Summary

This thesis investigates the use of data driven models to predict the size of the band gap for multi-element crystalline compounds. Models were developed using using linear and non-linear machine learning algorithms on a data set spanning over 2458 unique compound compositions. The algorithms explored in this thesis were, Random Forest and Cubist decision trees and Support Vector Regression. The best performing model achieved a R^2 value of 0.81, a root mean square error of 0.65 eV and an average absolute relative deviation of 47%. The descriptor set included 17 element and crystal properties, all of which were easily determined without the need for prior knowledge about the specific materials other than their composition. Although the models fail to achieve the same level of accuracy as density functional theory calculations, the speed with which the predictions can be made outperforms any other available methods. The thesis investigates possible sources of error in predictions and considers the emphasis placed by the models on the various properties. Two properties stood out as being particularly important for model predictions. These are the Pauling electronegativity difference for the crystals and the average atomic weight of the constituent elements. This thesis discusses potential explanations for why this is the case. Comparison to a similar model published in the *Journal of Physical Chemistry Letters* showed similar results and suggests that further development of both the descriptor set and the band gap data are needed in order to create robust models. The findings of this thesis have implications for researchers wanting to use predictive models in targeted material design.

Preface

This thesis has been written at the Institute of Chemistry at the Norwegian School of Science and Technology as part of the requirements for the fulfillment of my Master of Science in Nanotechnology. I have benefited greatly from the dedicated work of my many professors through my years as an NTNU student.

Special thanks is due to my supervisor Per-Olof Åstrand who has been more patient with me throughout my work on this and previous projects than anyone could expect. I cannot express how thankful I am for his guidance and mentorship, both as my supervisor on multiple projects, and when working for him as a TA in his class Statistical Thermodynamics.

Special thanks also to my co-supervisor Vishwesh Venkatraman whose door has been open at all times. Thank you for our many discussions on machine learning and its many pitfalls. Understanding that machine learning is not the new alchemy, but instead should be applied critically, opened a new conceptual universe for me. Thank you for guiding my way into this exciting field.

Thank you to my many friends in AdamsEplekor, at Studentersamfundet and in my class who have made my experience as a student in Trondheim, not only educational, but also delightful.

Contents

List of Tables

List of Figures

Abbreviations

DFT	=	Density functional theory
SVR	=	Support vector regression
PLS	=	Partial least squares
VEC	=	Valence electron concentration
RMSE	=	Root mean square error
AARD	=	Average absolute relative deviation
KS	=	Kohn-Sham
SVC	=	Support vector classification
RBF	=	Radial basis function
RSS	=	Residual sum of squares
LV	=	Latent variable
SVM	=	Support vector machines
fcc	=	Face centered cubic
MPEAs	=	Multi-principal element alloys

Chapter 1

Introduction

For as long as humans have been developing new technologies, there has also been a need for the development of new materials. In today's world, this constant search for new materials continuous stronger than ever before. Consider for example this excerpt from the Materials Genome Initiative's home page[1]:

> *"Advanced materials are essential to economic security and human well being, with applications in industries aimed at addressing challenges in clean energy, national security, and human welfare, yet it can take 20 or more years to move a material after initial discovery to the market. Accelerating the pace of discovery and deployment of advanced material systems will therefore be crucial to achieving global competitiveness in the 21st century"*

The ability to quickly design, from the bottom up, new materials with predictable properties would surly revolutionize both the way we search for new materials and how we develop new technologies.

One of the most fundamental properties of a functional material is the band gap. The size of the band gap determines a host of other properties, including electrical conduction and optical transparency[2, p. 8, ch. 10]. The application of materials with finely tuned band gaps includes areas like photovoltaic cells, light emitting diodes, microchip processing, etc[2, p. 8, ch. 10][3]. Predicting the band gap of a given composition is, however, a time consuming and often time expensive task[4]. Although precise experimental determination of band gaps is possible, an experimental approach makes material design challenging because of the vast possible composition space to explore. Computational prediction methods are often slow and not always accurate enough to precisely predict the band gap. One of the most popular approaches, based on density functional theory (DFT) investigations, is dependent on the correct choice of computation parameters that often are system specific[5]. This makes DFT a difficult approach to generalize. Clearly there is a need for faster approaches that make it possible for researchers to quickly determine a likely range for the band gap on any previously unexamined composition or class of materials.

The advent of machine learning has shown great promise in several fields of data science [6]. These methods provide the tools necessary for sifting through large data sets and extracting information about correlations that would otherwise have been impossible for researchers to spot. Patterns that the algorithms find in the data can be generalized and, when applied to unseen data, make predictions about likely outcomes. Examples of machine learning include a wide range of applications, such as face recognition, marketing, financial fraud detection and medical research[7, 8, 9, 10].

When applied to material science, machine learning might hold the key to quickly predicting properties of new materials. The central problem addressed by this thesis is whether it is possible to develop a model that sufficiently and accurately predicts band gap values for unseen material compositions. For this type of approach to be useful, the input parameters of such a model must be either easily accessible from already existing sources, or easily calculated. Information that implies already existing knowledge about the specific compound, such as structure information or other crystal properties which must be experimentally determined, like material resistivity or optical transparency, are therefore excluded.

The machine learning approach is also interesting from a theoretical point of view. Although the exact model developed by a machine learning algorithm functions a bit like a black box, the variables that are included in the model are weighted by level of model importance. The mechanism giving rise to the band gap is discussed in Section 3.4, but this explanation relies heavily on the structure of crystals. Creating a framework for understanding how more basic element properties like atomic radius and ionization energies might influence crystal properties would be interesting. This thesis provides some insight into this problem as well.

Zhuo et al.[11], in their paper *Predicting the Band Gaps of Inorganic Solids by Machine Learning*, explore this approach to predicting the band gap using a similar set of limitations. The work of this thesis builds on their work in an attempt to further our understanding about the possibilities and limitations of their approach. Zhuo et al.[11] focus their main attention on using support vector regression (SVR) to create the models. They also limit themselves to element properties and do not consider easily determined crystal properties.

This thesis explores several additional possible machine learning algorithms. The primary focus is on decision trees. Both the random forest algorithm and the cubist algorithm are tested. In addition, a linear approach using partial least squares (PLS) regression and the non-linear SVR are implemented. The set of element properties explored is a reduced version of that used by Ya Zhuo, et al.[11]. Instead, a new set of 4 compound properties is included. These are the valence electron concentration (VEC), atomic size difference, electronegativity difference and a configurational entropy approximation. These properties showed promise when applied to a related machine learning problem by Islam et al.[12].

The results indicate that machine learning performance can be drastically improved by reducing general applicability. However, pure machine learning based methods still have a long way to go before they can be reliably used in materials development efforts. Further, great care must be taken when developing these models such that overfitting does not occur. Machine learning can only perform as well as the data it is given. For the best results, this data should be as diverse and extensive as possible. Experimental or other

error sources in the data will drastically effect model performance. The best performing machine learning algorithm achieved a R^2 value of 0.81, root mean square error (RMSE) of 0.65 eV and average absolute relative deviation (AARD) of 47%.

Two of the properties that most heavily influenced the model prediction were atomic weight and the electronegativity differences. Atomic weight is likely important because it provides indirect information on the crystal structure. Electronegativity differences are likely important because they provide information about the type of bonds present in the structure. Both are key features in the crystal description of the origin of band gaps.

Chapter 2

Literature Review

Donald Michie writes in a 1968 nature memo *Functions and Machine Learning* that *"If computers could learn from experience their usefulness would be increased."*[13]. Since then, the development of machine learning algorithms have fundamentally changed the approach to data driven science.

Early work in machine learning focused on the development of neural network systems that could do simple pattern recognition. This includes algorithms like ADELINE and MADELINE developed by Bernard Widrow and Marcian Hoff, the second of which is still in use today to limit echo on phone calls[14]. In more recent times, several machine learning algorithms have been developed and applied to a large variety of fields, including face recognition, marketing applications, financial fraud detection and medical research [7, 8, 9, 10].

The last two decades of the 20^{th} century marked a rapid development of the chemoinformatics field because of the technological advancement of networks and computer systems combined with the creation of algorithms that are capable of dealing with chemical structures[15]. Using computer systems in chemical research made it possible to store, process and visualize chemical information in a way not previously possible. Combining the tools of chemoinformatics with machine learning algorithms lets researchers take a data driven approach, generating generalized descriptive models that can predict secondary properties from more basic descriptors[16].

Applying similar strategies for solid state materials as for chemical systems has taken longer than it did for chemical systems, even though there are several similarities. This is, in large part, because of the complexity of representing large solid state systems combined with the computational cost as compared to smaller molecules[17]. The development of computational methods such as DFT, Monte Carlo simulations and molecular dynamics, greatly improved the speed of materials design[4]. Examples include catalyst development for greenhouse gas conversion and materials discovery for energy harvesting and storage[18].

Machine learning as a tool for prediction in material science is a relatively recent development. The field faces some pretty serious challenges though. Finding large enough

high quality data sets for machine learning to reach its full potential is key[4]. Despite such hurdles, there have been several successful machine learning applications to material science, ranging from predicting phase stability to material properties and in combination with first-principle calculations[4]. Some examples include:

- Training a neural network to effectively distinguish chemical elements based on the topology of their crystallographic environment[19].
- Combining classification and regression machine learning to search for perovskites with high ferroelectric Curie temperature[20].
- The combination of machine learning and DFT first principle calculations in the search for new fast ion conductors for rechargeable battery applications[21].
- Using easily calculable crystal structure descriptors with a neural network to predict phase stability in multi-principal element alloys[12].

The advent of computational methods for predicting the properties of unseen materials has already changed the way researchers design materials. With modern crystal-growth techniques, crystal structures can be grown with very specific composition and structure[22]. If this ability is combined with a predictive ability, it is, in principle, possible to develop materials with very specific functional properties. Kiselyova et al. illustrated this approach already in 1998, using pyramidal networks to find new magnetic and electro-optical materials with specific crystal structures[23].

A key feature of crystals is the size of the band gap, the energy gap that electrons must excite in order to conduct electricity. Historically, accurately determining the electronic structure properties of crystals has been either computationally very expensive, inaccurate, or a combination of both[4]. For now though, DFT calculations as a tool for property predictions is a central part of the modern solid state physics tool box[24]. A central part of DFT calculations is the choice of exchange-correlation functional. This will effect both calculation time and performance. Borlido et al. have assembled a data set consisting of 472 unique compounds for large-scale benchmarking of exchange-correlation functionals when predicting band gaps[25]. The data set consists of nonmagnetic materials and includes a diverse group of covalent-, ionic-, and Van der Waals-bonded solids[25]. The best performing functionals produced an $R2$ value of 0.92 for the entire data set and a standard deviation of 0.7 eV.

DFT would be at a disadvantage if high throughput methods can be developed that either matched or improved upon DFT accuracy. One strategy moving forwards could be the application of machine learning. A combination approach, using both DFT and machine learning, has been attempted by several groups with good results. Lee et al.[26], using a data set of 270 inorganic compounds and non-linear SVR, predicted band gap values with a RMSE of 0.24 eV. The descriptor set included straightforward crystal and element properties such as cohesive energy and crystalline volume per atom, but also applied an estimate of the band gap using semi-high throughput Kohn-Sham (KS) DFT band gaps. KS DFT is known to significantly underestimate the band gap of inorganic solids[26], but when used as a descriptor in a SVR model, it greatly improved the accuracy.

Pilania et al.[27], uses a combination of low-fidelity (fast and inaccurate) and high-fidelity (slow, but better) DFT predictions to create a data set of 640 double perovskite

halides. What is unique about this approach is that low-fidelity band gap predictions are supplied for all compounds in the data set combined with a subset of high-fidelity predictions. Predictions made by this model match predictions made by high-fidelity DFT calculations for unseen compounds. Because the model is developed specifically for double perovskite halides, crystal structure is implicit in the model.

Zhuo et al.[11], developed a machine learning model using only element based descriptors, and no DFT. The advantage of this approach is that a model can be developed that is not dependent on preexisting knowledge about the crystal structure. This ensures maximum throughput. The descriptor set consisted of 136 variables derived from 34 element properties. Each property is represented for a given compound using the compound sum, difference, maximum and minimum value of the element property set. It is unclear from the description though, if the approach used element properties weighted for composition. The data set consisted of over 6000 compounds, 3896 of which had non-zero band gaps. However, this set included duplicate compositions for cases where band gaps had been measured using several experimental techniques. The band gap determination problem was divided into two:

1. A classification problem to determine if a given composition was likely to be a metal or a non-metal.

2. A regression problem to estimate the band gap for non-metals.

Zhuo et al. implemented several algorithmic approaches[11]. Best performance was achieved using support vector classification (SVC) for the classification problem, reporting 92% accuracy when using SVC in combination with a radial basis function kernel (RBF). For the regression problem Zhuo et al. achieved best performance using SVR, also in combination with the RBF kernel[11]. They report an R^2-value of 0.90 and a RMSE value of 0.45.

Chapter 3

Theoretical Foundation

This chapter will provide a summary of the basic underlying theory for the concepts discussed in the rest of this thesis.

3.1 Machine Learning

The following sections, 3.1, 3.1.1, 3.1.3 and 3.1.4, rely heavily on the summary of machine learning provided by Mueller et al. in the review book chapter *Machine Learning in Materials Science: Recent Progress and Emerging Applications*[6]. Machine learning is a term describing a collection of methods for the analysis of big data systems. The specific algorithmic approach may be classified into two main categories, supervised and unsupervised methods. Supervised methods attempt to find a function that, given a certain input, will produce an expected output. Consider that a supervised algorithm is given a data set that has within it n data rows and m columns. Each row represents some chemical, and each column represents a given property. Things like molecular weight, number of bonds, etc. These properties are called descriptors. They describe the chemical represented by the data. A researcher looking at the data might ask: Given that I know these things about a chemical, can I predict how this chemical is likely to react? A data set can be constructed that has in it the descriptor set along with experimental reaction results. A supervised algorithm is fed this data set X and attempts to create a model, a function $F(X)$, that determines if a chemical is likely to react or not. The reason such a strategy is thought of as supervised is because the machine learning algorithm is given a goal. The researcher has a fundamental hypothesis about the data presented: there is a correlation between the chemical properties, the descriptors, and how it reacts. Assuming this correlation exists, a supervised method will attempt to find it.

Unsupervised methods are also given a data set X. The researcher assumes that there are patterns within the data set, but they are unknown. It is up to the unsupervised algorithm to find and describe them. One example of this is data clustering analysis in which it is assumed that the data present can be sorted into clusters or groups.

3.1.1 Supervised Learning

Supervised learning methods can be further divided into two groups depending on the type of problem. Classification problems assume that there exists a discrete set of possible answers. Consider the problem of determining if a compound is metallic or not. The input data may be continuous, things like density, optical transparency and thermal conductivity, but the problem is discrete. The compound is either metallic or non-metallic. Regression problems assume the correct answer is a number on a scale. What is the band gap of a compound, given its properties? There is no finite set of possible answers.

Either way, all supervised methods operate on the assumption that there exists a function f that, given input data x, correctly predicts an expected value y. When developing a machine learning model, a data set will typically be divided into a training set and a test set. The program will be ”trained” on the training set and subsequently tested on the test set to check for the true predictive capabilities of the model.

The set of restrictions that is passed on a model is termed the hypothesis space. Consider for a moment the development of a model that can predict a person's height given only that person's age. To achieve this goal, a regression is performed, but there are several function types that are worth exploring. Figure 3.1 illustrates this concept.

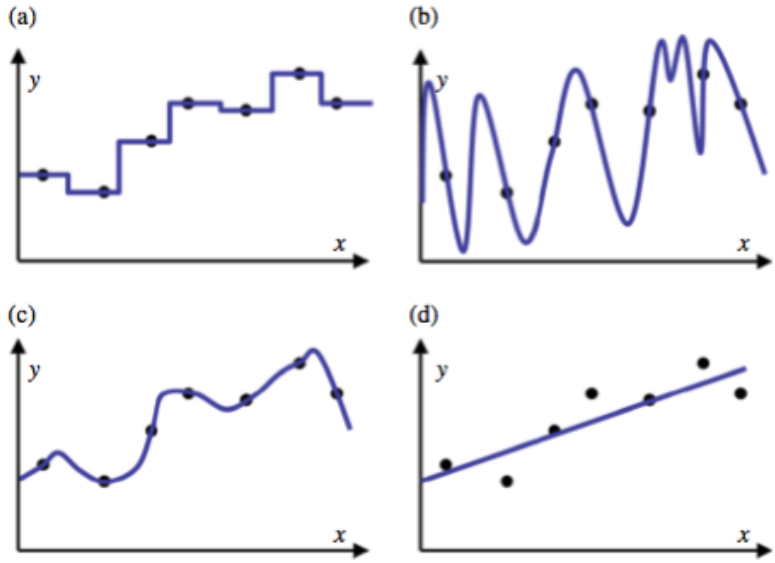


Figure 3.1: The figure illustrates how a large variety of functions may correctly map some data set x onto y. Figures a), b) and c) perfectly map the training data, but would not necessarily perform well when applied to the test data. Figure d) uses a linear approximation and does not provide a perfect fit, but may be more robust when applied to the test data.

The figure illustrates how, given a lack of external restrictions, any data set can be mapped to perfectly predict a given result. Sub-figures 3.1 a), b) and c) all fit the data perfectly, but the models are not likely to perform well when tested on unseen data about

a person's height. In contrast, the linear fit in Sub-figure 3.1 d) does not make a perfect fit, but may be more robust when tested. Overfitting is when a model or function learns how to predict the noise in the training data so well that it looses general applicability. Great care must be taken in any machine learning application to make sure this is not happening. In practice, the hypothesis space will be constrained to limit overfitting as a problem, combined with the use of a training and test set.

The performance of a model will depend on several limiting factors:

- The hypothesis space being too heavily constrained.
- The observed output or input values either being prone to experimental error or arising from other processes that are inherently non-deterministic.
- Some relevant input data is missing or unknown.

Although there are many different methods and algorithms for supervised learning, this thesis will discuss only three strategies. They are linear methods, decision trees and SVR. A brief introduction to these algorithmic approaches is presented below.

3.1.2 Linear Methods

Linear methods, or linear regression, is one of the most straightforward strategies in machine learning. The hypothesis space is reduced to the assumption that the relationship between the data x and the outcomes y can be modelled as a straight line described by Equation 3.1[28, ch. 12]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... \tag{3.1}$$

Although several parameterization techniques are possible, one standard approach is the least squares method. The parameters β_0, β_1, β_2, etc. are chosen to minimize the residual sum of squares (RSS), Equation 3.2[28, ch. 11].

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{3.2}$$

y_i denotes a given data point outcome, and $\hat{y_i}$ the corresponding predicted value.

If there are latent variables (LV) present in the data set, PLS may be more suitable[29]. A PLS model will create a new set of variables x_{LV} that are linear combinations of the old variables x. These new variables are then used as a predictor of the outcome y[29].

3.1.3 Decision Trees

Decision tree learning predicts an outcome by dividing the training set into smaller and smaller subsets according to a series of logical statements based on the data attributes. Understanding decision trees is easiest if considering a classification problem, but the basic idea remains the same for regression problems as well. Consider a set of questions, each with a finite set of possible responses. Each possible response corresponds to a specific

decision branch. The algorithm traverses the tree structure until a terminal node is reached. The concept is illustrated in Figure 3.2.

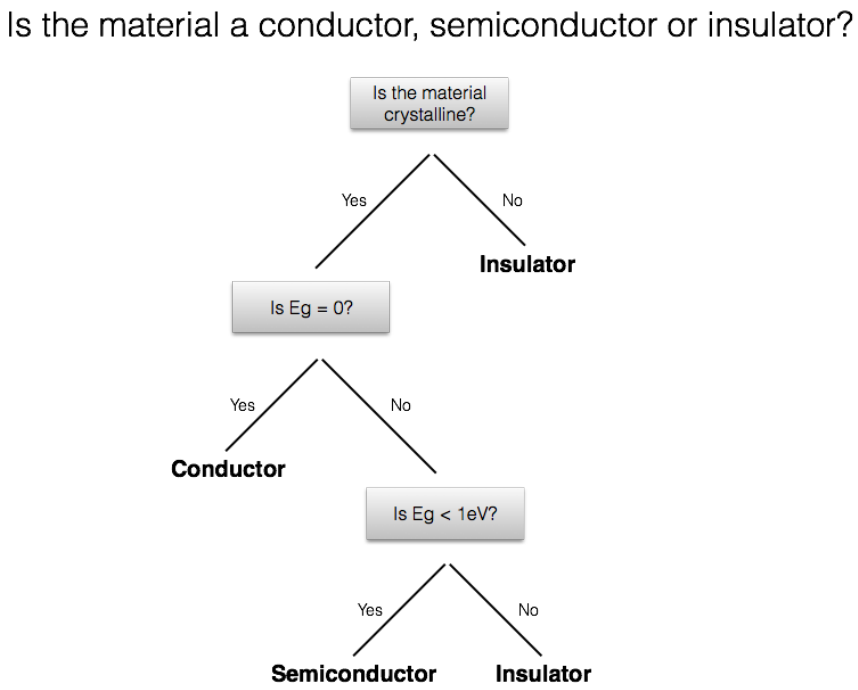


Figure 3.2: The figure illustrates how a basic classification decision tree might work to determine if a material should be classified as a conductor, semiconductor or insulator. A set of questions are asked in order of priority as one moves down through the branches to the nodes.

Figure 3.2 illustrates some important concepts of decision trees. First of all, the order in which the questions are asked matters. Consider a set of attributes that a material might have. Things like band gap, optical properties, thermal conductivity, etc. Each attribute can be converted into a question. Is the band gap greater or smaller than a given value?, Is the material transparent to visible light? and so on. Determining which of these questions to ask as a root (the beginning), and then which questions might be the most relevant at each node, will determine the accuracy of the model. Asking the questions in random order is likely to get a bad result. The root question should be the question that provides the highest information gain.

When decision trees are applied to regression problems, the order of questions is determined not by information gain, but instead by standard deviation reduction, a very similar concept. Consider an outcome set of n values. Equation 3.3 calculates the standard deviation, σ, of this set.

$$\sigma = \sqrt{\sum_i (x_i - \bar{x})^2 f(x_i)} = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}} \tag{3.3}$$

x_i and $\bar{x}$ denote the individual values of the set and the average value respectively. $f(x_i)$ is the probability distribution of picking a specific value x_i.

After a question has been asked the data will be divided into subsets. The total standard deviation for these new subsets, calculated using Equation 3.4, should be lower than before.

$$\sigma_{subset} = \sum_{answers} p(answer)\sigma(answer) \tag{3.4}$$

$p(answer)$ is the probability of a specific branch and $\sigma(answer)$ is the standard deviation of the corresponding branch set. The root question is determined by whatever attribute/question provides the highest decrease in the standard deviation of the subsets. This approach is then applied recursively until a terminal node is reached. If the attributes are continuous descriptors and not discrete, boolean questions of the type greater or less than are asked. In this way, continuous descriptors are converted into a discrete attribute.

For a classification problem, predicted outcome is reached by finding a terminal node. For regression problems the possible prediction outcomes exist within a continuous set. Instead of reaching a terminal node, a prediction is made one level up. The predicted value will be the average of the remaining subset or some other estimation method.

Overfitting is often a problem with decision trees. To limit this, decision trees can be pruned, either during construction or post-pruning. Principally, following the approach described above, a terminal node would be determined as having $\sigma = 0$. Pruning would prevent a node from branching further by defining a threshold deviation so that it terminates if $\sigma < \sigma_{threshold}$. Another cause for pruning might be if the number of instances remaining in a subset drops below a threshold. Applying these restrictions on branching will reduce overfitting.

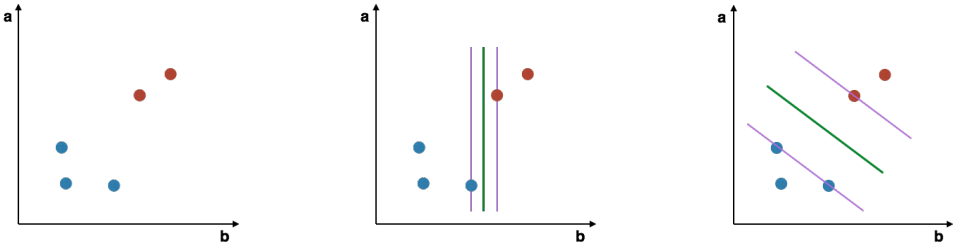
Random Forest

The Random forest algorithm, developed by Leo Breiman and Adele Cutler, seeks to reduce overfitting by creating multiple decision trees[30]. Each tree is created from a bootstrapped subset of the original training set[30]. Assume a training set has n data points and m descriptors. A bootstrapped data set, A, should also have n rows, but the rows are randomly selected and may repeat.

Another key feature of Random forest, is the method of tree construction. Although attributes are still chosen using the reduction in standard deviation criteria, the set of attributes that are considered at each node is a bootstrapped version of the full attribute set, including attributes that have already been selected at a previous node[30]. This way, the resulting decision trees A, B, C, etc. will be as independent from each other as possible. When performing a regression, the predicted output will be an average of all the decision trees, thereby minimizing overfitting[30].

Cubist

The Cubist algorithm is also based on averaging decision trees made from bootstrapped data sets and attributes. However, within each tree, the prediction is made at each terminal node, not from simply averaging the subset, but by applying a linear fit model to the subset. This concept is illustrated in Figure 3.3.



(a) The data points belong to either a red class or a blue class. To determine class, the data points are plotted in 2-dimensional space according to their attribute values a and b.

(b) There are several lines that might successfully classify the data points in the given training set correctly. One such line a simple vertical line (green). Note that the distance between the two boundary lines (purple) is in this case comparatively small.

(c) Support vector machine learning, seeks to find the hyperplane (green) that maximizes the width between the two boundary lines (purple).

Figure 3.4: The figure illustrates a classification problem and the concept of plotting n-dimensional data points in n-dimensional space. These data points can be separated into subgroups using a hyperplane (green). Depending on how this hyperplane is drawn, the data groups will be separated by the distance w between the two boundary lines (purple).

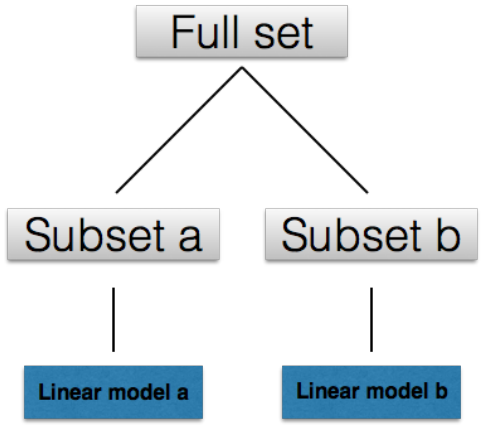


Figure 3.3: The figure illustrates how a linear model is applied at each terminal node of a decision tree, instead of just a simple average, when predicting the regression outcome.

3.1.4 Support Vector Machines

Support vector machines (SVM) seek to predict an outcome by dividing the training set into two groups separated by a hyperplane. SVM was initially developed to deal with two case classification problems. A data point belongs to only one of two categories. Consider a very limited case with only two attributes. Each data point is represented by a vector where the vector coordinates can be plotted in 2-dimensional space, illustrated in Figure 3.4a. The blue and red dots represent separate classes for determination.

Figure 3.4b shows the same data points separated by a vertical hyperplane (green), in this case a 1-dimensional line. The boundary lines (purple) are the parallel data point tangents. The goal of SVM learning methods is to maximize the distance between these two boundaries. Figure 3.4c shows the hyperplane that best achieves this goal of maximum boundary line separation.

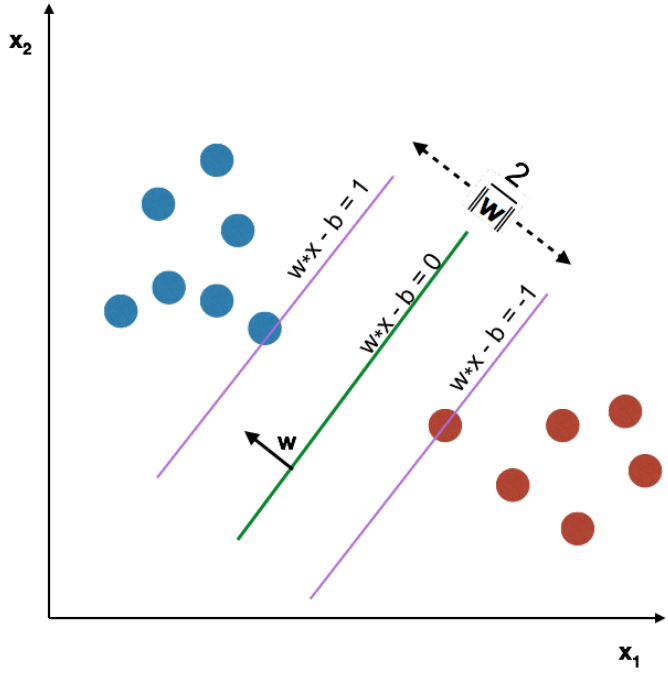


Figure 3.5: The figure shows the best separation of the data points using support vector machines. The vector $\boldsymbol{w}$ is perpendicular to the hyperplane. The offset b is defined so that $\boldsymbol{w}\cdot\boldsymbol{x} - b = 1$ for the positive support vectors and $\boldsymbol{w}\cdot\boldsymbol{x} - b = -1$ for the negative support vectors.

Consider a data set with n points. Each data point is a vector $\boldsymbol{x}$ and a corresponding outcome y_i. y_i can take the value $+1$ or -1. Any hyperplane that splits the data set so that all negative data points ($y_i = -1$) end up divided from the positive data points ($y_i = +1$) should obey Equation 3.5.

$$\boldsymbol{w} \cdot \boldsymbol{x} - b = 0 \tag{3.5}$$

where $\boldsymbol{w}$ is a vector perpendicular to the hyperplane and $\frac{b}{\|\boldsymbol{w}\|}$ is the offset distance from the plane.

Now consider that no data points can lie within in the boundary lines. Mathematically this is expressed so that the boundary lines must obey

$$\begin{aligned} \boldsymbol{w} \cdot \boldsymbol{x} - b &= 1 \\ \boldsymbol{w} \cdot \boldsymbol{x} - b &= -1 \end{aligned} \tag{3.6}$$

The concept is illustrated in Figure 3.5.

For the hyperplane that maximizes the distance between the two boundaries, one must maximize the width $\frac{2}{\|\boldsymbol{w}\|}$, subject to the constraints in Equation 3.6, using the Lagrange multiplication method. For mathematical convenience however, $\frac{1}{2}\|\boldsymbol{w}\|^2$ is instead minimized.

The case of red and blue dots in Figure 3.5 is reasonably straight forward to divide using a straight line, but the majority of cases will not be possible to separate in this way. Figure 3.6 illustrates this concept on a 1-dimensional problem. To solve such non-linear problems a kernel function is applied to the training data. A kernel function maps the data points to a higher dimension so that a linear hyper plane can successfully separate the data into two groups. There exists several possible kernel functions and the choice of kernel function will affect the robustness of the final model.

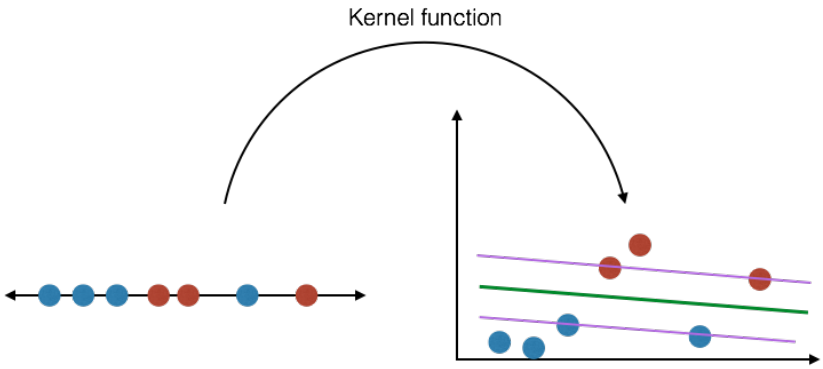


Figure 3.6: The figure illustrates how separation of the data points is not always possible using a straight line. These problems are called non-linear problems. By mapping the data to a higher dimension, the data can be separated in a meaningful way.

When SVM is applied to a regression problem, a tolerance error, ϵ, is set and a hyperplane is found that is as flat as possible within the limit of ϵ[31]. Consider a data set similar to the one considered for the classification case. The only difference is that y_i is now a continuous value. The boundary line equations given in Equation 3.6 are re-written in terms of this error and $\frac{1}{2}\|\boldsymbol{w}\|^2$ minimized accordingly[31].

$$\begin{aligned} y_i - < \boldsymbol{w}, x_i > -b &= \epsilon \\ y_i - < \boldsymbol{w}, x_i > -b &= -\epsilon \end{aligned} \tag{3.7}$$

$< \boldsymbol{w}, x_i >$ denotes the dot product of $\boldsymbol{w}$ and x.

3.2 Cross Validation

Cross validation is a way of addressing overfitting of a model to the training set so that when the model is later applied to unseen data it increases its chance of performing well[32]. Overfitting happens because the algorithms are rewarded for making correct predictions and model parameters are picked using the entire training set. One possible cross validation strategy is k-fold cross validation. Consider a training set with n data points. Every time the algorithm picks a set of model parameters, for example the number of trees to be included in a decision tree forest or the descriptors to be considered at a given node, the data set is divided into k random subsets. For each of the subsets, a new training set is created from the remaining $k - 1$ sets. The subset that is left out is used for validation. The parameter choice is made based on the reduced training set, but the choice is tested on the validation set. This is repeated k times so that each subset serves as a validation set. The parameters that performed the best when tested on the validation sets, are picked[32].

3.3 Model Performance Metrics

All statistical models are likely to be be prone to error. This error can be analyzed to give an indication of model performance.

3.3.1 Root Mean Square Error

The RMSE of a regression model is the standard deviation of the prediction error. Equation 3.8 calculates the RMSE for a sample of size n

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{i,\mathrm{pred}} - y_{i,\mathrm{obsv}})^2}{n}} \tag{3.8}$$

where $y_{i,\mathrm{pred}}$ is the predicted outcome for data point i and $y_{i,\mathrm{obsv}}$ is the true observed value for the same data point[33]. Because RMSE sums over the square of errors instead of the absolute value, the influence of the error is not proportional to the error size.

3.3.2 Average Absolute Relative Deviation

AARD is a way of assessing the magnitude of the error of the observed values and the sample values predicted by the model. Equation 3.9 calculates the AARD for a sample of size n

$$AARD = \frac{\sum_{i=1}^{n} \frac{|y_{i,\mathrm{pred}} - y_{i,\mathrm{obsv}}|}{y_{i,\mathrm{obsv}}}}{n} \tag{3.9}$$

where $y_{i,\text{pred}}$ is the predicted outcome for data point i and $y_{i,\text{obsv}}$ is the true observed value for the same data point[33]. AARD has the advantage over RMSE in that it uses the absolute value of the error instead of the square, making interpretability easier as the error influence is now proportional to the error size.

3.3.3 Coefficient of Determination

The coefficient of determination is a measure of model accuracy. One approach to calculating R^2, is to consider the Pearson correlation coefficient, r_{xy}. For two samples X and Y with n instances, we have

$$r_{xy} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.10}$$

where x_i and y_i are the sample instances and $\bar{x}$ and $\bar{y}$ are the sample means[28]. r_{xy} measures the linear correlation between the two samples and can range from -1 to +1. +1 indicates that the two samples are perfectly linearly correlated, -1 that they are perfectly negatively correlated. 0 indicates that there is no linear relationship between the two samples. The coefficient of determination is the square of this value[28].

$$R^2 = (r_{xy})^2 \tag{3.11}$$

A $R2$-value approaching 1 indicates an accurate model.

3.4 Crystals

A crystal is any material with long range periodicity, also known as translational symmetry[34]. Exactly what this symmetry looks like varies from material to material and depends on the constituent elements as well as things like phase stability determined by temperature, pressure, etc. The fundamental notion of long range periodicity is key to understanding how some properties arise in a crystalline material. One such property is the band gap[34].

Crystals form because constituent elements bind together in repeating blocks, sub-units, that define the chemical composition of the material. A simple case is the face centered cubic (fcc) structure of metallic copper. Consider marbles in a box. They will tend to organize in a closely packed repeating structure. Metallic copper replicates this structure on an atomic scale. Other crystalline structures exist as well. Some have the same basic structure as metallic copper, but the sub-unit is a little more complex. Table salt, NaCl, is an example of one such material[2]. Others have different crystal structures all together. The crystal structure is determined by the thermodynamics of the system and greatly affects a material's properties[34, 2, 35].

3.4.1 Orbitals and Orbital Overlap

To understand crystal bonding and the transfer of electrons within crystals, the reader should first understand the concept of orbitals. Atomic orbitals arise when applying the wave function to electrons around a nucleus[36]. In simple terms, they should be thought

of as the region around an atom in which an associated electron is likely to be found. The shape and orientation of an orbital depends on the orbital energy, and they form discrete energy ”levels” within an atom that the surrounding electrons can have[36]. There are, however, limitations. First of all, each orbital can only carry two electrons, one spin-up and one spin-down. This is a result of the Pauli exclusion principle[36], which states that no two electrons may be in the exact same quantum state. Secondly, according to the Aufbau principle, electrons will always organize in the lowest energy state possible if not otherwise excited[36].

When atoms bind together to form larger structures such as molecules, they experience orbital overlap[2]. One way of thinking about this is to consider the molecular orbital equivalent of atomic orbitals[37]. The interacting electrons within the molecule, those electrons that are part of bond formation, get organized into molecular orbitals and are thus ”shared” by the entire molecule[37]. The exact shape and orientation of these orbitals, and by extension the probability function describing electron location within the molecule, is once again determined by the electron wave function[37]. The same rules of organizing electrons into molecular orbitals apply as for atomic orbitals[37].

3.4.2 Bands and Band Gap

In crystalline materials, the long range periodicity means that orbitals may overlap throughout the entire crystal structure in a way that allows them to delocalize[37]. Delocalized electrons can be thought of as being ”shared” by the entire crystal structure and are therefore not bound to their parent atoms. In a simple model, this is thought of as a metallic bond and is used to explain why some crystalline materials can conduct electricity, but not others[37].

The delocalized electron is, nonetheless, an insufficient description of the physics of the metallic state. Consider the molecular orbitals that form around a molecule. An equivalent concept exists for crystal structures, only instead of each orbital corresponding to a discrete energy level, they form energy ”bands”[37]. In actuality the bands consist of several distinct energy levels, but there are so many of them and they are so close together that the continuous band approximation is applicable. Not all energies are allowed, however, and the bands tend to be separated by an energy gap. The origin of this gap is the diffraction of the electron wave function caused by the periodicity of the crystal lattice[34].

For a material to conduct electricity, some electrons must be ”free” to move around within the material[37, 34]. In quantum mechanical terms, this means that there must be unoccupied electron states that are accessible to a potentially conducting electron. The Pauli exclusion principle and the Aufbau principle apply here as they do for atomic and molecular orbitals[37]. If there are no accessible empty states, the electron is not ”free” and may not conduct. If the upper occupied band is only partly occupied, the electrons can easily access the available states within this band. The band is conducting and any material with this property is metallic[37, 2, 34]. If the uppermost band is fully occupied, it is termed the valence band. This band can not conduct electricity because there are no ”free” electrons. Above this band is located an empty band, separated by an energy gap, termed the band gap. This band is called the conduction band, and if an electron can make the ”jump” across the band gap, into the energy band, it becomes conducting. The

size of the band gap is important when assessing the usefulness of a material in functional technological applications[37, 2].

The band gap for an isoelectronic series of compounds, eg. ZnSe, GaAs and Ge, has been shown to obey equation 3.12

$$E_g^2 = E_H^2 + C \tag{3.12}$$

where E_g is the estimated band gap, E_H is the homopolar band gap and C is the bond charge transfer[38]. E_H is related to the highest molecular orbital energy of the constituent elements, effectively the band gap associated with a perfectly covalent bond. The bond charge transfer however, is directly correlated to the Pauling electronegativity difference of the elements involved[38]. Pauling electronegativity is discussed in Section 3.5.4.

3.5 Element and Crystal Properties

3.5.1 Atomic Weight

Atomic mass is the mass of the atom. Because most elements exist as several different isotopes, the atomic weight reflects the average mass[39].

3.5.2 The Van der Waals Radius

Due to the quantum nature of atoms, the atomic radius is difficult to define. However, several strategies exist. The Van der Waals radius, r_w, imagines a hard sphere atom, the radius of which is the distance of closest approach by another atom[40].

The atomic size difference, δ, for a crystal is described by Islam et al. as

$$\delta = 100 \cdot \sqrt{\sum_{i=1}^{n} c_i (1 - r_i/\bar{r})^2} \tag{3.13}$$

for a compound with element ratios c_i and n elements[12]. r_i and $\bar{r}$ are the radius of element i and composition weighted average radius for the compound respectively.

3.5.3 The Mendeleev Number

The Mendeleev number is an ordering number for elements. It attempts to order the elements in such a way that elements that behave in a similar fashion follow each other[41]. With a few exceptions, the Mendeleev numbers move from top to bottom within groups of the periodic table and then from left to right through the periods. Figure 3.7 shows the Mendeleev number for the entire periodic table of elements[41].

3.5.4 Pauling Electronegativity

Electronegativity describes the tendency of an atom to attract the shared electron pair when bonded to another atom[39]. The electronegativity difference between to bonded atoms

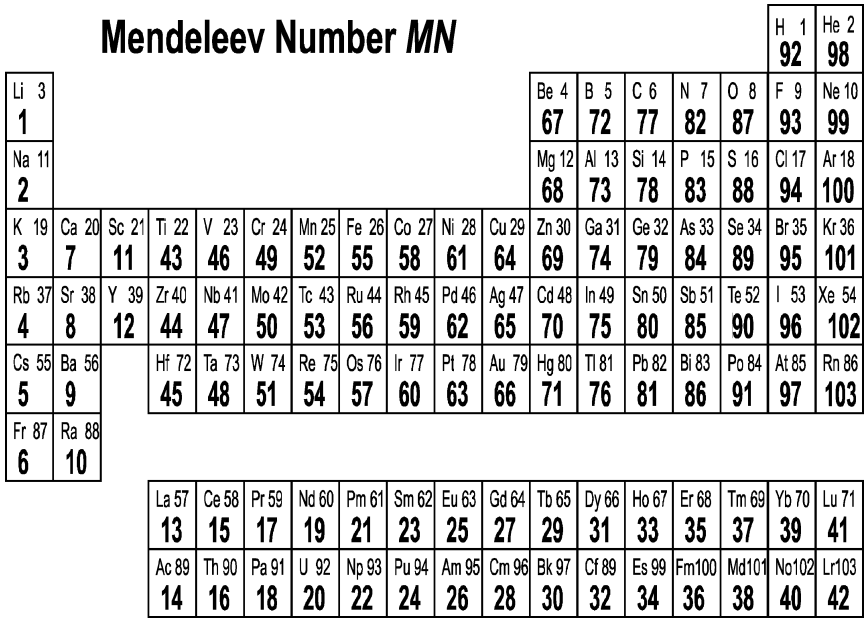


Figure 3.7: The figure tabulates the Mendeleev numbers for the entire periodic table of elements. With a few exceptions, the Mendeleev number increases from top to bottom within groups, and then from left to right through periods.[41]

will give an indication of the bond polarity and the ionic nature of the bond. Pauling introduced his relative scale to describe this tendency[39].

$$|\chi_A - \chi_B| = \sqrt{E_d(AB) - \frac{E_d(AA) + E_d(BB)}{2}} \tag{3.14}$$

Equation 3.14 defines the Pauling electronegativity for two elements A and B. χ_A and χ_B are the electronegativities of elements A and B and E_d is the dissociation energy of the associated bond[39].

The Pauling electronegativity difference $\Delta\chi$ for a crystal with n elements and weighted electronegativity average $\bar{\chi}$ is described by Islam et al. as

$$\Delta\chi = \sqrt{\sum_{i=1}^{n} c_i(\chi_i - \bar{\chi})^2} \tag{3.15}$$

for a compound with element ratios c_i and n elements[12]. r_i and $\bar{r}$ are the radius of element i and composition weighted average radius for the compound respectively.

3.5.5 Ionization Energy

The first ionization energy is a measure of the energy required to separate the first electron completely from the parent atom at ground state, leaving a positively charged ion[39]. A

high ionization energy indicates electrons are strongly bonded to their parent atoms.

3.5.6 Electron Configuration and the Angular Momentum Quantum Number

Electrons in an atom are organized into atomic orbitals as described in Section 3.4.1. The orbitals use a naming convention s, p, d, f to describe the various available orbital shapes[36]. The valence electrons of an atom will be organized into the set of outer orbitals, s, p, d, f, consistent with the Aufbau principle.

The angular momentum quantum number (ℓ) is the associated quantum number that determines an orbital's angular momentum. ℓ can take integer values 1, 2, 3, etc[36]. An orbital's angular momentum and its shape are uniquely paired so that the value of ℓ corresponds to the highest energy occupied orbital[36].

3.5.7 Valence Electron Concentration

An atom's valence electrons are those electrons that are located in the outermost occupied shell. Only these electrons are likely to take part in chemical reactions. Therefore, the valence electron count largely predicts an atoms chemical properties. The VEC is defined as the number of valence electrons per compound formula unit[12]. Although this is not an element property, it is easily derived by counting valence electrons. Equation 3.16 defines VEC for a compound with a set of element concentrations c_i and corresponding valence electron count VEC_i[12].

$$VEC = \sum_{i=1}^{n} c_i VEC_i \tag{3.16}$$

3.5.8 Dipole Polarizability

The dipole polarizability of an atom is a measure of how easily perturbed the electron cloud around an atom is when exposed to an electrical field[40]. The atomic energy shift, ΔW, is proportional to the square of the electrical field E[40].

$$\Delta W = -\alpha \frac{E^2}{2} \tag{3.17}$$

In Equation 3.17, the dipole polarizability, α is the constant of proportionality.

3.5.9 Thermal Conductivity

Thermal conductivity is a measure of how well a material conducts heat energy. As an atomic property in this thesis, it measures heat conduction in the pure element at 25°C. Consider a material with a temperature gradient $\boldsymbol{\nabla T}$. The heat flux $\boldsymbol{q}$ is defined so that

$$\boldsymbol{q} = -k\boldsymbol{\nabla T} \tag{3.18}$$

making the proportionality constant, k, the thermal conductivity of the material[34].

In materials with zero band gap, the Wiedemann-Franz law states that the ratio of thermal conductivity to electrical conductivity is proportional to the temperature[34]. This suggests that thermal conduction is strongly related to the process of electrical conduction.

3.5.10 Cohesive Energy

The cohesive energy of an element is defined as the energy required to form separated neutral atoms in their ground state from the elemental solid at 0 K and 1 atm[34]. The higher the cohesive energy, the stronger the electronic interactions between the atoms.

3.5.11 Configurational Entropy

The configurational entropy, S, of a system with multiplicity W is defined by Boltzmann so that

$$S = k_b \log_e W \tag{3.19}$$

where k_b is the Boltzmann constant[42]. For a compound with composition $A_aB_bC_c$... an equivalent expression of entropy, based on the probability, p, of a lattice site being occupied by a given element i is

$$\frac{S}{k_b} = -\sum_{i=1}^{t} p_i \log_e p_i \tag{3.20}$$

where t is the total number of unique elements in the compounds[42]. Assuming the compound elements are distributed at random throughout the lattice, p_i is simply the concentration of element i in the compound. For ordered crystals this is only valid as a first approximation and p_i would ideally be determined based on crystal structure information.

3.6 Linking the Theoretical Foundation to Model Development

Section 3.1 introduces the concept of machine learning and how various machine learning algorithms work. Section 3.4 discusses the concept of crystals and how a combination of orbital theory and crystal structure information can be used to understand how bands and band gaps arise in crystals. Lastly, in Section 3.5 several elemental and crystal properties are introduced.

Developing a machine learning model to predict band gaps includes two sets of choices: choice of descriptors for system representation and choice of machine learning algorithm. The choice of descriptors will impact the accuracy and robustness of the model. For a discussion on why the properties discussed in Section 3.5 were chosen as descriptors to represent the compound in the model, and how these descriptors might relate to the size of the band gap given the discussion in Section 3.4, the reader is referred to Section 4.1.1. The choice of machine learning algorithms further expands upon the literature discussed in Chapter 2. The implementation of these algorithms is discussed in Section 4.4.

Chapter 4

Method

This chapter includes a description of the method used for developing the machine learning models, the rationale of the choices made relating to which algorithms are explored, the properties included in the model, the creation of the descriptor set and an analysis of the input data used in model creation.

4.1 Method Rationale

The thesis seeks to explore and expand upon models that successfully predict the band gap property of crystals, eliminating the need for time consuming experiments when designing novel materials. The key question is whether it is possible to develop a robust data driven model that can reasonably predict the band gap of crystals. Only properties that are either already tabulated or easily determined using efficient calculations should be included. Element properties and certain composition properties, such as average crystal radial ratios and a configurational entropy approximation, are therefore studied in the approach. Expensive properties such as precise crystal structure information, that require either experimental data or time intensive DFT calculations are excluded from the study.

4.1.1 Choice of Descriptors

The thesis builds on the work done by Zhuo, et al.[11]. In that paper, a set of 34 element properties was used as the basis for forming the descriptor set. In this work, a reduced set of these descriptors was chosen for investigation, along with four new crystal properties. The element properties that have been chosen are:

- Atomic weight
- Van der Waals radius
- Mendeleev number

- Pauling electronegativity
- 1st ionization energy
- Electron configuration
- Angular momentum quantum number
- Dipole polarizability
- Thermal conductivity
- Cohesive energy

The 4 new crystal properties that have been added are:

- Configurational entropy
- Valence electron concentration
- Atomic size difference
- Aggregate Pauling electronegativity differences

The rationale behind the specific set of element properties chosen is presented below. The crystal properties are calculated using the methods described by Islam et al.[12]. The rationale for this choice is expanded upon in Section 4.2.3.

Atomic Weight and the Van der Waals radius

Indicators of size give indirect information about crystal structure. Band gap theory would suggest that crystal structure is a key factor in determining the size of the band gap.

Mendeleev Number

The Mendeleev number provides chemical information about an element. The chemical properties of an element are closely related to its electronic structure, and electronic structure is closely related to the band gap[2, 34].

Pauling Electronegativity

The electronegativity of an element provides information about the charge polarity of bonds that elements are likely to form. The charge polarity of the bonds in a crystal will effect the crystal's electronic properties[2].

Ionization Energy

The ionization energy is a measure of how strongly or weakly the electrons are bounded to parent atoms within a crystal. For an electron to conduct, it must be excited into the conduction band, loosing the locality of its parent atom.

Electron Configuration, the Angular Momentum Quantum Number and Valence Electron Concentration

The electron configuration, the angular momentum quantum number and the Valence electron concentration provide information about an element's electronic properties, thereby determining likely bonding configurations and band occupancy.

Dipole Polarizability

Dipole polarizability is an indication of how the electrons are likely to behave in the presence of an electrical field. This electrical field can be externally applied or be the result of local interactions between the atoms within the crystal structure.

Thermal Conductivity

Thermal conductivity is related to the electric transport properties of the crystal structure. The thermal conductivity might therefore provide information about how loosely bounded the electrons are.

Cohesive Energy

Cohesive energy provides information about the strength of the interatomic interactions between the pure elements. The descriptor has been successfully included in other attempts at using machine learning for predicting band gaps such as Lee et al.[26].

Configurational Entropy

The entropy is fundamental in understanding different crystal structures. Islam et al. applied this descriptor successfully to predict phase transformations in multi-principal element alloys[12].

4.2 The Data

4.2.1 Experimental Band Gap Data

The data set used for training and testing the model was a list of compounds and associated band gaps, spanning over 3896 reported experimental results and consisting of 2458 unique compositions. This is the same set as was used by Zhuo et al[11]. The compound list was parsed so that each compound was converted into a list of element composition ratios ranging from zero to one.

An element breakdown of the list showed a wide spread in the representation of the different elements in the list. The full list included 79 unique elements, and no element beyond element number 92. Table 4.1 summarizes some key information about the data. Table 4.2 shows the unique instance count for the five most common and five least common compounds in the data. For an exhaustive list of all elements present in the data, see Table A.1 in Appendix A.

Table 4.1: Table showing information about the total number of elements present (Tot), average element instance count (Avg), median count, and element count standard deviation (σ)

Property	Value
Tot	79
Avg	100
Median	40
σ	133.6

Table 4.2: Table showing the unique element instance count for the five most common and least common elements in the band gap data.

Element	Unique instances
Se	694
S	622
O	534
Te	401
Ga	312
⋮	⋮
Be	9
U	9
Re	6
Tm	6
Tc	4

The band gaps in the list spread from 0.02 eV to 11.70 eV. Key information about the band gaps is reported in Table 4.3. The histogram in Figure 4.1 shows the spread of band gaps in the data set. It is clear from the values in both Table 4.3 and Figure 4.1 that the majority of data points in the data set have band gaps in the range 0-3 eV.

Table 4.3: Table showing information about the average band gap, median band gap, and band gap standard deviation (σ)

Property	Value
Avg	2.09
Median	1.92
σ	1.49

4.2.2 Element Properties

The data used to construct the descriptors was retrieved from several sources. For an exhaustive list of element property sources, see Appendix B.1. Two approaches were tested for representing the element properties as descriptors in the training data. The first method attempted to create a composition based descriptor set. For each descriptor, the compound

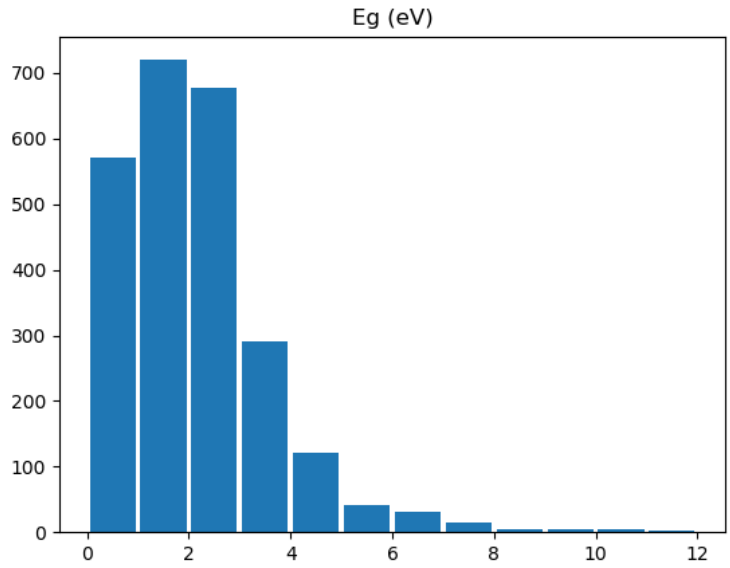


Figure 4.1: The histogram shows the spread of band gaps in the data set. Each bin spans over 1 eV.

is represented as a list of weighted property values. The weight is chosen as the relative composition of the column element. Consider the following example. Assume the data set consisted of only 5 elements: Na, Br, O and F. The atomic weights in g/mol of these elements are 22.99, 79.90, 16.00 and 19.00 respectively. Now consider the property vector for the compound Na_2O. The composition weights are 0.33, 0, 0.66 and 0 respectively; therefore the property vector is:

$$[22.99 * 0.33, 78.0 * 0, 16.00 * 0.66, 19.00 * 0] = [7.66, 0, 10.67, 0] \tag{4.1}$$

For the actual data set with 79 different elements, 2458 unique compositions and 12 different element properties, the resulting matrix has over 2 million input points, most of which are zeros due to each composition consisting of, at most, 5 different elements. This presents a problem for the machine learning algorithms which struggle when the descriptor columns have low variance and/or are highly correlated, both of which are the case for this type of representation.

The second strategy, based upon the work by Zhuo et al.[11] reduces each property to a set of four descriptors. It is assumed that each composition can be sufficiently described using the weighted sum, the weighted max, the weighted min and the difference between the weighted max and the weighted min. Zhuo et al.[11] used a similar approach, but it is not clear from the paper if they used weighted property values instead of true property values. The choice to use weighted property values was based on the understanding that the band gap must be dependent, not only on the elements present, but also on their relative ratios. For example, using the reduced data set from above, the vector in Equation 4.1 is

reduced to [7.66, 10.67] and passed through the four following functions in R:

```
sum()
max()
min()
max() - min()
```

producing the resulting description vector:

$$[18.33, 10.67, 7.66, 3.01] \tag{4.2}$$

This strategy scales up well, independent of the number of elements in the data set, and leaves much fewer mostly zero columns. Each element property is treated in the same way and the vectors are stacked horizontally so that each compound is now described by a total of $4n$ descriptors, where n is the number of properties chosen. The downside of this strategy is that a fair amount of information is lost. The functions chosen to represent the property do not necessarily always provide the relevant information.

4.2.3 Crystal Properties

The four crystal properties used, in addition to the element properties used by Zhuo et al.[11], are atomic size difference, crystal Pauling electronegativity differences, VEC and configurational entropy. The properties are calculated using Equations 3.13, 3.15, 3.16 and 3.20 respectively, presented in Section 3.5.

In addition to the choice rationale discussed in Section 4.1.1, these properties were chosen because they have all been applied successfully by Islam et al. to describe multi-principal element alloys (MPEAs) in a related problem[12]. Because these properties are numerical features of the compounds as a whole, they do not need further interpretation and are added to the descriptor set directly.

4.3 Pre-processing

Zhuo, et al. used the complete data set with 3896 compounds[11]. This set includes 1438 compounds that are composition duplicates, but with experimental band gaps determined using varying methods. In an attempt to minimize the risk of overfitting, the data set used in this thesis was reduced to a unique set. For compounds with duplicates, the band gap was averaged so that each compound is uniquely associated with only one band gap value. For comparisons sake, attention is also given to the full data set.

In order to maximize performance prior to choice of machine learning method, the full descriptor set was passed through two filters. Columns that were either strongly correlated or low-variance were removed. Out of 56 initial descriptors, 51 remained. For a list of the removed descriptors, see list in Appendix A.

The data was finally divided into a training set (80%) and a test set (20%). The division was done five times, creating five unique training/test sets, each using a randomly selected four digit seed. This was done in order to ensure two things. First of all, the elements are not all equally distributed across the full data set, as was discussed in Section 4.2.1. Creating multiple training/test set pairs, reduces the risk of certain elements not being accounted

for in either the training set or the test set. Secondly, when checking model performance, the average performance is a more accurate description of method performance than the results of any single trial.

4.4 Implementation of the Machine Learning Models

The machine learning models were created using built in R libraries, version 3.5.1. Each method was tried five times, one for each training/test set pair. The same sets were used for all machine learning methods. This makes comparison of the models at a later time more accurate. PLS was, in addition, tested on the composition based descriptor set.

Cross validation is performed in all cases by the built in R functions.

4.4.1 Testing and Refinement

The performance of each machine learning method was investigated using three statistical approaches: R^2, $RMSE$ and $AARD$. The values for all five trials were averaged and compared.

In addition to the unique set of compounds with averaged band gap values discussed in Section 4.3, the learning method that showed the best results after the initial investigation described above was also applied to the full set with duplicates used by Zhuo et al.[11] along with SVR for comparison.

The best performing model was tested on an external data set, borrowed from Borlido et al.[25].

Chapter 5

Results and Discussion

This chapter presents the performance metrics of the machine learning models and an analysis of the predictions. A discussion of the implications of these results follows.

5.1 Initial Results

Table 5.1: Table showing mean model performance metrics for all machine learning algorithms. PLS comp refers to the PLS method applied to the composition-weighted property table. PLS, Random Forest, Cubist and SVR, refer to the methods being applied to the Zhuo et al[11]. derived data set using sum(), min(), max() and max()-min() descriptors. R^2_{cv} and $RMSE_{cv}$ denote the Coefficient of Determination and the Root Mean Square Error for the cross validated training performance. R^2_{train}, $RMSE_{train}$, $AARD_{train}$ show the Coefficient of Determination, the root mean square error, and the Average Absolute Relative Error of the training set, and R^2_{test}, $RMSE_{test}$, $AARD_{test}$ show the Coefficient of Determination, the square root mean error, and the Average Absolute Relative Error of the test set.

	PLS comp	**PLS**	**Random Forest**	**Cubist**	**SVR**
R^2_{cv}	0.65	0.61	0.78	0.80	0.78
$RMSE_{cv}[eV]$	0.89	0.93	0.70	0.67	0.71
R^2_{train}	0.70	0.63	0.97	0.97	0.89
$RMSE_{train}$ [eV]	0.82	0.91	0.27	0.27	0.50
$AARD_{train}[\%]$	59	70	19	17	30
R^2_{test}	0.66	0.61	0.80	0.81	0.78
$RMSE_{test}$ [eV]	0.88	0.93	0.68	0.65	0.69
$AARD_{test}$ [%]	69	79	53	47	54

5.1.1 Linear Methods

PLS regression was applied to both the composition-weighted property descriptor set and the Zhuo et al.[11] derived descriptor set. The results of this investigation are reported in Table 5.1. Best performance using PLS regression was achieved for the composition-weighted property descriptors with an average R^2_{test} value of 0.66 for the test set. PLS applied to the Zhuo et al. derived descriptor set performed slightly worse for all metrics.

For both approaches, R^2 and $RMSE$ averages are similar for both the cross validated training set and the test set. However, the non-cross validated training set results are slightly higher. This suggests that cross validation successfully addressed overfitting.

It should be noted that, although one performed better than the other, both approaches yielded unsatisfactory results. Band gap is a crystal structure dependent property. Because of this, one should not necessarily expect a linear response between element properties and the associated band gap values. Due to the poor linear model performance, non-linear methods were applied instead.

5.1.2 Non-Linear Methods

The non-linear approaches are reported only for the Zhuo et al. derived data set. This is because attempts to use non-linear methods for a composition-weighted property descriptor set failed due to the columns for this descriptor set being mostly comprised of zeros which, in turn, resulted in highly correlated columns. Non-linear methods only work if this is not the case. The non-linear method that performed the best was the Cubist algorithm with an R^2_{test} value of 0.81 for the test set. The full results for the initial non-linear investigation are reported in Table 5.1. It should be noted that Random forest and SVR performed only slightly worse than Cubist, but Random Forest performs marginally better than SVR when only looking at the test set metrics. This indicates that decision trees may be better suited to approaching this problem than SVR. However, the differences in metrics reported for model application to the test set are so small that a definitive conclusion is difficult. The similar performance of the models on the cross validated training data and the test data is an indication that the training data is a representative selection of materials. The equivalent metrics for the non-cross validated training set are considerably better, but this is to be expected and is not an indication of model performance. The values are reported because they show that models have been created that successfully map the training descriptors to the observed outcomes.

Figures 5.1, 5.2, 5.3, 5.4, 5.5 and 5.6 show scatter plots for the predicted values as a function of the observed values. The diagonal line shows true prediction. Predictions below the line are under-estimated and predictions above the line are over-estimated. All scatter plots are reported for the same test set (red line)/training set (blue line) pair. The scatter plots show a clear majority of low band gap compounds present in both the training set and test set, consistent with the band gap data distribution reported in Section 4.2.2. Figure 5.5 indicates that SVR produces a significantly wider spread in training prediction error around the true line than the decision trees shown in Figure 5.1 and 5.3. This is consistent with the metrics reported in Table 5.1. The equivalent test set plots, Figures 5.6, 5.2 and 5.4 respectively, indicate that, when applied to the test set, error spread is similar for all approaches. This is also consistent with the model metric data reported in Table 5.1.

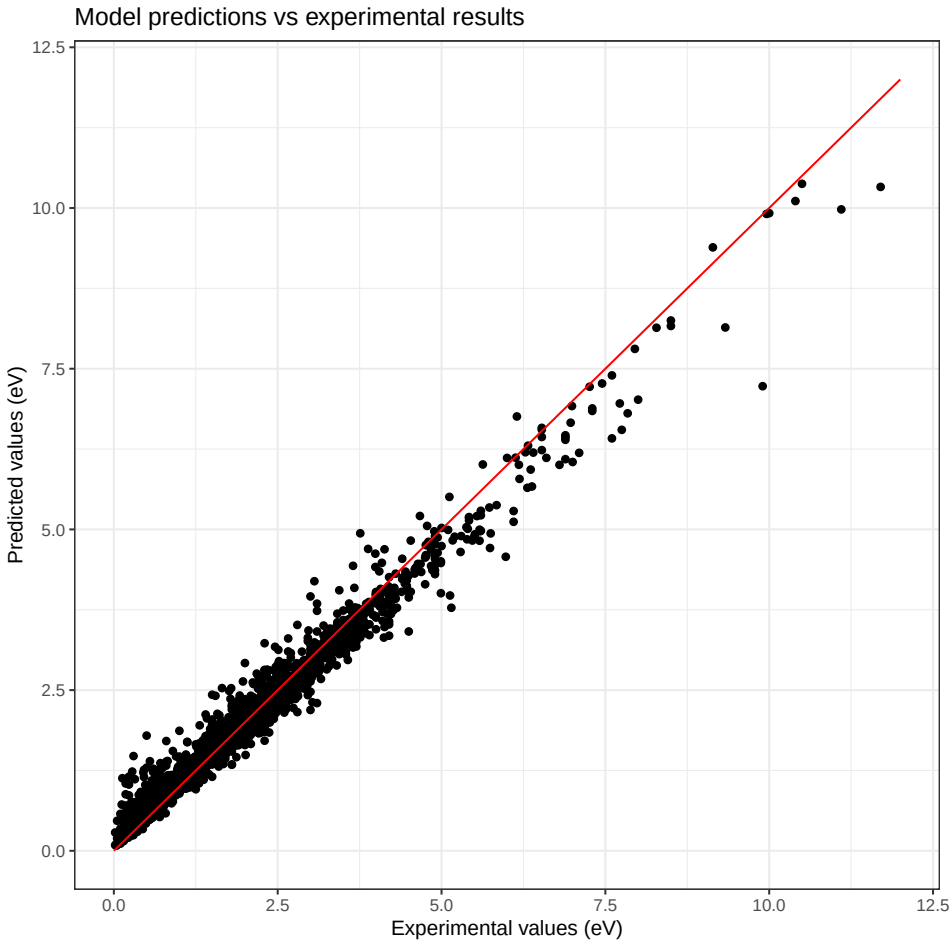


Figure 5.1: The scatter plot shows the predicted values by the Random Forest model relative to the experimental values. The data presented is from a training set.

Furthermore, all scatter plots show a tighter concentration of predictions along the central line in the lower band gap range (0eV - 5eV), relative to the higher band gap range, especially when predicting test set values. The most likely explanation is based on the band gap distribution of the data set. Because the majority of band gap data is located at the lower end of the spectrum, the algorithms are more likely to learn better for this range. There are so few high-range band gap compounds that it is possible that there is not enough data to learn how the elements behave in this range. The test set would be particularly vulnerable to this effect. This explanation is consistent with the observation that the effect is more visible when studying the test sets.

The scatter plots in Figures 5.7 and 5.8 show the prediction error as a function of the observed band gap. The red line shows a perfect prediction. The blue lines show the 0.5

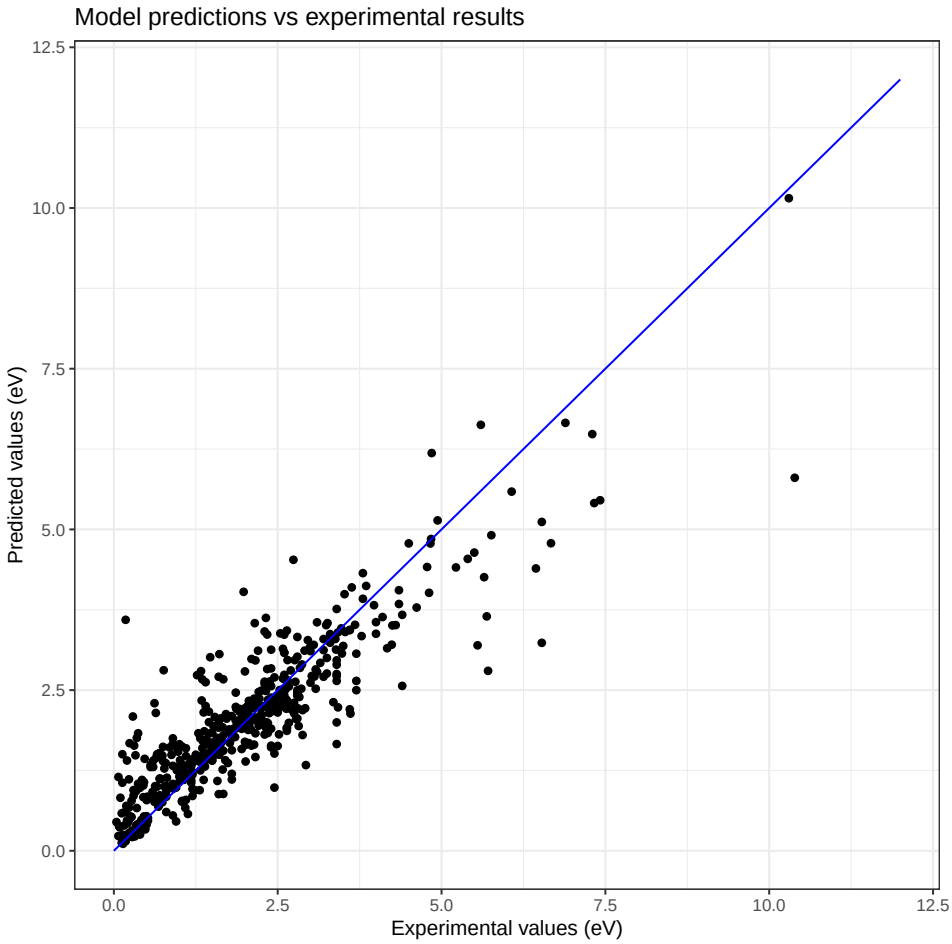


Figure 5.2: The scatter plot shows the predicted values by the Random Forest model relative to the experimental values. The data presented is from a test set.

eV error cutoff. Any prediction above the red line is an underestimation of the true value. A prediction below the red line is an overestimation. Both plots are produced using the same training set/test set pair as above. In addition to confirming the error spread trend discussed above, Figures 5.7 and 5.8 show that the errors are not symmetrically distributed around the perfect prediction line. For both Cubist and SVR there is a clear trend that low band gaps tend to be underestimated and high band gap s tend to be overestimated. The underestimation of high band gaps is likely also a consequence of the data band gap distribution. Because the majority of data points have a relatively small band gap, the algorithms learn to underestimate prediction for band gaps higher than the data majority. Similarly, the lack of zero-band gap data in the data set leads the algorithms to overestimate low band gap ranges, although not to the same extent. The most even error spread, for

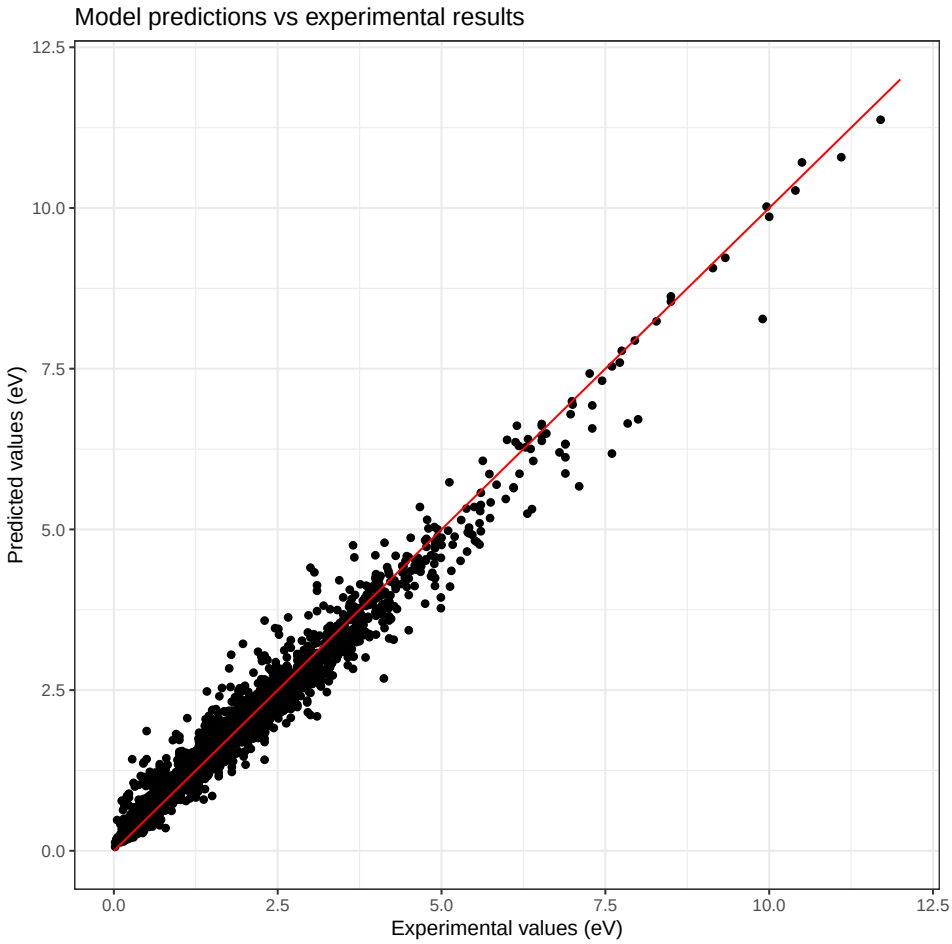


Figure 5.3: The scatter plot shows the predicted values by the Cubist model relative to the experimental values. The data presented is from a training set.

both Cubist and SVR, is achieved for mid-range band gaps between 2 and 2.5 eV, centered around the mean band gap of the full data set.

Because the results are so similar, irregardless of which non-linear method is chosen, the remaining discussion about errors uses data soly from the Cubist model. Of the 492 data points in the test data being studied 70% had errors within 0.5 eV. Seven compounds had a predicted error greater than 2 eV, listed in Table 5.2.

Of the seven compounds listed in Table 5.2, one has a hydroxide group. This is the only instance of (OH) in the data set and a high prediction error is to be expected. Of the remaining six, five compounds are oxides.

Table 5.3 shows the mean error and element count of the ten worst performing elements in the test set. Of these ten, oxygen ranks number nine, with an average error of 0.74 eV.

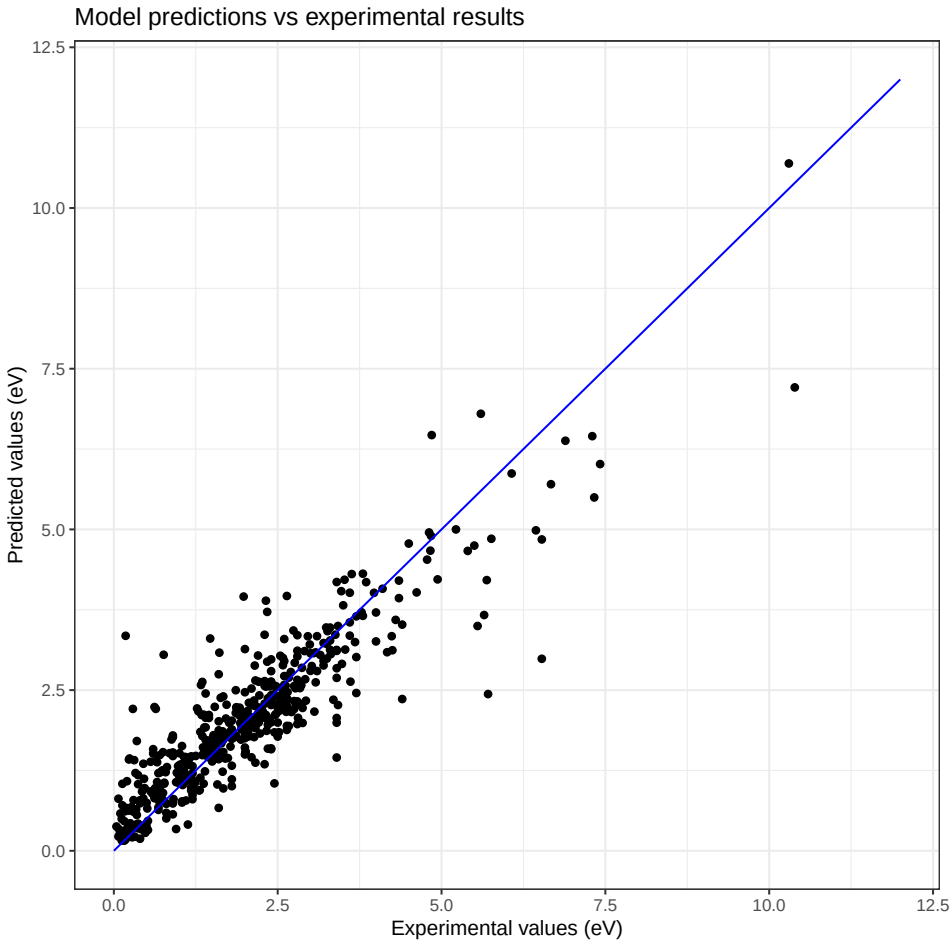


Figure 5.4: The scatter plot shows the predicted values by the Cubist model relative to the experimental values. The data presented is from a test set.

However, the remaining nine elements are poorly represented in the test set, with eight out of nine elements being represented only ten times or less. Of the twelve non-oxygen elements listed in Table 5.2, eight of them are also listed in Table 5.3. The results suggest that certain elements have properties that make them difficult to predict using the strategies presented in this thesis, specifically oxygen. Consider now the bar graph in Figure 5.9. In the test set, the four most common elements were oxygen, sulfur, selenium and tellurium, all elements in group 16 of the periodic table. Because there are so many instances of these elements, it can be assumed that the mean absolute error is a true indication of predictability. Figure 5.9 suggests that the higher up in group 16, the more difficult it is to predict.

Umebayashi et al[43] show in their paper *Band gap narrowing of titanium dioxide by*

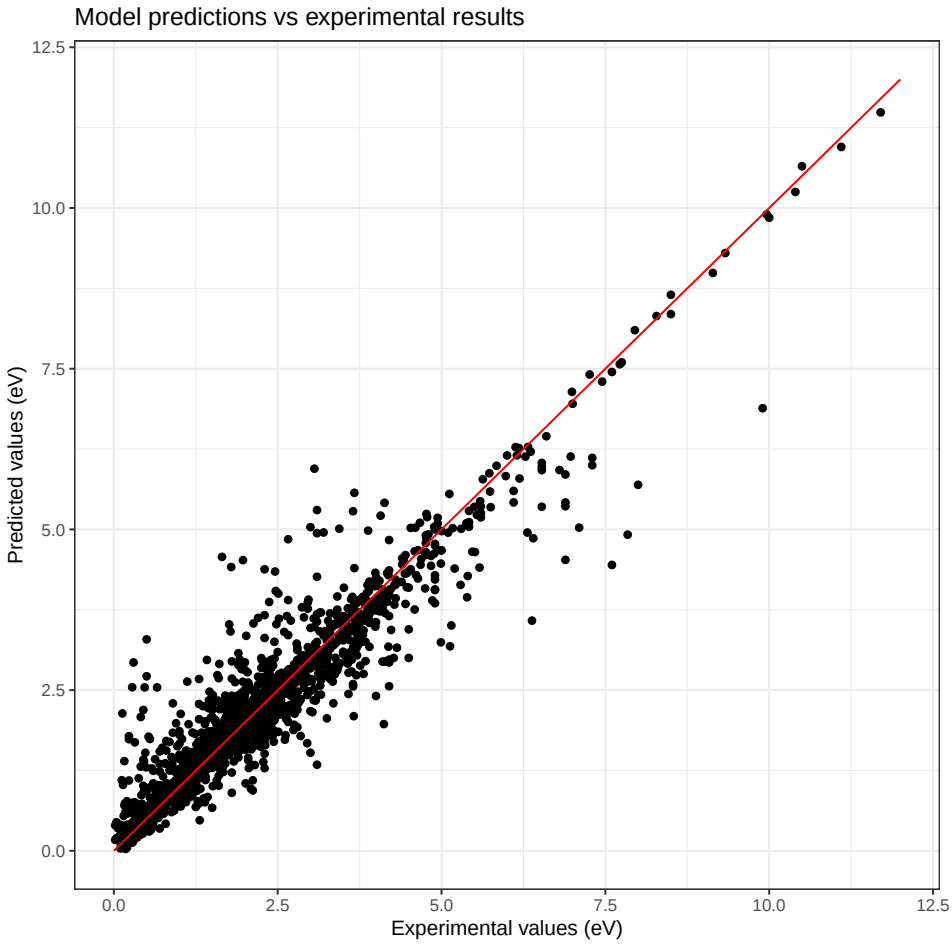


Figure 5.5: The scatter plot shows the predicted values by the SVR model relative to the experimental values. The data presented is from a training set.

sulfur doping that the band gap of a compound is highly sensitive to which group element is present in the compound, even though the electronic structure of sulfur and oxygen are similar. Both have five p-orbital electrons in the valence band. However, there is a significant difference between the two. The 2p orbital has far less shielding from the atomic nucleus than the p-orbitals located further out. One possible hypothesis is that the lack of shielding is making oxygen 2p-orbital electrons behave differently from the more shielded 3p-, 4p- and 5p-orbitals of sulfur, selenium and tellurium. The descriptors used do not make this distinction, however, and so the machine learning algorithm fails to capture this effect.

If the above hypothesis has any merit, a similar phenomenon should occur for the other p-group elements. Although the data is considerably more sparse for the other p-

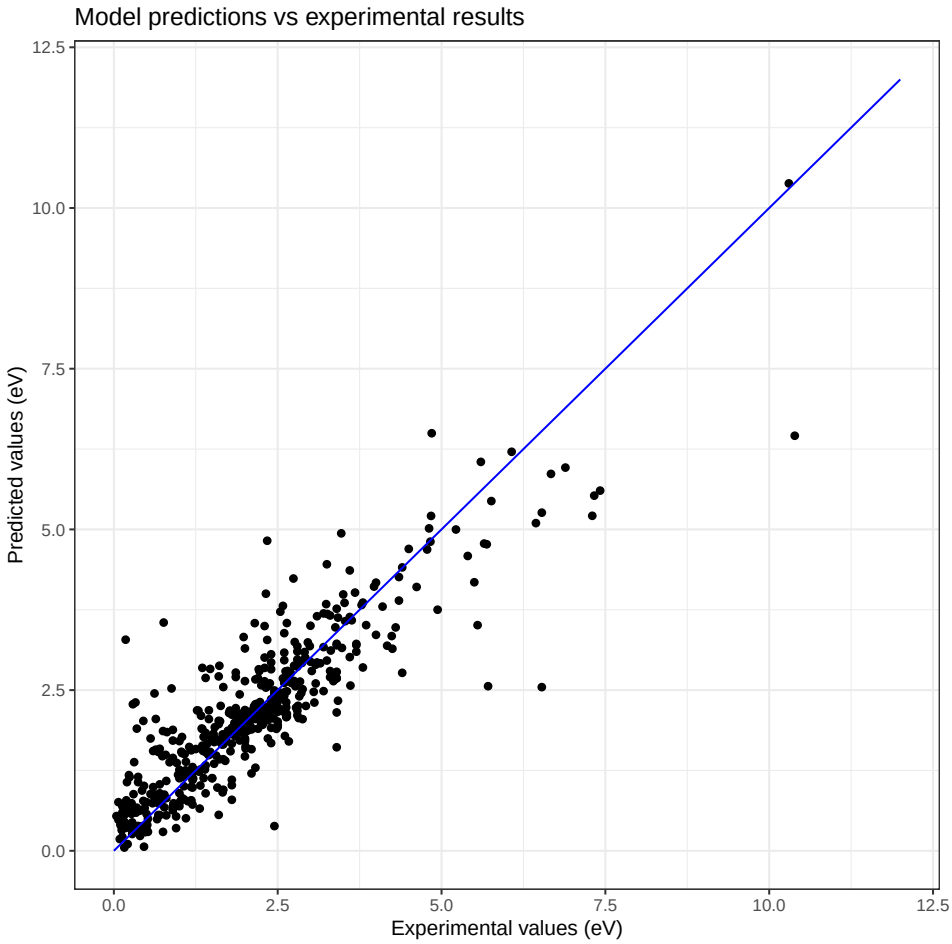


Figure 5.6: The scatter plot shows the predicted values by the SVR model relative to the experimental values. The data presented is from a test set.

group elements, the trend does repeat itself for groups 13, 15 and 17 as well. The data showing this trend is reported in Table 5.4. The group 14 elements seem to break the trend, but the error data for carbon is based on only one instance. Therefore, the group should probably not be considered. The remaining groups all show the same trend of predictability increasing when the p-orbital is more shielded. It should be pointed out, however, that because the number of instances was much lower for elements outside of group 16 the data is also less reliable.

To further support this hypothesis, the same trend is visible for s-orbital elements. Sodium and lithium are both more easily predicted than hydrogen, and magnesium is more easily predicted than beryllium. This is consistent, because 1s and 2s orbitals experience little shielding as compared to 3s-orbitals. The trend is not, however, visible for d-orbital

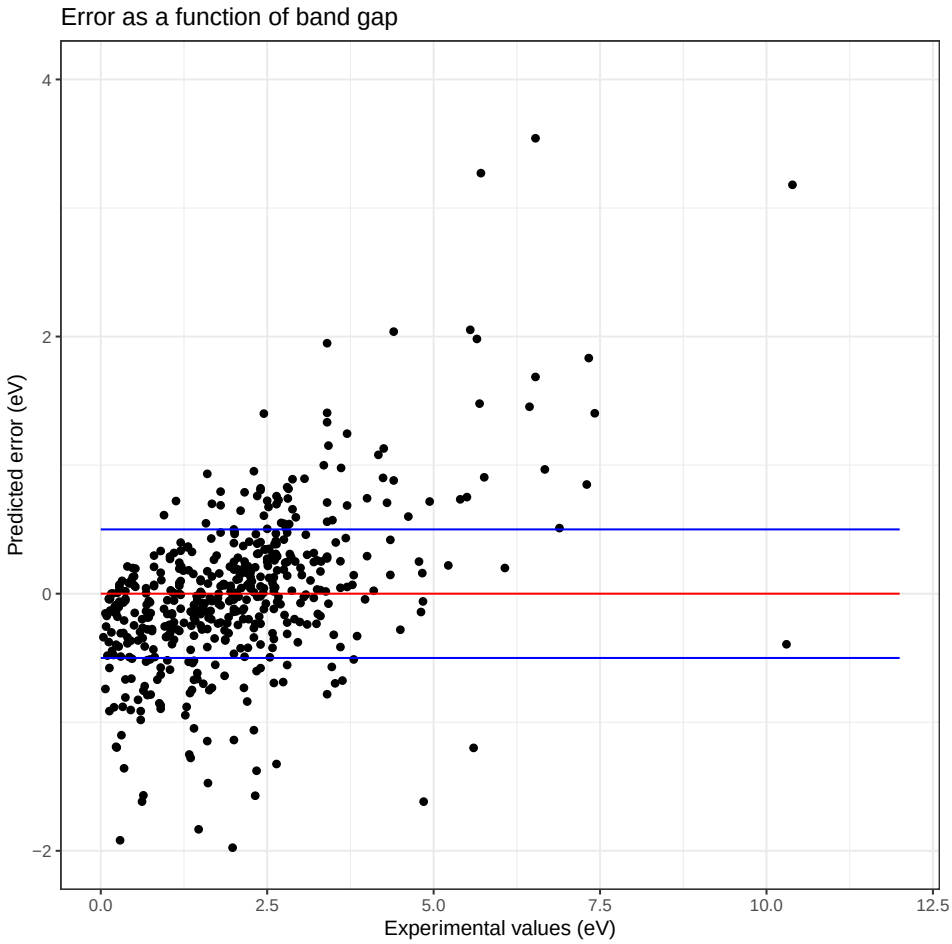


Figure 5.7: The scatter plot shows the true prediction error as a function of band gap as predicted by the cubist algorithm on a test set. The red line shows perfect prediction and the blue lines mark the 0.5 eV error cutoff. Any point below the red line is overestimated and any point above the red line is underestimated.

elements. This is also consistent with the hypothesis because all possible d-orbitals experience some degree of shielding. One would therefore not expect d-orbital shielding to be problematic for the machine learning models.

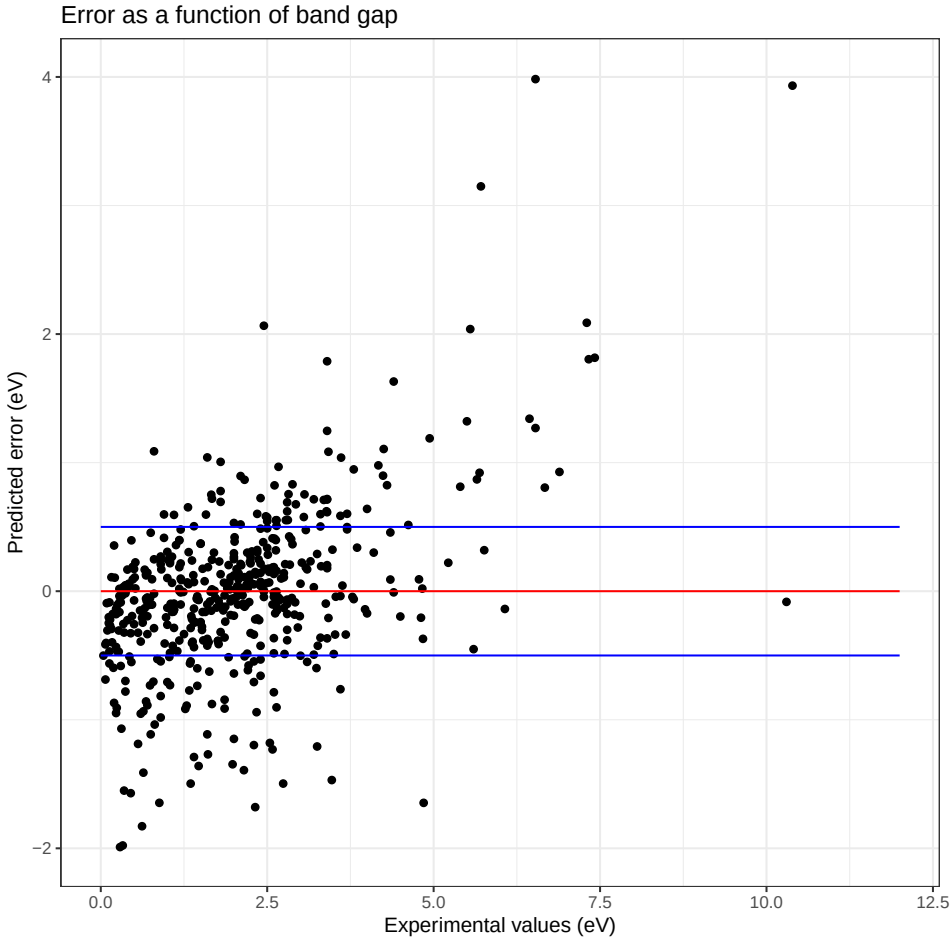


Figure 5.8: The scatter plot shows the true prediction error as a function of band gap as predicted by the SVR algorithm on a test set. The red line shows perfect prediction and the blue lines mark the 0.5 eV error cutoff. Any point below the red line is overestimated and any point above the red line is underestimated.

Table 5.2: Table showing the error [eV] of the seven compounds in the test set with errors greater than 2 eV

Compound	Error [eV]
HfO_2	2.05
K_2ReH_9	3.27
$Ba_2B_6O_9(OH)_4$	3.54
$LaGaO_3$	2.04
$V_3As_2O_9$	-2.29
BeO	3.18
$AgSO_4$	-3.16

Table 5.3: Table showing the mean error [eV] and element count of the ten worst performing elements in the test set. The predictions are made by a Cubist model.

Element	Mean error	Element count
Be	1.77	3
Re	1.71	2
Hf	1.12	3
H	1.12	7
Cr	0.98	3
F	0.91	7
B	0.9	32
V	0.78	10
O	0.74	108
Ta	0.64	8

Table 5.4: Table showing the mean absolute error [eV] associated with p-group elements.

Element	MAE [eV]	Element	MAE [eV]	Element	MAE [eV]
B	0.9	C	0.08	N	0.42
Al	0.3	Si	0.42	P	0.32
Ga	0.33	Ge	0.33	As	0.36
Element	MAE [eV]	Element	MAE [eV]		
O	0.74	F	0.91		
S	0.37	Cl	0.6		
Se	0.35	Br	0.16		

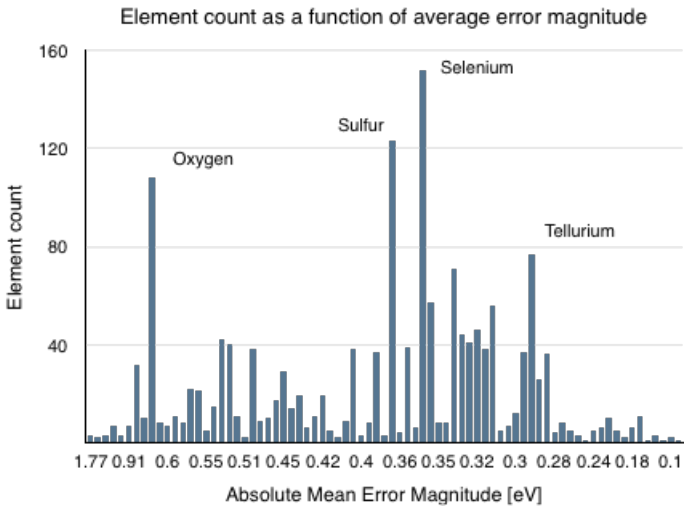


Figure 5.9: Bar graph showing the number of element instances as a function of absolute mean error. The bars corresponding to oxygen, sulfur, selenium and tellurium have been highlighted.

Figures 5.10, 5.11 and 5.12 are bar graphs showing the relative importance of the model descriptors. The importance is estimated using built-in functions in R. They indicate the extent to which a given variable is featured in the final model. Across all non-linear approaches, the Pauling electronegativity difference for the crystals and the average atomic weight are consistently considered significantly more important than all other available descriptors and are therefore worth paying extra attention to.

Equation 3.12 provides a direct relationship between the charge transfer in the bonds of two-component compounds and the resulting band gap. As the charge transfer increases so does the size of the band gap. This charge transfer is directly related to the electronegativity difference of the elements in the crystal. This explains not only why electronegativity is such an important variable in all of the non-linear models, but also why the representation provided by Equation 3.15 is preferred over the Zhuo et al.[11] derived element descriptors for electronegativity.

The average atomic weight for the compound, `sum(composition weighted atomic weight)`, is also prioritized. The average atomic weight was assumed to provide structural information. However, Van der Waals radius should provide similar information, yet the average atomic radius is consistently outperformed by the average atomic weight when considering variable importance. It is interesting that this is also true for the atomic size difference calculated using Equation 3.13. When studying the correlation, R^2, between these properties and the observed band gap, it seems clear why this might be the case. R^2 is significantly higher for the average atomic weight than for the two other descriptors. This information is presented in Table 5.5. Atomic weight is also associated with other properties such as degree of shielding of the valence electrons, atomic number, Z, as well as several of the properties included as descriptors, such as ionization energies, electronegativity and dipole polarizability. Figure 5.13 suggests that the band gap is ex-

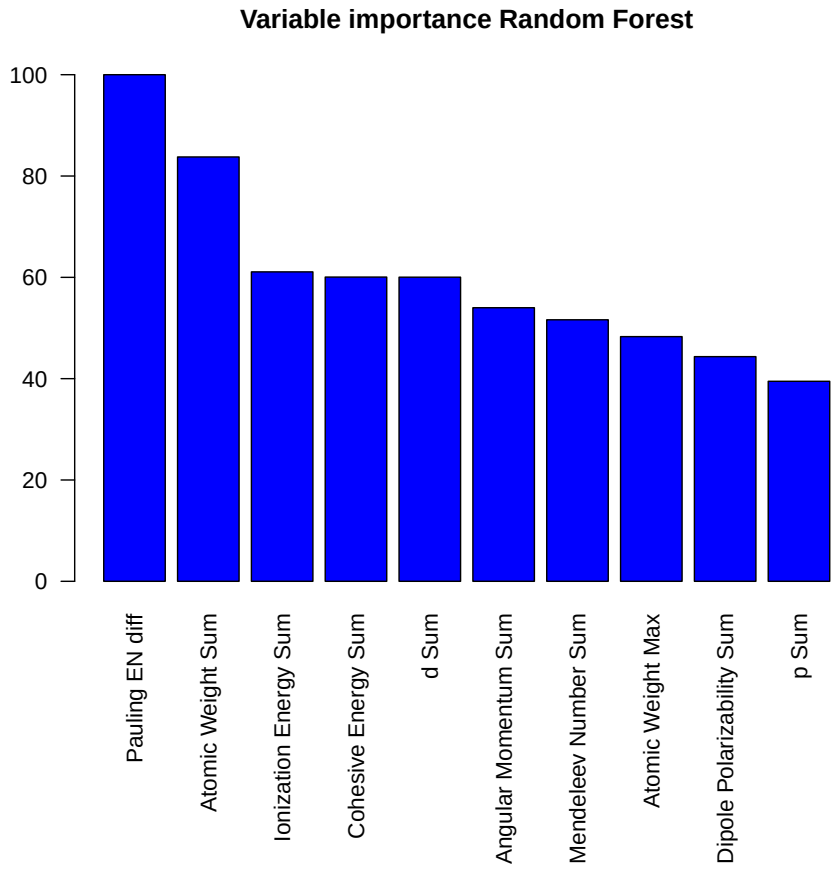


Figure 5.10: The bar plot shows the relative variable importance of the ten best descriptors in the Random Forest model. The top two variables are the Pauling electronegativity crystal difference, as defined by Equation 3.15 and the average atomic weight.

Table 5.5: Table showing the correlation, R^2, between the properties average atomic weight (avg weight), average atomic Van der Waals radius (avg VdW radius) and atomic size difference (δ).

Avg weight	δ	Avg VdW radius
0.35	0.25	0.17

pected to decrease as the average atomic weight increases. This is the case for many of the associated properties as well. Atomic weight might be expressing numerous of these associated properties in one variable, and is therefore picked more often by the decision tree models. This supposition is supported by Marikani[44, ch. 9.4]. He writes that the weight is associated with the atomic number which in turn is associated with a larger elec-

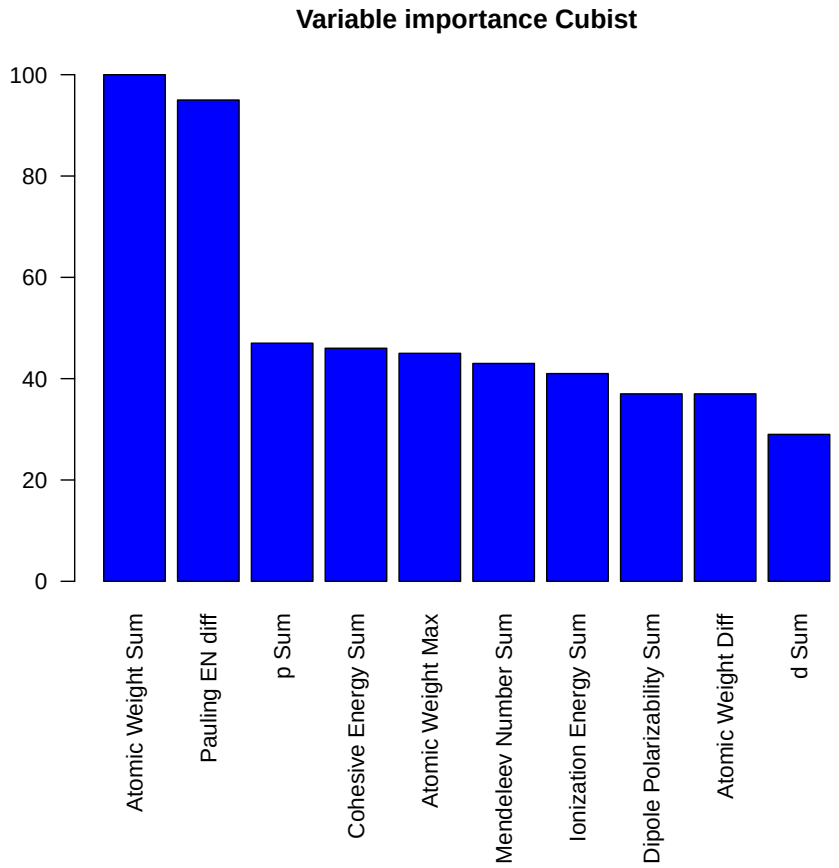


Figure 5.11: The bar plot shows the relative variable importance of the ten best descriptors in the Cubist model. The top two variables are the Pauling electronegativity crystal difference, as defined by Equation 3.15 and the average atomic weight.

tron cloud. The outer electrons will be less tightly bound for large molecules and the band gap therefore decreases. The average atomic weight is listed as a top two variable in the SVR model as well, but for the SVR model only the Pauling electronegativity crystal difference is markedly more important than the other variables. SVR models are created by considering all of the variables at the same time such that less emphasis is placed on the average atomic weight.

Some properties did not make it on to the top ten variable importance list in any of the models. They are electron count for outer f and s orbitals, thermal conductivity, VEC and configurational entropy. Thermal conductivity is not on the list, likely because the property is associated with elements in their elemental form. The thermal conductivity of oxygen

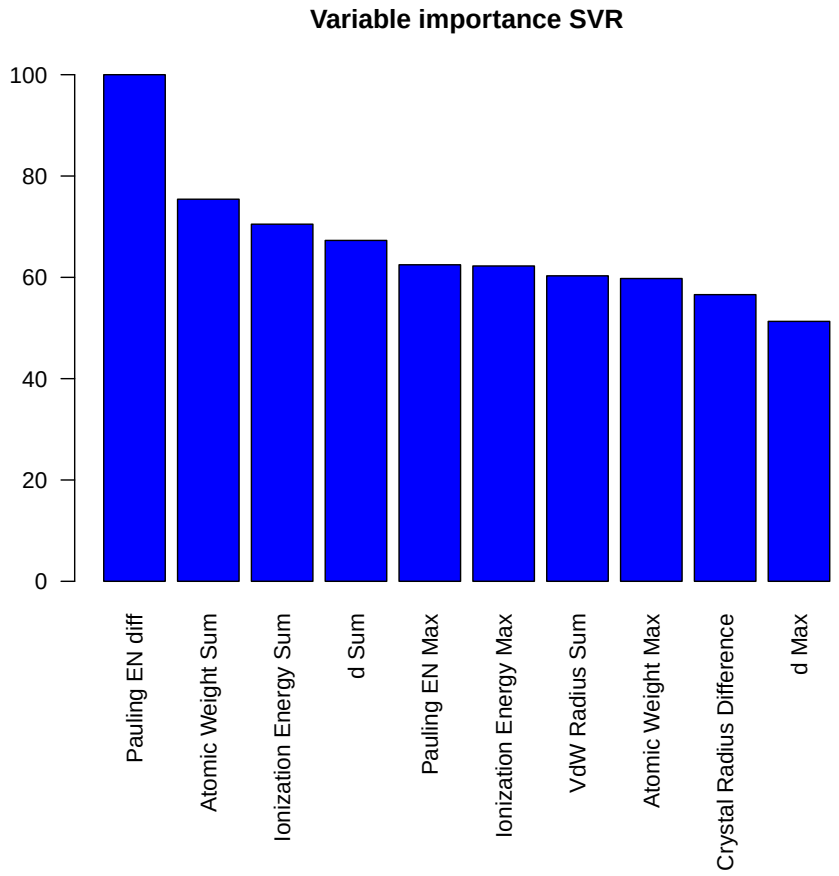


Figure 5.12: The bar plot shows the relative variable importance of the ten best descriptors in the SVR model. The top two variables are the Pauling electronegativity crystal difference, as defined by Equation 3.15 and the average atomic weight.

gas at room temperature is probably not associated with the equivalent crystal property. This is supported by the variable importance ranking. A similar observation is likely true for configurational entropy. The approximation that assumes atoms are free to occupy all crystal lattice sites does not provide accurate structural information. The f orbital is likely not included because of its low variance. Most elements do not have f-orbital valence electrons. After removing low variance and highly correlated columns, only the weighted max remained. As for s-orbital information, the number of s-orbital elements is relatively low when compared to d-orbital and p-orbital elements. This would explain why both p- and d-orbital descriptors are considered important, but not s-orbitals. The last property not represented in the top ten is the VEC, determined using Equation 3.16. The relevant

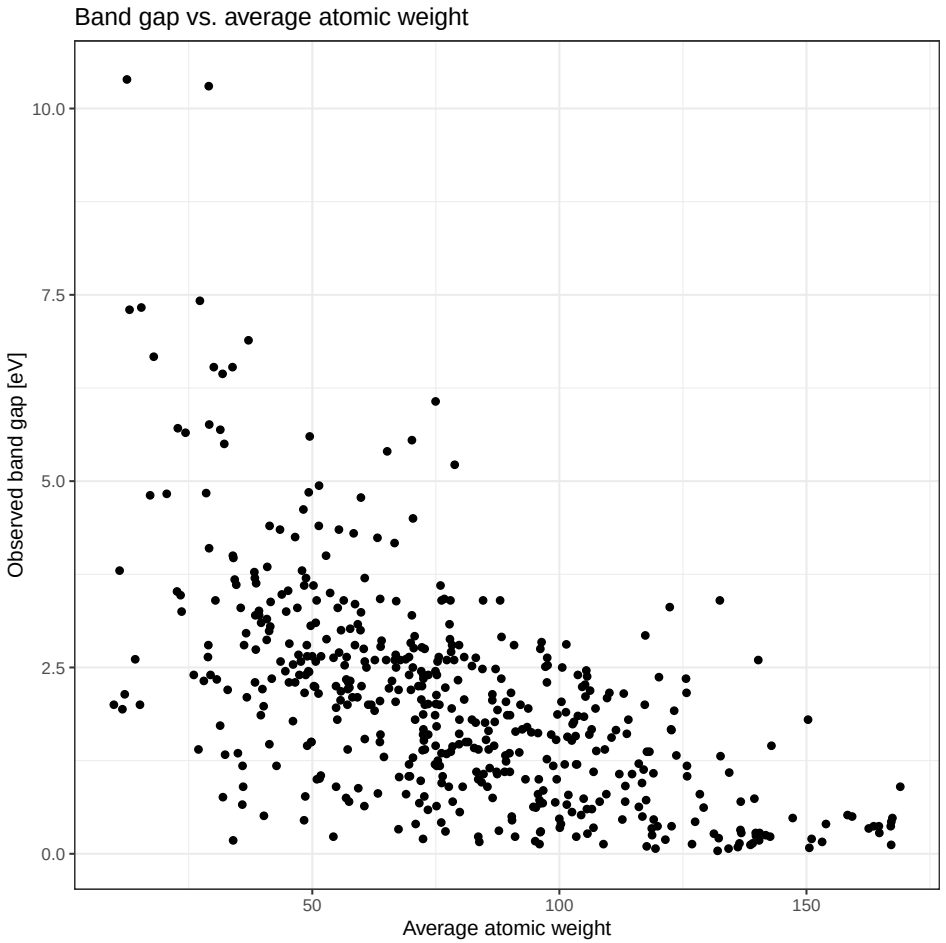


Figure 5.13: The scatter plot shows the observed experimental band gap [eV] vs. the average atomic weight [amu]. The band gap is observed to decrease as the average atomic weight increases.

information provided by this descriptor is likely captured by the valence orbital count p and d.

5.2 Applying the Cubist model to an external test set

Further model verification was done using an external data set assembled by Borlido et al.[25]. After removing single element compounds and phase duplicates, 69 compounds remained. The cubist model developed during initial testing was applied to this data set. The performance metrics of this investigation are reported in table 5.6 while Figures 5.14 and 5.15 show the predicted band gap and the prediction error respectively as a function of observed band gap. The performance, although slightly worse than the performance

reported during the initial phase, suggests that model performance is not greatly reduced when applied to completely unseen compounds. The decrease that is reported can be attributed to the small number of compounds. This makes metrics like R^2 and $RMSE$ less reliable.

Table 5.6: Table showing the correlation, R^2, between the properties average atomic weight (avg weight), average atomic Van der Waals radius (avg VdW radius) and atomic size difference (δ).

R^2	0.76
$RMSE$	1.30
$AARD$	38

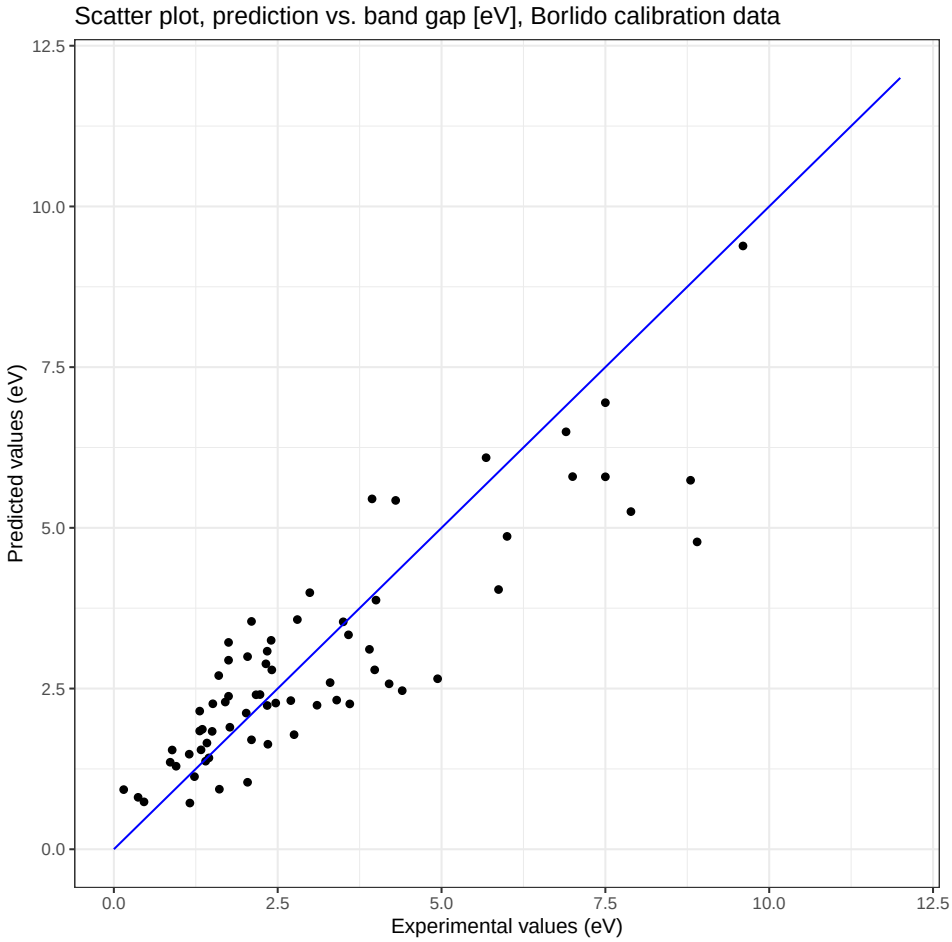


Figure 5.14: The scatter plot shows the predicted values by the Cubist model relative to the experimental values. The model was tested on an external data set provided by Borlido et al.[25].

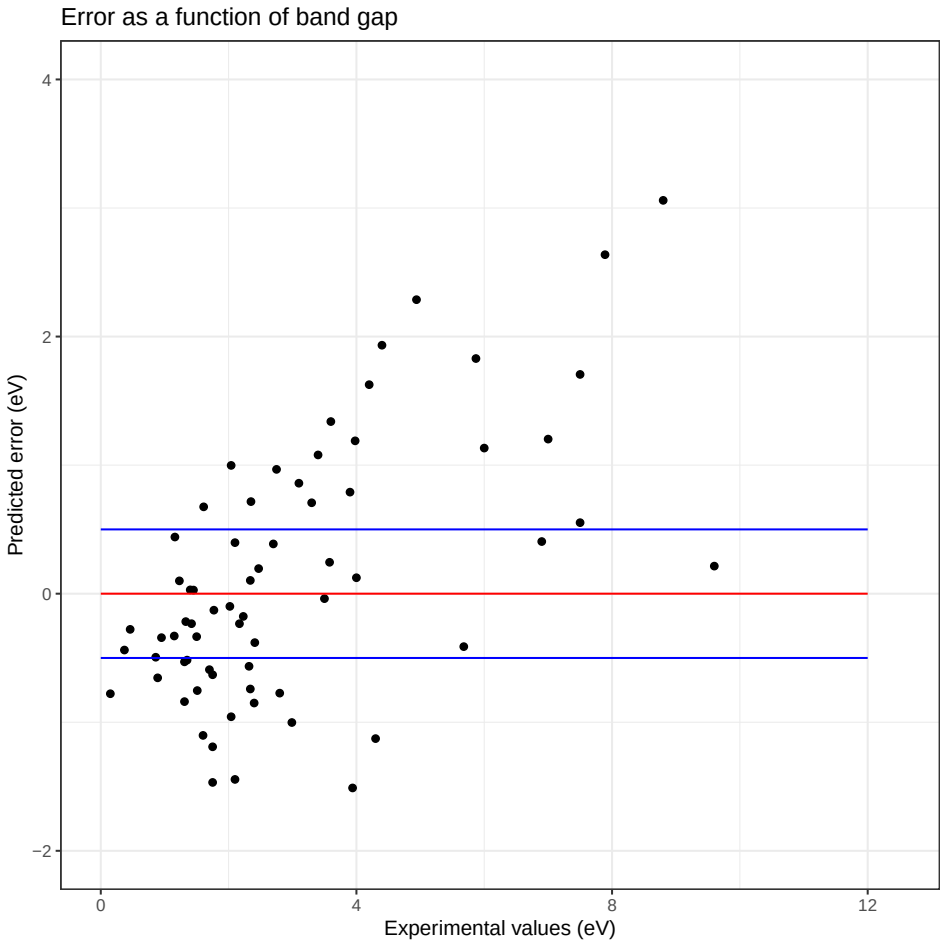


Figure 5.15: The scatter plot shows the true prediction error as a function of band gap as predicted by the cubist model applied to the Borlido et al. data set. The red line shows perfect prediction and the blue lines mark the 0.5 eV error cutoff. Any point below the red line is overestimated and any point above the red line is underestimated.

Further analysis of the error showed that five compounds, Al_2O_3, $AlPO_4$, LiH, SiO_2 and YF_3, had prediction errors grater than 2 eV. Of these five compounds, three are oxides and all have at least one element that is associated with a poorly shielded orbital. This is consistent with the p-orbital hypothesis presented in Section 5.2 to explain why some compounds are more difficult to predict than others.

5.3 Replicating Zhuo et al. Results Using the Cubist and SVR Algorithms

The model created by Zhuo et al.[11] was created using a data set with several duplicate compounds, many of which had very similar observed band gaps. For comparison, both the Cubist algorithm and SVR algorithms were applied to the full test set with duplicate compounds. Cubist performed the best, with an average R^2 value of 0.89 and an average $RMSE$ of 0.49 for the test set. The metrics are reported in Table 5.7. The results show that both models perform similarly with the Cubist model having results equal to the model developed by Zhuo et al.[11] There are, however, a few key differences in the approach chosen by Zhuo et al.[11] and the approach presented in this thesis. Zhuo et al.[11] uses 34 different element properties to form a descriptor set consisting of 136 descriptors. Of the 34 properties, nine were representations that provided similar information, including five versions of electronegativity and five versions of atomic radius. Because their results have been successfully replicated, this indicates that a large majority of the descriptors added to the descriptor set are contributing little or no information to the final model. Based on the variable importance analysis in Section 5.2, and the results presented in this section, it seems likely that reducing the descriptor set further would be beneficial.

There is a significant jump in performance when the machine learning algorithms are trained on data sets that include duplicate compositions. The data set utilized by Zhuo et al.[11] had 1438 duplicate compositions, roughly 1/3 of the entire data set. The band gaps reported for most of the duplicate compositions are so similar that they are likely the result of various experimental techniques and not the result of multiple composition phases, although this is not made clear. This is problematic because with multiple very similar data points present in the training sets, the machine learning algorithms are highly rewarded for overfitting, reducing the applicability of the model when applied to external compounds. The performance metrics for the test set are inflated because the test set likely has several of the duplicate compounds in it. The robustness of this type of machine learning model for property prediction should therefore likely be assessed using data sets that have had duplicates removed or averaged so that they are not rewarded for overfitting.

Table 5.7: Table showing mean model performance metrics for the Cubist and SVR algorithms applied to a data set with 1438 duplicate compounds as well as the results reported by Zhuo et al[11] using a similar approach. The descriptor set uses was the Zhuo et al[11]. derived descriptor set using sum(), min(), max() and max()-min() descriptors. R^2_{test}, $RMSE_{test}$, $AARD_{test}$ show the Coefficient of Determination, the square root mean error [eV], and the Average Absolute Relative Deviation [%] for the test set.

	Cubist	**SVR**	Zhuo et al.
R^2_{test}	0.89	0.86	0.90
$RMSE_{test}$	0.49	0.56	0.45
$AARD_{test}$	31	31	NA

5.4 Summary of Findings

The findings show that machine learning can, with some success, be applied to predicting the band gap of a wide range of compounds. The best performance was achieved using the Cubist algorithm applied to a previously published data set that contained multiple duplicate compounds, Zhuo et al.[11]. The performance indicated for these models is likely an overestimation caused by overfitting being rewarded due to the duplicates also being present in the test set.

The performance of the models dropped significantly when duplicates were removed from the data set, The Cubist algorithm still outperformed the other non-linear algorithms tested in this thesis, but the difference in performance was very limited. A R^2 value of 0.81, $RMSE$ of 0.65 eV and an $AARD$ of 47% was achieved. Both $RMSE$ and $AARD$ were significantly higher than equivalent values reported for DFT calculations. Therefore, the true robustness of machine learning models using the strategy proposed by Zhuo et al.[11] and further expanded upon in this thesis is likely better determined using a data set with only unique compositions.

The findings also show that some elements are harder to predict than others using the descriptor set presented in this thesis, indicating that more precise representations of the element properties could increase the performance of the models.

Chapter 6

Conclusion

The application of machine learning algorithms to the problem of band gap prediction in a wide variety of compounds has been investigated. Using only element properties and properties for crystals that are easily estimated, the best performing model achieved a R^2 value of 0.81, a $RMSE$ of 0.65 eV and an $AARD$ of 47%. Similar performance was also achieved when the model was applied to an independent data set. The performance of machine learning models using only element properties does not yet match the performance that has been reported for equivalent DFT calculations.

The spread of errors was studied to assess how the data band gap distributions effected the prediction made by the models. This assessment showed that the models tend to underestimate high band gaps and overestimate low band gaps. This problem can be explained by the fact that the machine learning algorithms did not have enough high band gap cases, nor any zero band gap cases to learn from. Best predictions were achieved for band gaps in the range of 1 - 2 eV, where the models neither preferentially overestimate nor underestimate the band gap value.

Some compounds were significantly harder to predict correctly than others. A large majority of these compounds were either oxides or compounds where one or more of the elements in the structure was poorly represented in the data set. It is hypothesized that oxides perform particularly poorly because of the 2p-orbital valence electrons. The orbital descriptors do not distinguish between 2p, 3p and 4p orbitals even though 2p orbitals experience far less electronic shielding than the 3p and 4p orbitals do. The trend is investigated for the remaining p-orbital elements as well as the s-orbital and d-orbital. This investigation provided consistent results with the proposed hypothesis. 2p-orbital elements were consistently more difficult to predict than 3p- and 4p-orbital elements. The same trend was true for s-orbital elements, but not for d-orbital elements. This is also consistent because the lowest d-orbital, 3d, experiences shielding, whereas the lowest s-orbitals, 1s and 2s, do not.

The variable importance was investigated and the results showed that Pauling Electronegativity crystal difference and the average atomic weight were consistently considered significantly more important than the other descriptors in the descriptor set. It is sug-

gested that the Pauling electronegativity is important because it provides information about the charge transfer between bonds in the structure. Equation 3.12 states that larger band gaps should be expected when the charge transfer between atoms in the crystal increases. No such clear explanation was found for the average atomic weight, but it is suggested that because atomic weight is associated with the atomic number, and the atomic number is indicative of the size of the electron cloud, the average atomic weight is providing relevant electronic information.

Several of the descriptors were not prioritized in any of the models. These descriptors were the f- and s-orbital valence electron count, the thermal conductivity, the VEC and the configurational entropy. A variety of reasons for why these properties are largely ignored are proposed. Future machine learning models should pay particular attention to what descriptor representations of physical properties are the most likely to provide the intended information.

The Cubist and SVR algorithms were finally tested on the full data set used by Zhuo et al. This set included a significant amount of duplicate compositions with very similar band gaps. A R^2 value of 0.89 and a $RMSE$ of 0.49 was achieved, matching the performance metrics reported by Zhuo et al. However, the model is likely overfitting the data because of the presence of duplicates in both the training and test data. Therefore, a more true estimation of a model's predictive abilities is achieved if the model is trained using only unique compounds.

Clearly, based on the performance metrics achieved, machine learning models based solely on element and simple crystal structure properties cannot yet compete with DFT calculations for accuracy. However, they vastly outperform DFT on speed and ease of use, making them a potentially useful tool for preliminary band gap estimations. There are several areas that should be investigated in order to further improve the machine learning models. There is a pressing need for descriptors that more precisely provide relevant information. Secondly, work should be done to reduce the number of descriptors further. This will make physical interpretation of the models easier.

Several of the problems that the models had were related to problems with the data set. Both the lack of data about compounds with high band gaps, as well as several elements being severely underrepresented seem to have caused problems. Compilation of a data set that more accurately represents the entire spectrum of compounds would increase the chance of successfully developing a model that can compete with DFT calculations.

Bibliography

[1] Materials Genome Initiative about the materials genome initiative. `https://www.mgi.gov/`. Accessed: 2019-08-24.

[2] A. R. West. *Solid State Chemistry and its Applications*. Wiley, United Kingdom, 2 edition, 2014.

[3] M. Quirk and J. Serda. *Semiconductor Manufacturing Technology*. Pearson Prentice Hall, United States of America, 1 edition, 2001.

[4] J. Schmidt et al. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):83, 2019.

[5] A. R. Leach. *Molecular Modelling*. Pearson Prentice Hall, England, 2 edition, 2001.

[6] T. Mueller et al. Machine learning in materials science: Recent progress and emerging applications. In Abby L. Parrill and Kenny B. Lipkowitz, editors, *Reviews in Computational Chemistry*, chapter 4, pages 186–273. John Wiley & Sons, 2016.

[7] A. Teller and M. Veloso. Algorithm evolution for face recognition: what makes a picture difficult. In *Proceedings of 1995 IEEE International Conference on Evolutionary Computation*, volume 2, pages 608–613 vol.2, 1995.

[8] G. Cui et al. Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *MANAGEMENT SCIENCE*, 52(4):597–612, 2006.

[9] E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559 – 569, 2011.

[10] L. Zhang et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, 22(11):1680–1685, 2017.

[11] Y. Zhuo et al. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.*, 9:1668–1673, 2018.

[12] N. Islam et al. Machine learning for phase selection in multi-principal element alloys. *Computational Materials Science*, 150:230–235, 2018.

[13] D. Michie. "memo" functions and machine learning. *Nature*, 218:19–22, 1968.

[14] S. Goyal and P. Benjamin. Object recognition using deep neural networks: A survey. *arXiv:1412.3684*, 2014.

[15] J. H. Noordik. *Cheminformatics Developments: History, Reviews and Current Research*. IOS Press, Nieuwe Hemweg 6B, Amsterdam, 1 edition, 2004.

[16] T. Engel. Basic overview of chemoinformatics. *Journal of Chemical Information and Modeling*, 46(6):2267–2277, 2006.

[17] T. Engel and J. Gasteiger. *Applied Chemoinformatics: Achievements and Future Opportunities*. John Wiley & Sons, Germany, 1 edition, 2018.

[18] K. T. Butler et al. Machine learning for molecular and materials science. *Nature*, 559:547555, 2018.

[19] K. Ryan et al. Crystal structure prediction via deep learning. *Journal of the American Chemical Society*, 140(32):10158–10168, 2018.

[20] P. V. Balachandran et al. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nature Communications*, 9(1):1668, 2018.

[21] R. Jalem et al. Bayesian-driven first-principles calculations for accelerating exploration of fast ion conductors for rechargeable battery application. *Scientific Reports*, 8(1):5845, 2018.

[22] F. Capasso. *Physics of quantum electron devices*. Springer, Berlin, 1 edition, 1990.

[23] N.N. Kiselyova et al. Computational materials design using artificial intelligence methods. *Journal of Alloys and Compounds*, 279:813, 1998.

[24] K. Burke. Perspective on density functional theory. *J. Chem. Phys*, 136:150901, 2012.

[25] P. Borlido et al. Large-scale benchmark of exchangecorrelation functionals for the determination of electronic band gaps of solids. *Journal of Chemical Theory and Computation*, 2019.

[26] J. Lee et al. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B*, 93:115104, 2016.

[27] G. Pilania et al. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Computational Materials Science*, 129:156–163, 2017.

[28] R. E. Walpole et al. *Probability & Statistics for Engineers and Scientists*. Pearson Education, United States of America, 9 edition, 2012.

[29] S. Wold et al. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130, 2001.

[30] L. Breiman. Random Forests. *Machine Learning*, 45:532, 2001.

[31] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199222, 2004.

[32] G. James et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, Great Britain, 8 edition, 2017.

[33] K. Paduszynski and U. Domanska. Viscosity of ionic liquids: An extensive database and a new group contribution model based on a feed-forward artificial neural network. *Journal of Chemical Information and Modeling*, 54:1311–1324, 2014.

[34] C. Kittel. *Introduction to Solid State Physics*. John Wiley % Sons, United States of America, 8 edition, 2005.

[35] W. D. Callister and D. G. Rethwisch. *Materials Science and Engineering*. Wiley, United States of America, 8 edition, 2011.

[36] P. Atkins and R. Friedman. *Molecular Quantum Mechanics*. Oxford, United States of America, 5 edition, 2011.

[37] R. J.D. Tilley. *Understanding Solids*. Wiley, United States of America, 2 edition, 2013.

[38] A. R. West. *Solid State Chemistry and Its Applications*. John Wiley & Sons, United States of America, 5 edition, 1991.

[39] I. Mills et al. *Quantities, Units and Symbols in Physical Chemistry*. Blackwell Science, Great Britain, 2 edition, 1993.

[40] J. R. Rumble. *CRC Handbook of Chemistry and Physics*. CRC Press/Taylor & Francis, United States of America, 100 edition, Internet Version 2019.

[41] P. Villars. Data-driven atomic environment prediction for binaries using the mendeleev number: Part 1. composition ab. *Journal of Alloys and Compounds*, 367:167175, 2004.

[42] K. A. Dill and S. Bromberg. *Molecular Dricing Forces, Statistical Thermodynamics in Biology, Chemistry, Physics and Nanoscience*. Garland Science, United States of America, 2 edition, 2011.

[43] T. Umebayashi et al. Band gap narrowing of titanium dioxide by sulfur doping. *Applied Physics Letters*, 81(3):454–456, 2002.

[44] A. Marikani. *Materilas Science*. PHI, Nieuwe Hemweg 6B, Amsterdam, 1 edition, 2017.

[45] mendeleev – a python resource for properties of chemical elements, ions and isotopes, ver. 0.3.6. https://github.com/lmmentel/mendeleev, 2014–.

Appendix A

Data Information

Table A.1: Table showing the unique element instance count for all elements in the band gap data. The table is ordered from most common to least common compound.

Element	Unique instances
Se	694
S	622
O	534
Te	401
Ga	312
In	289
Sb	266
P	242
Ge	218
Cu	217
As	203
Cd	191
Ba	190
Bi	190
Sn	184
K	177
Pb	175
Cs	165
Zn	158
B	125
Ag	124
Hg	121
Rb	113
I	101
Na	101

Li	99
Si	92
Tl	86
La	78
Cl	69
Al	64
N	60
Eu	56
Br	53
Ta	53
V	52
Sm	50
Sr	43
Nb	41
Gd	40
Ca	39
F	39
Mn	39
Mo	39
Ti	37
Ce	36
Fe	34
Mg	33
Y	33
Yb	33
H	32
Pr	31
Sc	30
Nd	29
Dy	27
Zr	27
Cr	24
Pd	23
Er	21
Co	20
Lu	19
Tb	18
Pt	16
Ni	14
Ru	14
Th	14
W	14
Ho	13
Ir	13
C	12

Os	12
Au	11
Rh	11
Hf	10
Be	9
U	9
Re	6
Tm	6
Tc	4

List of descriptors that have been removed, due to low variance or high correlation:

- f-orbital sum
- f-orbital max
- f-orbital min
- Dipole polarizability max - min
- Thermal conductivity max

Appendix B

Element Property References

Table B.1: Table showing the mean error [eV] and element count of the ten worst performing elements in the test set. The predictions are made by a Cubist model.

Property	Reference
Atomic weight	[45]
Medeleev number	[45]
Pauling electronegativity	[40]
Ionization energy	[40]
Electron configuration	[45]
Angular momentum quantum number	[45]
Dipole polarizability	[40]
Thermal conductivity	[40]
Cohesive energy	[34]