Band Gap Estimation Using Machine Learning Techniques

Anantha Natarajan S[1], R Varadhan[2], Ezhilvel ME[3]

[1,2,3] *Department of Metallurgical & Materials Engineering,*

[1,2,3] *National Institute of Technology, Tiruchirappalli*

Abstract— **The purpose of this study is to build machine learning models to predict the band gap of binary compounds, using its known properties like molecular weight, electronegativity, atomic fraction and the group of the constituent elements in the periodic table. Regression techniques like Linear, Ridge regression and Random Forest were used to build the model. This model can be used by students and researchers in experiments involving unknown band gaps or new compounds.**

Keywords— **Machine Learning, Prediction, Band gap, Regression, Random Forest, Ridge Regression.**

I. Introduction

Since the late 20th century the exponential growth of the electronics and communication sector has led to numerous innovations in the semiconductor industry. Everyday newer and more efficient semiconductors are being discovered. The main property the scientists are targeting is the conductivity of the material this is directly related to the band gap of the underlying material.

The concept of band gap pertains to solid state physics: it generally refers to the energy difference in electron volts between the top of the valence band and the bottom of the conduction band in insulators and semiconductors. It helps us understanding characteristics of conductors, semiconductors and insulators. The band gap is the minimum amount of energy required for an electron to break free of its bound state.

When the band gap energy is met, the electron is excited into a free state, and can therefore participate in conduction. The band gap being an intrinsic property, is calculated using experiments like UV spectroscopy or differential and cyclic voltammetry. These experiments require large, expensive equipment and are tedious to perform. Some materials are difficult to handle and require special environments to perform these experiments. Calculating band gaps accurately is still one of the unsolved problems in solid state physics. There is no equation or direct solution to calculate band gaps accurately.

To help overcome these difficulties we propose and compare machine learning models to recognize meaningful patterns in band gap values across thousands of compounds and their chemical properties. Predicting the band gap, or more specifically predicting the range in which the band gap of a new material would lie between would give researchers and students a good idea of what to expect in the experiments before going for more elaborate ways of band gap calculation.

II. Related work

Machine learning tools and data science have been here for decades, but it's use in fields like material science is yet to come. A study was conducted at Kyoto University [1] to come up with a prediction model of band-gap for AX binary compounds by combination of density functional theory calculations and machine learning techniques.

This study used Ordinary least square regression (OLSR), least absolute shrinkage and selection operator (LASSO) and non-linear support vector regression (SVR) methods are applied with several levels of predictor sets. When the Kohn-Sham band-gap by GGA (PBE) or modified Becke-Johnson (mBJ) is used as a single predictor, OLSR model predicts the G0W0 band-gap of a randomly selected test data with the root mean square error (RMSE) of 0.54 eV.

III. Features

We target building a model using basic physical and chemical attributes of compounds. An important factor in the selection of features is the availability of reliable and consistent data. In order to train a dataset of over 4000 binary compounds which also includes many rare compounds, it is important that we can obtain features for all of them. We have considered only the basic intrinsic properties like electronegativity, molecular weight, atomic fraction and the group of the constituent elements in the periodic table. The crystal structure of the compounds is a viable feature for band gap prediction, but unfortunately there is no reliable dataset available for binary compounds of all sorts.

A. Electronegativity Difference

We find an interesting correlation between the electronegativity difference of the constituent elements and the band gap of the compound.

This is quite intuitive: the electronegativity is the power of an atom to attract the electrons towards itself, hence it is difficult for the electrons in the valance band to move into the conduction band when the electronegativity of the compound is high.

As the electronegativity difference increases, the corresponding band gaps also increases [2].

The plot of electronegativity against band gap is represented in the graph below:

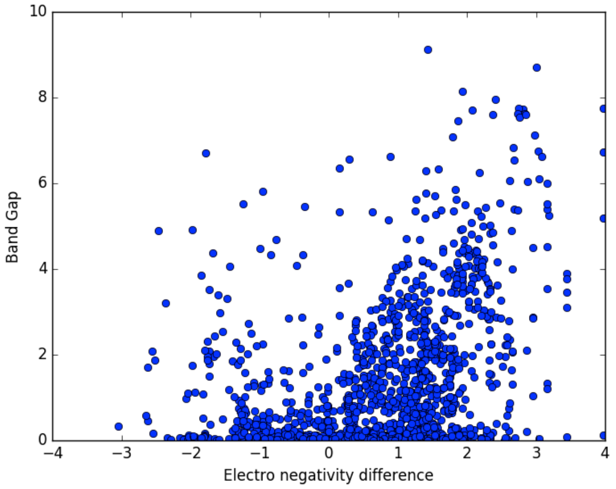

Fig. 1 Band gap vs Electronegativity difference

B. Molecular Weight

The molecular weight of the compound also contributes to the band gap. As the molecular weight increases, the number orbitals increases, and thus the force attraction on the valance shells is relatively smaller. Some cases (polymers) exhibit the opposite, and this is explained due to their alternating single-double bond structure, which gives rise to their semiconductor properties [3].

The plot of molecular weight against band gap is represented in the graph below:

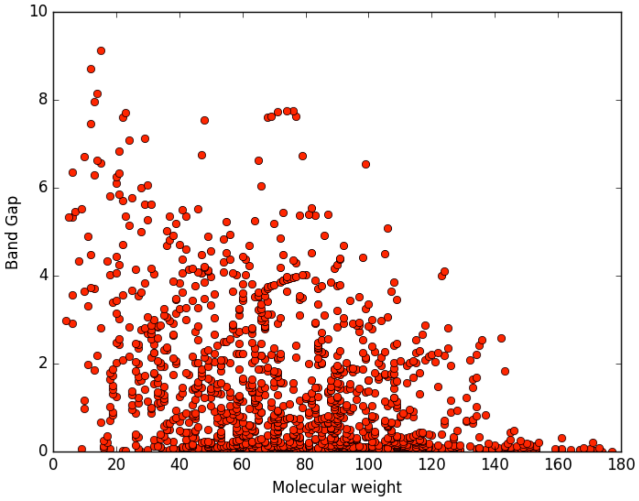

Fig. 2 Band gap vs Molecular Weight

C. Atomic Fraction

There is no clear correlation between the atomic fraction and the band gap energies. But when used in combination with the other features, the results obtained are more accurate.

The plot of atomic fraction against band gap is represented in the graph below:

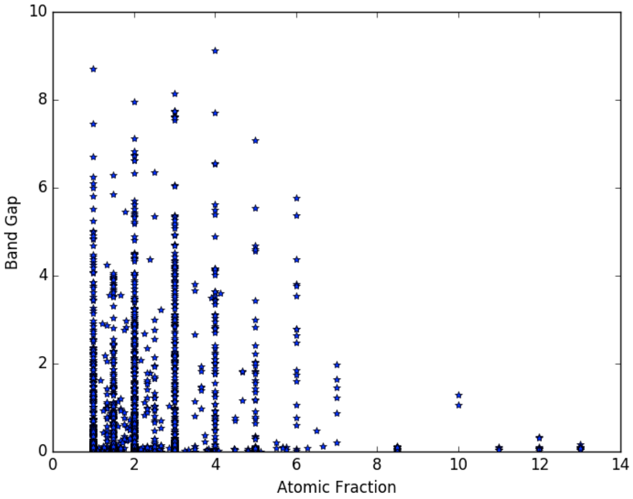

Fig. 3 Band gap vs Atomic Fraction

D. Periodic Group

The band gap decreases as we go down the group. This trend can be explained: the band gap is related to the energy splitting between bonding and anti bonding orbitals, and this difference decreases (and bonds become weaker) as the principal quantum number increases [4].

IV. Dataset Preparation

The data regarding the binary system and its bandage energy is obtained from an REST API of https://materialsproject.org [5]. The obtained subset data has 4096 rows of different binary systems. The data of binary compound systems and its band gap energy is collected in a csv format for further processing and usage. The generated csv file contains 4096 binary compound systems and its band gap energy.

To derive the features for each single binary compound, Pymatgen is used. Pymatgen (Python Materials Genomics) is a robust, open-source Python library for materials analysis.

The data is vectorized for easier processing. Python's numpy is used to read the csv data file for analysis and pre-processing. Before vectorizing the data, the dataset is cleaned by removing fields containing null values or noisy data.

V. Models

Before building machine learning models for prediction, we found the MASE (Mean Absolute Scaled Error) of a model which always predicts the average band gap energy of the 4096 compounds as the required band gap is 1.077 eV. Thus any model build must preform far better than the average prediction model.

A. Linear Ridge Regression

A Linear Ridge Regression model was developed using the 4 physical features: molecular weight, atomic fraction, electronegativity and group number. After training it with

90% of the dataset and using remaining 10% of the dataset for testing it, the MASE of the model is 0.8 eV.

B. Non Linear Random Forest

Random Forest is an ensemble method that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

Using the same features for non linear random forest, we achieved a MASE of 0.265 eV.

VI. Results

TABLE I

PREDICTED BAND GAP VALUES COMPARED WITH THE ACTUAL VALUES

Binary Compound	**Band Gap**		
	Ridge Regression	**Random Forest**	**Actual**
LiH	2.18 eV	2.62	2.981
HF	5.91 eV	6.36	6.781
H_3N	3.52	3.96	4.32
SiC	1.34	1.85	2.023
CO_2	5.92	6.38	6.633
BeS	2.62	2.85	3.143
Mg3N2	1.28	1.53	1.68
Li2Te	1.73	2.24	2.498

VII. Conclusions

The results obtained from each model is very interesting and we can learn a lot about he features and their variance from the error rates obtained.

The above table lists the predicted band gap of 8 binary compounds of varying band gap ranges.

Linear Ridge regression gives surprisingly good results, given that the features used are quite basic. We also have to consider the possibility of over fitting. The dataset contains a mere 4096 entries of binary compounds, and linear regression without any kernel could cause over fitting of the weights. It would be interesting to see how it would perform when a new binary compound is fed in, containing a element that is not present in the data set.

Linear Ridge regression shows a MASE of 0.8 eV. It appears that five degrees of freedom (four coefficients plus a constant) are too few to reasonably model band gaps with a linear model.

The Random Forest being an ensemble method performs better than the regression models. We observe that a random forest-based approach outperforms linear ridge regression model. The random forest trained on the physical features is actually decent at estimating band gaps of materials. While a single decision tree is often a poor classifier, a collection of many decision trees trained on different subsets of data can be very powerful for modeling data. Random forests have a number of tuning parameters but here we highlight only the number of trees in the forest; we chose 50 for computational expediency. Our code will thus constructed 50 independent decision trees and average the band gap predictions from each of them.

The MASE of the Random Forest model is 0.265 eV. This result quite comparable to the research done [1] by Joowi Lee et al., although our model is not just constrained to AX binary compounds. This error can easily be diminished by the use of more features. Features like crystal structure and co-efficient of conductivity would definitely improve the accuracy of the model.

However, predicting the band gap to 2 decimal points accuracy is a complicated task. Multiple researches [6] suggest and point out that band gaps when calculated at different temperatures, pressures, methods differ by up to +-1eV. The next step in this research would be to model this as a classification problem, and predict a range instead of a single value as the result.

References

[1] Prediction model of band-gap for AX binary compounds by combination of density functional theory calculations and machine learning techniques Joohwi Lee1, Atsuto Seko, Kazuki Shitara, Isao Tanaka1, Department of Materials Science and Engineering, Kyoto University, Kyoto, 606-8501, Japan Elements Strategy Initiative for Structure Materials (ESISM), Kyoto University, Kyoto 606-8501, Japan Center for Materials Research by Information Integration, National Institute for Materials Science (NIMS), Tsukuba 305-0047, Japan Nanostructures Research Laboratory, Japan Fine Ceramics Center, Nagoya 456-8587, Japan

[2] Kristen Dagenais1, Matthew Chamberlin, and Costel Constantin JYI | March 2013 | Vol. 25 Issue 3

[3] Small Bandgap Polymers for Organic Solar Cells (Polymer Material Development in the Last 5 Years) - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/41883422_fig3_Figure-4-Parameters-influencing-the-band-gap-E-g-molecular-weight-M-w-bond [accessed 20 May, 2016]

[4] http://chemwiki.ucdavis.edu/Textbook_Maps/Inorganic_Chemistry_Textbook_Maps/Map%3A_Inorganic_Chemistry_(Wikibook)/Chapter_10%3A_Electronic_Properties_of_Materials%3A_Superconductors_and_Semiconductors/10.4_Semiconductors%3A_band_gaps,_colors,_conductivity_and_doping#Contributors

[5] https://materialsproject.org/

[6] "Accelerating materials property predictions using machine learning"

[7] Ghanshyam Pilania, Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran & Ramamurthy Ramprasad