Supervised-learning Methods for Predicting Bandgap Energy of Promising Transparent Conductors

Yaqi Zhang (1006584288)
Department of Chemical Engineering, University of Toronto

Abstract: $(Al_xGa_yIn_z)_{2N}O_{3N}$ materials are promising for transparent conductors. This project presents a framework for predicting the bandgap energy of $(Al_xGa_yIn_z)_{2N}O_{3N}$ materials from geometrical properties and composition features, using supervised learning methods including ridge regression, neural networks, random forest regression and XGBoost regression, and resulting in satisfied accuracy.

Keywords: transparent conductors, $(Al_xGa_yIn_z)_{2N}O_{3N}$, bandgap energy, supervised learning

Introduction

Transparent conductors play a vital role and enable numerous applications in the fields of electro-optics, plasma, biosensing, medicine and "green energy". These materials characterized by high light transmission and high electrical conductivity, which are typically competing properties.[1] However, only a small number of compounds have been found to show sufficient transparency and conductivity that can be used as transparent conductive materials.

Due to the high conductivities and optical transparency over the visible range resulting from a combination of large bandgap energy, aluminum, gallium, indium sesquioxides (has the formula $(Al_xGa_yIn_z)_{2N}O_{3N}$, $x+y+z=1$) are some of the most potential transparent conductors materials. Bandgap energy is an important property for optoelectronic applications. To help improve the further discovery of transparent conductive materials, this report tries to propose and demonstrate a distinctive approach for predicting the bandgap energy using supervised learning methods by given the compounds' information including space group (a label identifying the symmetry of the material), total number of Al, Ga, In and O atoms in the unit cell, relative compositions of Al, Ga, and In (x, y, z), lattice vectors and angles.

Methods

Scheme of the project. The workflow of supervised learning of bandgap energy of transparent conductive materials are illustrated in Figure 1. In order to train the supervised learning models, a quantitative representation of the transparent conductive materials $(Al_xGa_yIn_z)_{2N}O_{3N}$ are required as input. There are 11 input features including space group (a label identifying the symmetry of the material), total number of Al, Ga, In and O atoms in the unit cell, relative compositions of Al, Ga, and In (x, y, z), lattice vectors and angles: lv1, lv2, lv3 (which are lengths given in units of angstroms (10^{-10} meters) and α, β, γ (which are angles in degrees between 0° - 360°). Dataset with 2400 representative $(Al_xGa_yIn_z)_{2N}O_{3N}$ materials are performed for the supervised learning methods.

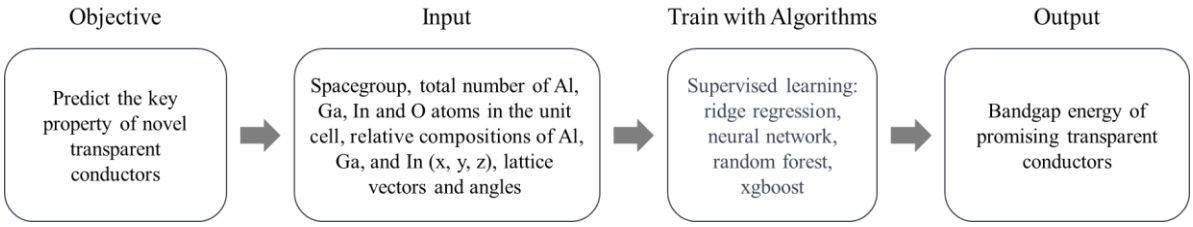

Figure 1. Workflow of the supervised learning of promising transparent conductors.

Exploratory data analysis. Before model implementation, I did some data exploration, combined with some domain material knowledge. 3 graphical figures that represent trends in the data are presented. Bandgap energy determines the electronic properties of materials and will be linked both to geometry and space group. Metals have zero bandgap energy, semiconductors have small bandgap energy, and insulators have large bandgap (may be 10 eV and higher). Figure 2 shows the distribution of bandgap energy which are in the range of about 5 eV. The peak value is about 1 eV and values above 5 eV are rare.

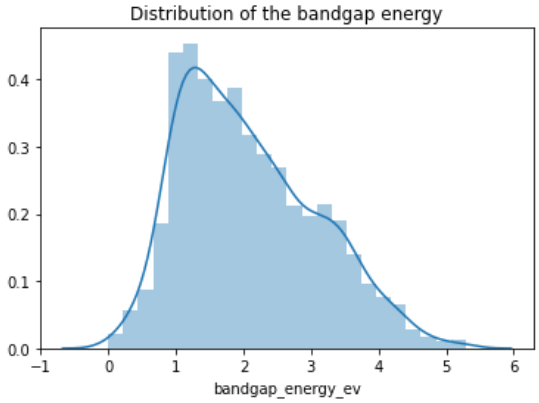

Figure 2. The distribution of bandgap energy

Space group is a geometrical feature which contains symmetry properties of the material. It only shows 6 different values in this dataset. Figure 3 gives some information about space groups. As the number becomes higher, space groups have a higher symmetry. Figure 4 visualize feature interactions. The number in the figure is Pearson correlation coefficient which measures linear correlation between two variables. The heatmap only shows feature with absolute value of Pearson correlation coefficient larger than 0.15 (correlation between the feature and target).

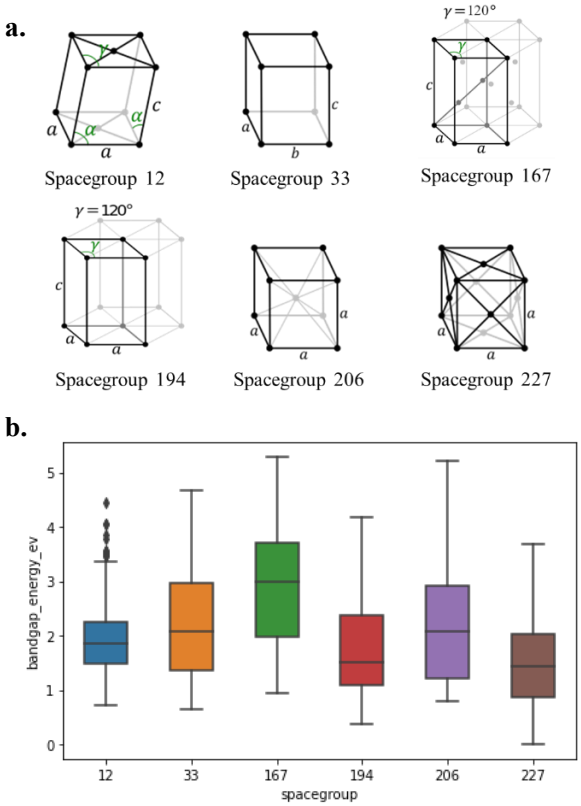

Figure 3. a. The geometrical structure of the showing space group in this dataset. **b.** Boxplot made by data grouped by space group feature.

Overall, the correlation between an individual feature and the target value is not quite significant. From the heatmap, bandgap energy correlates relatively stronger with the percentage of Al atoms and shows some anti-correlation with the percentage of In atoms. Since the feature number is small, all the 11 features will be considered as input to the models.

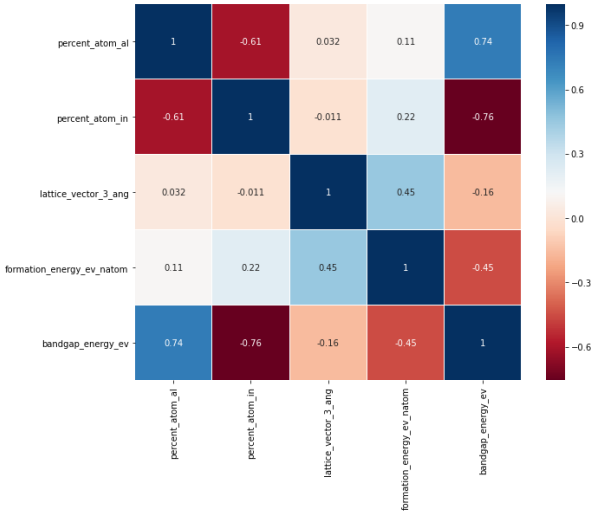

Figure 4. Heatmap with features correlation

Processing and organizing data. First, Nan was checked. There is no nan or infinity in the dataset. The input data features have different units and need to be normalized so that quantities with large values will not overwhelm the models. In this project, standard score normalization was used to weigh all the features equally, which subtracts the mean of the feature and divide by its standard deviation ($\frac{X-\mu}{\sigma}$). Then train_test_split function from SKLearn was used to obtain 80% training set and 20% test set.

Model implementation and model tuning. There are some related works from published papers giving me confidence in modeling. In 2004, S Yang, *et al*.[2] using linear regression models to predict conjugated polymers' bandgap. In 2013, Sajeev, R. *et al*.[3] compared many algorithms in predicting whether a molecule could be a semiconductor molecule and random forest got highest accuracy.

In my project, four kinds of supervised learning methods were implemented for predicting the bandgap energy of the given transparent conductors, including ridge regression, neural network regression, random forest regression and XGBoost regression. Implemented each of these four algorithms on the training data using 5-fold cross-validation and used the average scores of folds to evaluate model performance. For ridge regression, the L2 norm term is weighted by the regularization parameter alpha. Here, ridge regression models were tuned when alphas = 0.5, 1, 1.5, 2, 2.5, 3. For neural networks, sklearn's MLPRegressor was used to construct the models. Activation function and hidden layer sizes are two important hyperparameters. The hidden layer sizes were set to be (5,5,5), (5,5), (10,10,10), (10,10). Tanh and Relu function were tried for above hidden layers. For random forest regression, n_estimators represent the maximum number of iterations. If this number is too small, the model is easy to underfit, and if n_estimators is too large, model may overfit. In this project, "n_estimators" were chosen from 30 to 210 and incremented by 30. Another hyperparameter oob_score was set to True so that the model will use out-of-bag samples to estimate the R^2 on unseen data, which can help decrease the probability of overfitting. Lastly, the XGBoost model (stands for extreme gradient boosting) is an implementation of gradient boosted decision trees designed for speed and performance, which is similar to random forest regressor. The values of n_estimators were tuned from 30 to 150 and incremented by 30.

Model testing. After model tuning, four optimal models were selected. Applied these models to training and test sets separately. And calculated the mean absolute error, R-squared value, root mean squared error to measure model performance (see the code for specific calculation methods).

Results and Discussion

Table 1 shows comparisons for evaluation results in training and test sets of the four optimal models. On the training set, random forest and XGBoost models performed lower error and higher R square values, showing better performance. Compare the evaluation on training and test set, it is easy to find that these two tree-based models are overfitting.

Considered the overall fit of the models, the neural network model has better generalization ability.

Table 1. Evaluation results of optimal models*

		MAE	R^2	RMSE
Ridge	Train	0.3187	0.8039	0.4431
	Test	0.3308	0.8254	0.4725
Neural Networks	Train	0.1442	0.9737	0.2183
	Test	0.1526	0.9603	0.2366
Random Forest	Train	0.0750	0.9685	0.1231
	Test	0.1423	0.9332	0.2322
XGBoost	Train	0.0849	0.9632	0.1294
	Test	0.1370	0.9283	0.2236

* Optimal hyperparameters: Ridge: alpha=2.5; Neural Networks: hidden layer sizes = (10,10,10), activation = 'relu'; Random Forest: n_estimators=120; XGBoost: n estimators = 60.

In order to further illustrate the model performance in more detail, Figure 5. shows scatter plots of test predictions vs. true values. Generally, all four optimal models have high accuracy in predicting bandgap energy on the test set. The ridge model that erroneously predicted some small bandgap energy to be larger, had slightly bigger error than other models.

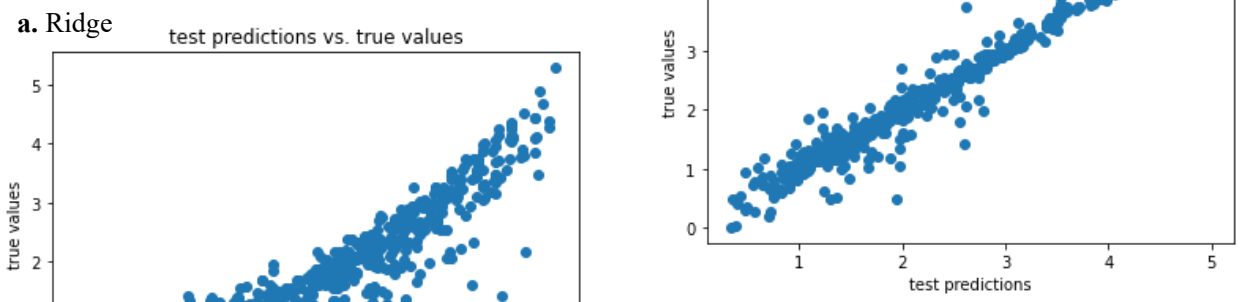

Figure 5. Test predictions vs. true values

In summary, these supervised learning models succeeded in predicting bandgap energy of promising transparent conductors. Next step of this project may try to improve model performance by using PCA as feature selection technique, GridSearchCV from sklearn.model_selection to perform better

hyperparameter optimization, model stacking to combine many models.

There are some other physical and chemical property related to performance of transparent conductors.[1] For the further applications of discovering transparent conductors, not only the bandgap energy, also the formation energy (an important indicator of the stability of a material) could be considered. Future extension of the project could attempt to predict the formation energy and construct a scheme that considers these energy properties together for exploring transparent conductors.

Acknowledgements

The dataset is from Kaggle's competition titled "Nomad 2018 Predicting Transparent Conductors". [4]

Thanks Professor Chandra Singh and teaching assistant Gurjot Dhaliwal for the hard work in MSE1065. This course is highly recommended.

References

(1) Gordon, R. G. (2000). Criteria for choosing transparent conductors. MRS bulletin, 25(8), 52-57.

(2) Yang, Shujiang, Pavel Olishevski, and Miklos Kertesz. "Bandgap calculations for conjugated polymers." Synthetic Metals 141.1-2 (2004): 171-177.

(3) Sajeev, R., Athira, R. S., Nufail, M., Raj, K. J., Rakhila, M., Nair, S. M., ... & Manuel, A. T. (2013). Computational predictive models for organic semiconductors. Journal of Computational Electronics, 12(4), 790-795.

(4) Predicting Transparent Conductors "https://www.kaggle.com/c/nomad2018-predict-transparent-conductors/data"