the recursion method of Haydock [3-8]. Even with the success of the recursion method, limitations exist within the method. By integrating data mining techniques with the recursion method logic, I address the limitations of the recursion method and “soft model” the DOS with high accuracy and limited computational requirements.

As a by-product of these objectives, this thesis better defines the linkage between the electronic and crystal structure of a material by being able to accurately classify crystal structures by inputting discrete but known quantum mechanical descriptors. I then investigate whether the DOS can be used to represent the quantum mechanical descriptors missing from the structure maps of Pettifor and others [9-16], while still classifying crystal structures. A traditional structure map has axes of different descriptors or properties, while here descriptors are derived directly from a complete description of the electronic structure. This methodology is significant because it relates the various length scales in a comprehensive manner.
1.2 Data Mining

This thesis employs primarily two data mining tools: principal component analysis (PCA) and partial least squares (PLS). Other data mining methods have been considered, and future work will continue to look at applying more data mining methods to this work.

PCA [17-20] is a projection technique for handling multivariate data that consists of interrelated variables. It inherently decomposes the covariance (or correlation) matrix by calculating the eigenvalues and eigenvectors of the matrix. This decomposition leads to the reduction of information dimensionality in a way that minimizes the loss of information and maximizes the variance of the linear combination of the variables and uncorrelated axes leading to the transformation (i.e. rotation) of the original coordinate system. The constructed axes, referred to as principal components (PCs) correspond with the eigenvectors of the original data’s covariance matrix and are orthogonal to each other. They consist of loadings, which are the weights for each original variable and scores containing information of original samples in a rotated coordinate system. Although the number of PCs equals the number of dimensions of the original data, a few PCs are usually sufficient to capture the majority of the system’s information. PCA is a powerful tool for understanding the underlying physics within materials science problems and has been used to address materials science issues for a variety of reasons and materials [21-25].

PLS [26-31] is used to make the predictions in this work and has the advantage over typical linear regression techniques of handling collinearity among properties. Like PCA, in PLS the data is converted to a data matrix with orthogonalized vectors. The relationships discovered in the dataset (training data) can then be applied to a test dataset based on the differences in known properties appearing in both the training and test sets. The accuracy of the prediction model improves with increasing number of conditions and responses, and thus all predictions shown in this paper can improve with a larger dataset including more systems and more properties/parameters. The necessary considerations and subtleties are discussed as appropriate with the predicted results.

The mathematics of both PCA and PLS are provided in the appendix, with further descriptions and interpretations of PCA and PLS provided throughout the thesis.
1.3 Novelty of Work

The application of statistical learning to electronic structure calculations is a new concept that has only been addressed a few times previously (Table 1.1). While data mining applied to properties calculated through electronic structure calculations, such as elastic constants or bulk modulus, have been attempted [32-34], they are very much different in objective than the work here. The most notable work combining data mining and electronic structure calculations include the work of Ceder [35, 36] and Zunger [37, 38]. In both cases, as in the case of Sluiter [39], Johannesson [40], and Mohn [41], the objective was to determine a stable phase that a material forms in, based on the relationship between the conditions and the energy of the system.

Another notable work is that of Alsberg [42, 43], which is unique in the literature by employing data mining on electronic structure descriptions. In that work, he analyzed the charge density of polymers via data mining tools trying to capture some trend connecting the chemistry with the electronic structure.

The concept of data mining the entire DOS curve is unique, as is the prediction of semi-empirical parameters via data mining. The objective of “soft modeling” the DOS is not a unique problem, but the addition of data mining tools to the logic is unique.

In contrast to the works of other groups, the questions being asked in this thesis are quite different. The first issue is examining if the semi-empirical parameters can be predicted for systems where the parameters are not currently available, thus allowing for the calculation of new systems. The data input into this analysis is the known parameters for some systems, and basic descriptions from the periodic table. The second issue is to see if semi-empirical parameters or descriptors from semi-empirical calculations can be reduced to a minimum number thus reducing computational requirements. The data for this analysis is different descriptors calculated through a semi-empirical calculation. The third problem is to classify the entire DOS spectra and to search for patterns within the DOS curve related to crystal structure or crystal chemistry, with the only input being different DOS curves. The final problem is to model the DOS using data mining, with DOS curves being the data considered with the addition of a database component. Therefore, as can be seen in
comparison to Table 1.1, the questions asked are very different and the data considered is also quite different.

**Table 1.1.** Examples of other groups employing statistical learning with electronic structure calculations, with the problem they are addressing and how they address the problem with data mining.

<table>
<thead>
<tr>
<th>Groups using statistical learning with DFT</th>
<th>What question are they addressing?</th>
<th>How is data mining applied to the problem?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceder <em>Nature Materials</em> 2006</td>
<td>What is the most stable structure for a given chemistry?</td>
<td>Using PCA/PLS, find a relationship between chemistry and energy so that the structure with lowest energy can be predicted.</td>
</tr>
<tr>
<td>Zunger <em>Physical Review Letters</em> 2006</td>
<td>What dictates the evolution of the building blocks of a system giving some property?</td>
<td>A genetic algorithm is used to find lowest energy crystal structure, with evolutionary search being performed on building block units.</td>
</tr>
<tr>
<td>Johannesson <em>Physical Review Letters</em> 2002</td>
<td>How to optimize material by screening DFT calculations for specific property?</td>
<td>Genetic algorithm is used to find most stable chemistries having some desired property.</td>
</tr>
<tr>
<td>Mohn <em>Computational Materials Science</em> 2009</td>
<td>How to find a disordered material with optimized property?</td>
<td>Genetic algorithm using a symmetry adapted crossover for finding stable phase of disordered systems and surfaces.</td>
</tr>
<tr>
<td>Alsberg <em>Computer Physics Communications</em> 2005</td>
<td>What bonding information from the charge density topology impacts the material properties?</td>
<td>PCA/PLS is used to analyze the trends within the charge density critical points.</td>
</tr>
</tbody>
</table>

The studies listed in Table 1.1 all employed data calculated through electronic structure calculations, but they all, with the exception of Alsberg, were not studying the electronic structure of the material, but rather were using electronic structure calculations to assess the material at an atomic level. While their approaches differed and the specific
questions asked varied, they all were essentially classification problems searching for some specific characteristic, namely stability.

These studies used only the energies or enthalpies calculated for each system. The data mining problem then becomes identifying which system has the lowest energy for a given configuration. The data in these cases is low-dimensional and data mining is used to search through the data to recognize the lowest energy. Relationships between most stable state and chemistry can then be expanded so that the most stable atomic configuration can be predicted. The exception to this discussion on input data is Alsberg, where he analyzes a charge density topology. However, in his work he still needed to make assumptions as to what in the charge density is most important, and his interpretation of the results did not provide a major linkage between the chemistry and the electronic structure.

Thus the characteristic of the data used in this thesis is quite different from the studies in Table 1.1. While they used limited values from each electronic structure calculation to assess the atomic configurations, I here use data of a very different nature. The data considered is either specific to a calculation approach and is analyzed to allow for the prediction of new systems, or else the data analyzed is a high dimensional output representing electronic structure which is analyzed without any prior assumptions.

Finally, a comment on the ability of these methods for analyzing the physics in terms of the math is necessary. The purpose of the genetic algorithm in these prior works was to assess the evolution of a material in terms of its most stable state, so that in the same manner some new chemistry can evolve following the same rules so that the stable structure may be predicted. Thus, the mathematics of the genetic algorithm captures the underlying physics involved in determining the ground state crystal structure of a material. A genetic algorithm is strong for finding some relationship within low-dimensional data so that the mathematics of the approach captures the physics of the system. The purpose for other groups in using PCA and PLS to find the most stable structure has a similar logic but different connection between the mathematics and the physics. The logic in these cases is to create a classification of materials based on their stable structures, and to then build a model to link the characteristics of the material to the lowest energy state, thus creating a mathematical
prediction of the physics. The connection between the mathematics and the physics of my approach are discussed throughout this thesis.

In this section I have demonstrated that this thesis provides the first significant example of using statistical learning to assess some component of an electronic structure calculation which then allows for the calculation to be performed more efficiently and for a more far ranging set of problems. Also, this work shows a first example of using a data mining methodology on an electronic structure output with no assumptions for more rapid calculation of the electronic structure. Of particular note is that the characteristics of the data are quite different from other studies and that the manner of interpreting the physics of the problem via mathematics is different.

1.4 Description of Density of States

The DOS fully describes the bonding within a material and also theoretically contains information on all electronic properties, as it represents the electronic structure [44-48]. The DOS in this case displays the probability that a state may be found at a certain energy, with a state being able to be occupied by two electrons. Due to the nature of quantum mechanics, there is a probability associated with an electron existing at a certain energy, and thus a spectral curve exists with numerous complex features.

Since the DOS reflects the probability a state exists at a certain energy, it provides a representation of the electronic structure. A change in the DOS indicates a change in the states available for an electron, and therefore this change relates to a difference in electronic structure and presumably electronic properties. The DOS curve captures a statistical distribution of the electronic states; however, the challenge is in determining the factors and their interactions representing this statistical distribution. These interactions result in a DOS curve of high complexity, thus making the interpretation difficult. Figure 1.1 shows the development of a DOS curve, from an initial system of one state to the DOS for a real material capturing these complex interactions.
Figure 1.1. The DOS curve captures a statistical distribution of the electronic states. This statistical distribution results from numerous complex factors and their interactions, so that a comprehensive interpretation of the DOS curve is challenging. This figure shows the development of a DOS curve, from a single state to a DOS curve capturing numerous interactions.

The first principles calculation approach used in this thesis is density functional theory (DFT) [49], which is based on the discovery that a relationship exists between a systems potential and the electronic density. As stated, the DOS is based on the interactions between various complex factors, some of which are shown in Figure 1.2. The power of DFT is that it finds the relationships between all of these factors and is able to calculate the electronic structure. An objective of this thesis is to use statistical learning to try to identify the statistical relationships that DFT utilizes and more efficiently model the DOS.
**Figure 1.2.** DFT requires the input of crystal structure and chemistry. The calculation then uses quantum mechanical approximations to convert these inputs into outputs such as the DOS. Thus, theoretically the DOS curve contains information on crystal structure, chemistry, quantum mechanics, and electronic structure descriptions. The power of DFT is that it is able to capture the interactions between these complex factors and to calculate the electronic structure.

An example of a property represented in the DOS is the bulk modulus, demonstrated in Figure 1.3, as first suggested in reference [50] and further developed here. The Fermi energy ($E_F$) indicates the maximum occupancy by electrons at ground state conditions. Therefore, the DOS curve at energies greater than $E_F$ represents the available states that are unoccupied by electrons. An additional feature within the DOS is the transition between bonding and anti-bonding states. Occupancy of a bonding state corresponds with an increase in strength, while additional occupancy of an anti-bonding state results in a decrease in strength. The transition of bonding to anti-bonding states is generally represented by a valley in the curve. The bulk modulus is shown to increase with number of valence electrons for transition metals, until a maximum value is reached for Fe, and then additional valence electrons result in a decrease in bulk modulus. The bulk modulus can then be found related
with the distance between the bonding – anti-bonding transition and $E_F$. The extraction of relationships such as these is discussed further in this thesis.

![Graph showing bonding-anti-bonding transition](image)

**Figure 1.3.** Demonstration of how bulk modulus can be extracted from the DOS. By identifying the distance between features in the DOS correlating with properties suggests that a methodology for analyzing the DOS can lead to the further extraction of such relationships.

The horizontal axis of the DOS curve, such as that in Figure 1.1, is energy, and the vertical axis is DOS or the probability that a state exists at the particular energy. Data mining is employed in this work to see how changing crystal structure, chemistry and properties are reflected in the DOS. As the DOS represents the electronic structure, analyzing changes in DOS due to crystal structure and chemistry leads to the development of improved electronic structure-crystal structure-chemistry relationships.

**1.5 Thesis Outline**

This thesis is organized as shown in Figure 1.4, addressing applications of data mining for semi-empirical calculations and for DFT calculations.
Figure 1.4. The logic of this thesis, with chapters 2 and 3 dealing with the application of data mining for semi-empirical calculations and chapters 4 through 7 addressing DFT calculations with data mining.

In chapter 2, data mining is used to estimate parameters within a semi-empirical approach that are either presently unavailable or are difficult to calculate, with the test bed being III-V Bi semi-conductors. In chapter 3, I demonstrate how data mining can be used to reduce the number of parameters needed from semi-empirical calculations by identifying redundant data from a semi-empirical model for Ti-Al alloys. The logic of this chapter is applicable to any data type.

In chapter 4, I classify DOS spectra from DFT calculations to demonstrate the ability of statistical learning classification methods (e.g. PCA) to detect structural and stoichiometric classifications of compounds. This chapter thus provides a linkage between crystal structure and electronic structure, with a test bed of both intermetallics and oxides, demonstrating that this approach is applicable to different systems.

Chapter 5 discusses the work of Haydock in modeling the DOS, with the logic compared with the approach proposed in this thesis. Chapter 6 then describes in greater detail my approach for modeling the DOS, while chapter 7 shows the results of my method, with results shown for different bonding and crystal structure problems.
Chapter 8 summarizes the work and makes suggestions as to the future direction of this work and the implications it has for alloy design. The appendix then contains some background information, including the mathematics of the data mining methodologies used--principal component analysis (PCA) for classification and partial least squares (PLS) for prediction--and the recursion approach of Haydock. Additionally, a numerical demonstration of the approach for modeling the DOS utilizing data mining is provided.
CHAPTER 2: ELECTRONIC STRUCTURE PREDICTION VIA DATA MINING
THE EMPIRICAL PSEUDOPOTENTIAL METHOD

This chapter serves to demonstrate an application of data mining to aid in expediting electronic structure calculations. In this example, data mining is used to predict form factors for the empirical pseudopotential method so that the electronic structure for new systems can be modeled. Combining data mining with the empirical pseudopotential method allows one to convert an empirical approach to a predictive approach. Here tetrahedrally bounded III-V Bi semiconductors are considered, and through the prediction of form factors based on basic elemental properties the band structure and charge density for these semi-conductors, for which limited results exist, is modeled. This work represents a unique approach to modeling the electronic structure of a material which may be used to identify new promising semi-conductors.

2.1. Introduction

The application of the empirical pseudopotential method (EPM) to semiconductors and insulators has been extremely fruitful. The fitting of the pseudopotential to experimental data has yielded a great deal of insight and new information about the electronic structure of these materials [51-55]. Among the enormous number of experiments that have been performed to probe the electronic structure of semiconductors and insulators, the bulk of the work has involved the exploration of the infrared and low-energy side of the visible spectrum to gain information about the fundamental band gap. Hence the measurements provide the main source of information for the adjustment of pseudopotential form factors.

The local part of the pseudopotential $V_L$ is expressed in terms of the symmetric and antisymmetric pseudopotential form factors. Once the pseudopotential form factors are chosen, the Hamiltonian

$$H \psi_{nk}(r) = E_{nk} \psi_{nk}(r)$$

(2.1)

with $H = \frac{p^2}{2m} + V_L$

(2.2)
can be solved for the energy eigenvalues and wavefunctions (band n, state k) at many points within the Brillouin zone. If the origin is taken halfway between atom A and B, $V_L$ is given as:

$$V_L = V_s(G)\cos(G.t) + iV_a(G)\sin(G.t) \quad (2.3)$$

$G$ is a reciprocal lattice vector and $t$ is a bravais lattice vector. The symmetric and antisymmetric form factors are given by:

$$V_s(G) = \frac{1}{2}[V_A(G) + V_B(G)] \quad (2.4)$$

$$V_a(G) = \frac{1}{2}[V_A(G) - V_B(G)]$$

For FCC structures, the relevant coefficients are $V_s(3)$, $V_s(8)$, $V_s(11)$, $V_a(4)$, $V_a(8)$ and $V_a(11)$, where the number in parenthesis is the corresponding value of $|G|^2$. These form factors are adjusted to the optical data until the calculated band structure and the dielectric function are in good agreement with experiments. Thus the construction of the pseudopotential relies mainly on the correctness of the experimental optical data, meaning this approach is purely empirical and not predictive, whereas, the \textit{ab initio} pseudopotential techniques predict material properties from the fundamental equations of quantum mechanics. These methods allow scientists to evaluate and pre-screen new materials “in silico” rather than through time-consuming experimentation, and in some cases even make suggestions for new and better materials [56, 57].

A new approach is proposed here that allows for converting the EPM from an empirical method to a predictive method. By analyzing the EPM results already collected on known systems, the trends and relationships between the calculation parameters and basic atomic and structural properties can be identified. These relationships will then allow for predicting the aforementioned form factors, which in turn allow one to perform EPM calculations for new systems and accelerate the electronic design of materials. In this work III-V Bi semiconductors are considered because they provide an interesting new system that
has not been extensively explored previously, while still allowing for validating the results based on the precious few calculations reported.

2.2. Classification of Form Factors

The role of this section is to better understand the relationship of the form factors with structural properties. These results will allow for better understanding the meaning of the form factors and the relationship between semi-conductor chemistry and the individual form factors. As discussed, from the form factors one can calculate electronic structure through additional calculations, namely the EPM. Data mining is used to explore if a simple linear relationship based on the various form factors is sufficient for predicting basic properties, such as lattice constant and bulk modulus. Beyond the value in calculating properties of new semi-conductors, the ability to predict these properties would suggest that form factors can be modeled, as this is simply the inverse of the prediction.

Therefore, this section provides a better physical basis for describing the form factors, uncovers trends between semi-conductor chemistry and the form factors, predicts basic semi-conductor properties, and determines the likelihood of being able to model the form factors.

Table 2.1 contains a portion of the dataset used for the initial analysis, where the dataset is comprised of 15 samples and 14 properties/parameters. The descriptors and their abbreviations are as follows: Symmetric and antisymmetric form factors (Vs and Va respectively), lattice parameter (a), bulk modulus (B), ionic character (F(i)), electronegativity (X), interatomic distance (L_{ab}), band gap (E_g), homopolar band gap (E_h), and heteropolar band gap (C).
Table 2.1. A part of the training dataset. The quantities given in the table are as follows: \(a\) is the lattice parameter in å, \(B\) is the bulk modulus in Kbar, \(V_s\) and \(V_a\) are the symmetric and antisymmetric form factors, respectively, \(L(ab)\) is the interatomic distance in å, \(f(i)\) is the ionicity factor and \(X\) is the electronegativity.

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(B)</th>
<th>(f(i))</th>
<th>(V_s) (3)</th>
<th>(V_s) (8)</th>
<th>(V_a) (4)</th>
<th>(V_a) (8)</th>
<th>(V_a) (11)</th>
<th>(L(ab))</th>
<th>(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlSb</td>
<td>11.5857</td>
<td>61</td>
<td>0.25</td>
<td>-0.20198</td>
<td>0.02</td>
<td>0.065836</td>
<td>0.039658</td>
<td>0.04</td>
<td>-0.009021</td>
<td>2.62</td>
</tr>
<tr>
<td>AlAs</td>
<td>10.699</td>
<td>66</td>
<td>0.274</td>
<td>-0.236869</td>
<td>0.02542</td>
<td>0.070999</td>
<td>0.061632</td>
<td>0.0625</td>
<td>-0.0075</td>
<td>2.44</td>
</tr>
<tr>
<td>AlP</td>
<td>10.3009</td>
<td>79</td>
<td>0.307</td>
<td>-0.21</td>
<td>0.04</td>
<td>0.08</td>
<td>0.13</td>
<td>0.08</td>
<td>0.015</td>
<td>2.36</td>
</tr>
<tr>
<td>GaSb</td>
<td>11.5668</td>
<td>59</td>
<td>0.261</td>
<td>-0.198866</td>
<td>0.005</td>
<td>0.048558</td>
<td>0.052258</td>
<td>0.03</td>
<td>0</td>
<td>2.62</td>
</tr>
<tr>
<td>GaAs</td>
<td>10.68852</td>
<td>65</td>
<td>0.31</td>
<td>-0.240326</td>
<td>0.0126</td>
<td>0.060392</td>
<td>0.067306</td>
<td>0.05</td>
<td>0.01</td>
<td>2.44</td>
</tr>
<tr>
<td>GaP</td>
<td>10.30051</td>
<td>74</td>
<td>0.327</td>
<td>-0.228235</td>
<td>0.03</td>
<td>0.076489</td>
<td>0.107601</td>
<td>0.07</td>
<td>0.02</td>
<td>2.36</td>
</tr>
<tr>
<td>InSb</td>
<td>12.23525</td>
<td>55</td>
<td>0.321</td>
<td>-0.2</td>
<td>-0.01</td>
<td>0.044</td>
<td>0.044</td>
<td>0.03</td>
<td>0.015</td>
<td>2.8</td>
</tr>
<tr>
<td>InAs</td>
<td>11.4156</td>
<td>61</td>
<td>0.357</td>
<td>-0.2313</td>
<td>0</td>
<td>0.055937</td>
<td>0.057972</td>
<td>0.05</td>
<td>0.03</td>
<td>2.62</td>
</tr>
<tr>
<td>InP</td>
<td>11.07541</td>
<td>63</td>
<td>0.421</td>
<td>-0.235</td>
<td>0</td>
<td>0.053</td>
<td>0.08</td>
<td>0.06</td>
<td>0.03</td>
<td>2.54</td>
</tr>
<tr>
<td>ZnS</td>
<td>10.2249</td>
<td>61</td>
<td>0.623</td>
<td>-0.257891</td>
<td>0.03</td>
<td>0.074164</td>
<td>0.207435</td>
<td>0.14</td>
<td>0.04</td>
<td>2.35</td>
</tr>
<tr>
<td>ZnSe</td>
<td>10.67851</td>
<td>58</td>
<td>0.63</td>
<td>-0.218</td>
<td>0.029</td>
<td>0.064</td>
<td>0.139</td>
<td>0.062</td>
<td>0.016</td>
<td>2.45</td>
</tr>
<tr>
<td>ZnTe</td>
<td>11.4723</td>
<td>54</td>
<td>0.609</td>
<td>-0.22</td>
<td>0</td>
<td>0.05</td>
<td>0.13</td>
<td>0.1</td>
<td>0.01</td>
<td>2.63</td>
</tr>
<tr>
<td>CdS</td>
<td>11.021</td>
<td>66</td>
<td>0.685</td>
<td>-0.24</td>
<td>0.03</td>
<td>0.04</td>
<td>0.23</td>
<td>0.13</td>
<td>0.05</td>
<td>2.52</td>
</tr>
<tr>
<td>CdSe</td>
<td>11.4324</td>
<td>60</td>
<td>0.699</td>
<td>-0.23</td>
<td>0.01</td>
<td>0.01</td>
<td>0.19</td>
<td>0.12</td>
<td>0.05</td>
<td>2.62</td>
</tr>
<tr>
<td>CdTe</td>
<td>12.11491</td>
<td>56</td>
<td>0.717</td>
<td>-0.2</td>
<td>0</td>
<td>0.04</td>
<td>0.15</td>
<td>0.09</td>
<td>0.04</td>
<td>2.8</td>
</tr>
</tbody>
</table>

A PCA analysis of the 15 different semi-conductors was performed. The scores plot is displayed in Figure 2.1 with Figure 2.2 being the loadings plot displaying the relationships between the 14 parameters. As labeled in Figure 2.1, a clear trend is evident between the semi-conductor chemistries, with the anion size decreasing mostly along the horizontal axis, while the cation size decreases primarily with increasing vertical axis. The percentage label of each axis is the amount of variance in the data set captured by the axis, which is a linear combination of the 14 parameters. As PC1 and PC2 capture similar amounts of information and neither labeled trend is solely along one axis, it is found that the impact of changing the anion as opposed to changing the cation is approximately the same and that one cannot primarily focus on only one chemical component of the semi-conductors.
Figure 2.1. PCA scores plot from a dataset of Table 2.1 containing the 15 semiconductors. From examination of the trends in semi-conductor chemistries within PC space, a clear relationship exists between the input parameters (including form factors) and the chemistry of the cation and anion.

Figure 2.2. PCA loadings values for the dataset containing the 15 semiconductors.
Of greater significance for the objectives here is relating the trends in samples with the form factors included in the analysis. The relationship of all parameters is clearly shown in the loadings plot. This plot combines form factors with properties, which appear to form roughly three clusters identified by eye, as labeled on the figure. Cluster 1 contains Vs(8), Vs(11), B and Eh, cluster 2 contains E_g, Va(4), Va(8), Va(11), C, X and F(i) and cluster 3 contains Vs(3), a and L_ab. Properties with similar PC values, as shown by defined clusters, are highly correlated, while inverse PC values indicate inverse correlations. Clusters 1 and 3 contain the symmetric form factors, while cluster 2 contains the anti-symmetric form factors.

As the points contained within the same clusters are correlated and clusters 1 and 3 are inversely correlated, one can identify several relationships. An increase in the lattice constant will lead to an increase of Vs(3) and a decrease in Vs(8) and Vs(11). The homopolar gap E_h is related with the symmetric form factors, which matches with the fact that E_h corresponds with the real portion of the Hamiltonian in the pseudopotential theory. Meanwhile, ionicity factor f(i), electronegativity X, and the heteropolar gap C are correlated and tightly bounded to the antisymmetric form factors.

Based on the trends labeled in the scores plot, one can estimate the impact that changing chemistry will have on the discussed descriptors. Increasing either the cation or the anion size will result in an increase of the parameters in cluster 3 and a decrease in the values of cluster 1. The relationship between cation and anion size with cluster 2 is less definitive. The three hypothetical semi-conductors which I will focus on in this paper are AlBi, GaBi and InBi. As all three have larger anions than AlSb, GaSb and InSb, respectively, it can be surmised that their PC value would be more negative PC1 and PC2, and thus have higher values of parameters in cluster 3 than AlSb, GaSb and InSb and lower values of parameters in cluster 1. Therefore increasing the anion size will increase L_ab and decrease B.

As was shown in Figure 2.1, a trend with cation and anion chemistry is evident within the semi-conductor chemistries when plotted on an axis system which is based in part on the form factors of the systems. Therefore, it is anticipated that combining the data of Table 2.1 with a description of the semi-conductor chemistry will be sufficient for the prediction of new form factors.
A fundamental property of any element in the periodic table is the covalent radius, as listed in Table 2.2. Combining these covalent radius values with the values used to create the PCA model (Table 2.1), a model to predict the form factors, as well as the lattice parameter and bulk modulus, is developed with the model tested on six additional semiconductors (AlN, AlBi, GaBi, GaN, InBi and InN). The predicted results for the lattice parameter and the bulk modulus are displayed in Figures 2.3 and 2.4, respectively. Reasonable agreement between the observed and the predicted values is seen, particularly for those semiconductor systems not included in the original dataset.

Table 2.2. The covalent radius in Å

<table>
<thead>
<tr>
<th>Element</th>
<th>Covalent Radius (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.82</td>
</tr>
<tr>
<td>N</td>
<td>0.75</td>
</tr>
<tr>
<td>Al</td>
<td>1.18</td>
</tr>
<tr>
<td>P</td>
<td>1.06</td>
</tr>
<tr>
<td>S</td>
<td>1.02</td>
</tr>
<tr>
<td>Zn</td>
<td>1.25</td>
</tr>
<tr>
<td>Ga</td>
<td>1.26</td>
</tr>
<tr>
<td>As</td>
<td>1.2</td>
</tr>
<tr>
<td>Se</td>
<td>1.16</td>
</tr>
<tr>
<td>Cd</td>
<td>1.48</td>
</tr>
<tr>
<td>In</td>
<td>1.44</td>
</tr>
<tr>
<td>Sb</td>
<td>1.41</td>
</tr>
<tr>
<td>Te</td>
<td>1.36</td>
</tr>
<tr>
<td>Tl</td>
<td>1.48</td>
</tr>
<tr>
<td>Bi</td>
<td>1.46</td>
</tr>
</tbody>
</table>

Figure 2.3. PLS results for the predicted versus the measured lattice constant.
The predicted results for the lattice parameter and bulk modulus for the new hypothetical semiconductors (GaBi, InBi and AlBi) are within 10% compared with those calculated by Ferhat [58]. Therefore, based only on the knowledge obtained though PCA results of known semiconductors and coupled with a fundamental parameter of the periodic table, such as the covalent radius, physical properties of new semiconductors have been predicted.

2.3. Prediction of Electronic Structure

In the previous section the relationship between form factors, properties, and semiconductor chemistry was discussed, with even the lattice constant and bulk modulus of hypothetical semi-conductors predicted. The developed PCA model showed that a clear relationship exists and can be extracted between chemistry and form factors. Thus, the PCA results were useful because they provide the confidence that the form factors can be accurately and efficiently predicted, as well as suggesting what new values (Table 2.2) should be included in the analysis to more accurately predict the form factors.
The PLS method (mathematics provided in appendix) was used to predict the form factors of GaBi, InBi and AlBi, of which some are known while others have not yet been investigated. Using PLS, the matrix of predictor variables was connected with the matrix of predicted (response) variables. The NIPALS algorithm and leave-one-out cross validation method were used. The goal in using the PLS method is to predict $V_s(3)$, $V_s(8)$, $V_s(11)$, $V_a(3)$, $V_a(4)$ and $V_a(11)$ for GaBi, InBi and AlBi.

The predictions were of varying accuracies. Predictions for $V_a(3)$ and $V_a(4)$ were highly accurate, while all other predictions followed the basic trend but had more error than ideal. However, the values of form factors are not crucial, but rather the accuracy of the EPM calculations using these form factors. Therefore, once the form factors are predicted for the new semi-conductors and the band structure and charge density calculated, the accuracy of the approach can be assessed. The values of the predicted form factors are given in Table 2.3.

**Table 2.3.** The predicted symmetric and antisymmetric form factors.

<table>
<thead>
<tr>
<th></th>
<th>$V_s (3)$</th>
<th>$V_s (8)$</th>
<th>$V_s (11)$</th>
<th>$V_a (3)$</th>
<th>$V_a (4)$</th>
<th>$V_a (11)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN</td>
<td>-0.4662</td>
<td>0.0312</td>
<td>0.3868</td>
<td>0.3245</td>
<td>0.1280</td>
<td>0.0204</td>
</tr>
<tr>
<td>InN</td>
<td>-0.1050</td>
<td>0.0556</td>
<td>0.0439</td>
<td>0.2046</td>
<td>0.2289</td>
<td>-0.0381</td>
</tr>
<tr>
<td>AlN</td>
<td>-0.0981</td>
<td>0.1098</td>
<td>0.0790</td>
<td>0.2821</td>
<td>0.2805</td>
<td>-0.0051</td>
</tr>
<tr>
<td>GaN</td>
<td>-0.1694</td>
<td>0.1016</td>
<td>0.1116</td>
<td>0.2905</td>
<td>0.2422</td>
<td>0.0265</td>
</tr>
<tr>
<td>AlBi</td>
<td>-0.1989</td>
<td>0.0036</td>
<td>0.0495</td>
<td>0.0292</td>
<td>0.0316</td>
<td>0.0022</td>
</tr>
<tr>
<td>InBi</td>
<td>-0.1868</td>
<td>0.0002</td>
<td>0.0418</td>
<td>0.0107</td>
<td>0.0254</td>
<td>-0.0054</td>
</tr>
<tr>
<td>GaBi</td>
<td>-0.1975</td>
<td>-0.0099</td>
<td>0.0470</td>
<td>0.0177</td>
<td>0.0291</td>
<td>-0.0141</td>
</tr>
</tbody>
</table>

On the basis of these predicted form factors the electronic band structures of GaBi, AlBi and InBi have been calculated within the empirical pseudopotential method. The resulting band structures are displayed in Figure 2.5, with the principal predicted band gaps listed in Table 2.4. All three semiconductors are narrow direct band gaps and InBi is a semimetal, with these values being coherent with the trend of the fundamental band gaps observed in III-V semiconductors (Table 2.5). Additionally the total valence charge densities
have been calculated for these semi-conductors along the (110) plane and the <111> direction in Figures 2.6 and 2.7, respectively.

**Table 2.4.** The calculated principal band gaps for AlBi, GaBi and InBi in eV.

<table>
<thead>
<tr>
<th></th>
<th>$\Gamma_v-\Gamma_c$</th>
<th>$\Gamma_v-X_c$</th>
<th>$\Gamma_v-L_c$</th>
<th>$X_v-X_c$</th>
<th>$L_c-L_v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlBi</td>
<td>0.436</td>
<td>1.472</td>
<td>0.934</td>
<td>3.346</td>
<td>1.72</td>
</tr>
<tr>
<td>GaBi</td>
<td>0.36</td>
<td>0.813</td>
<td>0.472</td>
<td>3.03</td>
<td>1.40</td>
</tr>
<tr>
<td>InBi</td>
<td>0</td>
<td>0.841</td>
<td>0.414</td>
<td>2.838</td>
<td>1.253</td>
</tr>
</tbody>
</table>

Despite the large amount of work in the computational materials science literature, there are relatively few simulation studies which have been developed solely for the purpose of exploring these new semiconductors [58, 59].

Table 2.5 displays the calculated fundamental band gaps for various III-V semiconductors, including the Bi semi-conductors. The results show that the predicted values correspond well with the theoretical predictions, with the prediction that InBi is a semimetal and that GaBi is a narrow band gap but not a semimetal being in agreement with Janotti [59] and Madouri [60]. It should be noted that the first principles LDA calculations underestimate the band gaps even if the author has added an atom dependent, spherical potential which has been fitted to GaAs and applied to GaBi and InBi. Francoeur [61] has found that adding Bi leads to rapid decrease in the band gap, but this study has been carried out only for rather small concentrations of Bi. The findings here would suggest that compounds based on Bi will behave differently than other III-V compounds, but the lack of experimental information on the band gaps of these compounds prevents validation of this finding.
Figure 2.5. The predicted electronic band structures of AlBi, GaBi and InBi.

Table 2.5. The fundamental band gaps of III-V semiconductors in eV.

<table>
<thead>
<tr>
<th></th>
<th>AlP</th>
<th>AlAs</th>
<th>AlSb</th>
<th>AlBi</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fundamental Gap</strong></td>
<td>2.45</td>
<td>2.16</td>
<td>1.61</td>
<td><strong>0.436</strong></td>
</tr>
<tr>
<td></td>
<td>GaP</td>
<td>GaAs</td>
<td>GaSb</td>
<td>GaBi</td>
</tr>
<tr>
<td><strong>Fundamental Gap</strong></td>
<td>2.26</td>
<td>1.42</td>
<td>0.725</td>
<td><strong>0.360</strong></td>
</tr>
<tr>
<td></td>
<td>InP</td>
<td>InAs</td>
<td>InSb</td>
<td>InBi</td>
</tr>
<tr>
<td><strong>Fundamental Gap</strong></td>
<td>1.34</td>
<td>0.36</td>
<td>0.176</td>
<td><strong>0</strong></td>
</tr>
</tbody>
</table>
The predicted charge densities follow the trend of other III-V semi-conductors [53] and confirm the high covalent character for these semi-conductors. The predicted ionicities are .25, .28, and .32 for AlBi, GaBi, and InBi, respectively. The results are in good agreement with those calculated by Janotti [59] but do not agree with the results of Ferhat [58]. However, this may be possibly since Ferhat did not take the contribution of d-electrons into account.

The general agreement of the predicted band structure and charge density with those cited in literature validates the approach introduced here. This work is unique in that it allows one to predict the electronic structure of hypothetical semi-conductors using an empirical approach. As a model for predicting form factors has now been developed, this method can be easily extended to predicting the electronic structure for numerous semi-conductors, and the general agreement of the final results with those published indicates that the model is sufficient for at a minimum suggesting semi-conductors warranting further consideration and may allow for the identification of promising new semi-conductors that would not be observed following a traditional trial-and-error approach.
Figure 2.6. The predicted total valence charge densities in the (110) plane in AlBi, GaBi and InBi
Figure 2.7. The predicted total valence charge densities along the <111> direction in AlBi, GaBi and InBi.

2.4. Summary

This chapter introduced the use of data mining for rapid prediction of semi-empirical parameters that are otherwise unknown. The empirical pseudopotential form factors of new semiconductor compound systems were predicted, which allowed for calculating the band structure and charge density of these systems. The results are generally in agreement with the few published calculations on these systems. The lattice parameter and bulk modulus of the new semiconductors were also predicted while also clearly identifying the impact of changing chemistry of these systems on properties and empirical pseudopotential parameters. This work represents a unique approach to modeling the electronic structure of a material which may be used to identify new promising semi-conductors.
CHAPTER 3: PARAMETER REDUCTION THROUGH STATISTICAL LEARNING-BASED SCREENING

A focus of materials design is determining the minimum amount of information necessary to fully describe a system, thus reducing the number of empirical results required and simplifying the data analysis. By screening parameters calculated through a semi-empirical model, I demonstrate how a data mining analysis can be used to address this issue with no prior assumptions. Using Ti-Al alloys of varying compositions and crystal chemistries as the test bed, dimensionality reduction techniques are used to show that only five of the 21 original calculated parameters are necessary to capture all of the information from the data set, thus reducing the structure-chemistry-property search space.

3.1. Introduction

Not only is the creation of data whether through calculation or experiment important, but a way to analyze the data in a comprehensive and robust manner is also necessary. Some of the challenges in searching through discreet data include the difficulty of analyzing large amounts of data, understanding the correlations amongst various properties, and using the correlations to better understand the underlying physics of the system. Utilizing a multivariate analysis, the data can be examined so that trends and correlations become apparent. Additionally, the number of properties required to describe a system may be reduced to a minimum number so that the problem of creating sufficient amounts of data and analyzing this data is reduced. This work develops the steps in the alloy design process of finding the minimum amount of information necessary to fully describe a system and then creating design maps of structure, property, and composition.

The multivariate analysis presented here uses data from the modeling of Xie et al [62-65]. In their works they calculated electronic structure, lattice constants, and energies based on the configuration of the atoms. Their model permits calculations in a structure at different compositions, so the impact of structure and composition can be analyzed conjointly.
Background of the calculations and some of the equations and methods used are included in their works. The calculations were performed for various Ti-Al binary compositions for the ordered FCC structure of TiAl$_3$ [62], TiAl$_2$ [63], TiAl [64], and Ti$_3$Al [65].

Semi-empirical calculations construct a potential combining parameters fit to data and parameters calculated self-consistently using \textit{ab initio} calculated data [66-70]. The power in semi-empirical calculations is once the fitting parameters reproduce the empirical result, thus constructing a good potential, the model can be coupled to molecular dynamic or monte carlo methods, which enable the same potentials and parameters to be used in predicting the properties at different pressure and temperature, without performing any additional fitting.

The data mining analysis employed in this work used Ti-Al alloys with compositions varying between 0 and 100% Al, in four different crystal chemistries.

Ti-Al alloys have been extensively explored due to their excellent high-temperature properties; however, they also exhibit a lack of ductility at ambient temperatures [71-73]. Ti-Al alloys are clearly an important intermetallic system, but the work presented here is applicable to any system. The goal of this chapter is not to design a new Ti-Al alloy nor to validate the calculation model used, but rather to discuss how a data mining analysis can be applied to solid state modeling and how these results permit better understanding of the alloy and can be used to guide future modeling. The data used in this work was selected because it computationally compares the properties of varying compositions within different crystal chemistries and thus provides a good test bed for developing this methodology, while being available in the literature. Figure 3.1 outlines the logic of reducing the number of descriptors, thereby reducing the structure-chemistry-property search space, which can make finding structure-property relationships easier.

Data mining is employed in this work because in order to fully uncover intercorrelations in the data, a multivariate classification tool is needed. Additionally, the mathematical relationship between descriptors must be developed before descriptor reduction is possible. Once the relationships are known then the model can be accurately extended to new data so that a virtual system can be fully described with limited data.
Figure 3.1. Finding structure-chemistry-property relationships in materials design is difficult due to the large dimensionality of the problem. Displayed are the four Ti-Al structures discussed in this chapter, which each have a series of associated properties. A change in chemistry results in a new set of structures each of which has different associated properties, resulting in a huge search space. To counter this “curse of dimensionality” the number of descriptors necessary is reduced by finding the relationship between descriptors so that a reduced property space can accurately describe the system of interest. Although I use a specific system and set of properties as a test bed here, the described methodology can be extended to any systems and properties.

3.2. Data Mining of Semi-Empirical Calculations

A problem in many engineering applications is that amounts of data are highly multivariate and are therefore difficult to analyze for trends and information. Although data mining algorithms can be used to find the “hidden” information, these analyses can often be cumbersome and inconclusive, while not reducing the amount of data that needs to be created. Another approach to solving the problem of high data dimensionality is to remove redundant descriptors by screening the data with an informatics analysis, which I demonstrate by
screening the results of a semi-empirical model. This new approach can be applied to any problem to reduce the number of descriptors, representing a unique approach to a traditional data mining problem.

Table 3.1 contains a portion of the dataset used. The total dataset is comprised of 44 samples and 21 properties. The properties and their abbreviations included in the analysis are as follows: number of free electrons in s-orbital ($s_f$), number of covalent electrons in s-orbital ($s_c$), number of covalent electrons in p-orbital ($p_c$), number of covalent electrons in d-orbital ($d_c$), number of non-bonded electrons in d-orbital ($d_n$), potential energy ($\epsilon$), atomic volume ($v$), lattice constant ($a$), and cohesive energy ($E_c$), where each property was calculated for the alloy and the Ti- and Al-component at each composition of the ordered FCC structure. The data is aligned as necessary for the multivariate analysis with rows of differing conditions (i.e. crystal chemistry and composition) and the columns as the responses (i.e. descriptors). Prior to analysis, the data was normalized to remove any effect due to changing units.

Table 3.1. A portion of the data set used in this analysis. The total dataset contains 44 rows and 21 columns.

<table>
<thead>
<tr>
<th>% Al</th>
<th>$s_f$</th>
<th>$s_c$</th>
<th>$\epsilon$</th>
<th>$a$</th>
<th>$E_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.673</td>
<td>0.539</td>
<td>-4.7789</td>
<td>0.41267</td>
<td>461.1</td>
</tr>
<tr>
<td>10</td>
<td>0.6252</td>
<td>0.6628</td>
<td>-4.7719</td>
<td>0.41085</td>
<td>460.42</td>
</tr>
<tr>
<td>25</td>
<td>0.5495</td>
<td>0.8516</td>
<td>-4.7952</td>
<td>0.40724</td>
<td>462.8</td>
</tr>
<tr>
<td>30</td>
<td>0.5505</td>
<td>0.9133</td>
<td>-4.7349</td>
<td>0.40649</td>
<td>456.85</td>
</tr>
<tr>
<td>40</td>
<td>0.5553</td>
<td>1.0438</td>
<td>-4.5977</td>
<td>0.40514</td>
<td>443.61</td>
</tr>
<tr>
<td>50</td>
<td>0.563</td>
<td>1.1793</td>
<td>-4.4421</td>
<td>0.40406</td>
<td>428.6</td>
</tr>
<tr>
<td>60</td>
<td>0.5731</td>
<td>1.3171</td>
<td>-4.2683</td>
<td>0.40281</td>
<td>411.83</td>
</tr>
<tr>
<td>75</td>
<td>0.5912</td>
<td>1.525</td>
<td>-3.9733</td>
<td>0.40294</td>
<td>383.37</td>
</tr>
<tr>
<td>80</td>
<td>0.5978</td>
<td>1.5942</td>
<td>-3.866</td>
<td>0.400305</td>
<td>373.01</td>
</tr>
<tr>
<td>90</td>
<td>0.6115</td>
<td>1.7323</td>
<td>-3.6376</td>
<td>0.40371</td>
<td>350.98</td>
</tr>
<tr>
<td>100</td>
<td>0.6255</td>
<td>1.871</td>
<td>-3.389</td>
<td>0.40496</td>
<td>326.99</td>
</tr>
</tbody>
</table>
3.3 Results and Parameter Reduction

The data mining results are shown in Figure 3.2, where (a) is the PCA scores plot, and (b) is the PCA loadings plot. The scores plot compares four different Ti-Al alloys (Ti$_3$Al, TiAl, TiAl$_2$, and TiAl$_3$ - represented in Figure 3.2(a) as a circle, diamond, square, and triangle respectively). Beyond the alloys, I also compare different compositions within these alloys, where the percentage labeled on the figure is the percent of Al in the binary alloy. Points in Figure 3.2 with opposite PC signs are inversely correlated, properties with similar PC values, for example as defined through the clusters, are correlated, while properties plotted perpendicularly are unrelated, such as cluster 2 and point 4.

PC1 captures 86.03% of the variance in the dataset, and PC2 captures 10.86% of the variance. No other PCs are included in any of the discussions because they do not provide new significant information. PC1 captures differences due to composition, as PC1 increases with increasing Al content. PC2 captures differences due to structure with differences between structure seen primarily along PC2, and the highest PC2 value corresponding with the expected stoichiometry. For example, at high Al content, TiAl$_3$ has the highest PC2 value, while at high Ti content Ti$_3$Al has the largest PC2 value. Since PC1 captures 86.03% of variance and PC2 captures 10.86% of variance, differences in the data due to composition are approximately eight times more important than changing alloy structure within this calculation.
Figure 3.2. Results of PCA on data table of 44 samples and 21 descriptors. (a) PCA scores plot of four different Ti-Al binary compounds. \(\text{Ti}_3\text{Al}, \text{TiAl}, \text{TiAl}_2, \text{and TiAl}_3\) are represented by a circle, diamond, square and triangle, respectively. The percentages shown on the figure correspond to percent Al in the alloy. For example, 25% means an alloy of 75% Ti-25% Al. (b) PCA loadings plot. Each point represents a different descriptor. Five clear clusters are displayed and assigned, as marked by the numerical labels.
The distance from the origin to the PC location of the property defines its impact on the functionality of the sample, and correlation between descriptors is captured through similarity in PC values. Therefore points within clusters are highly correlated. The descriptors within each of the five labeled clusters are as follows: (1) $s_{c_Ti}$, $E_{c_Ti}$, $d_{n_Ti}$, $\varepsilon_{Ti}$, $\nu_{Ti}$, $a_{Ti}$, $p_{c_Ti}$, $E_{c_Ti}$; (2) $a$, $s_{f_Ti}$, $d_{n_Ti}$, $\varepsilon_{Ti}$, $\nu_{Ti}$; (3) $s_{f_Ti}$; (4) $s_{c_Ti}$, $\varepsilon_{Ti}$, $d_{c_Ti}$, $E_{c_Ti}$, $s_{f_{Al}}$; (5) $s_{c_Ti}$, $\varepsilon_{Al}$, $d_{c_{Al}}$, $E_{c_{Al}}$, $\nu_{Al}$, $a_{Al}$. Once these correlations are identified and their relationships are mathematically known, only one descriptor from each correlation is necessary to build a descriptive or predictive model since the existing relationships can be extended to the other properties within the respective clusters.

The axes of the scores and loadings plots are the same so the information in the plots can be compared directly and the relationships between the plots can be used to understand the impact of changing the conditions on the responses. The clusters labeled 2 and 5 in Figure 3.2(b) contribute to the PC1 values of the plot and therefore are compositional effects, with negative PC1 values related with Ti-rich alloys and positive PC1 representing Al-rich alloys. Increasing Ti content in the alloy increases the values of the properties in cluster 2 and decreases the values in cluster 5, while increasing the Al content has the inverse effect on these properties. Obvious examples of this are lattice parameter and volume which have a negative PC1 value, while many of the other relationships between conditions and properties are less obvious.

In principle, only one descriptor from each cluster is needed as any additional information provided by including more properties is redundant. Using this logic, one value from each cluster should result in finding the same relationships between the alloys, which is verified in Figures 3.3 and 3.4. When the 21 original properties in Figure 3.2 are reduced to five properties, the scores plot of all alloys still appears similar to the results with 21 properties. Figure 3.3 displays the scores plot of the Ti$_3$Al structure based on (a) 21 properties and (b) five properties. Only Ti$_3$Al has been included in these figures for clarity, with the five properties included ($s_{c_{Ti}}$, $d_{n_{Ti}}$, $a$, $s_{f}$ and $s_{c}$) chosen because they are the most centrally located in each respective cluster. Figure 3.4 displays the loadings plot with (a) 21 descriptors and (b) five descriptors.
Figure 3.3. Ti$_3$Al scores plot with (a) 21 descriptors and (b) 5 descriptors. The figures are very similar with relationships between alloys remaining consistent, even though 16 of 21 descriptors have been removed.

Figure 3.4. Ti$_3$Al loadings plot with (a) 21 descriptors and (b) 5 descriptors. The relationships amongst the descriptors and between the descriptors and the alloys all remain consistent.
In Figures 3.3 and 3.4, it is seen that the relationships all remain consistent even while removing 16 of the 21 descriptors in the analysis. The reproducibility of the scores plot with 16 fewer descriptors validates the use of PCA for parameter reduction. While there are some minor changes in the position of the points, all conclusions drawn from Figure 3.2 remain consistent, and as the concern is with a relational analysis, the change in axis values is not problematic. I have shown for the first time that with PCA one can effectively determine the minimum number of descriptors necessary to create an accurate classification of the systems examined or the minimum number of parameters needed for a semi-empirical model with minimal loss of information.

3.4. Summary

In this chapter, I introduced a new approach for minimizing the number of parameters necessary for semi-empirical calculations, thus representing an approach to develop reduced order models. Based on classifications from a dimensionality reduction technique and with no prior assumptions, descriptor redundancy can be found through the inter-correlations identified in PC-space. Therefore, using only PCA and analyzing the variability amongst the descriptors, the necessary data can be reduced to only descriptors providing unique information, provided that the relationship between all descriptors of interest has been determined as was done here. The described methodology is not specific to the example shown here, but rather is an approach applicable for finding the minimum amount of information needed to describe any new system.
CHAPTER 4: CLASSIFICATION OF THE DENSITY OF STATES SPECTRA

The ability to model the density of states has been a long standing problem in condensed matter physics. The classical methods that have been used are based on a variety of approaches, ranging from maximum entropy methods to recursion methods involving high dimensional data. In this chapter, I assess the feasibility of modeling the density of states (DOS) by demonstrating that data mining can be used to classify the crystal structure of multicomponent systems based on the electronic structure, the inverse process of first principles calculations which calculate the electronic structure from crystal structure-based inputs. In addition to classifying multicomponent systems by crystal structure, I can classify alloys based on the degree of tetragonality and stoichiometry using solely the DOS spectra. This work seeks to describe the relationship between crystal and electronic structure based on a quantitative interpretation of the DOS.

4.1. Introduction

A methodology to analyze the output of electronic structure calculations is needed due to the large amounts of data created by these calculations. Two limitations with these calculations are that their complexity requires substantial time to perform the calculation and the amount of output data makes it difficult to analyze the data quantitatively and efficiently. Data mining can be beneficial in both of these areas by providing a methodology to quantitatively analyze the data and classify the materials based on the output so that predictions of new systems can help reduce the number of future calculations necessary. This work addresses the issue of analyzing the data from electronic structure calculations by integrating data mining techniques with first principles calculations.

In this chapter, the electronic structure is represented by the DOS. The object of this chapter is to find the relationship between crystal structure and electronic structure. Crystal structure is based on the long-range organization of atoms, while electronic structure describes which energy bands that the electrons of a system occupy. Electronic structure and
crystal structure have different properties associated with them, and finding a way to link electronic and crystal structure would permit a better understanding of how electronic properties are related with atomic properties. The results presented here suggest the future application of this methodology in creating a structure map linking electronic structure and crystal structure, by demonstrating that crystal structure is classified from discreet quantum mechanical parameters as contained within the DOS.

Traditional first principles methods calculate electronic structure descriptions from crystal structure level inputs (space group, lattice constant, and chemistry), thus representing a link across length scales. From the knowledge of the irreducible representation of the crystal structure and the pair wise interactions associated with the bonding between the constituent elements, a library of equilibrium properties can be derived (e.g. elastic constants, thermal and electronic properties).

This represents a traditional materials science problem: the connection between the electronic length scale and the atomic length scale. However, due to the complexity of the math involved in the first principles methods, the ability to understand trends between electronic and atomic characteristics is limited. In this regard, the design of materials through first principles typically follows a trial-and-error approach, much like experiments. The application of data mining techniques to search through the data and complex mathematics is necessary to guide the selection of future calculations.

The difficulty in assessing first principles calculations with data mining techniques is selecting the proper descriptors to analyze. As the inputs into the calculations are not sufficient for a proper analysis of electronic structure, a different approach to this problem is followed. By starting from electronic structure inputs, the relationship with atomic level characteristics is assessed, thus providing the inverse approach of first principles calculations.

The problem this chapter addresses is examining if the crystal structure can be captured and identified from the density of states (DOS) curve without having to work through the complexity of Schrödinger’s equation from an inverse problem approach. I show for the first time that by simply data mining the DOS spectra itself, the principal contributions (ergo principal components) that define the DOS, namely crystal structure, can in fact be captured.
The ability to capture clear trends from the DOS representing changes in atomic conditions provides confidence that the DOS can be modeled based on data mining methods. In the next section, the description of how the methodology represents an inverse problem approach is demonstrated, with subsequent sections applying the approach to intermetallics and to oxide systems. Therefore, in this chapter, the applicability of the developed methodology to model the DOS is assessed on different types of multicomponent systems.

4.2. Inverse Problem Approach

In this section I discuss how the applied methodology represents an inverse problem approach, which is necessary in order to assess the complex mathematics of first principles methods.

First principles calculations are based on approximating solutions to Schrödinger’s equation of a many-particle system, which serves as finding the statistical probability that an electron will have a specific energy. The solution to Schrödinger’s equation contains all measurable properties of the system. However, finding an exact solution to Schrödinger’s equation is nearly impossible for real problems. DFT overcomes this difficulty by developing a solvable equation based on density functionals which is used to approximate Schrödinger’s equation. The process of calculating the electronic structure starting from crystal structure is shown in Figure 4.1.
Figure 4.1. The process for calculating electronic structure. From the alloys chosen, the crystal structure is used to create the input parameters for the calculation, which are space group (structure), equivalent atom positions, and lattice parameters (in angstroms). Using only these inputs, numerous electronic properties can be calculated. DFT calculates a solution approximating the many particle Schrödinger equation by relating the energy to the electron density. This work will consider only the DOS. This process creates electronic structure data from crystal structures.

Space group is identified by the crystal lattice and the symmetry operations possible for the atoms, therefore describing the crystal structure. Although crystal structure is input into the calculation, none of the inputs are suitable for a data mining analysis which seeks to integrate information on properties. These inputs are used to approximate a solution to
Schrödinger’s equation, from which, based on quantum mechanics, all properties of the system can then be solved.

Following the description in [74], the DOS can be thought of as a histogram describing the distribution of eigenvalues of the Hamiltonian matrix. These eigenvalues are the basis in calculating the electronic portion of total energy in some calculation methods. The energy eigenvalues consist of a wide range of energies, and the DOS is used as a probability function describing the spread of the eigenvalues.

Additionally, the DOS shows the energy orbitals existing for a system because the energies where states are available would be expected to also be energies of available orbitals. As different structures have orbitals existing at different energies, it is anticipated that the DOS contains information on the structure of the system. Using this logic, by data mining the DOS the structural information should be extractable using solely the DOS. The difficulty in this extraction is selecting the proper descriptors. To overcome this difficulty, a multivariate analysis is being performed on the entire DOS, treating the DOS as spectral data.

PCA is used in this work to quantitatively analyze the DOS to classify materials based on crystal structure. This poses the inverse objective of first principles calculations, and this inverse problem is shown in Figure 4.2. In this work, the descriptors are all the same unit, and no scaling of each descriptor was necessary. However, the interest in this work initially is in the shape of the curves, as the shape of the DOS captures the information on the structure and some properties, and each curve was scaled so that the maximum DOS value of each curve is equal to unity. While data mining has been used on spectral data previously [75-78], in those cases the interpretation of the peaks was already developed and the alignment of the curves was obvious, while with DOS the meaning of changes in peaks is unclear and the alignment of data needs to be considered.
Figure 4.2. The inverse problem. From Figure 4.1, the process of first principles calculations was presented where from the crystal structure the electronic structure was calculated. In this work, data mining is used to determine the crystal structure from the electronic structure. The input parameters from Figure 4.1 cannot be determined exactly, but the relative values can be determined.

4.3. Analysis of Intermetallic Systems

As a first test bed for the analysis of alloy systems, intermetallics were used. The systems calculated were transition metal aluminides: TiAl, TiAl₃, CoAl, Co₃Al, NiAl, and Ni₃Al. The input crystal structure for each calculation was: TiAl (L1₀), TiAl₃ (DO₂₂), CoAl and NiAl (B2), and Co₃Al and Ni₃Al (L1₂). Transition metal aluminides are of interest because they constitute an exceptional class of structural materials for high temperature applications, with properties such as excellent high-temperature strength, corrosion resistance,
low density, and high melting point. However, most of these intermetallics exhibit brittle fracture and low ductility at ambient temperatures. This work can be extended to ternary compounds to study the effect of alloying and site occupancy, as successful attempts have been made in improving ductility with additional alloying of these compounds.

The data set of these six intermetallics contains six samples versus 783 energy values, although calculations can have many more points (calculations used in chapter 7 have 10,000 points). Energy values that do not provide new information or are only background have a PC loadings value equal to the origin and thus have minimal impact in the developed model. Therefore, when using PCA every point can be included without creating a daunting problem of too much data. The DOS was manually aligned by the tallest bonding peak, but other alignments such as focusing on the Fermi energy or using developed algorithms warrants consideration. This work provides a framework with which to better understand and interpret DOS curves and to use this information for materials design.

The organization of the data set and the PCA process are shown in Figure 4.3. Using PCA to analyze the data, 91.49% of the variance in the data is captured with two PCs. PCA is useful for analyzing spectral data because it can reduce large data sets to minimal amount of dimensions; in this case, PCA reduces the data from 783 dimensions to two dimensions while maintaining 91.49% of the information. This analysis is particularly useful because it provides a quantitative analysis of the DOS, while most analyses of the DOS are quantitative. This demonstrates the importance of integrating data mining with electronic structure calculations because this work makes a quantitative analysis of the entire DOS curve possible.

The classification of the DOS is done by treating the DOS as spectral data with each value at a different energy constituting a new descriptor. An approach developed for analyzing spectral data is used to classify the DOS based only on patterns dictated by differences in relative feature locations and feature shapes. Figure 4.4 shows a demonstration of how this approach operates by demonstrating the identification of differences in a peak between two spectra. The complexity in the case of the DOS is greatly increased due to the increased number of curves and the large number of poorly defined features. The importance of this approach is in finding the interrelationships between
features. With the identification of all relationships within the spectra, the features of the DOS curve and the changing conditions such as chemistry and symmetry can be related.

**Figure 4.3.** The composition of the matrix input into the data mining analysis. Using every data point, the DOS is treated as spectral data. The scores and loadings matrices capture the atomic level descriptions, while the input DOS captures the electronic structure.
Figure 4.4. A schematic of a PCA loading plot of two simple spectra. To track correlations between spectral peaks by PCA, the loadings plot of spectra gives new spectral patterns by changing the point of view to change in spectra. The axes of the right figure represent intensities of peaks from two original spectra (left plot). In the right plot, the starting and ending points of the coordinates have to be (0,0) since both peaks start at the background and return to the background while the trajectory has to have two “maxima,” one associated with each of the peaks. Note that peak difference of the two original spectra is shown differently in PCA loading space.

Figure 4.5 shows the scores plot of the analysis. PC1 and PC2 are comprised of a combination of descriptors (energy levels) that contain the most variance in the data. The combination of energy levels which most differentiate the structures are redefined as PC1 and the combination of energy levels which capture the most difference in an orthogonal direction to PC1 are redefined as PC2. The important feature of this figure is that clear distinctions between crystal structures can be seen. Additionally, the axes are capturing specific aspects of the crystal structure, as PC1 is seen to differentiate tetragonal distortion, or the change from cubic to tetragonal, while PC2 captures differences in stoichiometry.
Figure 4.5. PCA scores plot from data in scores matrix of Figure 4.3. PCA clusters points in the same structure, demonstrating that crystal structure is represented in the DOS. PC1 is capturing the tetragonality of a structure because the structures with negative PC1 values are cubic and the structures with positive PC1 are tetragonal. PC2 is capturing the differences in stoichiometry. The structures with positive PC2 have an AB stoichiometry and the structures with negative PC2 have stoichiometry of $A_3B$ or $AB_3$. This analysis provides a framework with which to analyze a large number of DOS quantitatively.

PC1 and PC2 are comprised of a combination of the descriptors (energy levels) that contain the most variance in the data. Using the loadings matrix from Figure 4.3, the contribution of each descriptor to the PCA model can be identified. Knowing how the energy levels contribute to the PCA model can be used to identify which features of the DOS are most important and with what those features correlate. NiAl, Ni$_3$Al, CoAl and Co$_3$Al all have negative PC1 values, and additionally all have the same space group, as shown in Figure 4.1. The alloys with negative PC1 values have cubic structure ($a=c$) while the alloys with positive PC1 values are not cubic. PC2 is capturing the differences in stoichiometry, with positive PC2 structures having an AB stoichiometry, and negative PC2 structures
having an $A_2B$ or $AB_3$ stoichiometry. From this interpretation, adding additional alloying elements will drive the system to a more negative PC2 value.

PCA is able to differentiate structure, tetragonal degree, and stoichiometries based only on the DOS. This classification is again emphasized when examining the eigenvector spectrum from the loadings plot (Figure 4.6).

**Figure 4.6.** Eigenvalue spectrum clearly showing the maximum (CoAl) and the minimum (Co$_3$Al) correspond to energies associated with the maximum in the density of states spectra for CoAl and Co$_3$Al but of opposite weights in both eigenvectors. Hence I have identified in which part of the DOS curve the separation/classification according to stoichiometry/chemistry is most effectively captured with regards to the systems analyzed. As anticipated from the scores plot, similar results exist when identifying which part of the DOS curve captures stoichiometry for all the other aluminides analyzed, or when identifying parts of the DOS curve due to the tetragonal distortion (PC1).

### 4.4. Analysis of Oxide Systems

In this section, I demonstrate that the approach for classifying the DOS curve is applicable to other systems, and was not just coincidental by analyzing a different system, namely oxide systems. The design of new oxides is slowed by the lack of electronic
description for new or “virtual” systems, and so an improved manner of analyzing these systems which may lead to suggested design pathways has far reaching implications.

Oxide systems are used as the test bed here, as they are a group that is of high interest in the microelectronic industry because of their use as a gate dielectric in metal-oxide-semiconductor devices [79, 80]. A specific example of an oxide of interest for gate dielectrics is HfO$_2$, which has numerous beneficial properties such as wide band gap size and high dielectric constant, as well as forming a stable electronic interface with Si [81-83]. Unlike HfO$_2$, ZrO$_2$ does not form a stable electronic interface with Si, although it has a nearly identical crystal structure as HfO$_2$, similarity at the electronic level besides the closed shell of f-electrons in Hf, and nearly identical size and electronegativity due to lanthanide contraction [84]. Both HfO$_2$ and ZrO$_2$ naturally form three low-pressure structural phases: monoclinic, tetragonal, and cubic, in order of formation with increasing temperature. The fact that two systems that are seemingly identical have one difference in characteristic that determines the suitability as a gate dielectric demonstrates the need for a data mining design strategy instead of following a trial-and-error approach, as an obvious substitute in chemistry does not necessarily lead to the intended characteristics. Meanwhile, a chemistry that has the desired characteristics may go undiscovered due to the enormous number of possible chemistries.

Thirteen DOS curves were included in the analysis: HfO$_2$ (cubic, tetragonal and monoclinic structures), ZrO$_2$ (cubic, tetragonal, monoclinic, brookite and cotunnite structures), Cd$_2$Nb$_2$O$_7$ (cubic), Y$_2$Ti$_2$O$_7$ (cubic), ReO$_3$ (cubic), CeO$_2$ (cubic) and MgO (cubic). The DOS curves are shown in Figure 4.7, where the curves are aligned so that the Fermi energy is equal to zero for all curves, and the curves have been normalized so that comparisons become easier to visualize, although for each structure HfO$_2$ has a larger maximum DOS value than ZrO$_2$ when looking at unscaled data. Details of the calculation of these DOS curves are presented in references [85-87].
Figure 4.7. DOS for all analyzed chemistries/structures. The horizontal axis is energy, where the Fermi energy has been set to zero, and the vertical axis is equal to density of states. Visually examining different chemistries and structures together allows for seeing obvious patterns and differences related with chemistry and structure.

Some insights into the DOS can be made from visual inspection of the curves in regards to patterns associated with crystal structure and chemistry. The first observation is that the curves for HfO$_2$ and ZrO$_2$ are similar for each structure, with the main difference being the larger band gap for HfO$_2$. Using ZrO$_2$ as an example, the number of peaks in the DOS are seen to increase with decreasing symmetry, with the number of peaks increasing
from cubic to tetragonal to orthorhombic (brookite and cotunnite) to monoclinic. $\text{Y}_2\text{Ti}_2\text{O}_7$ and $\text{Cd}_2\text{Nb}_2\text{O}_7$ show some similarity with cubic HfO$_2$ and ZrO$_2$ beyond the Fermi energy appearing at the higher band gap edge and additional peaks existing at higher energies, although $\text{Cd}_2\text{Nb}_2\text{O}_7$ appears more unique, anticipated since Cd has a full d-shell while Y, Ti, Hf and Zr all have one or two electrons in the d-shell. There are noticeable differences between ReO$_3$, MgO and CeO$_2$ with the other cubic structures. Of these three structures, ReO$_3$ is the only one with a transition metal-oxygen combination, but it has a unique stoichiometry, while CeO$_2$ has the same stoichiometry as HfO$_2$ and ZrO$_2$ but is comprised of a lanthanide-oxygen combination.

By visually comparing the DOS curves of the thirteen systems, patterns associated with structure and chemistry were recognized. The objective then becomes to clearly extract these conditions using pattern recognition techniques. Identifying these relationships with pattern recognition techniques allows for the identification of how properties are represented by the DOS and to tailor materials based on the electronic structure design pathways that are identified.

By analyzing HfO$_2$ and ZrO$_2$ in cubic, tetragonal and monoclinic structures, the differences between structures and between the chemistries of HfO$_2$ and ZrO$_2$ can be isolated. In order to fully capture the impact of structure the data was aligned so that the band gap edges aligned. The resulting scores plot (Figure 4.8), where the proximity between points corresponds to the similarity between compounds as defined by their DOS, shows that crystal structure is the dominant condition in differentiating these samples as opposed to the differences in chemistry. PC1, an axis comprised of a combination of energies in the DOS which most differentiate the oxides, separates monoclinic structure (low symmetry) from cubic and tetragonal structure (high symmetry), while PC2 captures features in the DOS related to an elongation of the c-axis in tetragonal versus cubic structure. Lower PCs more fully describe variations due to chemistry. This interpretation is notable because it means that not only is crystal structure clearly represented in the DOS curve, but that information on the different aspects of symmetry can be clearly extracted from the DOS.
Figure 4.8. Scores plot of HfO$_2$ and ZrO$_2$ in three structures, where it is shown that crystal structure can be clearly extracted from the DOS, with PC1 separating high symmetry structures from low symmetry structures, while PC2 differentiates cubic and tetragonal, representing the elongation of the c-axis. The two drawn triangles represent binary chemical phase diagrams, where the shift between triangles represents electronic dissimilarity. The percentage labeled on the axes is the amount of variability of the total data set captured by each respective axis.

To test the robustness of the methodology, the analysis was repeated with the additions of Y$_2$Ti$_2$O$_7$ (cubic), Cd$_2$Nb$_2$O$_7$ (cubic), and ZrO$_2$ (cotunnite and brookite). The results verified that structural information clearly exists within the DOS even when considering a wider variety of structures and ternary stoichiometries instead of only binary stoichiometries. In that scores plot, cotunnite and brookite (orthorhombic structures) had values similar to the monoclinic structures, following the trend related to lower symmetry structures, while all of the cubic systems were still plotted in close proximity to each other regardless of crystal chemistry.

For the purpose of isolating any patterns within the DOS curve related with chemistry, only the seven cubic structures from Figure 4.7 were analyzed, with the data aligned so that the Fermi energy was at the same energy for all curves while also including the band gap. The resulting scores plot (Figure 4.9) displays the existence of a trend related with chemistry,
although this trend is more subtle than that associated with crystal structure. PC1 is primarily related with stoichiometry, as labeled on the figure with ratio of metal to oxide increasing with increasing PC1 value, while other PCs are capturing variations related with changing chemistry. Since the curves were normalized for this analysis, the differences related with atomic size, chemistry and orbital structure are due at least in part to a change in the shape of the peaks. Also, it is concluded that stoichiometry has a bigger impact in defining the electronic structure for the oxide systems considered here as compared with the specific chemistries, agreeing with the previous findings on intermetallic systems.

**Figure 4.9.** Scores plot of seven cubic structures. The largest difference in the DOS beyond structure for the systems considered is due to stoichiometry, with AO₃, AO₂, A₂B₂O₇ and AO being clearly separated, while the other PCs capture chemistry related effects. This figure shows that differences due to crystal chemistry can be extracted from the DOS, in addition to structure.

The DOS is a combination of the associated contributions, whether crystal structure, chemistry, background or artifact from computational method, or other contributions. If it is known how every contribution is represented in the DOS, then the DOS curves can be
constructed using any combination of these contributions. A contribution from the original DOS curve can be removed and the DOS can be reconstructed with a new pattern. For example, for a cubic structure, the curve can be decomposed to its contributions, the crystal structure pattern switched from cubic to tetragonal, and the curve in the tetragonal structure constructed. A demonstration of following this approach for modeling the DOS of oxide structure using only PC values and relating them to physical meanings is presented in chapter 7.

4.5. Summary

In this chapter, a PCA analysis was used to classify different multi-component systems based only on the DOS curve. I have accurately developed structural and stoichiometric classifications of compounds. To verify the applicability of this approach to multiple data types, an analysis was performed on both intermetallics and oxides. The PCA maps created suggest design pathways to follow by relating trends in these maps based on changing DOS with desired characteristics. The success of the data mining approach to pick up clear physical significance from the DOS spectra provide confidence that the DOS can be modeled using data mining, which will be discussed in the ensuing chapters.
A method to model the density of states without encountering the complexities of traditional first principles calculations has been sought after for the past half-century. The most successful approach addressing this objective has been the recursion method pioneered by Haydock. While the Haydock method elegantly models the density of states (DOS) for many systems, it does have some limitations. Introduced in this chapter is the logic for a new approach, based on data mining principles, which overcomes the limitations of the Haydock method, while reducing the computational requirements.

5.1. Introduction

Numerous methods to model the density of states (DOS) curve while limiting the involvement of quantum mechanical calculations and approximations have been developed. Of these methods, the most notable is the recursion method pioneered by Haydock [88, 89], which permitted the calculation of systems without translational symmetry by emphasizing the resolvent elements (local quantities) of the Hamiltonian instead of the eigen-states. The other previously developed approaches to model the DOS without the limitations of first principles calculations include the characterization of the DOS in terms of its moments [90, 91], via equations of motion [92, 93], the utilization of Monte Carlo models [94, 95], and the maximum entropy approach [96, 97]. While first principles calculations are the most accurate, they are time consuming for difficult systems and have some limitations such as at non-zero temperatures and for systems without translational symmetry.

Figure 5.1 shows the timeline of when these approaches were introduced. The point of this figure is to show that this problem has been explored for the past half-century, with no breakthroughs for twenty years.
Figure 5.1. Timeline for the development of methods to model the DOS spectra. This problem has been researched for the past 40 years, with the work presented in this thesis being the first advancement on this problem in over twenty years.

The work presented in this thesis can be considered as a natural progression of these other methods. The moments approach began the concept of modeling the DOS without considering Bloch’s theorem by fitting a limited number of moments to calculated DOS in a Gaussian like manner. The recursion method advanced on this work by including a recursion formula, which permitted a more complete representation of the tight-binding Hamiltonian in tridiagonal form (mathematics presented in the appendix). The data mining approach introduced here uses a data mining formulism to calculate the DOS by capturing the relationship between material descriptors and electronic structure parameters, and then constructing the DOS of new systems based on these identified relationships. Therefore, I follow the logic of these other methods developed to address the limitations of first principles methods, but by changing the mathematics used in the process I have improved the accuracy and reduced the computational requirements.

The data mining approach to model the DOS operates by identifying patterns and trends within large data sets, and optimizing these patterns in predicting the DOS for new systems. In the previous chapter, I demonstrated that patterns associated with individual
contributions can be clearly extracted from the DOS curve, in that case capturing how crystal structure is represented in the DOS. Finding that patterns associated with crystal structure exist in the DOS was justified by recognizing that the main factors that govern the relative energies of different crystal structures are encoded within the sum of energies of the occupied states [98, 99]. Demonstrating that data mining tools can clearly extract the contributions suggested that a “soft modeling” approach to calculating the DOS was feasible because if all contributions can be represented then the DOS is simply the sum of these contributions.

This chapter compares the recursion method and the proposed data mining method for modeling the DOS. While the mathematics and individual steps are very different, I suggest that the overall logic is similar. The full logic of my approach is presented in the next chapter.

5.2. Haydock Recursion Method

Figure 5.2 shows a basic logic followed by the recursion method. Although the steps are not explicitly defined to be the same in Haydock’s papers as in this figure, the logic is properly summarized. The mathematics of each step is presented in appendix section A.4 of the thesis, with this chapter only presenting the logic of the approach.

Following the above logic, the density of states can be calculated while avoiding the electronic wave function. The approach is based on the finding of the moment approach that the one-electron Green’s function is related to the density of states. While Cyrot-Lackmann [90] converted the tight-binding Hamiltonian to a tridiagonal form by identifying the moments of the system, Haydock’s method is purely mathematical in nature, but efficiently converts the Hamiltonian to a tridiagonal matrix.

The input Hamiltonian matrix has dimensions of M x M, where M is the number of atoms multiplied by the possible orbitals. The diagonal of the matrix is the Hamiltonian of the atom and orbital combinations, referred to as the states \( u_n \), while the off-diagonals capture the interaction between states. If a one-nearest neighbor tight binding approach is used, then all non-nearest neighbor interactions would be equal to zero. The tight binding Hamiltonian is fit to the band structure, which is a limitation of the method since the band
structure of the system must be known. Additionally, all approximations or inaccuracies within the tight binding method are carried throughout the entire process so that the approximations exist within the final calculated DOS.

**Figure 5.2.** A simplified logic of the Haydock recursion method. By converting an input tight binding Hamiltonian to orthogonalized basis sets capturing local behavior, the DOS can be found through mathematical manipulation.
The Hamiltonian must be converted to a tridiagonal form, where a recursion is used to tridiagonalize the matrix, based on a localized basis. Defining a localized basis provides one of the primary benefits of Haydock’s approach, namely the capturing of localized effects. That is, if an impurity is to be modeled, the basis can be set near the impurity so that the electronic structure of this region can be modeled efficiently. The complexity of first principles methods is greatly increased when translational symmetry is broken.

The recursion formulation describes the projection of the Hamiltonian to the localized state. The portion of the Hamiltonian not projected onto the initial state is orthogonal, so that the remainder of the Hamiltonian projects onto a normalized state. The process continues, with the possibility for projections of the Hamiltonian onto \( M \) orthogonal states, with an additional consideration being the coupling of each state to the preceding states. This iterative process defines the recursion, where during the process the sequences of states and recursion coefficients are defined, with two recursion coefficients defined: \( a_n \) and \( b_n \), with \( n \) equaling the number of recursions performed. \( a_n \) is the energy in state \( u_n \), while \( b_n \) is the hopping integral between states. The Hamiltonian of this orthogonal basis of states has a tridiagonal form, where the diagonal elements are the \( a_n \) terms, while the first off-diagonal term is \( b_n \), and all other terms are zero. If the selection of initial state is not obvious, multiple initial states can be selected with the results averaged so that the impact of selection of initial state is reduced.

While the intent of this step is not to reduce dimensionality, effectively that is what occurs. The reason a dimensionality reduction occurs is because at some \( n \) the recursion coefficients start to converge on a value. Therefore, while \( n \) could be equal to \( M \), a number of recursions much less than \( M \) is necessary. Due to the orthogonality of the states, each recursion captures less of the projected Hamiltonian than the preceding state. While the tight-binding Hamiltonian does not have the terms ordered by information captured, requiring the inclusion of all terms, the tridiagonalized matrix is ordered based on the Hamiltonian, so that \( n \) less than \( M \) is sufficient for only minimal loss of the Hamiltonian. Therefore, the recursion step can be viewed as a dimensionality reduction step.

The number of recursions that can be efficiently calculated is only fifteen to twenty, while the possible number is typically hundreds. If the number of recursions is performed
equal to \( M \), then the entire Hamiltonian is represented. Each recursion not performed results in a lowering of the resolution of the Hamiltonian, so that while twenty recursions are sufficient to capture the general information of the Hamiltonian, resolution is still lost, even though as discussed the amount of information contained within each subsequent recursion is lowered.

Following this stage, a relationship between the dimensionally reduced terms and the DOS is used to model the DOS. A continued fraction of the recursion coefficients allows the calculation of the diagonal elements of the Green’s function. A truncation of the Green’s function is necessary to represent the recursion coefficients not included, since the number of recursions is less than \( M \). The truncation is chosen to represent the converged value of the recursion coefficients, while a truncation not chosen properly is a possible source of error, and has been a matter of debate \[100-102\]. The DOS spectra is then easily calculated from the imaginary part of the Green’s function.

5.3. Data Mining-Enabled Modeling

In this section, the logic of the data mining approach introduced in this thesis is presented. Figure 5.3 shows the logic of the approach, with steps of dimensionality reduction followed by predictions in a high-dimensional eigen-space.

The method requires two inputs: a matrix containing DOS curves for systems not including the system being modeled (Figure 5.4) and a matrix containing material descriptors for all of the systems of the first matrix as well as the system being modeled.

The input data set of training DOS is analyzed via principal component analysis (PCA) \[17-20, 22-24\]. The data set is decomposed into two matrices: the scores and the loadings matrices. The scores matrix describes how a sample is related to other samples; in this case, the scores relates the DOS curve of all of the alloys included in \( X \). The loadings matrix captures the features which most impact the scores matrix. From the loadings matrix, one can find the relationship between features in the DOS, and identify with which systems these features correspond. The value in each of the matrices is the respective principal component (PC) values, which defines the alloy system or descriptor in PC space. In a
scores or loadings plot, the proximity of points represents the degree of correlation, with the understanding that a difference in a PC capturing little variance is much less important than the same change in PC values for a latent variable capturing much more variation in the data. The mathematics of PCA is provided in the appendix.

The number of PCs necessary to capture all of the information is much less than the dimensionality of the input matrix. For the results shown in this thesis, generally ten to fifteen PCs were sufficient to capture all of the information of the data set.

A second input is necessary: a data set of material descriptors. Using partial least squares (PLS) [26-30, 103], a method for predicting properties while capturing co-linearity in the data, a relationship between the scores matrix and the material descriptors matrix can be developed. Based on this relationship, the DOS curve of a new system can be modeled accurately with the only input specific to the system being the material descriptors which are chosen so as to capture all of the variables comprising the DOS, as discussed in chapter 1. Modeling the DOS based on these descriptors would achieve the objective of modeling the DOS based on the interactions of the complex variables, following the logic of DFT calculations.

5.4. Comparing the Haydock Method with the Data Mining Method

This section serves to summarize the discussion of this chapter. Table 5.1 provides the comparison between the two approaches, with the mathematics of the two approaches displayed in the appendix. The full logic of the data mining approach is presented in chapter 6, with the results of the modeling presented in chapter 7.

There are a few clear limitations to the Haydock method. First, the input representing the system of interest requires knowing the band structure for that system. Secondly, the tight binding method requires several approximations so that the input Hamiltonian may have some error. The data mining approach addresses these limitations by not requiring any quantum mechanical input on the system of interest. Thus, for “virtual” systems, electronic structure calculations or experiments are not required. Additionally, the input DOS can be calculated through any approach. In this thesis, density functional theory calculations are
used, so that any error associated with the input data set is minimal. If the input DOS are calculated by another manner, any errors will be carried out through the process.

**Figure 5.3.** The logic of the data mining approach to modeling the DOS curve. Based on the relationship between the DOS curves and material descriptors, the DOS curves for new systems can be modeled.
Figure 5.4. The composition of the input matrix. All values from the DOS curve are included, so that a plotting of each row is the DOS curve for different sets of conditions. Each set of conditions represents a different row, while each column is a different energy value, as shown on the figure. The value within the matrix is the DOS at each energy.

The next source of error associated with the Haydock method is that the number of recursions that are realistic to perform is less than the number of possible recursions. While the importance of each recursion is less than the preceding one, the ability to capture fine structure within the DOS requires these additional recursions. Therefore, the number of recursions that can be effectively performed is sufficient for calculating the general shape of the DOS, but is lacking in resolution.

The data mining approach addresses this error by having fifteen to twenty parameters being sufficient to capture all of the possible information of the system. Therefore, the prediction of parameters can be done with minimal computational requirements while capturing the fine structure of the DOS. Additionally, as the Haydock method does not
calculate all recursion coefficients, a truncation of the continued fraction is required which can introduce error. As all parameters of the data mining approach can be determined, this source of possible error does not exist.

Table 5.1. Comparison between Haydock recursion method and data mining approach. This table serves to demonstrate that although the steps and mathematics are different, the logic between the two methods is similar.

<table>
<thead>
<tr>
<th>Input Data Matrix:</th>
<th><strong>Haydock Recursion Method</strong></th>
<th><strong>Data Mining Method</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tight-binding Hamiltonian matrix for system of interest</td>
<td>DOS curves of other systems — fully represents the multi-dimensionality of the problem</td>
</tr>
</tbody>
</table>

| Orthogonalize Matrix: | Calculate a finite number of recursive coefficients based on orthogonalized basis — due to computational limitations, not all coefficients determined | Via dimensionality reduction approach, the DOS curves are reduced to the common orthogonalized components — capture all components |

| Modeling DOS from orthogonal matrix | DOS is related to the imaginary part of an infinite continued fraction Green’s function, based on recursive coefficients | By recognizing that the orthogonalized components are related to a database description of the material, the components of new materials can be determined in high-dimensional eigen-space |

5.5. Summary

In this chapter, the logic of the Haydock method and the data mining method for modeling the DOS curve were presented and compared. While the mathematics are quite different, I suggested that the logic is similar. The limitations of the Haydock method were discussed, as well as how the data mining approach addresses these limitations.
In this chapter, I demonstrate how data mining can be used to extract clear patterns related to physical conditions from the density of states curves and through optimization of these patterns model new density of states curves in minimal time, thus representing a high-throughput approach for modeling the electronic structure of new materials under any set of conditions. As quantum mechanics is not explicitly defined in the prediction, this approach is applicable to any system without increasing the complexity of the model, while also allowing for the identification of electronic structure-property relationships beyond what is now possible.

6.1. Introduction

PCA is a mathematical tool for classifying systems based on a dimensionally reduced axis system, where the axes are termed as principal components (PCs) and are a linear combination of variables capturing the maximum amount of unique information. PCA operates by calculating the eigenvalues and eigenvectors of the covariance matrix of the input data set, with the eigenvectors defining the directions of maximum variance. Therefore, the eigenvector corresponding to the largest eigenvalue becomes PC1; that is, this eigenvector becomes the new axis capturing the most information within the dataset. The original dataset \( \mathbf{X} \) is decomposed into several matrices following equation (1), where \( \mathbf{T} \) is called the scores matrix, \( \mathbf{P} \) is called the loadings matrix, and \( \mathbf{E} \) is the residual matrix.

\[
\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad (6.1)
\]

The loadings can be understood as the weights for each original variable when calculating the PCs. The scores matrix contains the original data in a rotated coordinate system. The dimension of the scores matrix is equal to the number of PCs included, with typically a few PCs being sufficient to capture all relevant information. As the input data
matrix is spectral data, the loadings thereby becomes spectral as well, or what I term an “eigen-spectra.” If the numbers of PCs is chosen so as to capture all of the information in the dataset (i.e. \( E \) is equal to zero), then the sum of the product of the eigen-spectra and the scores is equal to the input matrix.

Having described this dimensionally reduced space, the prediction of properties via PLS is now considered. PLS can be considered related to principal component regression (PCR) which seeks to find the maximum variance within the predictor variables. Therefore, PCR can be thought of as a linear regression within the dimensionally reduced PC-space. PLS differs by finding the correlation factors between the predictor variables (\( x \)) and the predicted variables (\( y \)). In PLS, two linear combinations are generated from \( x \) and \( y \) respectively and the maximum variance between \( x \) and \( y \) is calculated. The matrix composed of predicted variables (\( Y \), with dimensions \( N \times M \)) can be decomposed in the same manner as the matrix of predictor variables was in equation (6.1).

\[
Y = UC^T + G
\]  

(6.2)

where \( U \), \( C \), and \( G \) are analogous to \( T \), \( P \), and \( E \) in equation (6.1).

6.2. Methodology Logic

A combination of principal component analysis (PCA) and partial least squares (PLS) was found to be sufficient to model the DOS with minimal computational requirements while having high-resolution results. Along with modeling the DOS, this work serves to demonstrate a new approach: the use of data mining as a simulation tool.

The development of the simulation tool utilizing data mining uses PCA to identify the primary patterns differentiating the samples and PLS to make predictions. This approach works by decomposing the input data into patterns which define the samples, then determining the relationship between these patterns and the descriptions related to chemistry and geometry, and finally constructing the new data by combining the patterns as determined by the newly identified relationships. In this regard, the approach here is a simulation tool
since the chemistry and geometry are used to model the electronic structure using only data mining without the direct use of classical quantum mechanical equations.

For the PLS portion, the predictor variables (d or $D$) are input descriptors (in the case of modeling DOS they include descriptors of chemistry and geometry), while the predicted variables (y or $Y$) are equal to the scores values ($T$) in equation (6.1). To help differentiate the matrices from equation (6.1), the decomposition of the predictor variables are given new nomenclature:

$$D = QR^T + H \quad (6.3)$$

Since the scores are good predictors of $Y$ in PLS,

$$Y = QC^T + F \quad (6.4)$$

where $F$ represents the error between observed values and the predicted responses. $Q$ has dimensions equal to the number of PLS components included ($A$). Combining equations (6.3) and (6.4) yields:

$$Y = DRC^T + F \quad (6.5)$$

The relationship between $D$ and $Y$ is described by the matrix $B$ containing the regression coefficients $\beta_{km}$, where $M$ is the number of dimensions in $Y$ and $K$ is the number of dimensions of $D$.

$$Y = DB + F \quad (6.6)$$

Comparing equations (6.5) and (6.6) indicates that the regression coefficients comparing the predictor (material descriptors) and the predicted (PC scores values) variables are based on a relationship between the loadings values of the predictor variables and the loadings values of the predicted variables.
The multidimensional space of $\mathbf{D}$ is reduced to the $A$-dimensional hyperplane, with the correlation of $y$ formed on this hyperplane. For clarification of the dimensions, Figure 6.1 shows a schematic of the various equations, with the number of columns and rows labeled. Of most interest are the dimensions of $\mathbf{T}$ and $\mathbf{P}$ for PCA ($M$ equal to the number of PCs included in the model) and $\mathbf{U}$, $\mathbf{C}$, $\mathbf{Q}$ and $\mathbf{R}$ for PLS ($A$ equal to number of PLS components included in the model) because they are both adjustable dimensions. Additionally, this agreement in dimensions between $\mathbf{T}$ and $\mathbf{P}$ is important because it allows the altering of the output data based on which PCs are to be included with the matrix dimensions still being in agreement and with the number of PCs canceling out since the transpose of $\mathbf{P}$ can be taken.

6.3. Methodology Specifics

Having described the data mining tools used for modeling spectral data, in this section the application to the DOS curves is discussed. Figure 6.2 shows the steps of the data mining methodology for predicting the DOS. Each box and arrow in the figure represents a different issue, and so the process of completing all of the steps is more significant than appears. This section serves to discuss all of the various steps. The numbers on the figure label the section in which that portion of the methodology is discussed (for example, (x) is discussed in section 6.4.x).
Figure 6.1. Schematic of equations (6.1) through (6.3). This figure serves to clarify the dimensions, and can be used as an aid in verifying that all the derived equations follow rules of matrix mathematics. Of particular importance are the dimensions M and A because these parameters are adjustable and depend on the amount of information included in the PCA and PLS models, respectively.
6.3.1. Mathematical Derivation

The input matrix $X$ used in the initial PCA analysis is comprised totally of DOS spectra. The rows of the matrix ($N$) is equal to the number of systems for which DOS curves have been calculated, and the columns of the matrix ($K$) is equal to the number of data points representing unique energies within the curve. The values within the matrix are the intensity of the spectra at each energy value. In this regard, the DOS is treated no differently than any other spectral data.

To summarize the mathematics of the previous section knowing that the scores values are to be predicted for new spectral data, the prediction of scores values for new systems is
based on a relationship between the loadings of the scores values of the spectral input data and the loadings of the descriptors for the new systems. $D$ in equation (6.4) contains material descriptors to differentiate the chemistries and $Y$ in equation (6.2) is equal to $T$ in equation (6.1). Therefore, for modeling the DOS, the equations in the previous section become:

$$T = DRC^T = DB$$  \hspace{1cm} (6.7)$$

and combining equations (6.1) and (6.7), an equation for the modeling of the DOS of new systems is derived, which will be labeled as $Z$. This equation uses $P$ calculated from equation (6.1) and $T$ calculated from equation (6.7). Knowing that the dimensions of $B$ are $K$ by $M$, using Figure 6.1 it is seen that the matrices dimensions are in agreement.

$$Z = DRC^TP^T = DBT^TX$$  \hspace{1cm} (6.8)$$

This equation implies that if the DOS predicted through data mining compares well with the DFT calculated DOS, then it is proven that material descriptors can be derived from the DOS. The larger implication is that this equation accurately approximates the complex quantum mechanics of DFT calculations, and therefore demonstrates that data mining can be used to simplify complex mathematics with limited loss of information.

### 6.3.2. Modeling Spectral Data

The number of input DOS spectra must be sufficient for creating a robust classification scheme; that is, the inclusion of an additional DOS curve should not result in a change in the spectral patterns or the scores values which are used to scale the patterns. Once clear patterns relating DOS to changes in chemistry and structure are identified, PLS is used to predict the scaling of the patterns due to the changes in conditions.

DFT requires an input of reduced atom position and chemistry, and results in electronic structure output such as DOS. This DOS then becomes an input into training the data mining model which then outputs multiple DOS curves. The mathematics of the data
mining methodology are much less computationally rigorous than the DFT approach, but as the output is the same, the data mining methodology mathematics approximates the DFT methodology, at least for the calculation of DOS curves. This result suggests that a mathematical description of limited complexity can be created to approximate quantum mechanical equations.

The concept of analyzing input data with data mining techniques, modifying the data to reflect changes in conditions, and then reversing the data mining analysis to create an output in the form of the input data while reflecting the physical changes is unique. Since the mathematics of PCA have reasonable computational requirements, the mathematical process can be reversed so that the pre-determination of PCA outputs allows for the modeling of complex electronic descriptions for new systems through identification of relationships between atomic and electronic length scales. Not only is this logic new, but the application of data mining techniques to the DOS spectra has never been performed before.

The determination of the differences between scores values based on changes in conditions is the step that creates a real link between quantum mechanics and chemistry or crystal structure, as differences in chemistry description are linked to changes in electronic structure. Therefore, the predicted scores values of new chemistries approximates numerous steps in the DFT mathematics, and creates electronic structure-chemistry-property relationships.

The logic for predicting a PCA scores value may not seem obvious, as the PCA axes which are being modeled do not necessarily have a clear physical meaning. However, these axes are comprised of a combination of descriptors, so the difference in predicting the PCA scores value as compared to predicting a single property is minor as a combination of descriptors is being predicted instead of just one descriptor. Therefore, the prediction of a few scores values is analogous to predicting the intensity of the DOS curve at all of the hundreds or thousands of energy values.
6.4. Methodology Minutia

The purpose of this section is to go into the details of the steps from Figure 6.2 and to provide the background information so that the implications of the results can be fully understood.

6.4.1. Training DOS Organization

The data set used for the classification of the electronic structures contained only the spectral values of the density of states (DOS). The calculated DOS values were combined into one worksheet, organized as shown in Figure 5.4, with rows of different systems and columns of different energy values (the horizontal axis of a DOS figure). The values are then the DOS value at the respective energy levels (the vertical axis of a DOS figure). This data set is thus organized like is done for traditional data mining analyses, with the rows being conditions (e.g. chemistry or crystal structure) and the columns being the responses, with the responses in this case being the change in electronic structure with changing condition. For each row, a plotting of the energy values (column labels) versus the respective values from the table would result in the original DOS curve for the specific system being reproduced. Therefore, this can be considered as a digitized stacking of the DOS curves for all of the systems included. Incorporating the curves into a single data set permits a quantitative comparison of the curves, while relating the comparison to the system conditions.

Since all the DOS curves are combined into one data set, it is necessary to have consistency within the data set. That is, the energy values for which values are taken should be consistent, with consistent spacing between energy values. This consistency is important because the comparison between values should be made with the same reference. Since each energy value is treated as a unique descriptor, inconsistency between energy values results in comparing systems using different sets of descriptors. Due to the uniqueness of the data type, namely data derived from quantum mechanical calculations, there is added complexity in the establishment of descriptors. While work using experimental spectra, such as FTIR curves, is simpler since the descriptors have physical meaning (e.g. in FTIR the wave numbers correspond to specific bonding) and each descriptor value is consistent between every system, for the DOS the descriptor values are not as clear since the values are not as important as the
relative values. Therefore, the actual values of the energies are not necessarily the same, and extra consideration of this concept is necessary.

Even without this consideration, the spacing between energy values must be the same. Once a point of reference is established, the separation between descriptors must be the same. If the energy of the point of reference for each system is labeled as zero eV and all energy values are shifted so that the energy values are relative to the points of reference, the energy values should then be in agreement. For the analyses shown in the next chapter, all the systems from the data sets were calculated within the same DFT program and in the same manner so that energy spacing was consistent, and no further consideration of this point was necessary. However, if calculation outputs from different sources are combined, this point becomes important. In that case, determining a way to standardize the points is required. A curve fitting which converts the DOS curves to standard axes is possible, and this has been done successfully for experimental spectral data, however this adds some approximation to the analysis as the new values are based on an averaging between contiguous points. Further consideration of this point will become necessary for future work which expands the data set by combining DOS curves from a variety of sources into a single library.

As mentioned, the determination of a point of reference for the alignment of the data is necessary, as the energy values output from the calculations are more relative than absolute since the values can be shifted during the calculation without impacting the interpretation. Additionally, some properties associated with the DOS do not correlate with any specific energy value but rather with different features. If the Fermi energy is determined to be the parameter of most importance, then all the curves should be aligned so that in the data set the Fermi energy value is the same for every column. However, other features may be the most important, resulting in the analysis being performed multiple times with different data alignments. For the analyses, the point of reference has been set to zero, all other values are shifted by the same as the point of reference, and the edges of the curves are deleted so that the number of descriptors is equal for every curve. For the analysis, the energy value of zero corresponds to a specific feature for every system, with an equal number of descriptors and consistent spacing between descriptors. This stage is necessary to capture the information of interest and to verify that there is consistency in the organization of every DOS curve.
Future consideration should be given to including an algorithm to align the data instead of manually performing this step.

A final point in the data set-up is the consideration of the normalization of the data. For some analyses, the shape of the DOS curve has been determined as being the most important consideration. To isolate the shape, the data set can be normalized by dividing the entire curve by the DOS value of the highest peak, and repeating this step for every DOS curve, resulting in every DOS curve having a maximum intensity of unity. Doing so causes the analysis to be based only on shape and not be influenced by total DOS. For the prediction of DOS curves, both shape and intensity are important and thus no normalization was performed.

6.4.2 PCA Classification

PCA was performed on the data set of DOS curves, organized as was described in chapter 5, with no normalization done as both the shape and the total DOS are both impacted by changing conditions. The number of eigenvectors (principal components) used in a PCA analysis is generally chosen so that all eigenvectors of the covariance matrix with corresponding eigenvalues greater than unity are included. For the modeling component of the work, to have a higher resolution, more eigenvectors were generally included than would be suggested by this guideline. The selection of the number of eigenvectors for the modeling was chosen so that all eigenvectors capturing more than .1% of the variance were included. In all cases, this resulted in approximately ten to twenty principal components.

A necessary consideration for the accuracy of the modeling of new DOS is the robustness of the PCA analysis. The robustness must be such that adding a new system into the analysis does not impact the existing relationships between the scores values and the loadings values. As the scores and loadings matrices are directly involved in modeling the DOS, a lack of robustness will result in error. For the single elements analyzed, ten DOS curves were sufficient for building a robust classification, as the addition of more DOS curves did not impact the existing relationships. Future work should consider utilizing a more robust technique for this stage of the methodology, but for the initial stages in developing the methodology, PCA was found to be sufficient. Figure 6.3 shows the process
used for the descriptor reduction with some example data. The full set of matrices is displayed in the appendix.

**Figure 6.3.** The process of the descriptor reduction. From an input set of DOS curves, a set of orthogonalized patterns can be determined. In total, 16 patterns are used to capture nearly all of the information of the DOS curves. Therefore, based on the sixteen patterns and the corresponding weights, all of the input DOS curves can be totally reconstructed.

### 6.4.3. Database Component Selection

Selecting the descriptors that represent the changing conditions of the samples is not trivial, nor is it concrete. By screening descriptors, dimensionality reduction techniques can be used to determine the minimum number of descriptors producing unique information. Based on classifications from a dimensionality reduction technique and with no prior assumptions, descriptor redundancy can be found through the intercorrelations in the descriptors as seen in PC-space. Therefore, using only PCA and analyzing the variability
amongst the descriptors, the necessary data can be reduced to only descriptors providing unique information. An example of this approach was presented in chapter 3, where the relationship between descriptors and the redundancy between descriptors was found.

The descriptors selected in the analysis were selected through a screening such as described in chapter 3. The screening was necessary not only to reduce the number of descriptors to a minimum number but also to reduce computational requirements and improve the predictive model by avoiding over-fitting the data.

Therefore, the selection of descriptors for this process was not trivial. Additionally, the descriptors selected can be altered. The final test of the accuracy of the descriptors is the final prediction, but by using a data mining screening method, the selection of descriptors can be greatly improved and expedited. The descriptors selected were chosen to reflect the variables discussed in chapter 1 as comprising the DOS curve. The descriptors included for this first modeling attempt were: atomic number, atomic weight, density, molar volume, number of valence electrons, atomic radius, pseudopotential radius, lattice constant, absolute electronegativity, first ionization energy, and occupancies of electronic orbitals.

6.4.4. PLS Prediction

PLS was used to predict the combination of spectral patterns for different chemistries. The prediction served to create a simple connection between chemistry and electronic structure. PLS is able to predict properties based on variables (even some of which may have only indirect impact), which collectively relate to the property of interest. PLS has the advantage over multiple linear regression because it can handle collinearity and missing data.

The PLS prediction requires two input matrices: a matrix which contains descriptors related to the input conditions and a matrix containing the values which are to be predicted, building a model between the input descriptors and the descriptor to be predicted. For the results shown in this thesis, the primary change in condition was the chemistry, so that the input matrix contained basic descriptors of the elements. A number of the systems were not included in building the model so that the accuracy of the model could be properly tested to ensure that the data was not being over-fitted. Additionally, only the systems used in
building the model were included in the PCA analysis of the previous section. The data was organized into three matrices: one matrix containing the elemental descriptors for the systems included in the model development, another matrix containing the scores matrix from the PCA analysis (the descriptors to be predicted), and a final matrix containing the elemental descriptors for the systems which the scores values are predicted.

The matrices of elemental descriptors and scores values were analyzed conjointly to build a model between PCA scores values and the descriptors, effectively linking chemistry and electronic structure. All descriptors were auto-scaled because they had different ranges and scales. An auto-scaled data matrix has a mean of zero and unit variance. Cross validation was used to choose an optimum number of latent variables (akin to PCs) for the calibration model. The NIPALS algorithm was used and the appropriate number of latent variables were chosen by the leave-one-out cross validation method. The leave-one-out method fits a PLS model to n-1 samples and makes a prediction for the value of the omitted sample [104]. This calculation is repeated for every sample in the data matrix. A few model parameters are then calculated: the root mean square error of calibration, the root mean square error of cross validation, the coefficient of determination, and the cross validated coefficient of determination. The calculation of each of these parameters is described in references [19, 105, 106]. A desirable model should have a balance between the coefficient of determination and the cross validated coefficient of determination [19], while minimizing the root mean square errors. Based on these considerations, four latent variables were selected in building the model. The development of the model should be reconsidered when changing the input data.

Having developed a model linking elemental descriptors with the PC scores values, the prediction of the scores values for different systems is possible. Although the scores value is not necessarily robust and does not have a definitive physical meaning, the prediction of scores value is justified by considering it as a combination of descriptors, and thus is akin to the prediction of multiple descriptors. Additionally, the accuracy of the final predicted DOS is found to have acceptable accuracy, giving justification to the approach. An example of part of the input and predicted data is shown in Figure 6.4.
Figure 6.4. Demonstration of the process for predicting how the patterns identified through the dimensionality reduction combine for new systems. Based on the relationship between the weights of the training systems and their elemental descriptors, the weights of new systems were predicted.

6.4.5 Construction of DOS Curves

From section 6.4.2, the loadings matrix was determined, with dimensions of descriptors (energy levels) versus PCs. From section 6.4.4, a scores matrix for new systems was predicted, with dimensions of systems (condition combinations) versus PCs. Thus, with these two matrices, the DOS for new systems can be modeled following equation 6.11.

\[ Y = U \cdot P^T \]  \hspace{1cm} (6.11)

where \( Y \) is a matrix containing the modeled DOS for new systems (dimensions of systems versus energy levels), \( U \) is the scores matrix predicted in section 6.4.4, and \( P \) is the loadings ...
matrix found in 6.4.2. As \( Y \) has the same dimensions as the input DOS matrix in section 6.4.1, the DOS for the new system based only on some training DOS and a few elemental descriptors has thus been modeled. \( Y \) can be plotted on the same axes as the original DOS, with horizontal axis of energy and vertical axis of DOS. The DOS for new systems were calculated with DFT after the prediction to validate the methodology.

6.5. Prediction Specifics

The PLS algorithm that was found to give the best results for the analysis in this thesis is non-iterative Partial Least Squares (NIPALS) [26, 29], within the framework of the mathematics presented in appendix A.3. The step-by-step process of NIPALS is as follows, where \( X \) is the predictor matrix, \( Y \) is the predicted matrix and the data is assumed scaled and centered, and \( U \) is the scores of \( Y \). The purpose of the process is to find the correlation factors between \( X \) and \( Y \) that have maximum variance.

\[
(1) \text{ take } u_{\text{start}} = \text{ some } y_j \text{ (for a single } y, u = y) \\
\text{ note: usually the column with greatest variance in } Y \text{ is chosen}
\]

In the \( X \) block:

\[
(2) \text{ calculate the weights, } w = \frac{X^T u}{u^T u}
\]

\[
(3) \text{ normalize, } w_{\text{new}} = \frac{w_{\text{old}}}{\|w_{\text{old}}\|}
\]

\[
(4) \text{ calculate } X \text{ scores, } t = \frac{Xw}{w^T w}
\]

In the \( Y \) block:

\[
(5) \text{ calculate the weights, } c = \frac{Y^T t}{t^T t}
\]

\[
(6) \text{ normalize, } c_{\text{new}} = \frac{c_{\text{old}}}{\|c_{\text{old}}\|}
\]
(7) calculate $Y$ scores, $u = \frac{Yc}{C^Tc}$

(8) Compare $t$ in step (4) with the one in the preceding iteration step. If they are equal, go to step (9), else go to step (2) and use $u$ calculated in step (7).

note: If $Y$ has only one variable, the procedure converges in a single iteration

(9) Calculate $X$ loadings, $p = \frac{X^Tt}{t^Tt}$

(10) Rescale the scores and weight

$$p_{new} = \frac{p_{old}}{\|p_{old}\|}, \quad t_{new} = \frac{t_{old}}{\|t_{old}\|}, \quad w_{new} = \frac{w_{old}}{\|p_{old}\|}$$

(11) Find the regression coefficient $\beta$, $\beta = \frac{u^Tt}{t^Tt}$

(12) Remove the present component from $X$ and $Y$ and use deflated matrices as new $X$ and $Y$ in the next component, $X = X - tp^T$, $Y = Y - tc^T$

(13) Continue with next component (step (1))

Utilizing the mathematics of the above process as well as the mathematics in the appendix, the necessary scaling between the DOS and material descriptors is found, where the scaling is the calculated values. The overall process used in the modeling of the DOS is described as follows.

**PCA: input -- training DOS spectra (form of Figure 2.3)**

(1) Decompose spectra into eigen-spectra ($P$ in Figure 4.1) and weights ($W$).

note: matrix size: $P = [A \times M]$ and $W = [N \times M]$, where $A$ is number of points in DOS spectra, $N$ is number of test DOS spectra, and $M$ is number of PCs.

note2: number of PCs selected should capture all variance in the dataset

**PLS: input -- $W$ and matrix of elemental descriptors**

(2) Select number of latent variables, with consideration of $R^2$ and $Q^2$, so that an accurate model is developed without over-fitting the data.

(3) Calculate regression coefficients $a$ and $b$, such as to maximize covariance between
predictor (training \( W \)) and predicted (test \( W \)) variables, following steps (1) – (13) in above process.

(4) Define matrix of coefficients \( A \) (size \( M \times B \), where \( B \) is number of properties used) and \( \mathbf{B} \) (size \( M \times 1 \)) to be applied to new system.

(5) Calculate the weights \( W \) for new system based on material properties.

\[
\text{ergo: } w_{mn} = a_1 \cdot \text{atomic number}_n + a_2 \cdot \text{atomic weight}_n + \ldots + a_B \cdot \text{descriptor}_B + b_1
\]

Perform calculation for all \( w \), so \( W \) for new system has \( M \) rows

(6) Multiply \( P \) from step (1) with \( W \) from step (4), resulting in matrix \( Z \) of size \( M \times A \)

(7) Calculate DOS spectra of new system by addition of all rows of matrix.

\[
\text{DOS} = Z_1 + Z_2 + Z_3 + \ldots + Z_M
\]

(8) To calculate DOS for another system using same model, go to step (4)

6.5. Summary

An approach for modeling the DOS using data mining techniques has been presented, addressing a problem that has been explored for the past half-century. This work utilizes data mining to extract complex patterns from large data sets, thus constructing a new simulation approach, namely the use of data mining as a simulation tool. Such an approach has the benefit of being computationally efficient, with the limitation being the dependency on the quality of the input data. This work also shows a data mining methodology to accurately approximate complex mathematics in a much simpler manner with limited loss of information. Demonstrating that a model can be developed based on data mining which can arrive at the same results as a quantum mechanical model has implications for the possibilities of creating approximations for highly complex mathematics without losing significant information but reducing computational costs. Beyond these implications, the application of a data mining methodology for addressing a long standing problem in condensed matter physics has been introduced. Several areas for future consideration to improve the accuracy and robustness of the methodology were mentioned.
Having introduced the methodology for modeling the density of states while avoiding Schrödinger’s equation in chapters 5 and 6, in this chapter I show the results of the data mining-enabled modeling of the density of states. The results are compared to results calculated through the recursion method. This work is shown to be applicable to different crystal structures, while the implications for future design discussed.

### 7.1. Introduction

The calculation of the electronic structure of a material requires complex quantum mechanical equations and numerous approximations. Through these calculations, outputs such as density of states (DOS), band structure, charge density, optical properties, energies, and other electronic parameters are created. Due to the complexity of these calculations and the required approximations of quantum mechanical calculations, the electronic design of materials has been slow. This limitation provides the motivation for developing an approach which uses data mining techniques to mimic the quantum mechanical parameters in a more efficient manner and with fewer constraints. As a first step in this process, an approach to model the DOS without any direct use of quantum mechanical equations has been developed.

The DOS curves used for training the predictive model and validating the predictions were calculated within the density functional theory (DFT), which approximates a solution to Schrödinger’s equation by solving an equation of electron density functionals. One issue with first principles techniques is that the results are dependant on the assumptions employed and the method of integration over the Brillouin zone. To verify the applicability of the approach, the work described here was carried out on multiple data sets calculated through different approximations, with the finding that the data mining methodology is applicable to every set of approximations, provided the approximations employed are consistent for all training data.
The DOS have been calculated using the full-potential linearized augmented plane wave (FP-LAPW) method [107] within the DFT approximation and implemented within the WIEN2K code [108]. The exchange-correlation term was determined within the generalized gradient approximation (GGA) using the scheme of Perdew and Wang [109].

7.2. Data and Methodology Description

The process for modeling the DOS spectra is shown in Figure 7.1, with the various issues considered listed and the overall steps described.

**Figure 7.1.** Specific logic followed in this chapter with the issues considered listed. This figure shows the input data used, and the various steps and considerations. These steps serve to explore the feasibility of using DOS in a new design strategy.
The input DOS curves are shown in Figure 7.2, with the DOS treated as a spectral dataset, with every energy value being considered a unique descriptor. The DOS curves shown are of transition metals in their naturally forming ground state structure, which is used as the dataset to demonstrate the data mining-enabled prediction methodology. By applying a dimensionality reduction of the training set, all transition metal DOS curves can be constructed by combining sixteen patterns, with the primary six patterns shown in the figure, where $p_1$ is the pattern which most differentiates the DOS curves. Each total DOS is then the sum of each eigen-spectra multiplied by the weights, defined as the corresponding scores values.

The first example required the development of a model which can predict the DOS curve based on properties describing the chemistry. Using the input DOS curves for Ta, Ag, Cd, Pt, Rh, Hf, Os, Re, Ru, Tc, Zn, Zr and Y, the relationship between elemental descriptors (atomic number, atomic weight, density, molar volume, number of valence electrons, atomic radius, pseudopotential radius, lattice constant, absolute electronegativity, first ionization energy, and occupancy of electronic orbitals) and scores values was identified via PLS. The model was tested on W, Au, Pd, Sc, Ti and Cu, finding that the predicted DOS were in excellent agreement with the DFT calculated DOS. The results are shown in the next section.
Figure 7.2. Identification of orthogonal spectral patterns differentiating the DOS. The DOS curves at the top of the figure are for transition metals, and are used to identify the unique patterns differentiating these curves. These patterns can then be recombined in various ratios to predict the DOS of new systems, such as those shown at the bottom. These DOS curves are for different transition metals, and are used to test the accuracy of the predictions, but are not included in building the predictive model.
The data mining approaches used include principal component analysis (PCA) for the classifications and partial least squares (PLS) for the predictions. It is notable that the complex mathematics used in the DFT calculations can be approximated using two of the most basic data mining approaches, although the way these methods are employed and combined is unique and has been developed specifically for this work. Although this work involves approximating a quantity which in itself is an approximation, the DFT approach for modeling the electronic structure is among the best available, and as such the concept of attempting to predict an approximated value is justifiable.

By decomposing the curves to the orthogonalized eigen-spectra associated with the individual contributions comprising the DOS, the impact of the contributions can be predicted to construct a DOS curve capturing not only the shape but also the detailed structure. Beyond the obvious impact of being able to create DOS curves quickly for “virtual” materials, the ability to represent quantum quantities by a simple parameterization combined with properties from the periodic table permits the development of more complex electronic structure-property relationships. As an additional benefit, the approach introduced here is applicable to other types of spectral data.

7.3. Modeling of Elemental DOS

As discussed in chapter 1, DFT operates by capturing the interactions between numerous variables that comprise the DOS, and uses the knowledge of these interactions to calculate the electronic structure. The objective of this section is to model the electronic structure by identifying the same interactions, but with a different mathematical approach, namely through statistical learning instead of quantum mechanics. PLS is used to define the regression of high dimensional correlations between the patterns identified from the DOS curves and the corresponding variables. The final model is then derived from an optimization scheme relating the patterns to the variables. As the predictive approach used captures the co-linearity of the data, the interactions between the variables is able to be represented.
While the approach uses two techniques which are considered linear by statistical learning standards, the final model is still highly complex. The final model used to predict the DOS curves in this section is comprised of 2.4 million terms and is shown in the following equation.

\[
DOS = \sum_{i=1}^{16} \sum_{j=1}^{10,000} (w_i e_{i,j})
\]  

(7.1)

where \(w_i\) is a function of the fifteen elemental descriptors and \(e_{i,j}\) is an individual point within one of the eigen-spectra identified through the dimensionality reduction. Therefore, each weight is predicted and has fifteen terms and then is multiplied by a spectral pattern of 10,000 terms, meaning the prediction of each component is based on an equation of 150,000 terms. The DOS is then the sum of sixteen components each predicted in the same manner. Thus, the model built to predict the DOS is deceptively complex.

To understand the impact of the approach, it must be compared to previous approaches. For this reason the results are compared with the recursion method of Haydock. Haydock’s method was described in chapter 5, but is summarized again here. By recognizing that the local DOS is related to the Green function, Haydock was able to accelerate the accuracy of the calculations. The recursion method makes use of a tight binding Hamiltonian matrix, which is based on the local atomic environment, thus permitting consideration of surfaces or isolated defects. From this Hamiltonian matrix, the basis coefficients can be defined using a recursion formula, and using these coefficients, the Green’s function can be expanded to an infinite continuous fraction. The amount of information necessary, as well as the consideration of resolution versus computational requirements, dictates to what extent the Green’s function is to be expanded.

The development of the Green’s function from basis coefficients involves more approximations and truncations than the data mining approach, which does not require choosing between accuracy and computational requirements. With the mathematics employed in the data mining prediction, all information from the input data can be captured
and utilized to predict DOS curves of a greater resolution than feasible with the recursion method.

With the optimum combination of patterns known for a few systems from the training data, weights for new systems can be predicted based on the co-variance existing between weights and a few descriptors known for every element from databases such as the Linus Pauling File [110]. These predicted weights can then be combined with the corresponding patterns to construct the DOS for the new system. In cases where the basic descriptors for a system are not available, such as if an impurity is introduced, then the trends between the scores and loadings matrices need to be considered. That is, the difference in the dimensionally reduced space between the ideal system and the system with the impurity can be translated to other systems to accurately predict the DOS for the non-ideal case in a high-throughput manner.

Using the predicted combination of the patterns partially shown in Figure 7.2 (a total of sixteen patterns were used), the DOS of the seven test samples was constructed. Figure 7.3 shows the DOS of Cu, with the data mining enabled prediction plotted with the DFT calculated DOS and the DOS predicted via the recursion method [111, 112]. The data mining predicted DOS is in remarkable agreement with the DFT calculated DOS, containing a high resolution, while additionally being less computationally expensive than the other approaches while remaining applicable to a wide range of systems. The result of Figure 7.3 is representative of all of the other predictions for the validation DOS curves.

To test the robustness of the approach, the DOS curves of elements in various crystal structures were modeled. FCC structure (Cu) is shown in Figure 7.3, HCP structure (Ti) is shown in Figure 7.4, and HCP structure (W) is shown in Figure 7.5. From these curves, the FCC structure was found to be the prediction with the highest accuracy, with the reason likely being the strong similarity with the DOS curve of Ag and the similarity in variance between the elemental descriptors. A larger initial dataset is required if any physical significance to the accuracy in prediction of the FCC structure is to be identified. This difference in accuracy demonstrates a possible application of this methodology, namely the critiquing of different DFT calculations. Figure 7.6 shows the first four predicted patterns of
the W DOS, where in general the lower PCs capture less of the general shape but more fine structure.

**Figure 7.3.** Prediction of DOS of Cu (FCC) with data mining, as compared with DFT and the recursion method. From this figure, it is seen how the recursion method accurately captures the general trends without capturing the fine structure, which is captured via data mining-enabled prediction. Once the data mining model is developed, all additional applicable predictions take minimal time. For this example, the model is only applicable for transition metals.
Figure 7.4. The modeled DOS of Ti (HCP) with data mining as compared to DFT.

Figure 7.5. The modeled DOS of W (BCC) with data mining as compared to DFT.
Figure 7.6. Modeled DOS of W with the first four predicted patterns. Sixteen total patterns exist, with the total DOS being a sum of these patterns. Each eigen-spectra is found by multiplying the loadings matrix from the dimensionality reduction of the training DOS with the predicted weights. The first PCs are primary in defining the shape of the curve with lower PCs capturing the fine structure missing from the recursion method modeling.

While this methodology has been tested on a simple system, it is applicable to any system. The system analyzed dictates the organization of the training data, as the contribution, such as surface impurity or elevated temperature, needs to be represented in a training DOS so that the pattern related to that contribution can be extracted and applied to other systems to allow for the prediction under these conditions. Multicomponent systems require the introduction of geometrical parameters representing changing crystal structure, akin to entering space group into the DFT calculations. While first principles methods may require modifications such as coupling with resolvent elements or molecular dynamics to calculate some contributions, the data mining approach does not require any additional steps in calculation once the model has been developed, thus permitting the calculation of numerous systems under abnormal conditions in a matter of minutes.
Of scientific note is the approximation of the quantum mechanical values. Even with the mathematical complexities involved in first principles methods with the approximation of Schrödinger’s or Dirac’s equations, the results associated with different conditions follow a clear pattern, which had never been identified before the inclusion of data mining techniques. The identification of these patterns in the calculation outputs, combined with the approximation of quantum mechanical values permits the rapid calculation of DOS curves. While other approaches represent the fundamental quantum mechanical equations through complex mathematics, the avoidance of any direct quantum mechanical considerations of the data mining approach permits rapid prediction with no limitations on the systems that can be represented.

The primary consideration in the data mining methodology is in regards to the training data, as the predicted results are limited in being able to improve on the quality of the training data. While spectra from any calculation type can be considered, if the input data has low resolution, the predicted DOS will also have low resolution. Additionally, the amount of input data should be large enough to form a robust classification scheme. That is, the classification aspect of the work must be robust enough so that any additional DOS curves included do not significantly affect the eigen-spectra patterns considered. On-going research is considering the development of additional data mining tools to address these considerations.

7.4 Modeling of Multi-component DOS

The results of the classification of the oxide systems in chapter 4 can be used to model the DOS for other oxide systems by combining the PCs which were shown to have physical significance. The logic for doing this for the oxide systems is shown in Figure 7.7, where the PCs related with chemistry and the PCs related with structure are combined so as to represent a new system.

The number of DOS curves that can be constructed is equal to the number of contributions represented in some form in the training set of DOS curves. There are thirteen DOS curves in the training set calculated through DFT. However, as the contributions of
crystal structure and stoichiometery have been identified, the data set can be extended to a much greater number. From CeO$_2$ in cubic structure, the DOS of CeO$_2$ can be constructed in tetragonal and monoclinic, as well as CeO and CeO$_3$ stoichiometries. This methodology cannot be used to predict the DOS curve for some chemistry that is not represented in any manner in the training set.

**Figure 7.7.** This figure presents the logic of modeling multicomponent systems based on the classification results of chapter 4. By identifying which conditions are reflected in the various PCs, a new DOS curve may be created based on the desired conditions.

It should be noted that although PCA is the only technique used to make this prediction, PCA is not a predictive technique, as the mathematical basis limits it to dimensionality reduction. However, as has been discussed, values are not being predicted with PCA but rather PCA is used to extract patterns from a data set, and then these patterns are combined in different manners.
A search of existing calculations shows a lack of quality calculated electronic structure data for a large number of oxides, emphasizing the power of the approach discussed here. Specifically, the possible systems that can be modeled with the training data are not sufficient to use as validation of the approach. To that end, ZrO$_2$ in tetragonal structure has been removed from the training data set and instead used for validating the prediction. The analysis done in the previous section was repeated without ZrO$_2$ in tetragonal structure to capture any change in crystal structure patterns resulting. It is found that the patterns of the crystal structures were similar, and the differences between chemistries within cubic and monoclinic structures were not altered significantly.

The ZrO$_2$ cubic DOS was then deconstructed into its orthogonalized components, and the cubic pattern was replaced with the tetragonal pattern. The results are shown in Figure 7.8, with the results compared with the DFT calculated data. The results are in reasonable agreement, at least to the point where qualitative conclusions will be the same. The major source in error is clearly defining the PC scores value for each structure, because as was shown in chapter 4, the structure and chemistry are not totally isolated.

![Figure 7.8. Predicted DOS for ZrO$_2$ in tetragonal structure. DOS curve created through data mining approach has reasonable agreement with DFT calculation, while being substantially less computationally expensive.](image)
This result demonstrates that patterns clearly exist within the DOS, and the DOS is a sum of these patterns. Therefore, although the DFT calculations entail complex mathematics, clear relationships exist between physical conditions and the patterns within the DOS. This interpretation provides confidence that the data mining methodology discussed in the previous chapters is extendable to multicomponent systems.

7.5. Developing a “Virtual” DOS Curve

The larger implication of this work for future alloy design is modeling the DOS curve for systems which have not yet been discovered. Other approaches for modeling the DOS, such as DFT, require the input of accurate material parameters in order to calculate the electronic structure. A benefit of the approach discussed in this thesis is that the electronic structure can be modeled based on any parameters input and does not require knowledge of any pair potentials. Therefore, if a certain value of some parameter is desired, the DOS for a system with that parameter can be predicted. Additionally, the PLS model can be amended so as to include terms related to physical properties. In that regard, the DOS of a material can be predicted with a change in the property reflected, and the change in electronic bonding responsible for the optimized bonding may be determined. Once the methodology is extended to multi-component systems, the DOS of an optimized system can be calculated, and the responsible electronic bonding can be identified. Therefore, this approach has implications for alloy design by creating a linkage between chemistry, crystal structure, electronic bonding, and properties.

7.6. Summary

A methodology has been developed for predicting the density of states curves using the data mining logic presented in chapter 6. The actual model developed for the prediction in this chapter contained 2.4 million terms so to capture the complexity of the problem. As this approach can lead to the calculation of DOS curves in a matter of minutes, it represents a high-throughput approach to calculating the DOS. The approach operates by identifying
clear patterns existing within quantum mechanical calculations and finding a relationship between quantum mechanical parameters and basic materials descriptors. This approach leads to a prediction capturing not only the general trends in the DOS, but the fine structure as well. Another demonstration following a different logic for oxide systems was presented. Therefore, I have demonstrated that the DOS can be modeled for different crystal structures and bondings.
CHAPTER 8: CONCLUSIONS

8.1. Summary

The primary objective discussed in this thesis was the modeling of the density of states (DOS) curve without directly considering quantum mechanical equations. Utilizing a carefully designed data mining methodology, the DOS of elements was modeled accurately based only on a database component and an input set of known training DOS curves. As this is a long standing problem in condensed matter physics, a comparison with other previous approaches to modeling the DOS was provided, most notably the recursion method pioneered by Haydock. While the approach introduced in this thesis and the recursion method follow similar logics, by first converting the data into a dimensionally-reduced space prior to orthogonalizing the data, the accuracy of the data mining approach exceeded that of the recursion method.

For the initial stage, only the DOS of single elements were modeled through data mining; however, a classification of DOS spectra of different types of multicomponent systems was performed. I demonstrated the ability of statistical learning classification methods to detect structural and stoichiometric classification of compounds. This detection gives confidence that the methodology can be extended to alloy systems.

The other portion of this thesis applied statistical learning to semi-empirical calculations. In the first example, the parameters necessary for performing the calculation were predicted for systems which the parameters are currently unknown. The calculation was then performed using these parameters for III-V Bi semi-conductors with reasonable agreement between the results possible with data mining and the limited number of results in literature.

The other example presented in this thesis was the analysis of parameters or descriptors from a semi-empirical calculation for the purpose of minimizing the number of parameters necessary, thus reducing the computational requirements and simplifying the data analysis. In this example, I showed that 21 descriptors could be reduced to five with minimal loss of information.
By exploring a diverse array of materials systems (metals, ceramics, different crystal structures) this work has laid the foundations for expanding the linkages between statistical learning and statistical thermodynamics. The results of this work provide exciting new opportunities in computational based design of materials that have not been explored before.

8.2. Future Work

As this thesis addressed new issues in combining statistical learning and electronic structure calculations, there are numerous future applications for this work. The future direction of the work was discussed in chapter 7, but other possible directions are briefly listed here.

First, a relationship linking the DOS curves for elements with the DOS curves for the alloys could provide a greatly accelerated design strategy for virtual alloys. While stoichiometrically the addition is simple, the change in DOS is not so clear. Once new DOS curves have been modeled following the approach discussed in this thesis, the property information from the modeled DOS curves can be extracted, which can then be used to predict the properties of a virtual material based on its modeled DOS.

The number of DOS curves used in the classification portion of the thesis can be greatly expanded so that design pathways can be identified within the classification space. Identification of such pathways can suggest ternary solute additions to the binary intermetallics, providing a new approach for identifying proper alloying elements. An example of this concept is in ceramics where dopant additions of chemistry can significantly change properties, for example due to grain boundary chemistry. This work can also be expanded to lower symmetry systems such as monoclinic or triclinic systems or other systems with high anisotropy.

The logic of the semi-empirical analysis can be applied to problems where pair potentials are limited or are imprecise so that the work may have a significant impact on both semi-empirical and DFT calculations. The work here should be applicable to any electronic structure calculation, so this work can be expanded to other calculation methods.
Beyond these future applications, other steps to improve this work are necessary. First, while the modeling of the DOS for elements was shown to be highly accurate, other approaches to this problem using data mining can be explored which may allow for a more robust analysis. The modeling of spectral data via support vector machine (SVM) for example should be further considered. SVM may allow for more robust modeling of the shape of the curve, with techniques such as partial least squares (PLS) being used to scale the curve and shift the axes. Also, all the data mining approaches used in this thesis are linear techniques by statistical learning standards, and as the data considered in this thesis is of high complexity, non-linear techniques such as kernel-PCA and kernel-SVM should be considered.

As the horizontal axis of the DOS curve is not easily compared between systems, a method for better aligning the data to compare the systems should be considered. In this thesis, the alignment was done manually based on the specific objective. A more systematic manner of doing this alignment should be developed.

A final objective is to better understand how properties are represented in the DOS curve. In the previous chapter, the concept of building a DOS curve based on an optimized set of properties was discussed. However, a classification method should also be capable of finding a linkage between features in the DOS curve and the properties. In either case, a better understanding of the physics involved and the relationship between electronic bonding and properties is needed.
A.1. Mathematics of PCA

To thoroughly describe PCA in an effective way, the treatment from different sources are combined here [113-116]. Let us consider the case of a vector \( x \) of \( p \) number of variables. With \( \alpha_1^T = [\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p}] \), the variance of the linear function \( z_1 = \alpha_1^T x \) is maximized in PCA. The linear function, \( z_2 = \alpha_2^T x \) which is uncorrelated with \( z_1 = \alpha_1^T x \), can then be calculated to capture the remaining variance. Therefore the \( k \)-th linear function, \( z_k = \alpha_k^T x \), is calculated to have maximum variance and to be uncorrelated with \( z_1 = \alpha_1^T x, \alpha_2^T x, \ldots, \alpha_{k-1}^T x \). Consider the case where the vector of random variables \( x \) has a known covariance matrix \( S \). \( \alpha_k \) is an eigenvector of covariance matrix \( S \) corresponding to its \( k \)-th largest eigenvalue \( \lambda_k \). If \( \alpha_k \) is chosen to have unit length (\( \alpha_k^T \alpha_k = 1 \)), then the variance of \( z_k \) is \( \text{var}(z_k) = \lambda_k \). To populate the first projection vectors \( \alpha \) in \( z_1 = \alpha_1^T x \), PCA finds maximum variance, such that

\[
\alpha_1 = \arg \max_{\alpha_1^T \alpha_1 = 1} [\text{var}(\alpha_1^T x)] = \arg \max_{\alpha_1^T \alpha_1 = 1} [\alpha_1^T S \alpha_1] \quad (A.1)
\]

With the constraint of unit length of \( \alpha \) and maximum variance of \( z_1 \), the method of Lagrange multipliers can be applied as

\[
\max(L) = [\alpha_1^T S \alpha_1 - \lambda(\alpha_1^T \alpha_1 - 1)] \quad (A.2)
\]

where \( \lambda \) is a Lagrange multiplier. Since differentiation gives the maximum value, equation (A.2) results in

\[
(S - \lambda I_p) \alpha_1 = 0 \quad (A.3)
\]

where \( I_p \) is a \((p \times p)\) identity matrix. This is known as the problem of eigenstructure for the covariance matrix. To avoid a trivial null solution, \((S - \lambda I_p)\) should be zero. \( \lambda \) and \( \alpha_1 \) should be an eigenvalue of \( S \) and the corresponding vector respectively. Therefore, the eigenvalue \( \lambda \) represents the variance because:

\[
\text{var}(\alpha_1^T x) = \alpha_1^T S \alpha_1 = \alpha_1^T \lambda \alpha_1 = \lambda \quad (A.4)
\]

Since variance should be maximized in PCA, the eigenvalue \( \lambda \) must be as large as possible. The vector \( \alpha_1 \) is the eigenvector corresponding to the largest eigenvalue \( \lambda_1 \) of \( S \). A graphical
representation of the eigenvectors and eigenvalues and the assignment of PCs is shown in Figures A.2 and A.3. The second principal component maximizes the variance.

\[ \alpha_2 = \arg \max_{\alpha \in \mathbb{R}^n} \alpha_2^T S \alpha_2 \]  

subject to the constraint, \( \text{cov}(\alpha_i^T x, \alpha_2^T x) = 0 \). Thus, it should be uncorrelated with \( z_1 = \alpha_1^T x \). Using the method of Lagrange multipliers,

\[ \max(L) = [\alpha_2^T S \alpha_2 - \lambda(\alpha_2^T \alpha_2 - 1) - \phi(\alpha_2^T \alpha_2 - 0)] \]  

where \( \lambda \) and \( \phi \) are Lagrange multipliers. The following relations result in \( (S - \lambda I_p)\alpha_2 = 0 \). The vector \( \alpha_k \) is called the loadings for the \( k \)-th principal component (PC). The algorithms for calculation of principal components are mainly based on the factorization of matrices. Singular vector decomposition (SVD) and eigenvalue decomposition are the main techniques for factorization of matrices. For any \((I \times I)\) matrix \( A \) and \( P \) which are non zero orthonormal matrices, the eigenvalue problem can be expressed as

\[ AP = \Lambda P \]  

where \( \Lambda \) is an eigenvalue matrix and its components are \( \Lambda = \text{diag}\{\lambda_1, I \} \). Then matrix \( A \) by eigenvalue decomposition is

\[ A = P \Lambda P^T = \sum_{i=1}^I \lambda_i p_i p_i^T \]  

Here, the property \( P^T P = P^{-1} \) was used from the fact that \( P \) is orthonormal. If a covariance matrix \( S \) of \( X \) is a matrix \( A \), the data manipulation involves decomposition of the data matrix \( X \) into two matrices \( V \) and \( U \), and \( V \) is orthonormal,

\[ S = X^T X = VU^T U V^T = V \Lambda V^T \]  

The columns of \( U \) are known as scores and those of \( V \) are called loadings. PCA is a technique to decompose eigenvalues of a covariance matrix, \( S \), of a given data matrix. The loadings can be understood as the weights for each original variable when calculating the principal components. The matrix \( U \) contains the original data in a rotated coordinate system. The mathematical analysis involves finding these new “data” matrices \( U \) and \( V \). The dimensions of \( U \) (i.e. its rank) that capture all the information of the entire data set of \( X \) (i.e. # of variables) is far less than that of \( X \) (ideally 2 or 3). One now compresses the \( N \) dimensional plot of the data matrix \( X \) into 2 or 3 dimensional plot of \( U \) and \( V \). While the eigenvalues
geometrically represent the length of each of the principal axes (i.e. scores), the eigenvectors of the covariance matrix represent the orientation of principal axes of the ellipsoid (i.e. loadings). By using just a few latent variables, the dimensionality of the original multivariate data sets are reduced and visualized by their projections in 2D or 3D with a minimal loss of information. Therefore, PCA is a process of dimensionally reduced mapping of a multivariate data set.

**Figure A.1.** A graphical representation of the data points and their eigenvalues
Figure A.2. Determination of two principal components (PC1 and PC2) in a new scaled coordinate, $x_1$ and $x_2$.

### A.2. Mathematics of PLS

Partial least squares (PLS) finds the maximum variance in the predictor variables ($X$) and finds the correlation factors between $X$ and the predicted variables ($Y$) that have maximum variance. In PLS, two linear combinations are generated from the $X$ and $Y$ respectively and the maximum covariance between $X$ and $Y$ is calculated. Consider an $X$ matrix of size $N \times K$ and an $N \times M$ matrix $Y$.

The following descriptions are mainly based on [26, 29, 30, 116]. The scores of $X$, $t_a$ ($a=1, 2, \ldots, A=$ the number of PLS components) are calculated as linear combinations of the original variables with the weights $w_{ka}^*$. The mathematical expression is

$$ t_a = \sum_k w_{ka}^* x_{ik} \quad \text{or} \quad T = XW^* $$  \hspace{1cm} (A.10)

where $k=(1, \ldots, K=$ the number of $X$ variables). The predictor variables, $X$, are expressed as:

$$ x_{ik} = t_{i1}p_{ik}^T + t_{i2}p_{2k}^T + L + t_{ia}p_{Ak}^T + e_{ik} = \sum_a t_{ia}p_{ak} + e_{ik} \quad \text{or} \quad X = TP^T + E $$  \hspace{1cm} (A.11)

where $e_{ik}$ is the $X$ residuals.
Similarly, for predicted variables $Y$, if the scores of $Y$ are $u_a$ and the weights $c_{am}$:

$$y_{im} = \sum_a u_a c_{am} + g_{im} \text{ or } Y = UC^T + G$$

(A.12)

Since scores $X$ are good predictors of $Y$ in PLS, then:

$$y_{im} = \sum_a t_{ia} c_{am} + f_{im} \text{ or } Y = TC^T + F$$

(A.13)

where $F$ represents the error between observed values and the predicted response. Using equation (A.10), the equation (A.13) is also expressed as

$$y_{im} = \sum_a c_{am} \sum_k w_{ka} x_{ik} + f_{im} = \sum_k b_{mk} x_{ik} + f_{im} \text{ or } Y = XW^*C^T + F = XB + F$$

(A.14)

From equation (A.14), the PLS regression coefficients $\beta_{mk}$ is written as

$$\beta_{mk} = \sum_a c_{am} w_{ka} \text{ or } B = W^*C^T$$

(A.15)

Geometrically, all the above parameters are shown in Figure A.3. As discussed before, the multidimensional space of $X$ is reduced to the $A$-dimensional hyper plane. Since the scores are good predictors of $Y$, the correlation of $Y$ is formed on this hyper plane. As in PCA, the loadings of $X$ ($P$) represent the orientation of each of the components of the hyper plane.

According to the approach of Phatak and de Jong, after $n$ dimensions have been extracted the following equations are available.

$$T_n = XW^*, \quad P_n = X^T (T_n^T T_n)^{-1}, \quad W^* = W_n (P_n^T W_n)^{-1}$$

(A.16)

The prediction of $y$ then has a general form given by equation (A.7)

$$\hat{y}_{PLS} = T_n^T y$$

(A.17)

From the equations (A.16) and (A.6), equation (A.17) is written as:

$$\hat{y}_{PLS}^n = X\hat{\beta}_{PLS}^n = XW^* (W_n^T X^T X W_n^*)^{-1} W_n^T X^T \hat{\beta}_{OLS}$$

(A.18)
A.3. Schematic of Modeling the Density of States

The diagrams in Figure A.4 show the matrices used in three different portions of the methodology for modeling the DOS. In the training DOS, the values are equal to DOS, and the values in the scores and loadings matrices are scores values and loadings values, respectively. This serves to provide a numerical example of the work and to better show the composition of the matrices. These figures were partially shown in chapter 6.
Scores (Y)

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ta</td>
<td>-60.23</td>
<td>0.0342</td>
<td>-11.04</td>
<td>9.776</td>
<td>-3.842</td>
<td>6.767</td>
</tr>
<tr>
<td>Ag</td>
<td>-23.412</td>
<td>-124.31</td>
<td>-4.662</td>
<td>-5.205</td>
<td>2.187</td>
<td>0.746</td>
</tr>
<tr>
<td>Pt</td>
<td>-81.043</td>
<td>-1.0138</td>
<td>15.402</td>
<td>8.541</td>
<td>-12.789</td>
<td>0.2521</td>
</tr>
<tr>
<td>Rh</td>
<td>-83.863</td>
<td>3.491</td>
<td>15.770</td>
<td>6.592</td>
<td>4.3887</td>
<td>18.627</td>
</tr>
<tr>
<td>Hf</td>
<td>-129.423</td>
<td>0.055</td>
<td>-3.689</td>
<td>-21.10</td>
<td>-33.39</td>
<td>6.3093</td>
</tr>
</tbody>
</table>

Database Component (X)

<table>
<thead>
<tr>
<th>at #</th>
<th>at wt</th>
<th>density</th>
<th>mol volume</th>
<th>valence</th>
<th>e #</th>
<th>atomic rad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ta</td>
<td>73.0</td>
<td>180.9</td>
<td>16.6</td>
<td>10.9</td>
<td>5.0</td>
<td>147.0</td>
</tr>
<tr>
<td>Pt</td>
<td>78.0</td>
<td>195.1</td>
<td>21.5</td>
<td>9.1</td>
<td>10.0</td>
<td>139.0</td>
</tr>
<tr>
<td>Rh</td>
<td>45.0</td>
<td>102.9</td>
<td>12.4</td>
<td>8.3</td>
<td>9.0</td>
<td>134.0</td>
</tr>
<tr>
<td>Au</td>
<td>79.0</td>
<td>197.0</td>
<td>18.9</td>
<td>10.3</td>
<td>11.0</td>
<td>144.0</td>
</tr>
<tr>
<td>W</td>
<td>74.0</td>
<td>183.9</td>
<td>19.4</td>
<td>9.6</td>
<td>6.0</td>
<td>141.0</td>
</tr>
<tr>
<td>Cu</td>
<td>29.0</td>
<td>63.5</td>
<td>8.9</td>
<td>7.2</td>
<td>11.0</td>
<td>128.0</td>
</tr>
</tbody>
</table>

PLS (equation A.18)

Predicted Scores (Y_PLS)

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ti</td>
<td>-149.23</td>
<td>6.6679</td>
<td>5.8010</td>
<td>4.0573</td>
<td>-3.6110</td>
<td>-6.9278</td>
</tr>
<tr>
<td>Cd</td>
<td>-60.2151</td>
<td>-6.4077</td>
<td>-10.2017</td>
<td>26.9086</td>
<td>0.241809</td>
<td>5.575461</td>
</tr>
<tr>
<td>Au</td>
<td>-53.1761</td>
<td>-76.2847</td>
<td>-31.4391</td>
<td>-1.7058</td>
<td>9.10604</td>
<td>-0.63924</td>
</tr>
</tbody>
</table>

Apply relationship
Figure A.4. Demonstration of the process for predicting the DOS of new materials. Shows a portion of the numerical contents of the matrices. (a) PCA process of converting the input DOS matrix into a scores and a loadings matrix (b) Prediction of scores values for new systems based on database component (c) Modeling of new curves by combining loadings matrix from (a) and the predicted scores matrix from (b).
A.4. Mathematics of Haydock’s Recursion Method

In chapter 2, the logic of the recursion method was provided. In this section, a brief explanation of some of the mathematics involved in the method is given. This section combines the explanations from a variety of sources [6, 88, 89, 117-124].

The first step is to tridiagonalize a tight-binding Hamiltonian. A nearest-neighbor tight-binding Hamiltonian may have the following form

\[ H_{ij} = \sum_i|\phi_i\rangle e_i \langle \phi_i | + \sum_{ij} |\phi_i\rangle t \langle \phi_j | \]  

(A.19)

with \( e_i \) being the site energies following a probability distribution and \( t \) being the hopping matrix elements between states \( \phi_i \) and \( \phi_j \). The recursion method generates a set of orthogonal functions \( u_n \), which can be used to create a tridiagonal matrix with orthogonalized elements. These functions can be thought of as projections of the Hamiltonian onto state \( n \). This step, as I suggested in chapter 5, effectively reduces the dimensionality of the Hamiltonian of equation A.19. The recursion relation is then:

\[ \hat{H} u_n (r) = a_n u_n (r) + b_{n+1} u_{n+1} (r) + b_n u_{n-1} (r) \]  

(A.20)

with \( a_n \) and \( b_n \) being the recursion coefficients. \( a_n \) and \( b_n \) have the form:

\[ a_n = \int u_n (r) \hat{H} u_n (r) dr \]  

(A.21)

\[ b_{n+1}^2 = \int |(\hat{H} - a_n) u_n (r) - b_n u_{n-1} (r)|^2 dr \]  

(A.22)

The recursion process has then created from the tight-binding Hamiltonian an orthonormal basis set \( (u_n) \) and a matrix of recursion coefficients \( (a_n \) and \( b_n \)). \( a_n \) sit on the diagonal of the matrix and \( b_n \) sits on the first off diagonal. If the matrix containing these coefficients is \( m \times m \), then \( b_n = b_{m+1} \) and \( a_n = a_m \). All other values in the matrix are equal to zero. Having defined the recursion coefficients, the recursion method operates by recognizing that the DOS

\[ n(E) = \sum_m (u_\alpha | \psi_m \rangle)^2 \delta(E - E_m) \]  

(A.23)

is related to the imaginary part of Green’s function based only on the tridiagonalized matrix. The Green’s function is expressed as a continued-fraction.
\[ G(E) = \frac{1}{E - a_o - \frac{b_1^2}{E - a_1 - \frac{b_2^2}{E - a_2 - \ldots}}} \]  
\quad \text{(A.24)}

Assuming that \( n \) in equation A.20 is less than \( i \) in equation A.19, a truncation of the continued fraction is necessary. The equation for the DOS in equation A.23 can then be rewritten as:

\[ n(E) = -\frac{1}{\pi} \lim_{\epsilon \to 0} \text{Im}[G(E + i\epsilon)] \]  
\quad \text{(A.25)}

To summarize this process as was done in chapter 5, equation A.20 converts a high-dimensional Hamiltonian matrix into a matrix which is effectively reduced in dimensions. The possible dimensions of this matrix is equal to the dimensions of the Hamiltonian matrix of equation A.19, but an \( n \) much less than \( i \) is necessary as the values for the recursion coefficients begin to converge on a single value. Thus, if \( n \) is equal to \( i \), the entire tight-binding Hamiltonian is recovered and the DOS can be calculated entirely. Equation A.24 effectively converts equation A.20 to a real space representation, and equation A.25 then converts the Hamiltonian to a spectral representation.
REFERENCES


