Computational Materials Science 147 (2018) 304–315

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Accelerating band gap prediction for solar materials using feature selection and regression techniques ☆

Fadoua Khmaissia [a], Hichem Frigui [a,*], Mahendra Sunkara [b], Jacek Jasinski [b], Alejandro Martinez Garcia [b], Tom Pace [c], Madhu Menon [d]

[a] University of Louisville, Department of Computer Engineering and Computer Science, Louisville, KY 40202, United States
[b] University of Louisville, Department of Chemical Engineering, Louisville, KY 40202, United States
[c] University of Kentucky, Department of Physics and Astronomy, Lexington, KY 40506, United States
[d] University of Kentucky, Center for Computational Sciences, Lexington, KY 40506, United States

ARTICLE INFO

Article history:
Received 2 October 2017
Received in revised form 16 January 2018
Accepted 3 February 2018
Available online 22 February 2018

Keywords:
Machine learning
Band gap engineering
Chalcopyrites
Feature selection
Regression

ABSTRACT

We present a novel approach to apply machine learning techniques to build a more robust prediction model for band-gap energies (BG-E) of chalcopyrites, a class of materials for energy applications in the fields of solar energy, photocatalysis, and thermoelectrics. Guided by knowledge from domain experts and by previous works on the field, we aim to accelerate the discovery of new solar materials. 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 the band gap energy of chalcopyrites materials. Compared to the results reported in Zeng et al. (2002), Suh et al. (1999, 2004), and Dey et al. (2014), our approach shows that learning and using only a subset of relevant features can improve the prediction accuracy by about 40%.

1. Introduction

Solar energy provides around 2% of the world's total energy [4]. 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 made of semiconductors like silicon. When it comes to designing new solar material, a key step 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. It is the energy difference (in eV) between the top of the valence band and the bottom of the conduction bands in semiconductors and insulators [5]. To perform this task, scientists have mostly relied on appropriate *ab initio* techniques [6–8]. 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 get computationally expensive when system size gets very large.

Since the launch of the Material Genome Initiative (MGI) in 2011 by the US government, more efforts have been invested to address most of the above-mentioned challenges. This initiative aims mainly to accelerate the discovery, development and deployment of new materials at a fraction of the cost [9]. This is made possible by providing the necessary policies, resources and infrastructure for collaborative researches.

As we enter this MGI era, it has become crucial to quickly and accurately predict the band gaps 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.

Our work builds on previous works [1–3] aiming to predict band gaps of new chalcopyrite compounds using statistical learning approaches such as Ordinary Least Squares, Partial Least Squares and Lasso regression methods combined with well adjusted pre-processing techniques. The used data set comprises atomic and crystallographic properties of ternary chalcopyrites semiconductors which are $CuFeS_2$-like compounds that crystallizes in the tetragonal form (ABC_2 formula) [10].

☆ This work was supported in part by U.S. Army Research Office Grants Number W911NF-13-1-0066 and W911NF-14-1-0589.

* Corresponding author.
E-mail address: h.frigui@louisville.edu (H. Frigui).

https://doi.org/10.1016/j.commatsci.2018.02.012
0927-0256/

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 prediction is still a very challenging task, and it can be related to several aspects and properties of the compound in question. The main problem for such applications is that important features that can lead to reliable predictions are still unknown. Thus, we need to include as many features as possible. However, a regression model learned from a training data where a large number of attributes are irrelevant cannot be robust and may not generalize well to new test data.

Our goal is, as a first step, to investigate and determine if all of the previously chosen descriptors [1–3] 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 investigate 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.

2. Related works

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 [7]. The main interests are to use these materials for energy applications in the areas of solar cells photocatalysis, and thermoelectrics [11]. These chalcopyrites exhibit band-gaps that can be tuned to absorb light of different wavelengths in multi-junction cells [12], which optimizes the usage of the solar spectrum. The best example is Cu(In,Ga) Se_2 (CIGSe) [13], for which, the solar cell efficiency has recently been demonstrated at 22.6% [14].

Several studies have started investigating the possible models that can describe the relationship between the band gap and the chemical stoichiometries and fundamental properties of the constituents of these chalcopyrites [1–3,15–17].

The pilot study carried out by Zeng et al., in 2001 [1], has laid the foundation of our work. The authors used artificial neural networks to estimate the correlation between band gap energies (and lattice constants) of chalcopyrites and their respective chemical 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.

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 1.

The band gap (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}.

3. Proposed 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 studied application. To this end, four major tasks are performed as detailed in Fig. 1.

Table 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 developed algorithms are based on *ab initio* 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. Various feature selection algorithms are applied in order to extract the top performing set of features for the predicted variable. We started by analyzing the original set of features [1–3] and added more features that were likely to correlate with band gap energy according to experts in the field. Regression analysis is performed afterward to estimate the underlying models. We trained single models for the top selected features subsets which were then evaluated in terms of prediction errors for compounds with different confidence values in order to assess the accuracy.

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 in our research group. Our training data, on which the relevant descriptors are learned and the regression models are built, consists of both theoretical and experimental data.[1] Descriptors are extracted based on fundamental atomic and crystallographic properties of the studied materials and according to their physical significance to the target variable.

We started from the reported data by previous studies [1–3]. We constructed similar data sets in order to replicate the previous results and explore the possibilities of enhancement. Earlier work relied mainly on the five elementary descriptors outlined in Table 1. In this paper, we propose using additional features that capture the interactions between the compounds' elements. We added binary descriptors like Bond Dissociation Energy (BD) and Bond Length (BL) as outlined in Table 2.

The Bond Dissociation energies were obtained using the *ab initio* density functional theory (DFT) computations employing the hybrid exchange-correlation functional, Heyd-Scuseria-Ernzerhof (HSE06) [18]. All our *ab initio* calculations were performed using the Vienna *Ab Initio* Simulation Package (VASP) electronic structure computer code [19–21]. A kinetic energy cutoff of 500 eV was found to be sufficient to achieve a total energy convergence of the energies of the systems to within 1 meV. The optimization of atomic positions was allowed to proceed without any symmetry constraints until the force on each atom is less than 5 meV/Å.

The Bond Length measures were obtained using *CrystalMaker* software [22]. This software estimates the length of the bonds based on relaxation of the crystallographic structure of the

[1] Experimental data refers here to the known band gap energies as reported in previous works [1–3].