Computational Materials Science 83 (2014) 185–195

Contents lists available at ScienceDirect

Computational Materials Science

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

Informatics-aided bandgap engineering for solar materials

Partha Dey [a,1], Joe Bible [b,1], Somnath Datta [b], Scott Broderick [a], Jacek Jasinski [c], Mahendra Sunkara [c], Madhu Menon [d], Krishna Rajan [a,*]

[a] Iowa State University, Materials Science and Engineering, Ames, IA 50011, United States
[b] University of Louisville, Department of Bioinformatics & Biostatistics, Louisville, KY 40202, United States
[c] University of Louisville, Department of Chemical Engineering, Louisville, KY 40202, United States
[d] University of Kentucky, Center for Computational Sciences, Lexington, KY 40506, United States

ARTICLE INFO

Article history:
Received 20 May 2013
Received in revised form 9 August 2013
Accepted 10 October 2013
Available online 30 November 2013

Keywords:
Compound semi-conductors
Bandgap
Chalcopyrites
Informatics
Solar materials

ABSTRACT

This paper predicts the bandgaps of over 200 new chalcopyrite compounds for previously untested chemistries. An ensemble data mining approach involving Ordinary Least Squares (OLS), Sparse Partial Least Squares (SPLS) and Elastic Net/Least Absolute Shrinkage and Selection Operator (Lasso) regression methods coupled to Rough Set (RS) and Principal Component Analysis (PCA) methods was used to develop robust quantitative structure – activity relationship (QSAR) type models for bandgap prediction. The output of the regression analyses is the predicted bandgap for new compounds based on a model using the descriptors most related to bandgap. Feature ranking algorithms were then employed to: (i) assess the connection between bandgap and the chemical descriptors used in the predictive models; and (ii) understand the cause of outliers in the predictions. This paper provides a descriptor guided selection strategy for identifying new potential chalcopyrite chemistries materials for solar cell applications.

1. Introduction

Semiconducting chalcopyrites are a material system of interest due to their non-linear optical properties and technological applications [1,2]. *Ab initio* methods have previously been used to calculate the bulk electronic band structure of some chalcopyrite compounds. The perturbation method to the density functional has been used to calculate structural and dynamic properties of the lattice of some chalcopyrites [3]. But these methods are computationally intensive, which explains why only a limited number of the possible compounds have been studied until now. Moreover, the values obtained must be calibrated to match with actual experimental bandgap values because *ab initio* calculations most often underestimate bandgaps. Other groups have proposed approaches to rapid quantum mechanical calculations [4–6]. We instead utilize statistical learning to allow high throughput computation by predicting properties for compounds that have yet to be measured and which serve to guide future material science discoveries.

The present investigation refines and improves the predictive model proposed by Suh and Rajan for predicting the bandgap of chalcopyrites as a function of several basic elemental properties [7]. In that paper, the model fit the training data well, but did not match with the data used only for validation. Of the 205 new bandgap predictions for additional chemistries, 77 had predicted negative bandgap values. Further, the model was not applicable for understanding underlying physics and causes of outliers. To address these problems, we have made several modifications to the mathematical logic. (i) We expand beyond using partial least squares (PLS) as the only regression method to an ensemble approach, improving model predictivity and robustness. (ii) The dimensionality of the input data is increased by including the elemental descriptors for each component, as opposed to a single value based on a weighting scheme. By doing this, we remove any error associated with the scaling and enable physical interpretation of the model since the change in bandgap is linked to the individual components. (iii) We apply feature ranking approaches to delve into the underlying physics and identify the cause of outlier compounds [8]. (iv) We qualitatively define the uncertainty in new bandgap predictions so that the results can be more effectively used for new material design.

Discovery of new materials for solar energy conversion to fuels through photoelectrochemical energy conversion is regarded as one of the grand challenges in materials science, chemistry and renewable energy [9]. The key underlying aspect for enabling such discovery is the availability of tools and methods for rapid prediction of a large number of compounds with bandgaps in the range of 1.0–2.4 eV range. Specifically, the chalcopyrites (chemical formula ABC_2) have been widely used in processes such as catalysis, chemisorption or bioleaching, but the most promising application is

* Corresponding author. Address: 2240 Hoover Hall, Ames, IA 50011, United States. Tel.: +1 515 294 2670; fax: +1 515 294 5444.
E-mail address: krajan@iastate.edu (K. Rajan).
[1] Equal contributions.

0927-0256/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.commatsci.2013.10.016

perhaps their use in solar cells. While silicon-based solar cells still dominate the market, thin-film solar cells have nearly tripled in market share over the last five years [10]. The most successful in terms of application is the Cu(In,Ga)(S,Se)$_2$ compounds (CIGS) exhibiting bandgaps which are ideal for photovoltaic and photoelectrochemical solar energy conversion. These chalcopyrites provide an array of bandgaps that can be tuned to absorb different energy bands in multi-junction cells, utilizing as much of the solar spectrum as possible [11–13]. The majority of chalcopyrites considered for solar cell applications may be expressed as having a I-III-VI$_2$ composition or stoichiometry, though II-IV-V$_2$ compositions have also been studied. The most widely reported I-III-VI$_2$ compounds have Cu as the group I element, with Ag being chosen in a few cases.

Some other informatics based approaches [14–16] have utilized the physical parameters of known compounds to develop suitable mathematical models to predict bandgaps and other physico chemical properties of undiscovered compounds. Jackson et al. [14,15] proposed the rough set approach to explore the dependence of bandgap on several parameters like nonlinear second-order optical coefficient $\chi(2)$, ratio of the lattice constants c/a or the lattice dislocation parameter u. Zeng et al. [16] have used artificial neural network to relate bandgap energy and lattice constant of chalcopyrites to their chemical stoichiometries and properties of the constituent elements. We build on these works by developing multiple predictive models to ensure mathematical robustness, and use an input data set which is general enough that the number of chemistries that can be modeled is maximized. A key part of this work is the appropriate selection of descriptors. This is an issue that we have explored for numerous systems using a variety of techniques. The details of these works are provided in the literature [17–23]. While the criteria for descriptor selection is robust, the selection of descriptors utilized here does not preclude that other descriptors are needed.

The number of reported chemistries and corresponding properties of compound semiconductors is limited, with many of the reports instead focusing on processing issues. For example, the direct bandgap for only 44 chalcopyrite chemistries was found in literature. We use this limited knowledge base to develop our quantitative structure-activity relationships (QSAR) linking chalcopyrite chemistry and bandgap. This QSAR is then used to predict the bandgap for 227 chalcopyrite chemistries. The compounds modeled here overlap with those from the previous paper [22], but we are not considering those values here due to the issues discussed in terms of their questionable accuracy. Fig. 1 demonstrates the sporadic nature of increasing the knowledge base.

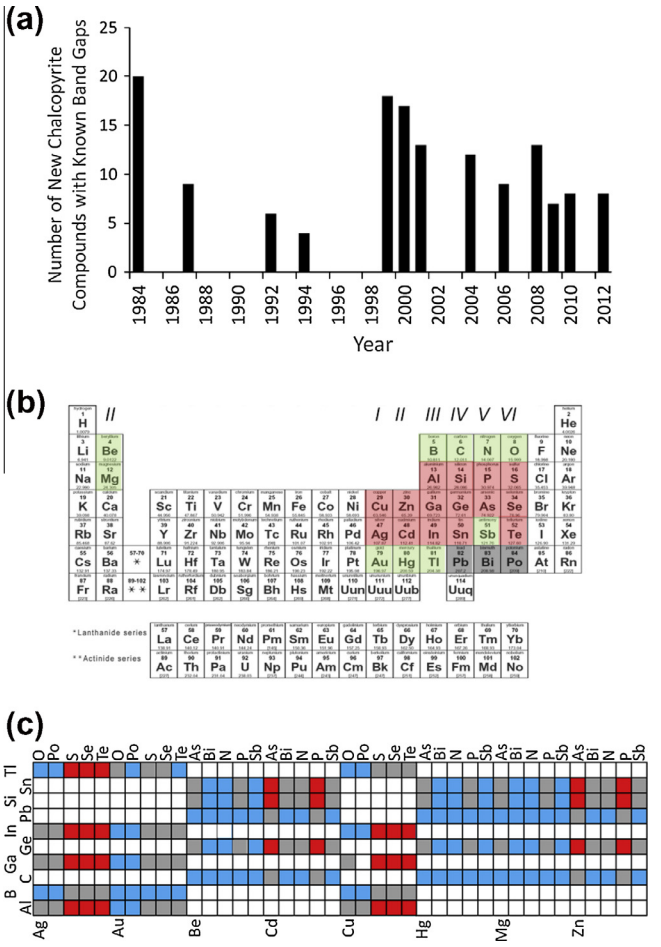

Fig. 1. (a) "Knowledge" spectrum of the pace of chalcopyrite semiconductor discovery starting with reports over the last two decades. The vertical axis represents the number of distinct chalcopyrite chemistries with known bandgaps (including both experimental and computed data). The data reported prior to 2004 for I-III-VI$_2$ or II-IV-V$_2$ compounds (please see Appendix) was used for developing the QSAR type model, which we then use to develop a virtual library of chalcopyrite compound semiconductor bandgaps. (b) Periodic table showing the elements used in the QSAR modeling as compound chemistries, with the group numbers shown. The red shaded elements are included in the training data, the green boxes are elements which were not used in the training data but have been recently reported via first principles calculations, while the gray boxes are elements for which the bandgap of compounds containing these elements has not been reported. The QSAR presented here is applicable to all of these elements, expanding the chemical design space for compound semiconductors. (c) Uncertainty in predictions of possible compounds, with bottom corresponding to A element, left corresponding to B element, and top corresponding to C element for ABC_2 compounds. Red indicates low uncertainty (comprised of elements in red in (b)), gray indicates some uncertainty, and blue indicates high uncertainty. White squares correspond to compounds that are not I–III–VI$_2$ or II–IV–V$_2$ compounds, although we can predict the bandgap for these with high uncertainty. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2. Methods

The present analysis is based on different techniques working in synergy with one another. The scheme of the analysis may be viewed as two separate inter-linked modules. The dependence of bandgap on the basic elemental input variables is modeled with (1) Ordinary Least Squares (OLS), (2) Sparse Partial Least Squares (SPLS) and (3) Elastic Net/Lasso. The importance of attributes and outlier compounds are assessed by: Rough Set (RS) Theory and Principal Component Analysis (PCA). These techniques are briefly summarized here, with the connection between the techniques shown in Fig. 2.

2.1. Regression techniques

The regression function $m(x) = E(Y|X = x)$ explains the relationship between a univariate response variable Y (also termed a dependent variable) and a p-variate input vector X (also called covariates or independent variables). Typically a parametric model $m(x) = (x, \beta)$ *beta* $\in \Re^p$, for a known function f (often linear) is postulated and the model is fit on a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$. Depending on the regression techniques used, the unknown parameter (state of the relationship) is estimated by $\hat{\beta}$ obtained by either a closed form expression, or by solving an estimating equation or by optimizing an objective function often subject to certain constraints. Heuristically, β and $\hat{\beta}$ can be thought of as the true coefficients which explain the physical relation between the descriptors and bandgap and the associated estimates thereof respectively. The hat notation is intended to serve as a disambiguation indicating that $\hat{\beta}$ is acquired using available information and thus may incompletely explain said relation. A major characteristic