5-2017

Data driven discovery of materials properties.

Fadoua Khmaissia
University of Louisville

Recommended Citation

Khmaissia, Fadoua, "Data driven discovery of materials properties." (2017). *Electronic Theses and Dissertations.* Paper 2700.
https://doi.org/10.18297/etd/2700

DATA DRIVEN DISCOVERY OF MATERIALS PROPERTIES

By

Fadoua Khmaissia
B.E., Telecommunications Engineering, Higher School of Communications of Tunis, 2015

A Thesis
Submitted to the Faculty of the
J. B. School of Engineering at the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science in Computer Science

Department of Computer Engineering and Computer Science
University of Louisville
Louisville, Kentucky

May 2017

DATA DRIVEN DISCOVERY OF MATERIALS PROPERTIES

By

Fadoua Khmaissia
B.E., Telecommunications Engineering, Higher School of Communications of Tunis, 2015

A Thesis Approved On

April 21st, 2017

Date

By the following Thesis Committee:

Hichem Frigui, Ph.D., Thesis Director

Mahendra Sunkara, Ph.D.

Olfa Nasraoui, Ph.D.

ACKNOWLEDGEMENTS

"The teacher who walks in the shadow of the temple, among his followers, gives not of his wisdom but rather of his faith and his lovingness. If he is indeed wise he does not bid you enter the house of his wisdom, but rather leads you to the threshold of your own mind." Khalil Gibran. That being said, I would like to convey my special regards to my distinctive advisor, Dr. Hichem Frigui for his perpetual guidance, friendly assistance and enthusiastic encouragements during both planning and fulfillment of this work.

I thank Dr. Olfa Nasraoui and Dr. Mahendra Sunkara for accepting to serve in my thesis committee and being a part of this special milestone.

I address, likewise, my thanks to my colleagues in the Multimedia Research Laboratory, and the Computer Engineering and Computer Science Department for their support and friendship.

Finally, I want to convey my most heartfelt thanks to my family for their continuous support and unconditional love.

ABSTRACT

DATA DRIVEN DISCOVERY OF MATERIALS PROPERTIES

Fadoua Khmaissia

April 21st, 2017

The high pace of nowadays industrial evolution is creating an urgent need to design new cost efficient materials that can satisfy both current and future demands. However, with the increase of structural and functional complexity of materials, the ability to rationally design new materials with a precise set of properties has become increasingly challenging. This basic observation has triggered the idea of applying machine learning techniques in the field, which was further encouraged by the launch of the Materials Genome Initiative (MGI) by the US government since 2011.

In this work, we present a novel approach to apply machine learning techniques for materials science applications. Guided by knowledge from domain experts, our approach focuses on machine learning to accelerate data-driven discovery of materials properties. Our objectives are two folds: (i) Identify the optimal set of features that best describes a given predicted variable. (ii) Boost prediction accuracy via applying various regression algorithms.

Ordinary Least Square, Partial Least Square and Lasso regressions, combined with well adjusted feature selection techniques are applied and tested to predict key properties of semiconductors for two types of applications. First, we propose to build a more robust prediction model for band-gap energy (BG-E) of chalcopyrites, commonly used for solar cells industry. Compared to the results reported in [1–3], our approach shows that learning and using only a subset of relevant features can improve the prediction accuracy by about 40%. For the second application, we propose to determine the underlying factors responsible for Defect-Induced Magnetism (DIM) in Dilute Magnetic Semiconductors (DMS) through the analysis of a set of 30 features for different DMS systems. We show that 8 of these features are more likely to contribute to this property. Using only these features to predict the total magnetic moment of new candidate DMSs has reduced the mean square error by

about 90% compared to the models trained using the whole set of features.

Given the scarcity of the available data sets for similar applications, this work aims not only to build robust models but also to establish a collaborative platform for future research.

TABLE OF CONTENTS

LIST OF TABLES

TABLE **Page**

LIST OF FIGURES

FIGURE **Page**

CHAPTER 1

INTRODUCTION

The discovery of semiconductors marked a key milestone for the digital age. They cover a wide range of applications and they are, most prominently, considered as the building blocks of modern electronics [4]. They are used for computers' and high speed chips manufacturing, space research, medical science, energy efficient lighting and solar energy production. Each particular application requires the identification and production of the semiconducting materials having the most suitable properties. The high pace of nowadays industrial evolution is creating an urgent need to design new cost efficient materials that can satisfy both current and future demands. However, time has always been the main constraint to achieve this goal. It has become almost intolerable to waste it synthetizing and testing thousands of new materials candidates while only few of them are actually needed. Even though this exhaustive experimental process has contributed to the discovery of several new materials, it still depends on chance and luck to succeed as fast as possible. This reflects a general oddity of materials design which "still depends on serendipity" [5, 6].

The classical approach to overcome this problem was to use first principle ab initio methods for predicting electronic properties. For instance, computational quantum mechanical modeling techniques, such as Density Functional Theory (DFT) calculations have been widely used for this purpose. DFT provides the ground state characteristics of a given compound by modeling every interacting system of intrinsic fermions via its density and not via its many-body wave function [7, 8]. It can predict a great variety of molecular properties: vibrational frequencies, molecular structures, atomization and ionization energies, electric and magnetic properties, reaction paths, etc. [9].
Due to the intricate nature of the relationships between particles such methods are computationally expensive which makes a large scale investigation of interesting materials rather infeasible [10]. Hence, the idea of applying machine learning techniques in the field has emerged.

1.1 A new vision for Materials Innovation

Materials properties can be seen as the result of the interactions between more than one factor, each with a different weight depending on the final target. This includes the atomic composition, the material's morphology and microstructure, the physical state, and many other intrinsic and extrinsic parameters that can be related to the preparation conditions [11]. This basic observation has triggered the idea of designing new tools that are able to model these interactions in a way that facilitate the prediction of the properties an eventual new compound. The created models should benefit from the already established theoretical background in order to adjust the search space and to limit the set of materials candidates when looking for a new discoveries.

Materials design is, therefore, an interdisciplinary field that requires the collaboration between more than one research area to get the best results within the least possible amount of time. The "Materials Genome Initiative" (MGI) launched by the US government since 2010 has emphasized this vision [9]. It is a multi-agency initiative rolled out to encourage the creation of shared resources and infrastructures to support national institutions in their effort to design and expedite the synthesis of new advanced materials both rapidly and cost-efficiently [12–14]. As a result, combinatorial and high-throughput (CHT) experimentation [11] in materials science has been acclaimed as a new scientifically-efficient approach to generate new knowledge [11, 13].

Exploiting these unique opportunities established a new infrastructure for materials innovation. As illustrated in Figure 1.1, this infrastructure is based on the inter-operability between three main procedures: Theory, Experiments and Computer simulations, each involving different areas of studies and expertise [13].

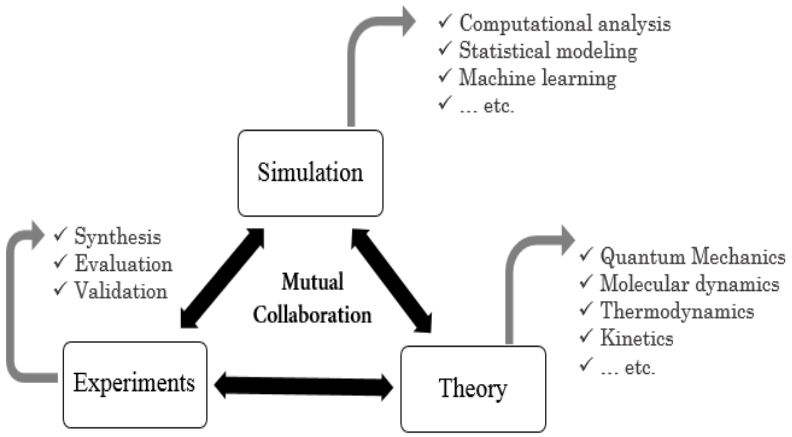

Figure 1.1: Materials Innovation infrastructure

1.2 Predictive analytics

The interdisciplinary aspect of the new materials innovation infrastructure has been associated with the raise of data science as a major contributor. In fact, machine learning has proven to be a promising field through the past decades. It is a flexible area of study that, based on data analysis and patterns extraction, adequate models are learnt to describe the inherent structure and behavior of a given training set of observations. Such models, if judiciously validated, can predict the response of any future unlisted observation.

Even though the integration of machine learning techniques for materials discovery purposes is still new, it is beginning to show enormous promises [15–19]. Moreover, data-centric approaches can provide valuable insights into the fundamental rules and aspects underlying materials behavior which have been difficult to apprehend for decades [15]. Both supervised and unsupervised algorithms could be applied to predict materials properties, depending on the availability of the training sets. However, due to data scarcity, and to the complexity of generating new measurements, supervised techniques are used more often. For instance, statistical learning, such as regression techniques are very popular in the field [15, 20].

1.3 Contributions

As we enter this MGI era [12], it has become crucial to quickly and accurately predict the properties of new materials that have yet to be synthesized. Applying machine learning techniques to develop an efficient computational tool for solving this specific problem is a new yet promising research area.

This thesis finds its roots within this context. Different regression algorithms combined with well adjusted pre-processing techniques will be applied and tested on two different applications. These applications involve predicting two key properties of semiconductors for different types of applications as described below.

The first application is based on a previous work aiming to predict the bandgap of new chalcopyrite compounds using statistical learning approaches such as Ordinary Least Squares, Partial Least Squares and Lasso regression methods coupled with Principal Component Analysis [3]. The used data set comprises atomic and crystallographic properties of ternary chalcopyrites semiconductors which are CuFeS2-like compounds that crystallizes in the tetragonal system (ABC2 formula) [21]. The band-gap is, in fact, a key property for solar cell materials design. It refers to

the energy gap (in eV) between the top of the valence band and the bottom of the conduction band in semiconductors and insulators [22].

Our replication and analysis of the previous results indicated that the predictor's performances can be enhanced. Our contribution herein is based mainly on features analysis. In fact, Band-gap engineering is a delicate task and obtaining an accurate and reliable prediction of totally new compounds requires a robust modeling that has to take into account not only the intrinsic characteristics of the included elements but also the interactions between them. This work will explore this possibility by predicting the band-gap of the same compounds after adding the duly chosen binary descriptors.

The second application is designed to model the physical properties of Dilute Magnetic Semiconductors (DMS). These are materials that exhibit both ferromagnetic and semiconductor properties. These materials are widely studied within the newly emerged field known as Spintronics (Spin Transport Electronics) [23–25]. If implemented within electronic devices, these materials offer the exciting prospect of combining classical semiconductor electronics with non volatile magnetic storage by providing a new type of conduction control. The intrinsic electronic spin as well as its associated magnetic moment are the key features to assess the level of applicability of a given new DMS compound [24]. This work will focus on the analysis and modeling of the magnetic properties that DMS materials exhibit upon co-doping with extrinsic defects at room temperature. For different sets of hosts, features' selection and regression algorithms will be investigated in order to predict the total magnetic moment resulting from the co-doping process. Features' analysis will be emphasized during this process in order to unveil the synergistic action of the different descriptors.

The remainder of this thesis is organized as follows. Chapter 2 provides a review of the existing methods used for band gap engineering and DMS materials design, as well as the application of machine learning techniques for knowledge discovery. Chapter 3 introduces our proposed learning approach. Chapter 4 provides experimental results and analysis of the proposed methods. Finally, chapter 5 provides conclusions and potential future work.

CHAPTER 2

LITERATURE REVIEW

This chapter will depict some aspects of this work's background. The key concepts will be described in details.

2.1 Application 1:Band-Gap engineering

Solar energy provides around 2% of the world's total energy [26]. But it has the potential to provide much more than that if the true challenges behind its industry are well addressed. Overcoming the barriers to boost solar power generation requires several engineering innovations in different fields starting from capturing solar energy and converting it to useful forms, ending by storing it for later use.

The main challenge here, is therefore to design powerful, cost-efficient solar cells most often made of semi-conductors like silicon. Given their manufacturing costs, modules of today's solar cells incorporated in the power grid would produce electricity at a cost roughly 2 to 6 times higher than current electricity prices [22].

A key step to designing new solar material is that of predicting the electronic properties of the prospect compound before manufacturing it.

An important property of any new solar material is its band gap. Heuristically, the band gap of a material can be defined as the amount of energy needed to change the conductive properties of a semiconductive material. Based on the photovoltaic effect, solar cells convert the absorbed light into electricity. Their basic principle can be easily understood by considering P-N junctions - based solar cells. Two types of materials are put next to each other, one has abundance of free negative charge carriers (electrons), called n-type material, and the other one has many free positive charges (holes), called p-type material. Upon absorption of an incident radiation, electrons from the p-type layer are excited, jump across the barrier into the upper n-type layer and then escape out into the circuit [27]. Efficiency is the most important characteristic of solar cells. It is calculated as the ratio of the created electricity by the absorbed light (Equation 2.2).

There is a very important trade-off that should be made in order to guarantee an acceptable efficiency. In fact, a good efficiency is tightly coupled to the properties of the used materials, especially the band gap [28]. And this is why band gap engineering is a key process within the manufacturing cycle.

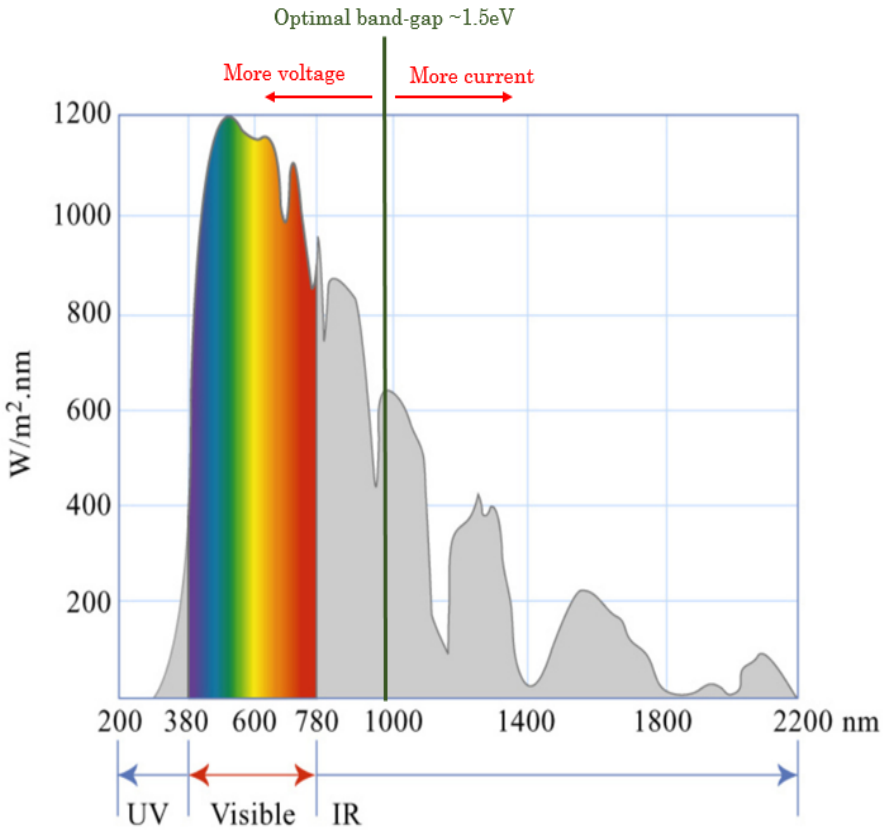

Figure 2.1: Solar spectrum and semiconductors band gap.[1]

Figure 2.1 shows the solar energy intensity on the Earth surface versus the radiations' wavelength. The band gap energy of the solar cell material should be chosen as low as possible (higher wavelength) in order to ensure more absorption of light with higher energy, which will excite more electrons and thus more current will be generated. At the same time, the band energy should be kept as high as possible to retain a high enough output voltage. In fact, the output voltage is directly related to the band-gap of the cell, and even with very high current, if the output voltage is low, then the output power will be too low which will drastically deteriorate the cell efficiency (Equations 2.1 and 2.2). Theoretical calculations have shown that the efficiency for a single band

[1] Image adapted from [29].

gap semiconductor, is maximum (around 33%) at a band gap $1.5eV$ for standard conditions [27,29].

$$\mathbf{Power}(W) = \mathbf{Voltage}(V) * \mathbf{Current}(I), \tag{2.1}$$

$$\mathbf{Efficiency} = \frac{\mathbf{Voltage} * \mathbf{Current}}{\mathbf{IncidentEnergy}}, \tag{2.2}$$

Satisfying the various constraints and defining the optimal range of band gap is still a very challenging task to perform. Yet, even for a given band gap range, finding the adequate materials is much more strenuous. To perform this task, scientists relied most on *abinitio* techniques. Standard Density Functional Theory (DFT) methods, for instance, were the workhorse of computational materials science for a good while. They provided acceptable results. However, they are still computationally expensive which justifies the limited set of compounds that has been studied till now.

Materials scientists have been considering chemo-informatic alternatives to estimate band gaps for years. The work of Zeng et al. (2001) [1] and that of Suh and Rajan (2004) [30] and the extensions that were based on it (2014) [3] have laid the basic framework for our investigation. They attempted to estimate the band gap of 28 known chalcopyrite compounds through the implementation of different regression techniques. They used five elementary descriptors for each atom present in the studied ternary compounds (ABC_2 formula, where A, B, and C are three atoms.); The atomic number(AN), the electronegativity (EN), the valency (VL), the melting point (MP) and the pseudo-potential radii (PR). Therefore, the band-gap (BG) of the compound ABC2 was predicted as a function of MP(X), AN(X), EN(X), VL(X), PR(X) with $X\epsilon\{A, B, C\}$. This work will focus on this choice of features. It aims, mainly, to determine if all of the previously chosen descriptors are relevant and have meaningful contribution to the prediction process while investigating the possibility of adding new ones.

2.2 Application 2: Dilute Magnetic Semiconductors

Spintronics (Spin-based electronics) research aims to investigate new applications and functionalities to microelectronic devices by engineering the carrier's spin, instead of or in addition to its charge. Metal-based spintronic devices have been widely studied and integrated in circuits. Expanding these studies to semiconductor devices can open wider horizons into achieving the full potential of spintronics. This is why the new trends in the field are focusing in developing magnetic semiconductors, Dilute Magnetic Semiconductors for instance [31].

Dilute magnetic semiconductors (DMS) are a subclass of magnetic semiconductors where a fraction of the cations in the lattice are substituted by magnetic ions. They are typically constracted

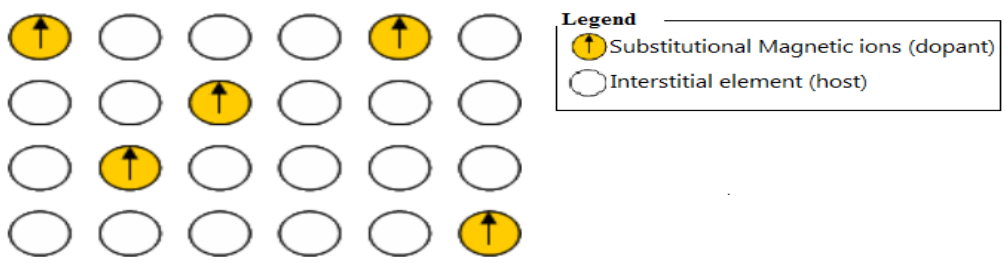

Figure 2.2: Semiconductor host doped with magnetic ions.

as alloys between a nonmagnetic semiconductor (host) and a magnetic element deriving from the doping compound. Figure (2.2) illustrates the typical structure of a DMS material [31]. The exchange interaction between the spin of the dopant atoms and the carriers in the semiconductor host can alter the global ferromagnetic properties yielding an extremely interesting characteristic.

Understanding the origins of magnetism in these materials is still a very challenging task, it was subject to various studies starting from the 1980s [32]. Several mechanisms have been suggested that describe the origin of magnetism in DMS like Mean Field Theory [33] and Bound Magnetic Polaron [34]. Most theories attempt to identify the various spin coupling energetic concurrent in a system, and by plugging it in the material parameters, attempt to estimate if the energetics lead to ferromagnetic, antiferromagnetic or spin-glass like interactions between individual atomic spins. Even though the exact underlying process is not yet modeled, the idea of electrically tuning DMS magnetism remains a fascinating prospect. This can create many revolutionary functionalities [32].

Semiconductors, unlike ordinary metals, offer the interesting opportunity that their properties can be tailored to fit the target applications. Spintronics is centered around three main process; injection, manipulation and detection of the carrier's spin. [35]

2.3 Informatics-aided materials science

Materials informatics is an emerging field that aims to accelerate the development cycle through high speed and robust acquisition, management, analysis, and dissemination of diverse materials data. This field of studies includes the research, development, and application of information about materials properties (including both physical data, theoretical and empirical models) and the software tools for querying and mining those databases.

2.3.1 Machine learning for knowledge discovery in materials science

Machine learning schemes have already impacted multiple areas such as cognitive game theory, pattern recognition, event forecasting, and bioinformatics. They are beginning to make major inroads within materials science and hold considerable promise for materials research and discovery [15]. Some examples of successful applications of machine learning within materials research in the recent past include accelerated and accurate predictions of phase diagrams [36], crystal structures [2, 37] and materials properties [38, 39].

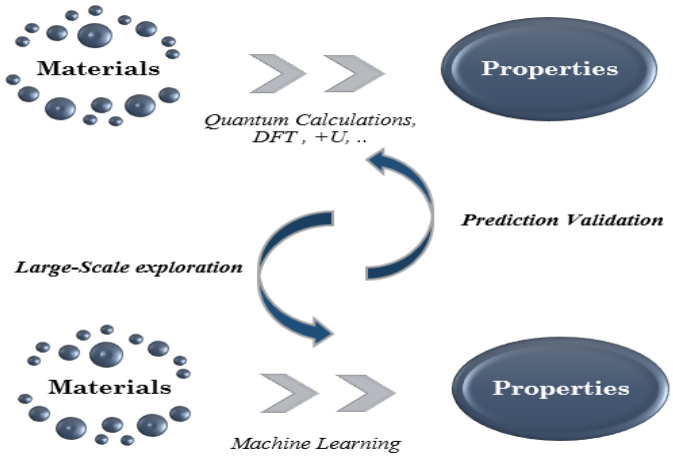

Figure 2.3: Role of machine learning in accelerating quantum mechanical computations.[1]

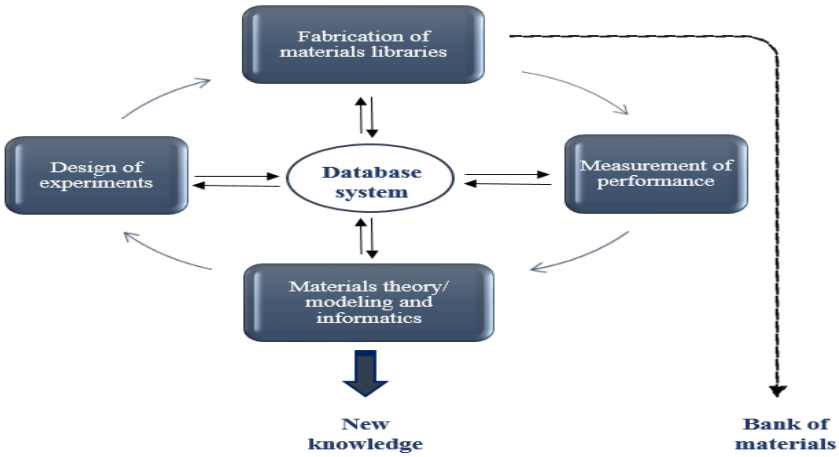

Figure 2.4: Concepts of the combinatorial materials-development.[2]

[1]Image adapted from [15].
[2]Image adapted from [11].

As already highlighted, the process of designing new materials can profit enormously from the available machine learning techniques. A given property could be predicted using past knowledge from other similar known materials. Data scarcity is a common issue when it comes to this kind of application.

Researchers typically resort to performing parallel computations to provide the necessary inputs needed to train a robust model. Even if this might seem contradictory, the amount of computations to be performed when integrated with a machine learning process will be noticeably reduced. The additional computations will focus mainly on balancing the training data set in order to build a model that can be easily generalized afterward. Figure 2.3 highlights this new perspective. Machine learning applications in materials science, help not only avoiding superfluous time-consuming computations, but also in accelerating the pace of new discoveries.

A new scheme of combinatorial materials development cycle has recently emergerged as illustrated in Figure 2.4 [40]. Informatics-aided modeling has become a core feature to extend the existing known materials library. Combinatorial high throughput work-flow combines both human input and automated modeling to respond to the compelling market's needs.

2.3.2 Materials Genome Initiative (MGI)

Several computational materials science projects have been carried in the last decades. Informatics researchers have created frameworks to acquire and store data, fuse complex and disparate data, and add theoretical and computational models. Digital libraries of materials property information and existing computational tools for predicting material properties are important resources in informatics; their development has laid much technical groundwork for informatics approaches. The structured environment developed from measurement or computation is no longer simply a single data point; it is a step in an information-based learning process that uses the collective power to achieve greater efficiency in new materials exploration.

Applying data mining techniques for knowledge discovery in materials science is, however, still at its beginning. Data acquisition and preprocessing were the main concerns till the last few years. This has resulted in the development of good cyberinfrastructures such as the National Science Digital Library Materials Digital Library (NSDL-MatDL) which contains both digital and human resources in a collaborative platform for shared results and data dissemination. [41]

Since the launch of the Materials Genome Initiative (MGI) by the US government, those works have witnessed a relatively huge expansion [13, 14]. The main focus was on establishing

the basic frameworks for this field including collaborative databases and toolboxes. The Lawrence Berkeley National Laboratory (LBNL) has been one of the main lead performers of this initiative. They have introduced the *The Materials Project*, a core program of the Materials Genome Initiative that uses high-throughput computing to uncover the properties of all known inorganic materials [42]. They aim to remove guesswork from materials design in a variety of applications and to accelerate innovation in materials research and in cleantech [13,42]. Their work has been focusing on automating quantum mechanical characterization of known and new materials and their properties. They combined this so-called high-throughput computing with existing data mining approaches for the discovery of new materials for Li-ion batteries, photocatalysts, thermoelectrics, piezoelectrics, and other functional materials.

Several other national laboratories (e.g., the National Renewable Energy Laboratory (NREL), Sandia National Laboratories (Sandia Labs), the SLAC National Laboratory (SLAC), etc.,) have been carrying similar investigations. A collaborative framework is being established and actively engaged in both the development of data mining and high-throughput methodologies and the application of these techniques to a range of physical problems and materials discovery [13, 14].

Most of the previous projects focused mainly on computational thermodynamics and on modeling key materials properties for different applications. Even though several investigations were focusing on solar enrgy applications, they usually tackeled specific compounds [43–45]. Not enough studies were carried for a large scale discovery of new chalchopyrites which explains the few amount of available data for our first application.

2.4 Relevant machine learning algorithms

2.4.1 Feature selection

Feature selection is the process of identifying and eliminating the maximum subset of irrelevant and redundant features. This helps reducing the dimensionality of the data and improving the performance and generalization of regression algorithms.

Several factors are involved in the success of a machine learning task. One of the most important factors and commonly ignored is the representation of the data as well as the quality of its instances [11]. The presence of noise or unreliable information makes learning a very challenging task and can induce a huge drop in efficiency. Real-world data is often represented using too many features. Yet, only a few of them may be related to the target concept. Moreover, there might be

redundancy, where some features are correlated so that is not necessary to include all of them in modeling.

In general, feature selection algorithms have two components: a selection algorithm that generates the candidate subsets of features and attempts to find the optimal one among them; and an evaluation algorithm that determines how good a candidate feature subset is by returning some measure of goodness to the selection algorithm. A stopping criterion should be set to avoid exhaustive search through the space of subsets. This could be: when addition (or deletion) of any feature does not produce a better subset; or when an optimal subset, according to some evaluation function, is obtained. Different strategies have been proposed over the last years for feature selection. These include filter, wrapper, embedded [46], and more recently ensemble techniques [47].

2.4.1.1 Filter-based feature selection techniques

These techniques are independent of the algorithm that will use the selected subset at the end. They assess the discriminative power of features based only on the intrinsic properties of the data. In general, a relevance score is used herein to estimate the optimal subset based on a predefined threshold. Most filter methods consider the problem of FS as a ranking problem. The solution is provided by selecting the top scoring features while the rest are discarded [48].

Let $[X, Y] = \{(x_{ij}, y_i), i = 1..N \text{ and } j = 1..d\}$ be the input data; a matrix X of N rows, corresponding to the number of observations and d columns which represent the used features. $Y = \{y_i, i = 1..N\}$ is the output variable, where $y_i \in \{1..k\}$ is the label associated with i^{th} observation, k being the number of classes. Let $X_j = \{x_{ij}, i = 1..N\} \in \mathbf{R}^d$ be the jth feature of the given data.

In the following subsections, we outline commonly used ranking methods.

Correlation criteria

This method consists of using the Pearson correlation coefficient [49] [50] defined as:

$$R(j) = \frac{cov(X_j, Y)}{\sqrt{var(X_j) * var(Y)}} \tag{2.3}$$

for $j \in \{1..d\}$

The correlation ranking helps detecting the linear dependencies between each feature and the target variable Y.

Mutual information criteria:

The information theoretic ranking criteria uses the measure of dependency between two

variables [3, 49–52]. The mutual information measure (I) of two random variables is a measure of the mutual dependency between the two variables. It is derived from Shannon formulas of the entropy and the conditional entropy as given by the following expression:

$$I(Y, X) = H(Y) - H(Y|X) \tag{2.4}$$

where $H(Y)$ is the entropy of the variable Y given by:

$$H(Y) = -\sum_{y} p(y)log(p(y)) \tag{2.5}$$

$H(Y)$ represents the information content, more specifically the uncertainty, in the output variable Y. In (2.4), $H(Y|X)$ is the conditional entropy which represent the entropy of the variable Y given the known observations of a variable X, i.e.,

$$H(Y|X) = -\sum_{x}\sum_{y} p(x, y)log(p(y|x)) \tag{2.6}$$

Introducing the information about the observed values of X into the entropy of the output Y reduces the amount of uncertainty which gives the mutual information between X and Y and it can be interpreted as follows:

$$\begin{cases} I = 0 \text{ if } X \text{ and } Y \text{ are independent} \\ \text{I} \nleq 0 \text{ if } X \text{ and } Y \text{ are dependent} \end{cases}$$

For continuous variables the same formulas are applied while replacing summations with integrations.

For mutual information based feature selection methods, I is computed between each features and the target variable. After sorting the obtained values, the top d features will be selected, where $d < D$ is a threshold to be selected independently. The fact that the inter-feature mutual information is not taken into account for this method may give poor prediction results [51].

2.4.1.2 Wrapper techniques

These methods rely on the performance of the inductive algorithm as the selection criterion. Wrapper methods wrap the feature selection around the learning algorithm to be used, using cross validation to predict the benefits of adding or removing a feature from the used feature subset.

Wrapper selection algorithms aim, therefore, to evaluate the different possible subsets of features on the learning algorithm and keep those that perform the best. Wrapper techniques are broadly classified into Sequential selection algorithms and Heuristic search algorithms as outlined below.

Sequential Selection Algorithms

This category of algorithms starts with a set and adds (or removes) iteratively new features until the required number of features is obtained or the required performance is reached. Different sub-categories can be defined depending on the way the required subset of features is constructed.

Sequential Forward Selection (SFS) is the simplest method. It starts with an empty set and greedily adds attributes one at a time. At each of the remaining steps, FS adds, permanently, the attribute that yields the learned structure that generalizes best when added [53, 54].

This process can be considered, a naive sequential feature selection algorithm since it doesn't take into account the dependency between the features. However, it can determine small effective subsets quite rapidly since the first evaluations involving relatively few variables are fast.

Sequential Backward Selection (SBS), is similar to SFS. The only difference is that it starts from the complete set of features, on each pass it removes one feature whose removal results in the lowest decrease in the predictor's performance, until the stopping criterion is satisfied. In SBS interdependencies are well handled, but early evaluations are relatively expensive [55].

There is also Sequential Forward Floating Selection (SFFS) and Sequential Backward Floating Selection (SBFS) which are more flexible than naive sequential feature selection previously described since they introduce an additional backtracking step. They are characterized by the changing number of features included or eliminated at different stages of the procedure [56].

The main problem with sequential forward selection approaches is that they can produce nested subsets since the forward inclusion is usually unconditional which means that two highly correlated variables might be included if they give the highest performance in the SFS evaluation [41] .

To avoid this nesting effect, an adaptive version of the SFFS was proposed [57]; The Adaptive Sequential Forward Floating Selection (ASFFS) algorithm uses a parameter to specify the number of features to be added in the inclusion phase. This parameter needs to be calculated adaptively. The ASFFS also uses another parameter to remove the maximum number of features that increase the performance during the exclusion phase.

The ASFFS can therefore obtain a less redundant subset than the other algorithms depending on the objective function and the distribution of the data [41, 57].

Heuristic Search Algorithms

This method consists of evaluating different feature subsets that can be generated either as a solution for an optimization problem or by searching around in a search space using an adequate configuration [46, 53, 54]. Generally, this category of algorithms starts the search from a random

subset. In this case, a solution is typically a fixed length binary string representing a feature subset, the value of each position in the string indicates the presence or absence of a particular feature. It is is an iterative process where each new generation is the result of applying genetic operators like crossover and mutation to the members of the current generation [53, 54].

2.4.1.3 Embedded techniques

Embedded techniques interact with the learning algorithm in a different way. They include variable selection as part of the training process without splitting the data into training and testing sets [49, 56, 57]. Their main advantage is that they are computationally more efficient than wrapper techniques.

2.4.1.4 Ensemble methods

This approach was proposed to cope with the instability of feature selection methods caused by perturbations that can occur in the training set. They consist of fusing the subsets obtained by applying different features selection method to the input data based on a consensus function [47, 49, 50, 58].

2.4.2 Regression analysis

Regression analysis is a statistical process for estimating the relationships between a dependent variable and a set of independent features also called "predictors". A regression model involves three main variables: The unknown parameters, denoted as β , which can be a scalar or a vector, the independent variables that can be represented by a matrix X containing the set of descriptors and finally the dependent variable Y.
The model relates Y to X and β as $Y = f(X, \beta)$. This approximation is usually formalized as $E(Y|X) = f(X, \beta)$, where $E(Y|X)$ is the expected value of Y given X.

Depending on the used regression technique, the unknown parameter, estimated by $\hat{\beta}$ is obtained by either a closed form expression, by solving an estimating equation, or by optimizing an objective function often subject to certain constraints [46] .
Heuristically, β and $\hat{\beta}$ can be thought of as the true coefficients which explain the physical relation between the descriptors and the target output and afterward, a regression model is generated.

A good regression model $f(X, \hat{\beta})$ is not only characterized by its ability to fit the training set, but also by how accurately it can predict a future response given a new unlabeled test data [55].

Cross validation is often used as a reliable mean to judge the robustness of the predicted regression models.

Let $Y = \{y_1, y_2, ..., y_N\}$ be the N dependent variable, $X = \{(x_{ij}), i = 1..N$ and $j = 1..p\}$ be the N p-dimensional feature vectors, and the model's parameter coefficients $\beta = \{\beta_1, \beta_2 ... \beta_p\}$. β can be estimated by different ways depending on the used regression technique. Following are the most commonly used methods.

2.4.2.1 Ordinary Least Square Regression (OLS)

OLS, also known as linear least square, is one of the simplest form of regression. Under appropriate assumptions, OLS regression consists of minimizing the sum of the squares of the differences between the target values and the predicted values by a candidate linear function. The estimates of the parameters of an OLS model are obtained as the solution to the following optimization problem [59, 60].

$$\hat{\beta}^{OLS} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} \beta_j x_{ij}) \tag{2.7}$$

It can be shown that $\hat{\beta}^{OLS}$ is given by the closed form solution [47]:

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T Y \tag{2.8}$$

This can be easily explained by uni-dimensional variables. OLS aims to estimate the parameters of the linear model that best relates the input variables x_i to the corresponding observed responses y_i for $i = 1..N$. This is done by minimizing the resulting least square errors $|\epsilon_i|$ as illustrated in Figure (2.5) [59].

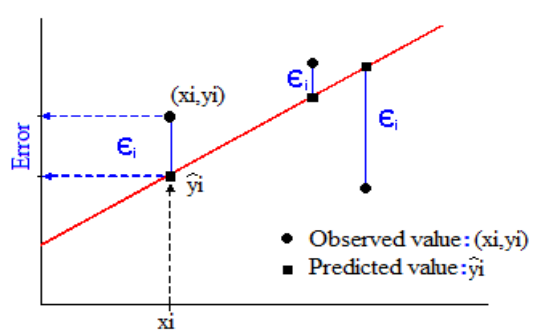

Figure 2.5: OLS regression - Geometrical interpretation for 1-dimensional data.

2.4.2.2 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO [59, 60] performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. LASSO models are carried by imposing an L_1 constraint on the sum of the β coefficients as given by the following formula:

$$\hat{\beta}^{LASSO} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \tag{2.9}$$

$$\text{Subject to } |\beta|_1 < t$$

The solution to (2.9) can be expressed as a penalized least squares optimization [60]:

$$\hat{\beta}^{LASSO} = argmin_{\beta} (Y - X^t\beta)^T (Y - X^t\beta) + \lambda_1 |\beta|_1 \tag{2.10}$$

In (2.10), λ_1 is a tuning parameter which controls the amount of shrinkage.

LASSO performs L_1 shrinkage, so that there are "corners" in the constraint region. For a 2-dimensional case, as illustrated in Figure (2.6) [61], this region can be assimilated to a diamond-like area corresponding to $|\beta_1| + |\beta_2| < t$. If the sum of squares (i.e. the red ellipses) "hits" one of these corners, then the coefficient corresponding to the axis is shrunk to zero [61].

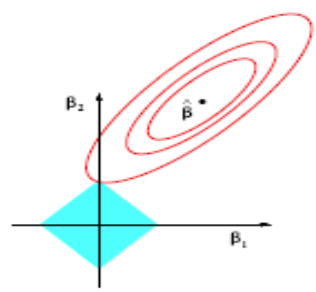

Figure 2.6: Lasso regression - Geometrical interpretation for 2-dimensional data

2.4.2.3 Partial Least Square Regression (PLS)

PLS regression takes into account the latent structure in both, the set of features and the response variable. It uses predictors derived as linear combinations of the original features in order to predict the response. The latent structure is obtained by maximizing the covariance between the

derived descriptors and responses [59, 60]. It is recommended mainly when the number of variables is high, and when it is more likely that the explanatory variables are correlated.

PLS generates a linear model. It is based on the same mathematical foundations as OLS regression (Section (2.4.2.1)). However, PLS considers the input data structure differently. Both the predictors, X matrix, and the target variables, Y matrix, are decomposed into latent structures in an iterative way. The latent structure that corresponds to the most variation of Y is extracted and explained by a latent structure of X that explains it the best.

Figure (2.7) illustrates this process for 3-dimensional input data sets. u_1 represents the direction of most variation for the target, Y. t_1 is the direction that explains u_1 the best within the set of observations X. It is not necessary that t_1 explains the most variation in X as well. The regression problem is reduced therfore to estimate the linear model that best relates u_1 to t_1 using ordinary least squares [62].

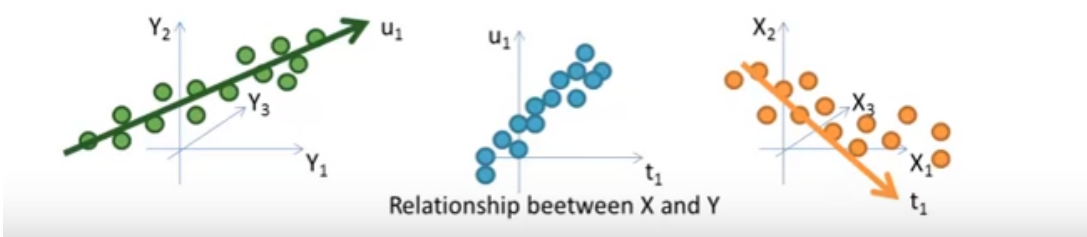

Figure 2.7: PLS regression - Geometrical interpretation for 3-dimensional data.[1]

2.4.2.4 Cross-validation

Cross validation is usually used as a convenient technique to assess the performence of a learning algorithm including regression methods. Cross validation is very important to assess the generalization ability of the statistical analysis results to independent data sets [59].
Different cross-validation approaches can be used according to the available data.

The holdout method is the most common cross validation. The data set is split into two sets, a training and a testing sets. The predictor model is built based on the training set only. Then it is tested on the new observations of the testing set. The generated errors are accumulated to give the mean absolute test set error, which is used to evaluate the model's performance. This method could be advantageous since it is fast to perform. However, its evaluation can have a high variance, especially when the used sets are not well balanced [58].

[1] Image adapted from [62]

K-fold cross validation is one possible way to improve over the holdout method [58]. The data set is split into k subsets, and the holdout method is repeated k times. In each iteration of the learning process, one of the k subsets is used as the test set and the other k-1 subsets are combined to form the training set. Afterward, the average error across all k trials is calculated. This method gives a model that is more adapted to the inherent data distribution since the learning takes into account all instances within the data which reduces the variance of the resulting estimate. The variance of the resulting estimate is reduced as k is increased. Nevertheless, this method is still more computationally expensive compared to the previous method. The training algorithm has to be rerun from the beginning k times, which means it takes k times as much computation to make a single evaluation [58].

Leave One Out (LOO) cross validation is a k-fold cross validation taken to its logical extreme, with k equal to N, the number of observations [58]. In each observation, an observation is removed from the trained set, the model is fitted for the incomplete set, and then an error estimate is computed for the removed instance. The same process is repeated until an error estimate has been obtained for all the instances within the data. Mean Squared Cross Validated Error (MSCVE) is then obtained as the mean of the squared errors [58]. LOO cross-validation is typically used when N is small to maximize the number of training samples. This is almost the case for our applications [58].

For our analysis, (MSCVE) will be used for most cases as the performance measure for comparison between different models.

CHAPTER 3

DATA DRIVEN DISCOVERY OF MATERIALS PROPERTIES

In this chapter, we describe our novel approach to apply machine learning techniques for materials science applications.

The objectives of our work can be summarized by two main tasks:

- Determine the optimal set of features that best describe the given predicted variable.
- Boost prediction accuracy via applying various regression algorithms.

Our approach will be applied to two applications. The first one consists of Band gap prediction for chalcopirites, while the second one aims to predict the magnetic moment of dilute semiconductor materials.

3.1 Predicting Band Gaps of Chalcopyrites

Semiconducting chalcopyrites (chemical formula ABC_2) have a special interest for material scientists due to their several technological applications as well as their non-linear optical properties [63]. Yet, their most promising application is, probably, their use for solar cells industry. These chalcopyrites exhibit band-gaps that can be tuned to absorb various energy bands in multi-junction cells, which optimizes the usage of the solar spectrum.

For our application, we focus on chalcopyrites made of the combination of elements from $I - III - VI$ and $II - IV - V$ groups of the periodic table as illustrated in Figure 3.1. The green shaded elements were included in the training data, the yellow boxes are elements with known compounds' band gaps which were used only for testing to validate the created models, while the red boxes are elements with unknown compounds' band gaps.

As mentioned in Section 2.1, different studies have started investigating the possible models that can describe the relationship between the band gap and the chemical stoichiometrics and fundamental properties of the constituents of these chalcopyrites. The pilot study carried by Zeng et al in 2001 [1], has laid the foundation to our work. The authors used artificial neural network to estimate the correlation between band gap energy (and lattice constant) of chalcopyrites and their

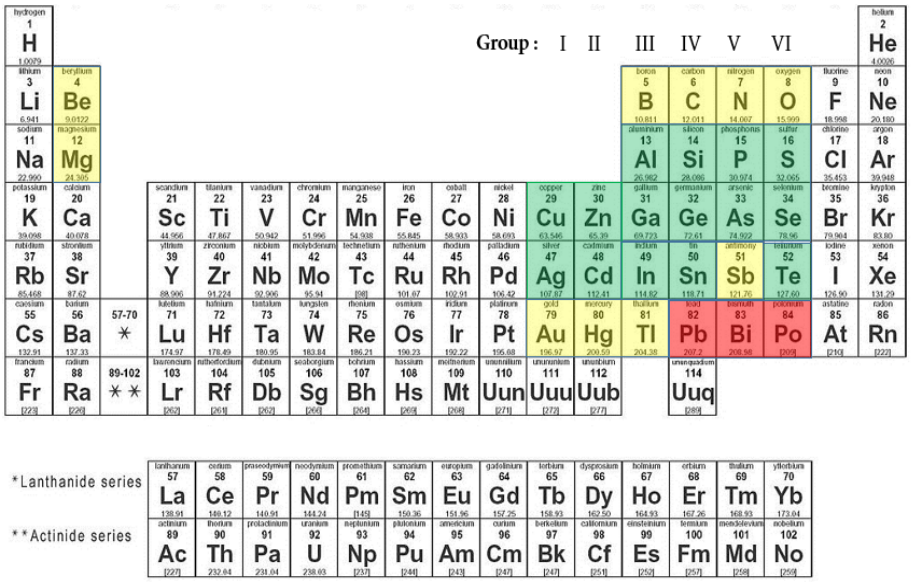

Figure 3.1: Periodic table of the elements used in the QSAR modeling as compound chemistries

respective chemical stoichiometries and elementary properties. They proved that the dependency can, actually, be modeled linearly which oriented future research towards the use of linear regression techniques. Using the same descriptors as in Zeng et al. study, Suh and Rajan (2004) [2] exploited PLS regression to estimate the underlying linear model. In 2014, [3] went further, and used more regression techniques (OLS and LASSO for instance) in order to build a more robust model.

TABLE 3.1

Description of the features used for Band gap prediction

Variable Name	**Description**
Atomic Number (AN)	The number of protons in the nucleus of an atom, which determines the chemical properties of an element and its place in the periodic table
Electronegativity (EN)	Measure of the tendency of an atom to attract a bonding pair of electrons.
Melting Point (MP)	The temperature at which a given solid will melt.
Valency (VL)	Measure of the element's combining power with other atoms
Pseudo Radii (PR)	A measure of the crystal lattice

The choice of features, however, remained the same throughout all these different studies. The included chemical properties were basically; the Electronegativity (EN) ($eV^{1/2}$), the Atomic Number(AN), the Melting Point (MP) (K), Zunger pseudopotential radii sum (PR) (atomic units,

au) and the number of Valence electrons (VL) as explained by Table 3.1.

The bandgap (BG-E) of the compound ABC_2 was predicted as a function of $MP(X)$, $AN(X)$, $EN(X)$, $VL(X)$ and $PR(X)$, where X refers to any of the three atoms within the compound formula: A, B and C.

Table 3.2 presents the list of variables, and their values for the elements that form the studied chalcopyrites as reported in [1–3]. In this table, Grp refers to the group number within the periodic

TABLE 3.2

Elements and their features values forming the chalcopyrites of our training set.

$I-III-VI_2$ Compounds							$I-III-VI_2$ Compounds						
Grp	Elm	EN	AN	MP	PR	VL	Grp	Elm	EN	AN	MP	PR	VL
I	Cu	1.08	29	1358.0	2.04	11	II	Zn	1.44	30	692.7	1.88	12
	Ag	1.07	47	1235.0	2.375	11		Cd	1.40	48	594.3	2.215	12
III	Al	1.64	13	933.5	1.675	3	IV	Si	1.98	14	1687.0	1.42	4
	Ga	1.70	31	302.9	1.695	3		Ge	1.99	32	1211.0	1.56	4
	In	1.63	49	429.8	2.05	3		Sn	1.88	50	505.1	1.88	4
VI	S	2.65	16	388.4	1.1	6	V	P	2.32	15	317.3	1.24	5
	Se	2.54	34	494.0	1.285	6		As	2.27	33	1089.0	1.415	5
	Te	2.38	52	722.7	1.67	6							

table and Elm refers to Element.

All the possible combinations of elements in Table 3.2, except those where both Zn and Sn are present, make the 28 chalcopyrite compounds with known band gap energies (BG-E) that were used for the training set. The BG-E of these 28 compounds is given in Table 3.3.

TABLE 3.3

Experimental Band gap Energy (BG-E) of the training set's chalcopyrites (eV)

$I-III-VI_2$ Compounds									
$AgAlS_2$		$AgAlSe_2$	$AgAlTe_2$	$AgGaS_2$	$AgGaSe_2$	$AgGaTe_2$	$AgInS_2$	$AgInSe_2$	$AgInTe_2$
3.13		2.55	2.27	2.64	1.8	1.32	1.87	1.24	0.95
$CuAlS_2$		$CuAlSe_2$	$CuAlTe_2$	$CuGaS_2$	$CuGaSe_2$	$CuGaTe_2$	$CuInS_2$	$CuInSe_2$	$CuInTe_2$
3.49		2.67	2.06	2.43	1.68	1.12	1.53	1.04	1.06
$I-III-VI_2$ Compounds									
$ZnSiP_2$	$ZnSiAs_2$	$ZnGeP_2$	$ZnGeAs_2$	$CdSiP_2$	$CdSiAs_2$	$CdGeP_2$	$CdGeAs_2$	$CdSnP_2$	$CdSnAs_2$
2.07	1.74	2.05	1.15	2.33	1.55	1.72	0.57	1.17	0.26

Our original goal of investigating this data set, is to assess the relevance of the different features. In fact, band gap prediction is still a very challenging task, and it can be related to several aspects and properties of the compound in question. The goal is, as a first step, to determine if all

of the previously chosen descriptors are relevant and have meaningful contribution to the prediction process. To this end, feature selection and ranking algorithms are applied.

In a second step, we investigated the possibility of adding new features to improve the system's prediction accuracy. The main focus was on binary descriptors that reflect the interactions between each pair of the elements present in the studied compounds. Bond dissociation energy and bond length measure were selected as prominent candidates based on their physical signification. The best subset of features, along with the best regression models are then used to predict the band gaps of over 150 compounds.

3.2 Modeling Magnetism of DMS materials

Even though it has been subject to various studies lately, modeling the magnetism of DMSs still needs more focus. Conflicting results are still being reported which urges an in-depth understanding that goes into the microscopic origins. Advanced DFT calculations implementing local force theorem for magnetic exchange were a key step to prove that there is a correlation between magnetism and defect concentration, which is also refered to as Defect-Induced Magnetism (DIM) [64,65]. DFT has led therefore to providing a better modeling to the exchange interactions in DMS. Theoretical investigations have justified that the presence of both intrinsic and extrinsic defects (e.g., holes and impurities like magnetic or non-magnetic atoms), has a dual role in DMSs. They not only generate unpaired electrons, hence, the necessary magnetic moment but also, might contribute in establishing the ferromagnetic coupling (FMC) among the magnetic moments [66]. The question that arises at this point is how exactly defects affect FMC properties in DMSs and how could this be modeled?

The recent pilot investigation by Andriotis and Menon (2016) [67] has succeeded to correlate the defect-induced magnetism (DIM) of DMSs with some key factors. Knowing that DIM in DMSs and related materials is tightly coupled to co-doping and the synergistic action between the co-dopants, they demonstrated that defect synergy is the result of the exchange among correlated spin-polarization processes that takes place at the co-dopant's neighborhood within the host. These processes were shown to have a direct effect in enhancing the FMC among the magnetic co-dopants. The proposed FMC was demonstrated using *ab initio* calculations of the electronic properties of co-doped ZnO, GaN and TiO_2 hosts. They included features related mainly to the atomic properties of all the elements of the DMSs in question in addition to features related to the DMS-host lattice as detailed in Table 3.6.

Our application herein aims to provide a computational support to this proposed theory which could later on constitute an unyielding foundation for a unified theory of magnetism in DMS materials. We will exploit the observed correlation between DMS magnetism and the atomic features of their constituent atoms further in order to build a more conclusive model. To this end we will use a predictive machine learning approach based on the idea of virtual combinatorial screening (Section 2.4). This will lay a map for material scientists to guide their search for the appropriate host and dopant materials having specific properties in order to check their magnetism.

TABLE 3.4

Used dopant atoms according to their nature

Nature of dopant	**Example of atoms**	**Substitutional impurities type**
Magnetic atoms	From $3d$ and $4d$ transition metal series (e.g., V and Pd)	Cationic
Non-magnetic atoms	Cu, Ag, Zn, Li, Al, etc.	Cationic
Non metals	N, S, C, Si, O, etc.	Anionic

TABLE 3.5

Used DMS systems

Host	**Co-dopants (A_2B formula)**
GaP	Co_2Cu, Mn_2Cu, Ni_2V, Co_2Cu, Co_2Z, Cr_2Mn, Fe_2Cr, V_2Cu, Fe_2Cu, $F2_2V$, Mn_2Ni, Mn_2Cr, Co_2Mn, Ti_2Mn, Ti_2Co, V_2Ti
GaN	Co_2Cu, Mn_2Cu, Mn_2Co, Co_2Ni, Co_2F2, Cr_2Co, V_2Co, Ti_2Co, Co_2Cr, Co_2V, Co_2Ti, Mn_2Ti, Mn_2V, Mn_2Cr, Mn_2Fe, Mn_2Ni, Mn_2Zn, Mn_2Mo, Mn_2Ag, V_2Pd

The data we are using is based on reported *ab initio* calculations and existing experimental results. We will use different hosts with dopant pairs of different natures as detailed in Table 3.4. This will help us gain better insight into the impurities roles for the magnetism in the studied systems.

The reported results herein were carried for two hosts: GaP and GaN. These hosts were co-doped by a dopant pair (A_2B formula, A and B are atoms) as detailed in Table 3.5.

For each host, we used descriptors related to the atomic properties of all atoms present in the system,

i.e., properties of the two atoms present in the co-dopant (A and B), as well as properties of the atoms forming the host, in our case, Ga, P and N. The features we used are detailed in Table 3.6.

TABLE 3.6

Description of the features used in the DIM application

Variable Name	**Description**
Atomic Number (AN)	The number of protons in the nucleus of an atom, which determines the chemical properties of an element and its place in the periodic table
Electronegativity (EN)	Measure of the tendency of an atom to attract a bonding pair of electrons.
E(s)	Energy of of the atomic s orbital (eV).
Valency	Measure of the element's combining power with other atoms
E(d)	Energy of of the atomic d orbital (eV).
Covalent Radius (CR)	Measure of the size of an atom that forms part of one covalent bond ($\AA$)
d-band center(d_c)	Energy at the center of the electronic d-band (eV)
Magnetic Moment (Mo)	Measures an object's tendency to align with a magnetic field (SI). For an atom, it is the vector sum of its orbital and spin magnetic moments.

3.3 Computational approach

We propose a standard statistical learning approach with a physicist and a computer scientist in the loop. The prime goal is to accelerate the discovery of new materials for the two studied applications. To this end, four major tasks are performed as detailed in Figure 3.2.

The developed algorithms are based on *abinitio* calculations and experiments to analyze various descriptors of electronic and crystal structure parameters of the considered materials.

The generated input data is pre-processed in order to remove outliers and normalize all the features to be within the same dynamic range of values. Regression analysis is performed afterward in order to estimate the underlying models. The obtained models are then evaluated in terms of prediction errors for compounds with different confidence values in order to assess the accuracy.

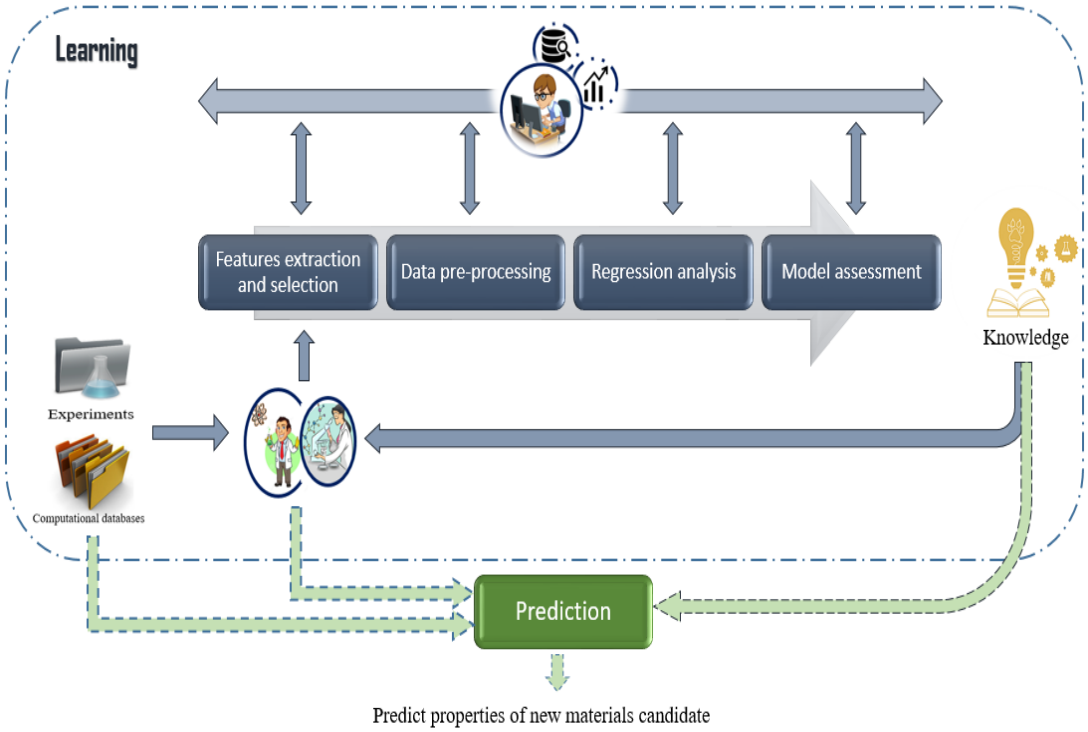

Figure 3.2: Proposed Learning approach with a physicist and a data scientist in the loop.

3.3.1 Data acquisition

The first step in our approach is to build our data sets. This step relied strongly on the expertise of the material scientists we are collaborating with. The procedure was almost the same for both applications. Our training data, on which the relevant descriptors are learned and the regression models are built, consists of both theoretical and experimental data. Descriptors are extracted based on fundamental atomic and crystallographic properties of the studied materials and according to their physical significance to the target variable.

For band-gap prediction, we relied mainly on the reported data from previous studies. We constructed similar data sets in order to replicate the previous results and explore the possibilities of enhancement. When faced with the need to add more descriptors (e.g., binary descriptors like Bond length and Bond Dissociation energy) specific computations and software simulations were carried for the compounds already present in our sets.

For magnetism modeling in DMS materials, we started by choosing the parent materials which could act as a host for developing possible DMSs upon co-doping according to previous studies, GaN and GaP for our case. For each of these hosts we produced a set of hypothetical co-doped systems for different dopant pairs as detailed in Section (3.3). Afterward, we collected the

reported data claiming DMS functionality upon doping which lays the foundation of more advanced processing as described in the following paragraphs.

3.3.2 Data pre-processing

The representation of the input data as well as the quality of its instances are of a great importance for the success of any machine learning process. Before tackling any advanced learning step, the data needs to be cleaned and prepared in order to remove noisy and unreliable information. Afterward, actual pre-processing steps including outliers detection, normalization and missing values handling are performed. This is a crucial step that helps avoiding performance deterioration due to error propagation and model mis-specification.

3.3.3 Features analysis

For both applications, feature analysis represent the core task of our learning approach. In fact, several factors can influence materials properties, starting from the preparation condition (i.e., temperature, humidity, pressure, etc.) to the more complex microscopic and macroscopic interactions and transformations.

We focused on building a set of descriptors that best describe our target variables for a wide variety of materials. This relied on both expertise in materials science and computational analysis. We relied on expertise, to identify all features that are likely to affect the predicted outputs according to existing studies. A large set of descriptors is therefore extracted for both applications. Since we have no accurate *apriori* knowledge about their exact relevancy we went further and applied computational feature selection techniques as detailed in Section (2.5.1).

We exploited mainly the correlation criteria, PCA weights and sequential feature selection in order to rank the extracted descriptors according to their corresponding weights for every technique. According to this ranking, the set of features that gives the least prediction error on the training set as well as the set of the top performing features for all the techniques combined, are kept to train the final prediction models using regression methods.

3.3.4 Regression analysis

After building our training data sets and identifying the optimal set of features for our two applications, we tackle the actual learning process. Three different regression techniques are considered for this work: Ordinary Least Squares (OLS), Partial Least Squares (PLS) and LASSO

as explained in Section (2.5.2). The goal was to build a robust prediction model that takes into account the nature of the used data as well as the specifications of the application. Using various regression methods is a key choice since every technique has its own anatomy and its own way of handling the input in order to recognize the hidden patterns and recognize knowledge.

3.3.5 Model assessment

In order to assess our models and evaluate our predicted results we resorted to cross-validation. We tried both the holdout and Leave One Out methods. The cross-validated errors were computed in order to compare the obtained models and investigate their prediction accuracy. We carried tests for data of three confidence levels (High, Medium or Low) according to the level of presence of the tested compound on the training set. Our results were compared as well to the previously reported works.

CHAPTER 4

EXPERIMENTAL RESULTS

The experiments were ran on a computer equipped with a 3.6 GHz Intel Xeon processor and a 24 GB RAM.

4.1 Band-gap prediction for chalcopyrites

4.1.1 Approach

The objective of our experiments for Band Gap Energy prediction is two fold. First, we will investigate the relevance of the features originally used in Dey et al. (2014) experiments [3]. As detailed in Section (3.2), the original data consists of five elementary descriptors for each atom present in the ternary compounds in question, (ABC_2) which are, the Atomic Number (AN), the Electronegativity (EN), the Valency (VL), the Melting Point (MP) and the Pseudopotential Radii (PR), thus, a total of fifteen features. We will apply several feature selection technique and compare our results to what they have obtained for OLS, PLS and Lasso regressions.

Second, we will investigate the possibility of adding more binary features describing the interactions between the elements of the given chalcopyrites. Based on our experts view, the binary Bond Dissociation energy as well as the Bond Length measures represent good candidates for our application. We added five new descriptors (BD_{AC}, BD_{BC}, BD_{CC}, BL_{BC}, BL_{CC}), each describing a given physical bond. The total number of features for this set of experiments is, therefore, 20.

Table 4.1 shows the distribution of the data that we used for both experiments. Prior to performing any task, our data was normalized using "max-min" technique in order to fit the features within the same dynamic range.

TABLE 4.1

Number of instances within the different data sets used for Band Gap Energy prediction

Experiment	**Impact of Feature Selection**	**Impact of adding new features**
Available Data	Original Features	Original Features + Bond Dissociation (BD) + Bond Length (BL)
Training	28	14
Test - Labeled (Reported)	13	7
Test - Labeled (New Experiments)	4	4
Test - Unlabeled	159	21
Total	**204**	**46**

4.1.2 Effect of feature selection on the original data set

The first goal of our investigation for this application was to assess the relevance of the already chosen features. To this end we applied both wrapper and filter selection techniques to the original input data for the 15 elementary descriptors.

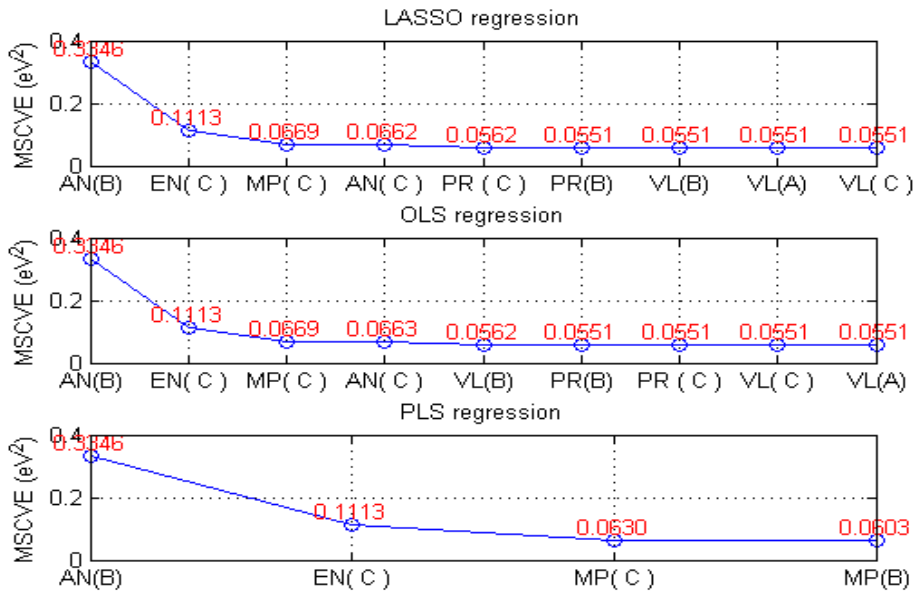

Figure 4.1: Evolution of the sequential forward feature selection error for the three regression methods. The x-axis represents the selected features ahile the y-axis represents the errors.

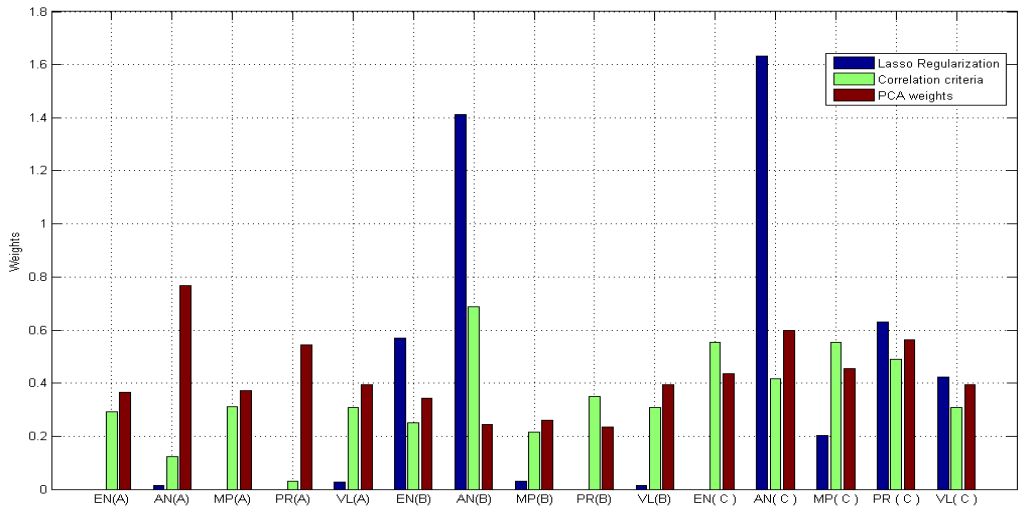

Figure 4.2: Ranking of the importance of the different features on for band-gap prediction applications using different criteria.

TABLE 4.2

Selected sets of features for Band gap prediction

Features	**Method** ($MSCVE$)				
	Correlation (0.0551)	Regularization (0.0646)	SFS		
			OLS(0.0551)	PLS(0.0603)	Lasso(0.0551)
EN(A)					
AN(A)		✓			
MP(A)					
PR(A)					
VL(A)		✓			
EN(B)		✓			
AN(B)	✓	✓	✓	✓	✓
MP(B)		✓		✓	
PR(B)	✓		✓		✓
VL(B)		✓	✓		✓
EN(C)	✓		✓	✓	✓
AN(C)	✓	✓	✓		✓
MP(C)	✓	✓	✓	✓	✓
PR(C)	✓	✓	✓		✓
VL(C)		✓			

We used sequential forward feature selection technique for the three regression methods. As detailed in Section 2.5, SFS algorithm starts with an empty set of features and iteratively adds the feature that minimizes the prediction error for the given learning method. Figure 4.1 illustrates the evolution the SFS output for LASSO, OLS and PLS regressions. Mean Square Cross-validated error was used as the objective function. SFS output shows that the original set of fifteen features can be reduced to at least five features including AN(B), EN(C), MP(C), AN(C) and EN(C) with

a $MSCVE \approx 0.05eV^2$.

TABLE 4.3

Testing compounds for band gap energy prediction

Compounds	**Band Gap (eV)**		**Confidence**
	Old	**New**	
$HgGeAs_2$	0.2	-	Med
$GaAnP_2$	2	-	Med
$AgAlO_2$	3.6	-	Med
$AgGaO_2$	4.1	-	Med
$CuAlO_2$	3.5	-	Med
$CuBO_2$	2.2	-	Low
$CuBS_2$	3.61	-	Med
$CuBSe_2$	3.13	-	Med
$CuGaO_@$	3.37	-	Med
$CuInO_2$	3.9	-	Med
$MgGeAs_2$	1.6	-	Med
$MgSiAs_2$	2	-	Med
$MgSiP_2$	2.3	-	Med
$ZnGeN_2 2$	2.67	-	Med
$AgBO_2$	-	2.21	Low
$AuBS_2$	-	2.55	Low
$AuBSe_2$	-	1.53	Low
$AuBTe_2$	-	1.33	Low

In addition to SFS, we applied filter selection techniques to rank the features according to their weights. We based our ranking on Pearson correlation between the fifteen features of the training set and the target variable. We also considered the features weights obtained from Lasso regularization (weights of the Lasso coefficient) as well as the Principal Components weights (Norms of the features for the first three principal components (87%)). Figure 4.2 illustrates the features' weights according to each of the three criteria.

Overall, both Lasso regularization and the correlation criteria agree that features related to the third element, i.e., C atom, as well as AN(B) and EN(B) have higher weights compared to the rest. These two criteria take into account the target variable which makes them more reliable for assessing the features' relevance. Whereas, PCA weights reflect the norms of the features within the data set, separately from the learning algorithm. They serve as an indicator of the normalization effect.

Table 4.2 shows the different sets of features that were selected using the different techniques as well as their corresponding MSCVE for the training data. The top performing set of features for

every method as well as the set obtained by majority vote, labeled "Overlap" in our experiments, were used to train separate regression models.

In order to assess the performances of the trained regression models, we used a labeled testing set containing a total of 18 compounds. None of these compounds is present in the training set. They were chosen with different confidence levels. Table 4.3 shows the used testing set. The column labeled "New" contains newly computed bandgaps which were not used in the 2014 paper. These can be used as an additional validation criterion that serves to quantify the uncertainty in the predictions. The column labeled "Confidence" takes values High, Medium and Low. A 'High' value indicates that the compound contains only elements that are used in different compounds in the training data, a 'Med' indicates that the compound includes one element not used in any training compound, and 'Low' indicates all others.

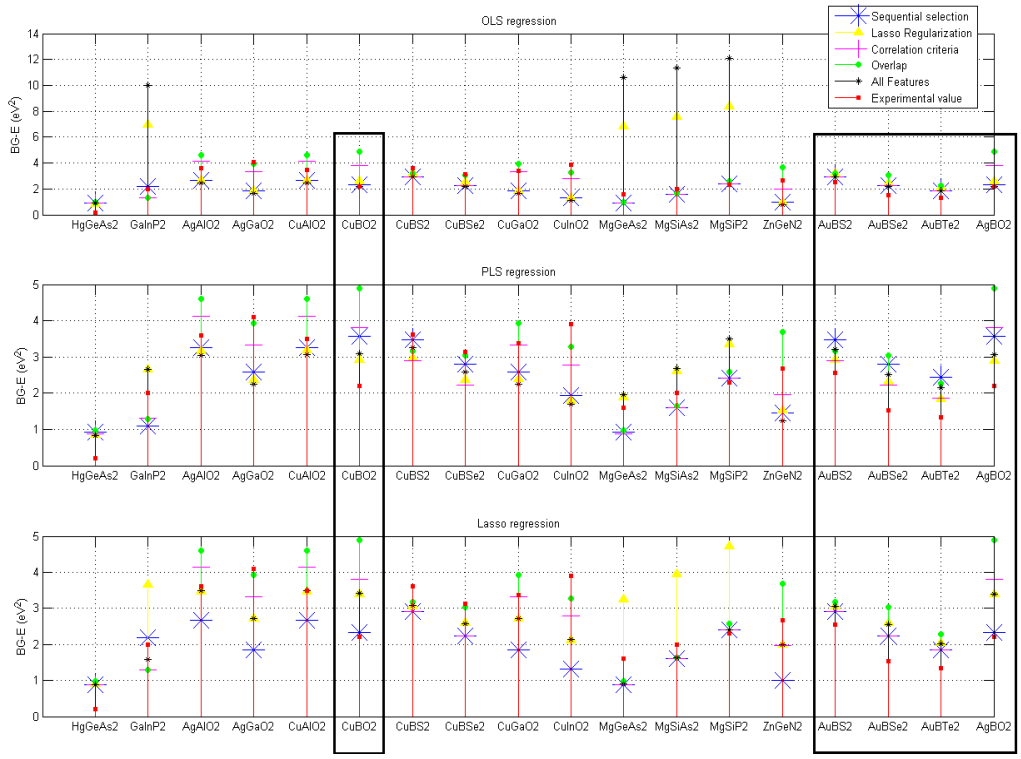

Figure 4.3: Plot of the Predicted band gaps vs. Experimental band-gap using various subsets of selcted features and 3 regression methods.

Figure 4.3 shows the plot of the predicted band gaps by each of the trained models using OLS, PLS and Lasso regression. The red square corresponds to the experimental value (i.e., ground truth) to which the predicted values should be compared. The graph shows that, for the different

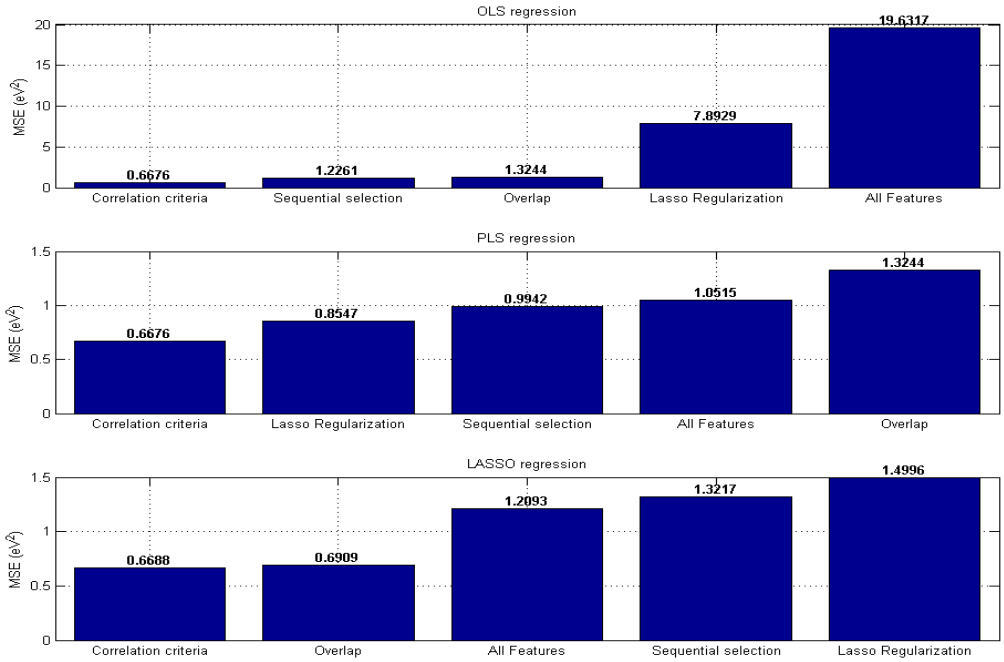

Figure 4.4: MSE of different models for the testing Set.

regression techniques, most of the selected features subsets are giving better predictions that that obtained using all the original features (black marker). For instance, the models trained using the top performing features selected based on both the Pearson Correlation Criterion (magenta marker) and on Sequential Search (blue marker) gave the closest estimation to the reference values for almost all three regressions.

Furthermore, all regression techniques agree that feature selection can significantly improve the prediction of compounds with low confidence (compounds within the black boxes in the plot:$CuBO_2$,$AgBO_2$,$AuBS_2$,$AuBSe_2$ and $AuBTe_2$) which we are most interested in achieving.

Figure 4.4 showing the Mean Squared Error (MSE) of all the tested models for the three regression methods validates our observations. The computed MSE labeled "All Features" shows the same MSE as reported by Dey et al., 2014 [3]. Our experiments prove that feature selection can enhance the prediction accuracy with more than 40% for the different regression techniques. Features selected based on correlation and sequential search are the best performing so far.

4.1.3 Boosting performances by adding new features

The second fold of our experiments consists of assessing the impact of adding new features to the original data set that take into account the interactions between pairs of elements. As aforementioned, we focused mainly on two new features, Bond dissociation energy (BD) and Bond length measure (BL). New DFT +U calcualtions were carried to provide the BDs of the available compounds, whereas, BL measures were obtained using *CrystalMaker* software [68]. This software estimates the length of the bonds based on the relaxation of the crystallographic structure of the compounds.

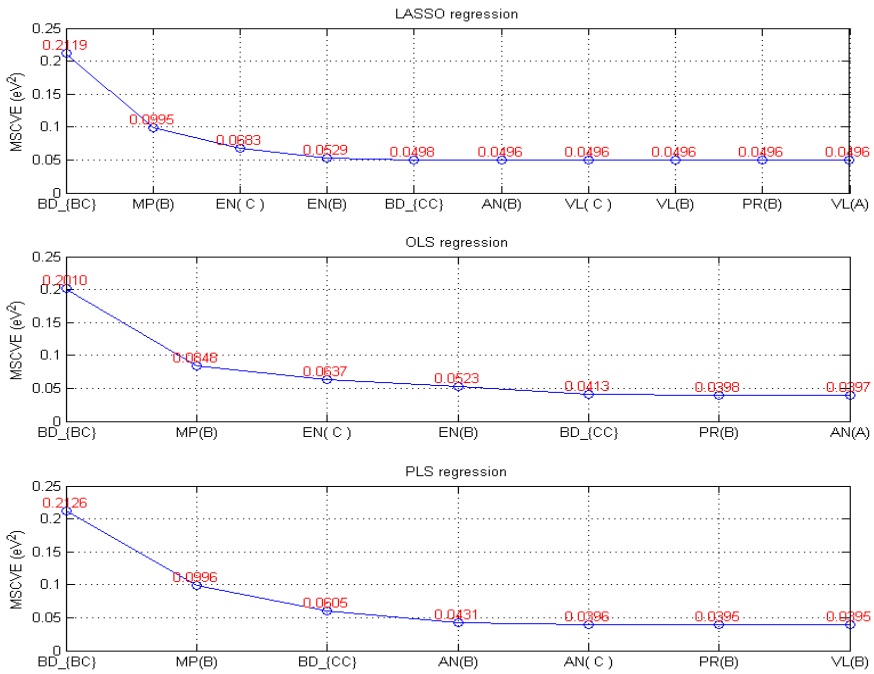

Figure 4.5: Error Evolution for sequential feature selection after adding binary descriptors.

We carried several tests to seize the impact of the new features. First we applied feature selection and ranking technique to the whole set of the 20 descriptors. This step aims to compare the behavior of the new features to that of the original features.

As illustrated in Figure 4.5, BD_{BC} and BD_{CC} were selected in addition to the originally selected set of features using sequential feature selection with an even smaller error ($MSCVE = 0.0395eV^2$) for PLS regression). Moreover, both BD and BL have a strong correlation with Band gap energy as shown in Figure 4.6. This preliminary test proves again that descriptors related to the first element of the compounds have small contribution to the prediction process.

A more detailed observation of the error evolution for PLS regression is provided by Figure 4.7.

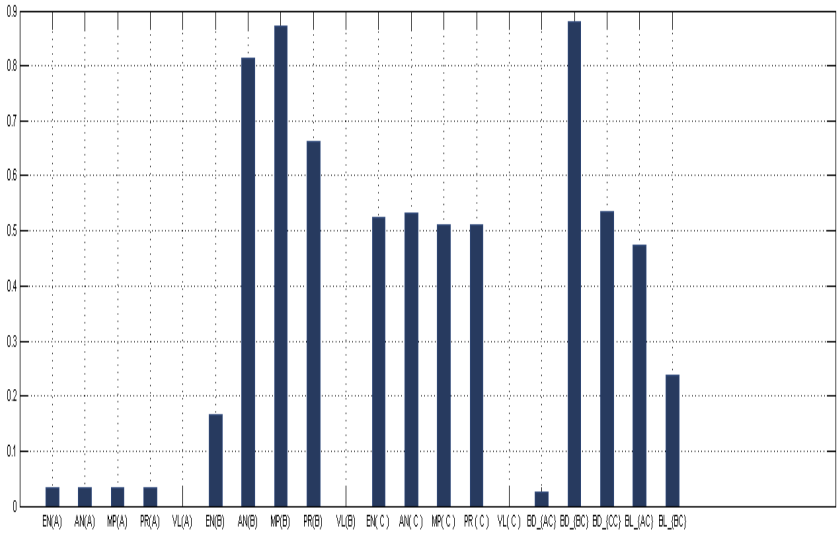

Figure 4.6: Correlation of the new set of 20 features to the target variable.

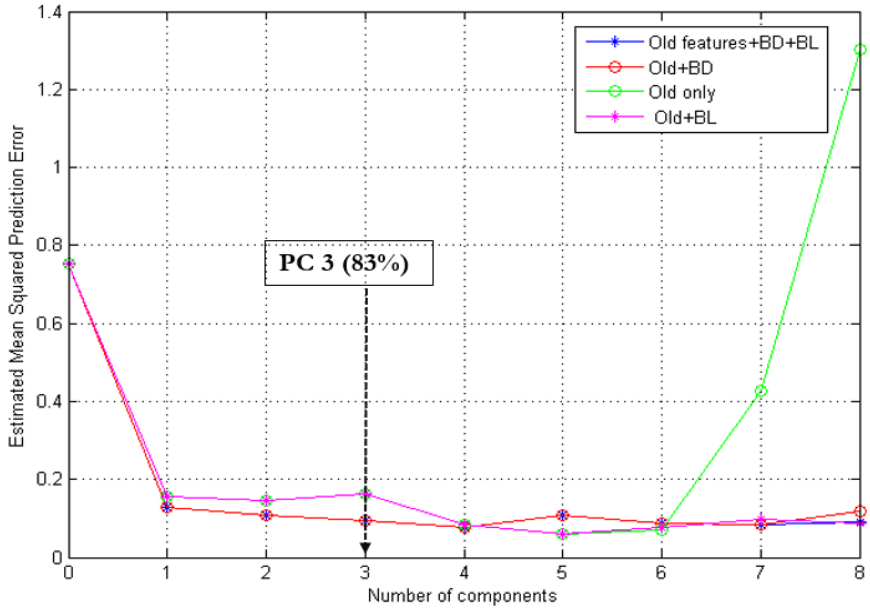

Figure 4.7: Error Evolution vs. number of principal components for PLS regression for different sets of feature.

We considered four features' sets: Only old features, Old features and BD, Old features and BL, and finally Old features with BD and BL. PLS derives latent structure in a manner that maximizes the covariance between the derived descriptors and responses while reducing the dimensionality. Using both data sets containing BD provides a smaller prediction error for the first three principal components reflecting 83% of the information within the training set. These tests can induce that

BD descriptors would have a better contribution to our prediction process compared to BL's.

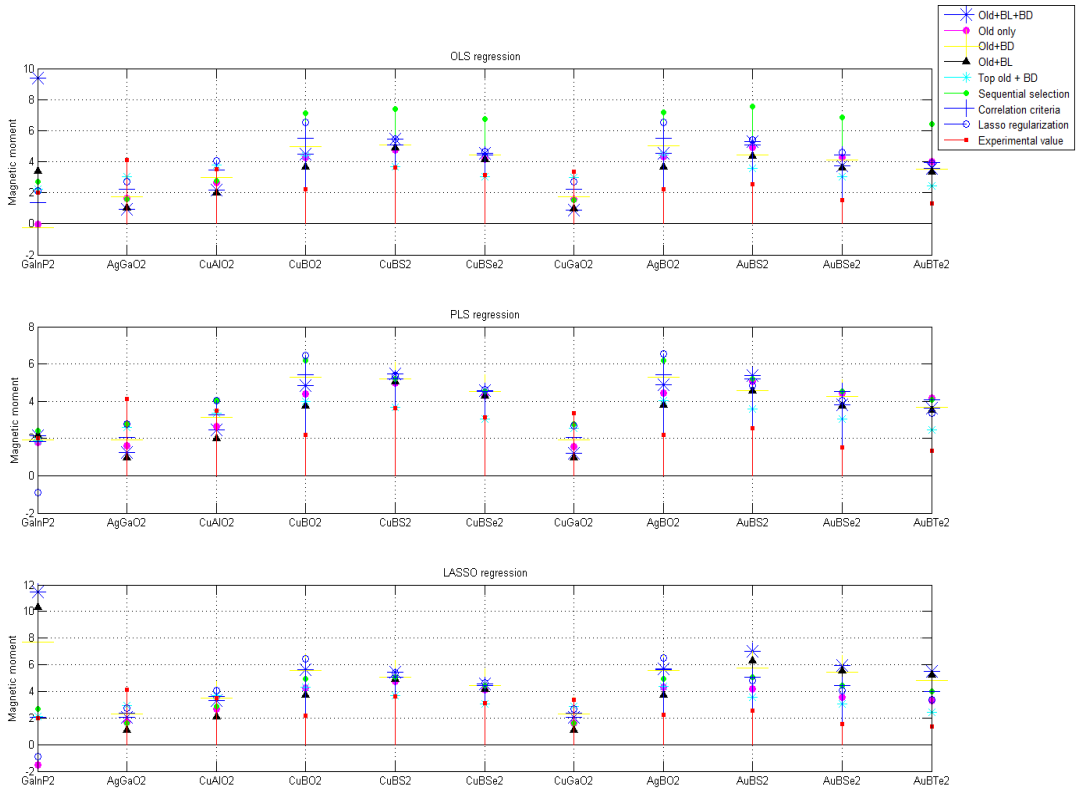

Figure 4.8: Plot of the Predicted band gaps vs. the Experimental band-gap after including binary features for different regression models

Based on these observations, we trained different regression models using Lasso, OLS and PLS regressions. We used the same features' sets as in Figure 4.7 and added the subsets selected by sequential selection, Lasso regularization and correlation criteria. We also considered the subset of features combining the top performing features from our first set of experiments (Section 4.1.2) with the bond dissociation energy which seems to correlate better with our output.

We tested the obtained models on the 11 labeled compounds which were not included in the training data (c.f. Table 4.3). Figure 4.8 gives the predictions obtained for these 11 compounds. The results are close in most cases for the different regression techniques. The cumulative MSE as illustrated by Figure 4.9 and Figure 4.10 shows that the top performing models were obtained for the set that combines the best features as selected from the original data with the bond dissociation energy. This set enhanced the prediction accuracy with more than 70% for the Lasso, OLS and PLS regressions.

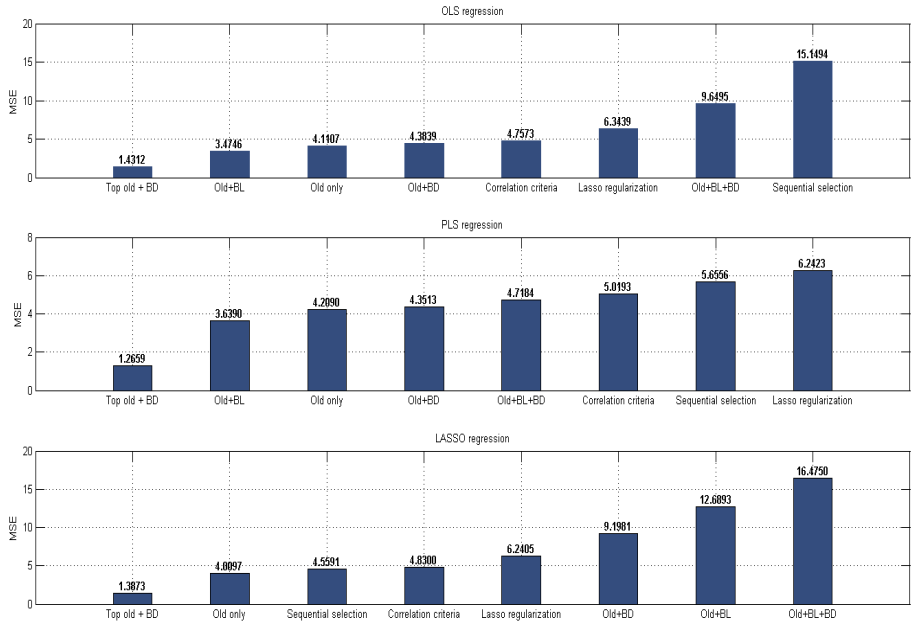

Figure 4.9: MSE of different models for the testing Set after including the binary descriptors.

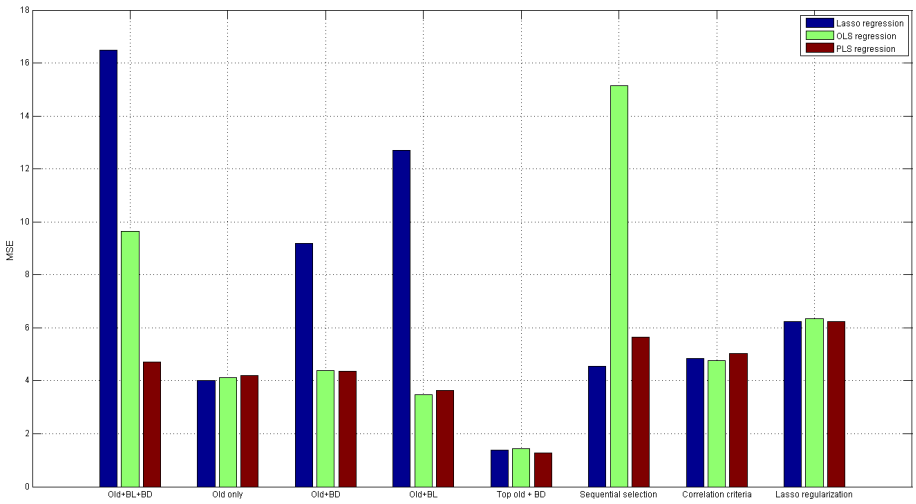

Figure 4.10: Summary of the trained models MSEs for the testing Set after including the binary descriptors.

4.1.4 Discussion

Our experiments yielded a number of useful findings. First, we showed that the original set of fifteen features can be reduced to at least six features including AN(B), PR(C), EN(C), MP(C), AN(C) and EN(C). Furthermore, features related to the last two elements of the chalcopyrite seem to be more relevant for band gap energy prediction. Reducing the original set of features improved the predicton error with about 40% compared to the previously reported results.

Second, we investigated the impact of adding new binary features. Our preliminary experiments demonstrated that the bond dissociation energy describing B-C and C-C bonds can be a good candidate to boost our application performance. Combining BD to the optimal set from the first set of experiments yielded 70% reduction of the prediction error.

Regarding the used regression techniques, overall, PLS and Lasso gave better performances compared to OLS regression. These two methods provide readily interpretable models in terms of the original input set. They take into account the inherent feature's dynamics.

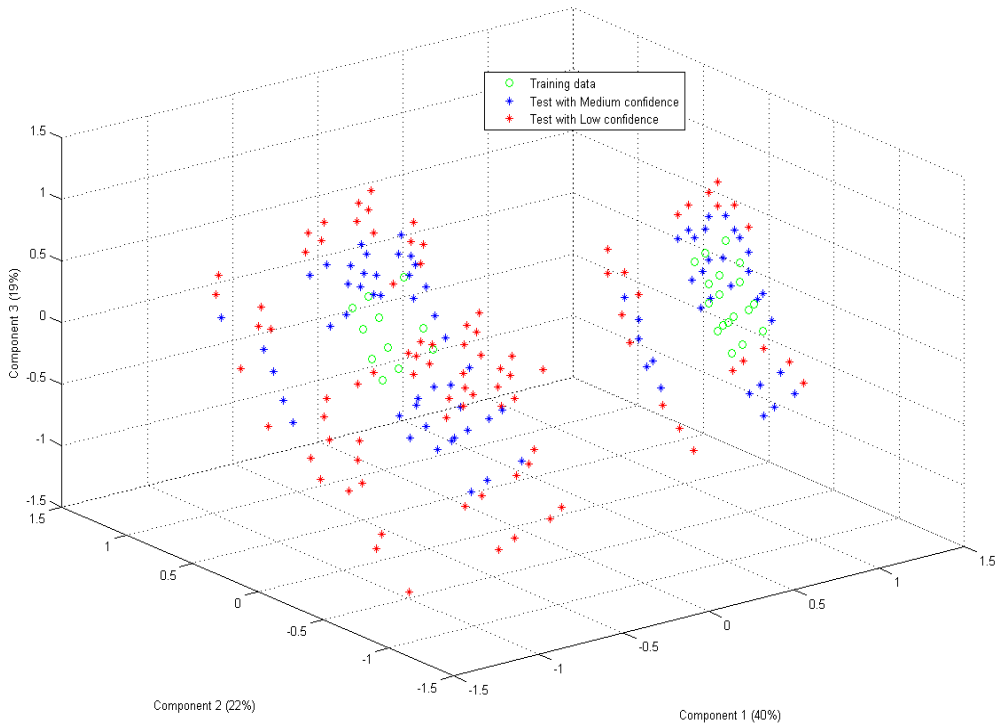

Figure 4.11: Compounds distribution based on their confidence value.

Data scarcity was the main challenge for this application. The available band gap values for chacopyrites were very limited due to its expensive calculations. This resulted in a small set of labeled data that was to be used for both model training and validation. We focused our effort on building robust models that can be generalized for bigger data sets with totally new compounds, and thus we worked on improving the accuracy for compounds with low confidence level.

Figure 4.11 shows the 3-dimensional distribution of both training and testing sets color coded by confidence. Using the best performing OLS, PLS and Lasso models, we predicted the band gap of the 204 compounds we have. Only 28 of these compounds were included in the training,

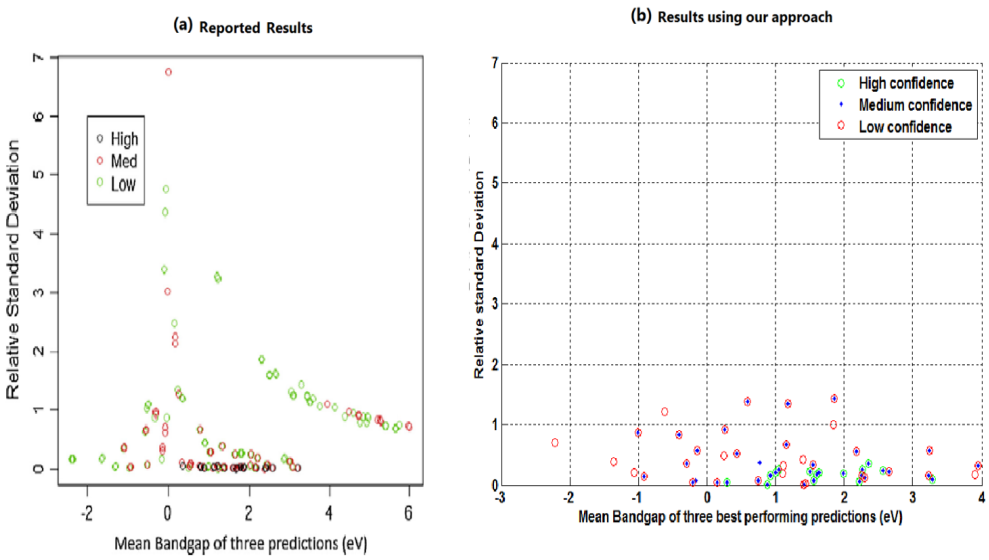

Figure 4.12: Consistency of the top performing models for OLS, PLS and Lasso regressions for the whole data set.

whereas the rest were excluded with different confidence level. Figure 4.12 shows the consitency of our predictions versus the results reported by Dey et al. (2014) [3]. It gives the plot of the variability of prediction as captured by the relative standard deviation of the three top predicted values for each compound (using OLS, PLS and Lasso) versus their means.

An overall inspection of Figure 4.12 reveals that the predictions are most consistent for "High" and "Medium" confidences, and least consistent for the "Low" confidence compounds, as expected. Our predictions were obtained using the models that were trained using the set of six features selected by Correlation criteria (AN(B), PR(B), EN(C), MP(C), PR(C), AN(C) and EN(C)) which has proved to be the most accurate when tested on the new labeled compounds. Using this reduced set of features has improved both the consistency and the reliability of our models compared to the previously reported results. We obtained less negative predictions with a noticeable decrease of the relative standard deviations.

4.2 Predicting the magnetic moment of Dilute Materials Semiconductors

The data used for the magnetic moment prediction for DMS materials consists of descriptors corresponding to the binary and elementary properties of 36 DMS systems based on two hosts: GaP and GaN hosts. We constructed our data such that each row contains all the available elementary descriptors of the elements present in both hosts (H_1H_2 formula) and the co-dopant compound (A_2B

formula). To this end, we exploited all the features described in Table 3.6. These features were roughly chosen based on previous studies and on the already existing conjectures about magnetism origins in DMSs.

We formed a data set that comprises a total of 36 observations (for both hosts), and 30 features. The data was divided into two subsets (a training and a testing sets). Each subset contains observations for both hosts as detailed in Table 4.4. We kept a balanced representation of both hosts within the subsets.

TABLE 4.4

Data distribution for DMS application.

	Host 1: GaP	**Host 2: GaN**	**Total**
Training Set	10	12	**22**
Testing Set	6	8	**14**

Our main goal herein was to assess the relevance of the features candidate and to identify the ones that have the highest contribution to the DMSs' magnetic moment. We applied features selection and ranking techniques and used the top performing set of each of these techniques to train OLS, PLS and Lasso regression models.

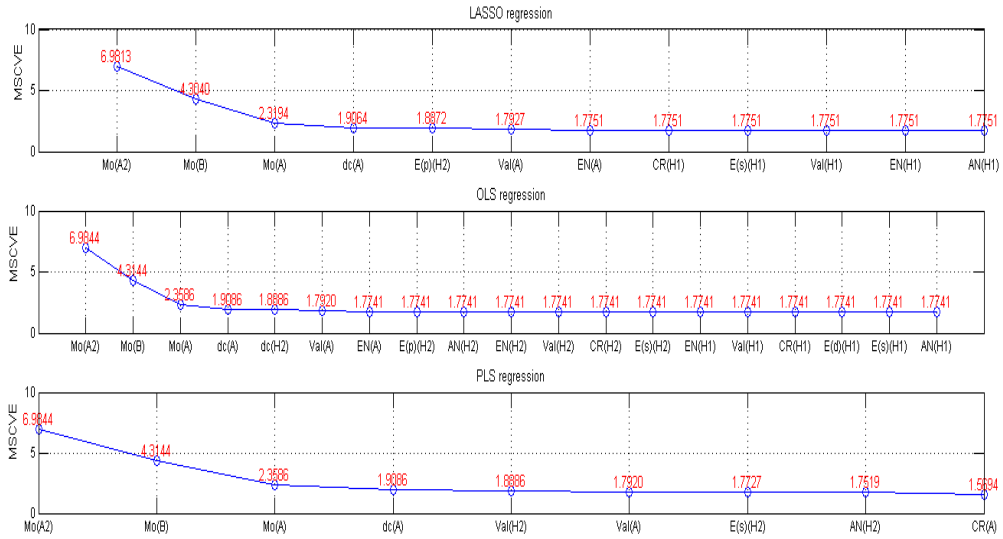

Figure 4.13: Sequential forward feature selection error evolution for DMS magnetic moment prediction.

Figure 4.13 shows the evolution the MSCVE of the the sequential forward feature selection on the training set. According to this graph, the three descriptors related to the elementary magnetic

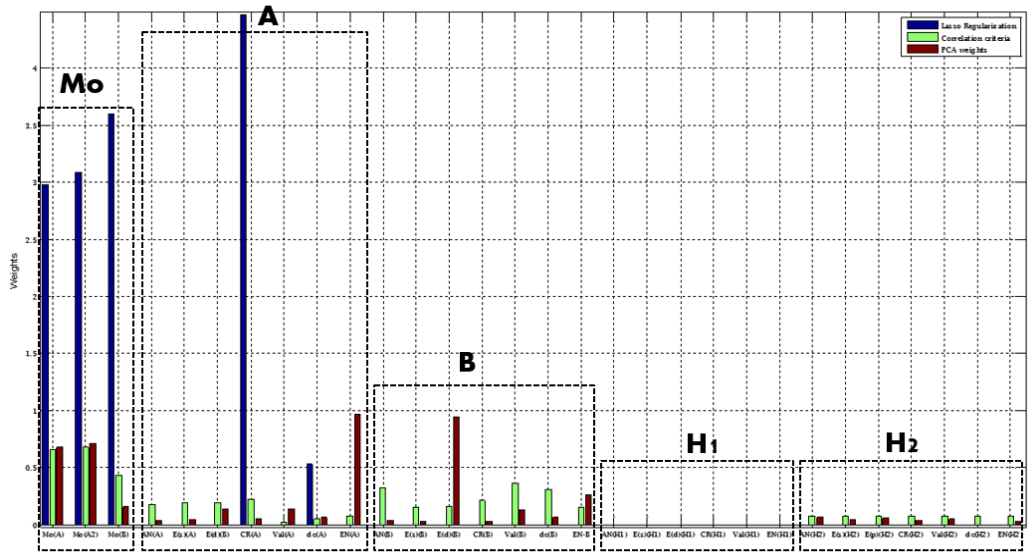

Figure 4.14: Features' weights for DMS magnetic moment prediction.

moment are the most important descriptors. Once included in the features' sets, the MSCVE drops byabout 70%. For the remaining features, those related to the properties of the elements present in the host (i.e., H_1 and H_2) and the first element of the co-dopant (i.e., A) seem to be more relevant compared to the rest. This observation was confirmed further by the features ranking we obtained based on the correlation criteria, Lasso regularization coefficients' weights and the first three principal component weights weights. Based on this ranking, we can notice also that the second element of the host (H_2) has more contribution than the first element (H_1).

TABLE 4.5

Selected features subsets.

Technique		**Selected subset**	**MSCVE**
Correlation criteria		$Mo(A)$, $Mo(B)$, $Mo(A_2)$, $d_c(B)$, $Val(B)$, $AN(B)$	1.893
Lasso regularization		$Mo(A)$, $Mo(B)$, $Mo(A_2)$, $d_c(A)$, $CR(A)$	2.878
Sequential forward selection	**Lasso**	$Mo(A)$, $Mo(B)$, $Mo(A_2)$,$d_c(A)$, $E(p)(H_2)$, $Val(A)$, $EN(A)$	1.775
	PLS	$Mo(A)$, $Mo(B)$, $Mo(A_2)$, $d_c(A)$, $d_c(H_2)$, $E(p)(H_2)$, $Val(A)$, $EN(A)$	1.774
	OLS	$Mo(A)$, $Mo(B)$, $Mo(A_2)$, $dc(A)$, $E(s)(H_2)$, $Val(A)$, $AN(H_2)$, $Val(H_2)$	1.569

We used the output of these different techniques for the training set to select the top per-

forming descriptors subset as highlighted in Table 4.5.

The interpretations based on the training subset can not be trusted unless validated using a testing set including new materials. In order to draw more insightful conclusions about the features that are more likely to contribute to magnetism for DMSs, we tested our models, five models per regression technique depending on the selected features set, on the fourteen compounds that we have excluded from our training set.

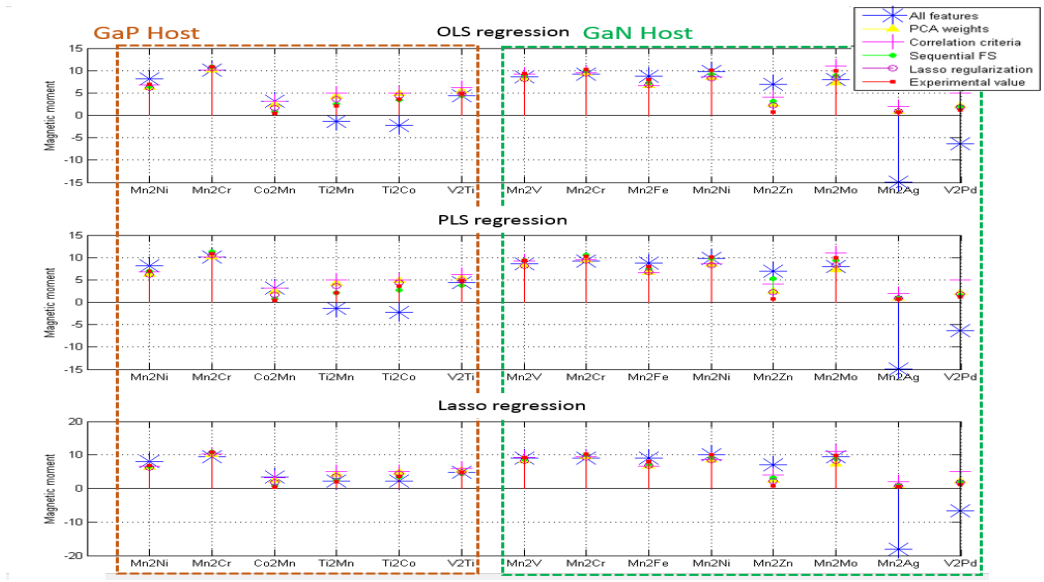

Figure 4.15: Plot of the Predicted total magnetic moment vs. the Experimental magnetic moment for different regression models.

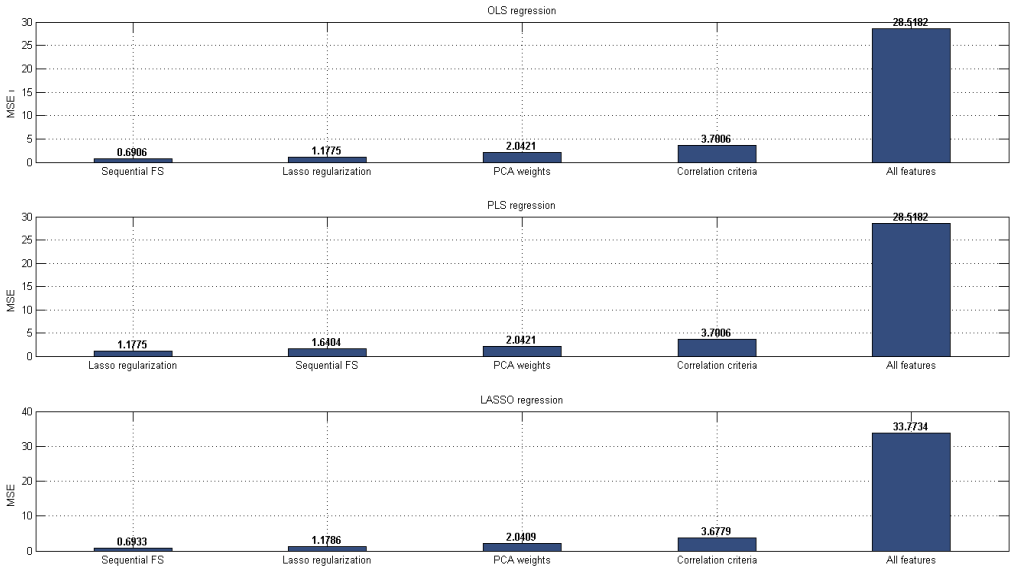

Figure 4.16: Models' MSE for the testing DMS systems.

The predictions we have obtained are given by Figure 4.15. Each model's output is plotted with a different marker. The reference value is marked by the red dot. Generally, all the markers are close to each other and to the reference value as well. However, the blue star marker corresponding to the estimations obtained using the whole set of features (30 features)is out of range in many cases for OLS, PLS and Lasso regressions.

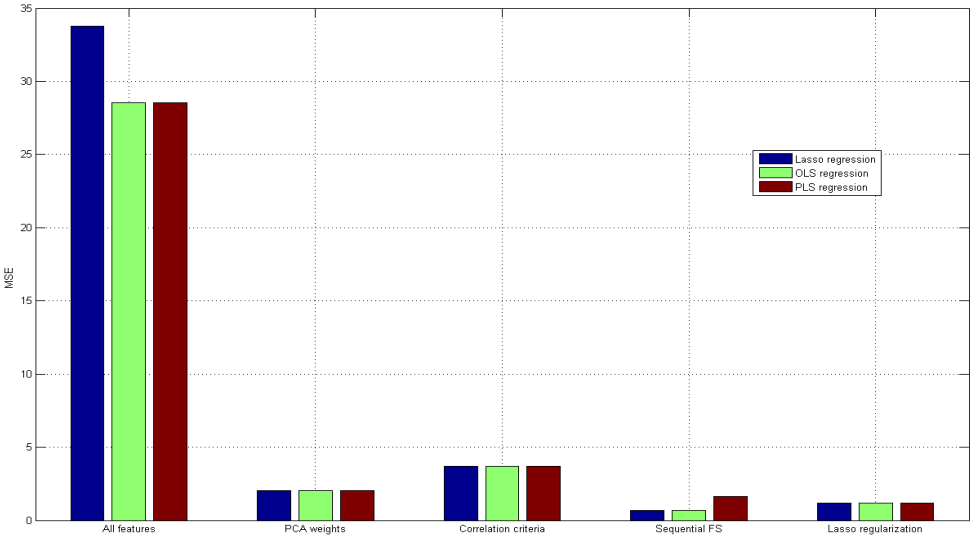

Figure 4.17: Summary of the models' MSE for the testing DMS systems.

We computed the Mean Square Error, for all the estimations for the different models as illustrated in Figures 4.16 and 4.17. Applying features selection techniques have considerably improved the prediction performances. The models that were trained using all the set of features gave the largest error which was reduced by more than 90% after applying feature selection. Sequential selection gave the best performances for both Lasso and OLS regressions. The other methods gave similar estimations with tiny deviations. The models behavior is similar for the three regression techniques. PLS and OLS seem to give better result for all the features. Whereas, Lasso performance increases for the selected features subsets.

CHAPTER 5

CONCLUSIONS AND POTENTIAL FUTURE WORK

5.1 Conclusions

In this thesis, we have introduced a new approach that aims to accelerate data-driven discovery of materials property. We developed statistical learning algorithms supervised by fundamental materials science principals to predict and model key properties of different types of semiconductors. We focused on two main components of the machine learning process: (i) feature extraction and selection; and (ii) learning algorithms. We combined different feature selection techniques to Lasso, Ordinary Least Square and Partial Least Square regressions to build robust regression models that take into account the intrinsic properties of the training data as well as the impacts of the different descriptors.

Our approach was successfully applied and tested to enhance the prediction of chalcopyrites' band gap, and to identify the top factors responsible for Defect-Induced Magnetism in Dilute Semiconductors by predicting several systems' total magnetic moment.

We have significantly improved upon prior results in informatics-based prediction of band gap by reducing the dimensionality of the training data. Our experiments showed that our approach can boost band gap prediction accuracy by more than 40% for the same testing set while keeping a high consistency between the different regression techniques. Furthermore, we found that the band gap energy of ABC_2 compounds, is highly correlated to the atomic number and the pseudopotential radius of the B and C elements and to the melting point of the C element. We also showed that the bond dissociation (BD) energy describing B-C and C-C bonds can be good additional features to boost our application performance. In fact, combining BD to the previously selected set of features yielded 70% enhancement of the prediction error.

For DIM modeling in DMS materials, our approach yielded a number of interesting findings. We showed that the elementary magnetic moment as well as the elementary properties of the first atom in the co-dopant formula and the second atom in the host are highly correlated to the system's total magnetic moment. This helped us reduce our original set of 30 features to just 8 features while

improving the prediction accuracy by around 90%.

5.2 Potential Future Work

Although our approach has shown promising results, there is still room for improvement. The following sections list three main areas that could be explored in the future to build upon the proposed work.

5.2.1 Expanding the data sets

Data scarcity has been the main challenge during our work. Expanding the labeled training data can improve the prediction accuracy. For DIM modeling in DMSs, our conducted work is still at its preliminary phase. We have explored the possibility of expanding our data sets by including new hosts, new features and new co-dopants. Our experiments were carried for two hosts only (GaN and GaP). More hosts including ZnO, TiO_2, MoS_2 and SnO_2 could be used to expand the training data. This can help in building more robust models and especially drawing more insightful conclusions.

5.2.2 Clustering

Generalizing a given regression model is not an easy task, especially when the inherent structure of the data is not favorable. When the available labeled data is limited, which is the case for our applications, applying clustering techniques can be very helpful to perform. It gives more insights about the validity of the generalized model by investigating distributions of both the training and the testing data sets. It helps also detecting outliers prior to the data mining process which allows avoiding performance deterioration due to error propagation and model mis-specification. Integrating clustering to our learning approach for DMS modeling can be very helpful. In fact, data corresponding to different hosts can form different clusters. We can use this observation to investigate the possibility of building separate single regression models for different hosts and even to explore the possibility of integrating multiple instance learning to the process.

5.2.3 Ensemble learning

Another key area worth investigating in future works is ensemble learning techniques. In fact, building a robust regression model that takes into account the intrinsic property of the training data as well as the impacts of the different perturbations it can undergo requires aggregating

multiple regression models. So far, we used three different regression techniques and investigated their consistency. However, each technique has its own advantages and flaws depending on the way it considers the inherent data structure. Integrating these different techniques using appropriate fusion methods can improve the ensemble's performance by pruning the base models' weaknesses.

REFERENCES

[1] Yingzhi Zeng, Soo Jin Chua, and Ping Wu, "On the prediction of ternary semiconductor properties by artificial intelligence methods," *Chemistry of materials*, vol. 14, no. 7, pp. 2989–2998, 2002.

[2] C Suh, A Rajagopalan, X Li, and K Rajan, "Combinatorial materials design through database science," in *MATERIALS RESEARCH SOCIETY SYMPOSIUM PROCEEDINGS*. Warrendale, Pa.; Materials Research Society; 1999, 2004, vol. 804, pp. 333–342.

[3] Partha Dey, Joe Bible, Somnath Datta, Scott Broderick, Jacek Jasinski, Mahendra Sunkara, Madhu Menon, and Krishna Rajan, "Informatics-aided bandgap engineering for solar materials," *Computational Materials Science*, vol. 83, pp. 185–195, 2014.

[4] Md. Atikur Rahman, "A review on semiconductors including applications and temperature effects in semiconductors," *American Scientific Research Journal for Engineering and Technology and Sciences*, pp. 2313–4410.

[5] Mark E Eberhart, Dennis P Clougherty, et al., "Looking for design in materials design," *Nature materials*, vol. 3, no. 10, pp. 659–661, 2004.

[6] Martin Jansen and J Christian Schön, ""design" in chemical synthesis—an illusion?," *Angewandte Chemie International Edition*, vol. 45, no. 21, pp. 3406–3412, 2006.

[7] RG Parr and W Yang, "Density functional theory of atoms and moleculesoxford univ," *Press, New York*, 1989.

[8] Pierre Hohenberg and Walter Kohn, "Inhomogeneous electron gas," *Physical review*, vol. 136, no. 3B, pp. B864, 1964.

[9] Walter Kohn and Lu Jeu Sham, "Self-consistent equations including exchange and correlation effects," *Physical review*, vol. 140, no. 4A, pp. A1133, 1965.

[10] Peter G Schultz, "Commentary on combinatorial chemistry," *Applied Catalysis A: General*, vol. 254, no. 1, pp. 3–4, 2003.

[11] Radislav Potyrailo, Krishna Rajan, Klaus Stoewe, Ichiro Takeuchi, Bret Chisholm, and Hubert Lam, "Combinatorial and high-throughput screening of materials libraries: Review of state of the art," *ACS combinatorial science*, vol. 13, no. 6, pp. 579–633, 2011.

[12] U.S.D. of Energy, "Computational materials science and chemistry for innovation," 2010.

[13] "Materials genome initiative for global competitiveness," 2011.

[14] Prachi Patel, "Materials genome initiative and energy," *MRS bulletin*, vol. 36, no. 12, pp. 964, 2011.

[15] Tim Mueller, Aaron Gilad Kusne, and Rampi Ramprasad, "Machine learning in materials science: Recent progress and emerging applications," *Reviews in Computational Chemistry*, vol. 29, pp. 186, 2016.

[16] Bryce Meredig, Ankit Agrawal, Scott Kirklin, James E Saal, JW Doak, A Thompson, Kunpeng Zhang, Alok Choudhary, and Christopher Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," *Physical Review B*, vol. 89, no. 9, pp. 094104, 2014.

[17] Ghanshyam Pilania, Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran, and Ramamurthy Ramprasad, "Accelerating materials property predictions using machine learning," *Scientific reports*, vol. 3, pp. 2810, 2013.

[18] Felix Faber, Alexander Lindmaa, O Anatole von Lilienfeld, and Rickard Armiento, "Crystal structure representations for machine learning models of formation energies," *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1094–1101, 2015.

[19] Koji Fujimura, Atsuto Seko, Yukinori Koyama, Akihide Kuwabara, Ippei Kishida, Kazuki Shitara, Craig AJ Fisher, Hiroki Moriwake, and Isao Tanaka, "Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms," *Advanced Energy Materials*, vol. 3, no. 8, pp. 980–985, 2013.

[20] Arun Mannodi-Kanakkithodi, Ghanshyam Pilania, and Rampi Ramprasad, "Critical assessment of regression-based machine learning methods for polymer dielectrics," *Computational Materials Science*, vol. 125, pp. 123–135, 2016.

[21] Joseph Leo Shay and Jack Harry Wernick, *Ternary Chalcopyrite Semiconductors: Growth, Electronic Properties, and Applications: International Series of Monographs in The Science of The Solid State*, vol. 7, Elsevier, 2013.

[22] Jason M., Crowley, Jamil Tahir-Kheli, and William A. Goddard, "Resolution of the band gap prediction problem for materials design," *Materials and Process Simulation Center, California Institute of Technology, Pasadena, California 91125, United States*, vol. MC139-74, March 2016.

[23] SA Wolf, DD Awschalom, RA Buhrman, JM Daughton, S Von Molnar, ML Roukes, A Yu Chtchelkanova, and DM Treger, "Spintronics: a spin-based electronics vision for the future," *Science*, vol. 294, no. 5546, pp. 1488–1495, 2001.

[24] Stuart A Wolf, Almadena Yu Chtchelkanova, and Daryl M Treger, "Spintronics—a retrospective and perspective," *IBM Journal of Research and Development*, vol. 50, no. 1, pp. 101–110, 2006.

[25] SA Wolf, DD Awschalom, RA Buhrman, JM Daughton, S Von Molnar, ML Roukes, A Yu Chtchelkanova, and DM Treger, "Spintronics: a spin-based electronics vision for the future," *Science*, vol. 294, no. 5546, pp. 1488–1495, 2001.

[26] U.S. Department of Energy, *DOE Solar Energy Technologies Program, Overview and Highlights*.

[27] Martin A Green, "Solar cells: operating principles, technology, and system applications," 1982.

[28] Martin A Green, "The path to 25% silicon solar cell efficiency: history of silicon cell evolution," *Progress in Photovoltaics: Research and Applications*, vol. 17, no. 3, pp. 183–189, 2009.

[29] Marilyne Andersen, "Wavelengths of solar radiation," 2004.

[30] Richard LeSar, "Materials informatics: An emerging technology for materials development," *Statistical Analysis and Data Mining*, vol. 1, no. 6, pp. 372–374, 2009.

[31] Tomasz Dietl, José Menéndez, and Chris G Van de Walle, "Spintronics and ferromagnetism in wide-band-gap semiconductors," in *AIP Conference Proceedings*. AIP, 2005, vol. 772, pp. 56–64.

[32] JK Furdyna, "Diluted magnetic semiconductors: Issues and opportunities," *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, vol. 4, no. 4, pp. 2002–2009, 1986.

[33] Tomasz Dietl, H Ohno, F Matsukura, J Cibert, and D Ferrand, "Zener model description of ferromagnetism in zinc-blende magnetic semiconductors," *science*, vol. 287, no. 5455, pp. 1019–1022, 2000.

[34] M Herbich, A Twardowski, D Scalbert, and A Petrou, "Bound magnetic polaron in cr-based diluted magnetic semiconductors," *Physical Review B*, vol. 58, no. 11, pp. 7024, 1998.

[35] Gordon E Moore, "Cramming more components onto integrated circuits, reprinted from electronics, volume 38, number 8, april 19, 1965, pp. 114 ff.," *IEEE Solid-State Circuits Society Newsletter*, vol. 20, no. 3, pp. 33–35, 2006.

[36] Krishna Rajan, "Materials informatics," *Materials Today*, vol. 8, no. 10, pp. 38–45, 2005.

[37] CJ Long, J Hattrick-Simpers, M Murakami, RC Srivastava, I Takeuchi, VL Karen, and X Li, "Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis," *Review of Scientific Instruments*, vol. 78, no. 7, pp. 072217, 2007.

[38] Geoffroy Hautier, Christopher C Fischer, Anubhav Jain, Tim Mueller, and Gerbrand Ceder, "Finding nature's missing ternary oxide compounds using machine learning and density functional theory," *Chemistry of Materials*, vol. 22, no. 12, pp. 3762–3767, 2010.

[39] Geoffroy Hautier, Christopher C Fischer, Anubhav Jain, Tim Mueller, and Gerbrand Ceder, "Finding nature's missing ternary oxide compounds using machine learning and density functional theory," *Chemistry of Materials*, vol. 22, no. 12, pp. 3762–3767, 2010.

[40] Joseph J Hanak, "The "multiple-sample concept" in materials research: Synthesis, compositional analysis and testing of entire multicomponent systems," *Journal of Materials Science*, vol. 5, no. 11, pp. 964–971, 1970.

[41] Laura M. Bartolo, "Notes aterials informatics lab," *Kent State University*, October 2006.

[42] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al., "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *Apl Materials*, vol. 1, no. 1, pp. 011002, 2013.

[43] Melinda Y Han, Barbaros Özyilmaz, Yuanbo Zhang, and Philip Kim, "Energy band-gap engineering of graphene nanoribbons," *Physical review letters*, vol. 98, no. 20, pp. 206805, 2007.

[44] Nicolai Lehnert, Mary Grace I Galinato, Florian Paulat, George B Richter-Addo, Wolfgang Sturhahn, Nan Xu, and Jiyong Zhao, "Nuclear resonance vibrational spectroscopy applied to [fe (oep)(no)]: The vibrational assignments of five-coordinate ferrous heme nitrosyls and implications for electronic structure," *Inorganic chemistry*, vol. 49, no. 9, pp. 4133, 2010.

[45] Michael N Blonsky, Houlong L Zhuang, Arunima K Singh, and Richard G Hennig, "Ab initio prediction of piezoelectricity in two-dimensional materials," *ACS nano*, vol. 9, no. 10, pp. 9885–9891, 2015.

[46] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[47] P. Yang et al., "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, 2010.

[48] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, vol. 9, no. 4, July/August 2012.

[49] Leo Breim, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[50] Yong Liu, Xin Yao, and Tetsuya Higuchi, "Evolutionary ensembles with negative correlation learning," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 4, pp. 380–387, 2000.

[51] Juan J. Rodriguez, Ludmila I. Kuncheva, and Carlos J. Alonso, "Rotation forest: a new classifier ensemble," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 3, pp. 271–302, 2005.

[52] Nicolas Garcia-Pedrajas, Cesar Hervas-Martinez, and Domingo Ortiz-Boyer, "Cooperative coevolution of artificial neural network ensembles for pattern classification," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 4, pp. 380–387, 2000.

[53] Guyon I and Elisseeff A, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, pp. 1157–1182, 2003.

[54] Blum AL. and Langley P., "Selection of relevant features and examples in machine learning," *Artificial Intelligence Journal*, vol. 97, pp. 245–270, 1997.

[55] J. M. Moreira, C. Soares, A. M. Jorge, and Jorge Freire de Sousa, *Ensemble Approaches for Regression: a Survey*, Elsevier, University of Porto, Rua Dr. Roberto Frias, Porto, Portugal, 2007.

[56] Fabio Roli, Giorgio Giacinto, and Gianni Vernazza, "Methods for designing multiple classifier systems," *International Workshop on Multiple Classifier Systems*, vol. LNCS 2096, pp. 78–87, Springer,2001.

[57] Cristopher J. Merz, "Classification and regression by combining models, phd thesis,," *University of California*, vol. U.S.A.

[58] T. Mitchell, *Machine learning*, McGraw Hill, 1997.

[59] Alan O. Sykes, "An introduction to regression analysis," *University of Chicago*.

[60] P. Dey, J. Bible, S. Datta, S.Broderick, J. Jasinski, M. Sunkara, M.Menon, and K.Rajan, "Informatics-aided bandgap engineering for solar materials," *Computational Materials Science*, vol. 83, pp. 185–195, 2014.

[61] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[62] Hervé Abdi, "Partial least squares regression and projection on latent structure regression (pls regression)," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97–106, 2010.

[63] Walter RL Lambrecht and Sergey N Rashkeev, "First-principles calculations of second-order optical response functions in chalcopyrite semiconductors," *Journal of Physics and Chemistry of Solids*, vol. 64, no. 9, pp. 1615–1619, 2003.

[64] VP Antropov, "The exchange coupling and spin waves in metallic magnets: removal of the long-wave approximation," *Journal of magnetism and magnetic materials*, vol. 262, no. 2, pp. L192–L197, 2003.

[65] Kazunori Sato, Lars Bergqvist, J Kudrnovskỳ, Peter H Dederichs, Olle Eriksson, Ilja Turek, Biplab Sanyal, Georges Bouzerar, Hiroshi Katayama-Yoshida, VA Dinh, et al., "First-principles theory of dilute magnetic semiconductors," *Reviews of modern physics*, vol. 82, no. 2, pp. 1633, 2010.

[66] S Picozzi, M Ležaić, and S Blügel, "Electronic structure and exchange constants in magnetic semiconductor digital alloys: chemical and band-gap effects," *physica status solidi (a)*, vol. 203, no. 11, pp. 2738–2745, 2006.

[67] Antonis N Andriotis and Madhu Menon, "Successive spin-correlated local processes underlying the magnetism in diluted magnetic semiconductors and related magnetic materials," *Computational Approaches to Materials Design: Theoretical and Practical Aspects: Theoretical and Practical Aspects*, p. 13, 2016.

[68] D Palmer, "Crystalmaker crystalmaker software ltd," *Yarnton, Oxfordshire, England*, 2009.

CURRICULUM VITAE

NAME: Fadoua Khmaissia

ADDRESS: Computer Engineering & Computer Science Department
Speed School of Engineering
University of Louisville
Louisville, KY 40292

EDUCATION:

M.S., Computer Science & Engineering
May 2017
University of Louisville, *Louisville, Kentucky*

B.Eng., Telecommunications Engineering
June 2015
Higher School of Communications of Tunis, *Tunis, Tunisia*

HONORS AND AWARDS:

1. J. B. Speed School CECS department Arthur M. Riehl Award, May 2017.
2. Golden Key International Honour Society Member, September 2016.
3. Higher School of Communications of Tunis Travel Award, June 2014.
4. Tunisian National Scholarship for Engineering Studies, September 2012.