Capstone Project 1 Final Report: Predicting the Key Properties of Materials Using Machine Learning

Meng Zhao

1. Research Contexts, Motivations and Goals

Machine learning and data analytics can accelerate materials innovation. Among the various categories of the materials, transparent conductors are an important class of compounds that are both electrically conductive and having a low absorption in the visible range. Transparent conductors have a wide variety of applications including solar cells, flat-panel displays, touch-screens, and energy-conserving windows due to their ability to reflect thermal infrared heat. Currently, only a limited number of compounds meet the criteria of high conductivity and low visible light adsorption displayed by transparent conductors. Conventional high-throughput theoretical calculations and experiments could provide guidelines for identifying new transparent conductors. However, they are computationally expensive and time-consuming, due to a large materials searching space generated by the numerous possible compositions and configurations. Therefore, building data-driven models to accurately predict materials stabilities and light adsorption properties becomes an alternative and helps with an efficient search on promising transparent conductors.

The goal of this project is to accurately predict two chemical properties: formation energy and bandgap. Therefore, two regression models are needed as two targets are independent and the values of both targets are continuous. Three regression algorithms will be used to train the model and their performances are benchmarked. The prediction accuracy is evaluated by both root mean squared logarithmic error (RMSLE) and mean absolute error (MAE), where the lower both the values the better the prediction accuracies. The RMSLE is calculated using the equation below:

$$RMSLE = \sqrt{(1/n)\sum_{i=1}^{n}(log(pi - corr + 1) - log(ai - corr + 1))^2}$$

Where n is the total number of observations for the test data, pi is the predicted value, ai is the actual value, *corr* is the minimum among actual and predict test values. As the dataset contains negative values, *corr* term is making sure log function always conducting on positive values.

The structural-properties relationships and any promising transparent conductor candidates found from the model are valuable to the companies manufacturing electronic devices. They can use these information as a starting point/shortcut to conduct further advanced experiments,

performance tests and manufacturing scale-up , which tremendously reduces their time of trial-and-error.

2. Data collection and preprocessing

Four data sources are used for this project: the first is provided by kaggle competitions named Nomad 2018 Predicting Transparent Conductors (https://www.kaggle.com/c/nomad2018-predict-transparent-conductors). Data is acquired by downloading the *train.csv* file which contains a set of materials with target values (the bandgaps and formation energies) and their corresponding spatial information from *geometry.xyz* files saved under the directory of {train}/{id}/. This dataset has been confirmed having 2400 entries and 14 columns in total without missing values. Each entry represents a material. Within these 14 columns, 12 of them are materials features and the rest of two columns, formation_energy_ev_natom and bandgap_energy_ev, are the targets. Although the Cartesian coordinates (positions) for each atom and a list of elements in a chemical system can be derived from these 12 materials features, a more efficient way of retrieving comprehensive structural information is to transform the *geometry.xyz* files into *pymatgen.core.structure.Structure* objects. The structural information stored in the *pymatgen.core.structure.Structure* object can further be utilized to generate structure-based fingerprints ready to train the machine learning model. The processed Kaggle dataset contains three features: formula, structure, space group and two target values of formation energy and bandgap energy.

The Kaggle repository only contains oxides containing Al, Ga and In. To generalize and extend the chemical space, the other three data sources are included. The second data source from Materials Project (https://materialsproject.org/) database is employed. Materials Project database is an open-source materials genome database containing both experimental and theoretical information of materials throughout the periodic table. A subset of 1,284 entries representing oxides of Al, Ga, In, Mo, Zr ,W, Ta, Sb, Zn, Sn, Ti, Ce, Fe, Co, Cu, Ni, Mn, Pt, Pd, Ir, Ru are collected using Materials Project API (Application programming interface) called MPRester and a HTTP (HyperText Transfer Protocol) Python library called request. Its API responses are JSON strings containing target values and structural information in a CIF format. Crystallographic Information File (CIF) is the internationally agreed standard file format for information exchange in crystallography. Similar to the way of processing the Kaggle dataset, lattice vector, coordinates and a list of elements, are extracted. The data extraction is realized by the function read_pymatgen_cif(). This function requires lattice vectors, coordinates and a list of elements as inputs and produces *pymatgen.core.structure.Structure* objects as a output for each material entry. The processed Materials Project dataset has the same columns as those in the

processed Kaggle dataset, which includes three features of formula, structure, space group and two target values.

The third and fourth parts of data are the subsets in OQMD (The Open Quantum Materials Database) and ICSD (Inorganic Crystal Structure Database), from which ~18,000 and 9,357 entries are collected, respectively. Similar data preprocessing approach is applied to generate the processed data with the same features. The complete dataset is the concatenation of all four datasets, in which over ~31,000 entries are available.

3. Data exploratory analysis

A boxplot of two targets in the complete dataset is generated for initial data exploration. Figure 1 demonstrates that formation energies vary from - 4 to ~6 eV/atom, and bandgaps range from 0 to 6 eV. Although both have a significant amount of data lie in the outlier region, those data points may not actually be outliers as both targets have wide ranges of formation energies and band gaps due to the materials diversity. The data points lie in the outlier region are going to be carefully examined before training machine learning models.

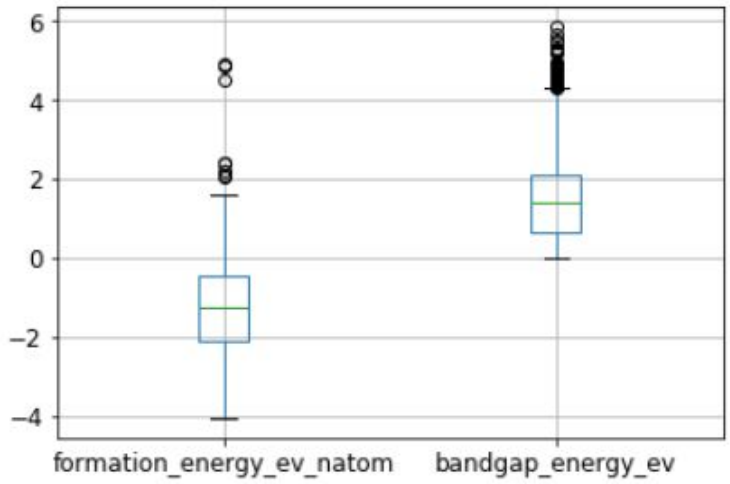


Figure 1. Boxplot for two targets in the complete dataset

Summary statistics in Table 1 has shown detailed information of the complete dataset. It has 31,543 entries in which their formation energies ranging from -4.05 to 4.91 eV/atom with the standard deviation of 1.03 eV/atom. The negative formation energies indicate that the formation of a compound is exothermic (the amount of energy took for breaking bonds is less than that released when making the bonds) and positive formation energies indicate that the formation of a compound is endothermic (the amount of energy it takes to break bonds is greater than the amount of energy that is released when making the bonds). Formation energy is one of the most important properties of a compound as is directly relates to its stability. The more negative the

formation energy, the more stable the compound is likely to be. Bandgaps in complete dataset range from 0 to 5.87 eV with the standard deviation of 0.87 eV. By definition, bandgap is the energy difference between the valence band and the conduction band of a solid material where no electron state can exist and only positive bandgaps are physically meaningful. Therefore the bandgap between 0 to 5.87 eV in the complete dataset indicates that the collected materials are either conductors or semiconductors.

	spacegroup	formation_energy_ev_natom	bandgap_energy_ev
count	31543.000000	31543.000000	31543.000000
mean	107.592081	-1.298717	1.440845
std	81.400168	1.029570	0.866672
min	1.000000	-4.053535	0.000000
25%	15.000000	-2.107609	0.665000
50%	99.000000	-1.257404	1.382000
75%	186.000000	-0.459231	2.132000
max	230.000000	4.910340	5.853700

Table 1. Summary statistics on the complete materials dataset

To gain an in-depth understanding of the entire dataset, summary statistics analyses have been conducted on each data sources separately. It has been found that dataset from kaggle repository has positive formation energies from 0 to 0.66 eV/atom with the standard deviation of 0.10 eV/atom; positive bandgaps are from 0 to 5.29 eV with the mean of 2.08 eV and the standard deviation of 1.0 eV. Formation energies from materials project include negative values and range from -3.95 to 2.4 eV/atom with the standard deviation of 0.88 eV/atom. Its bandgaps are from 0 to 5.85 eV with the standard deviation of 1.16 eV. The interesting observation here is that dataset from materials project has ~350/827(~42%) entries with bandgap close to 0 eV, indicating almost a half of the compounds collected from materials projects are conductors.

Dataset from ICSD (Inorganic Crystal Structure Database) has a large range of formation energies (-4.01 to 4.49 eV/atom with the standard deviation of 0.88 eV/atom) and bandgaps from 0.12 eV up to 3 eV with the standard deviation of 0.78 eV. The clean cutoff of bandgap at 3 eV is due to fact that insulators or semiconductors with large bandgaps are intentionally neglected, to align with the project goal of targeting promising photocatalysts which become catalytically active when absorbing lights. This bandgap cutoff also applies to data from OQMD (Open Quantum Materials Database), resulting in its bandgaps ranging from 1.33 eV to 3 eV.

Distributions of formation energies and bandgap energies are shown in Figure 2 for the complete dataset as well as each data resource separately.

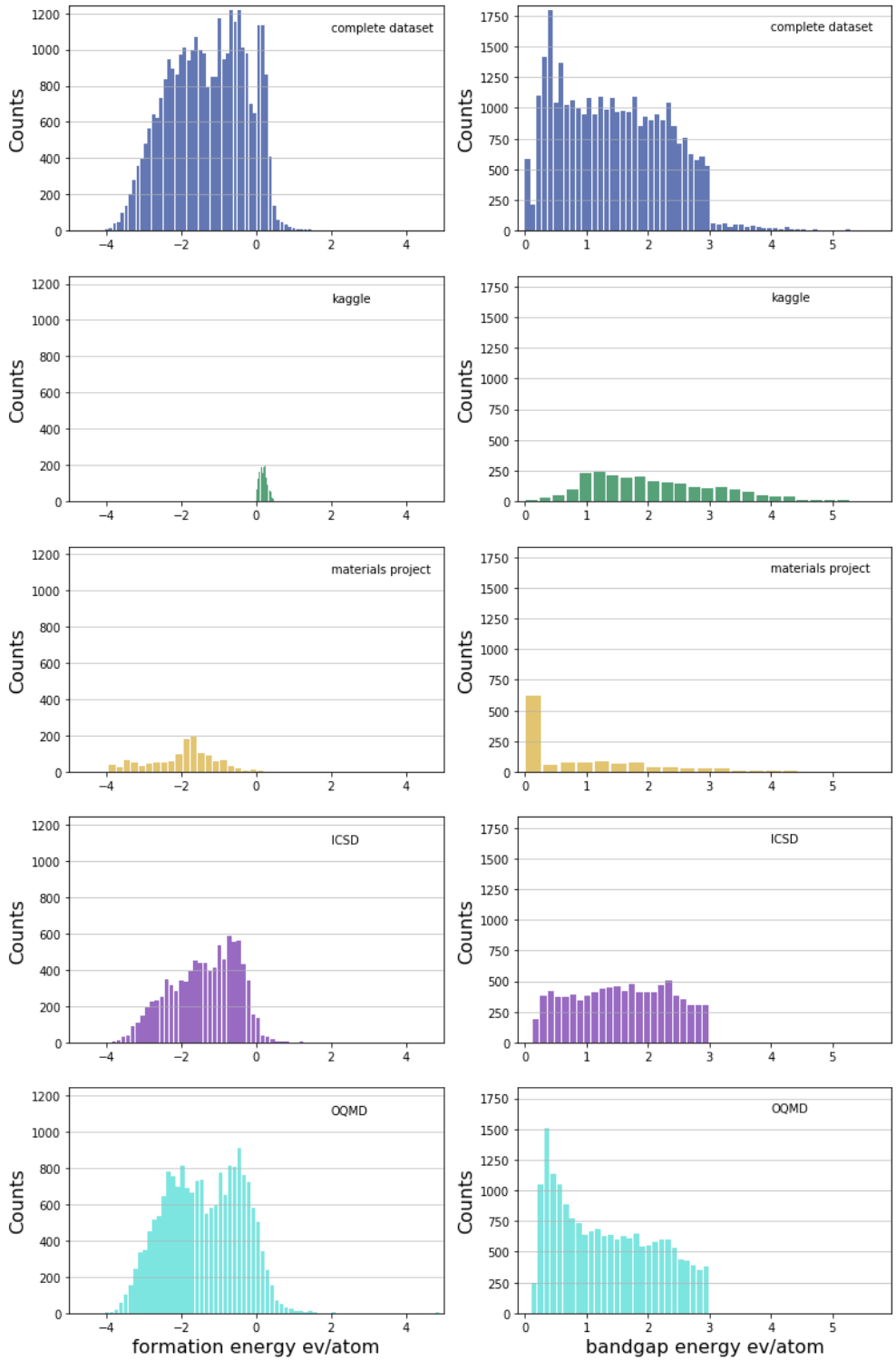


Figure 2. Distributions of formation energies and bandgap energies for the complete dataset as well as each data resource separately

Anomalies have been detected in the dataset from OQMD. While the minimum formation energy of this dataset seems normal (- 4.05 eV/atom), the maximum formation energy shoots up to 384.23 eV/atom. Based on chemical intuitions, formation energy larger than 5 eV/atom is generally unreasonable. Therefore a scatter plot of formation energy vs. bandgap from OQMD is generated to locate the anomalies. The scatter plot in Figure 3 clearly shows two anomalies. The detailed information of these two anomalies has been shown in Table 2 in which one anomaly is $YbHfO_3$ having the formation energy of 294.55 eV/atom and the other is $YbZrO_3$ with formation energy of 384.23 eV/atom. These two anomalies are dropped from the complete datasets to improve data quality.

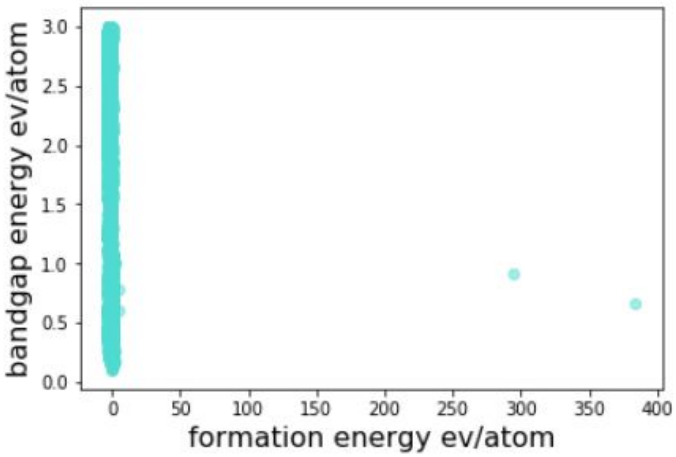


Figure 3. The scatter plot of formation energy vs. bandgap from OQMD

	formula	structure	spacegroup	formation_energy_ev_natom	bandgap_energy_ev
12289	YbHfO3	Full Formula (Yb1 Hf1 O3)\nReduced Formula: Yb...	99	294.545550	0.918
12372	YbZrO3	Full Formula (Yb1 Zr1 O3)\nReduced Formula: Yb...	99	384.229115	0.668

Table 2. Details of two anomalies identified

Composition-based and structure-based features(~140) are generated. A selective set of features (spacegroup, mean Electronegativity, density, vpa, packing fraction,mean CovalentRadius) are examined for correlation. The correlation heatmap in Figure 4 obviously indicates that mean CovalentRadius and mean Electronegativity are highly correlated. The scatter plot in Figure 5 confirms the negative correlations between these two features. In other words, as the mean electronegativity increases, mean covalent radius decreases.

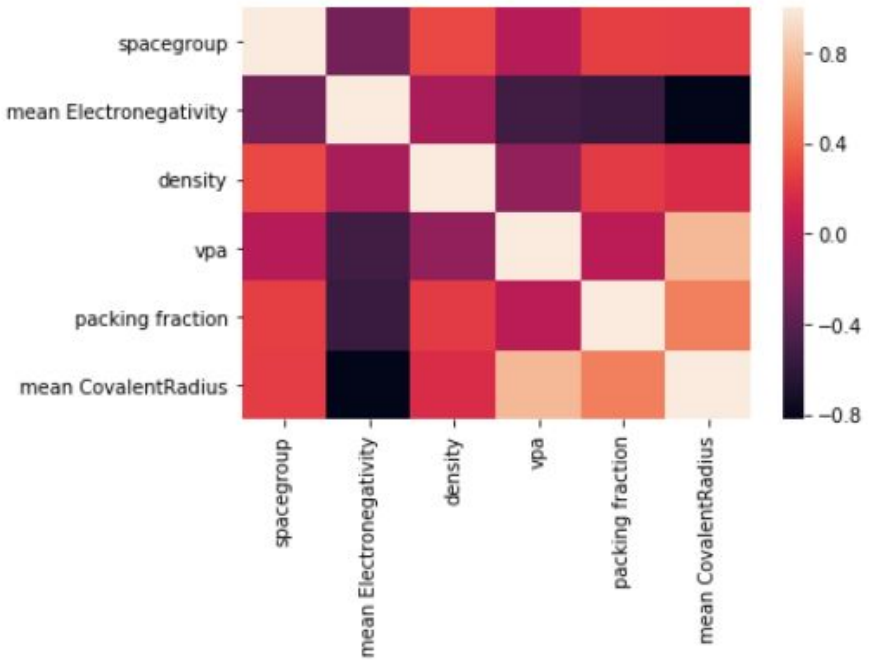


Figure 4. The correlation heatmap for selective features of spacegroup, mean Electronegativity, density, vpa, packing fraction,mean CovalentRadius

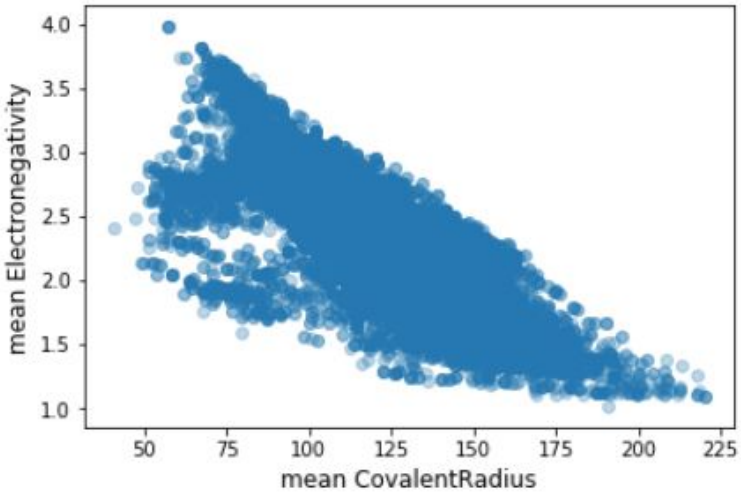


Figure 5. The scatter plot between mean CovalentRadius and mean electronegativity

chi-squared test is conducted to determine if there is a statistically significant correlation between mean CovalentRadius and mean electronegativity. The null hypothesis is that two variables are independent. The computed p value of the chi-squared test is 0 which suggests to reject the null hypothesis. In other words, mean CovalentRadius and mean electronegativity are statistically correlated. The chi-squared test also applies to check the correlations between other variable pairs. Statistics analysis above has shown that two targets in the complete dataset have a

wide range of values, some of the features generated from materials structural information have statistically significant correlations, which might lead to the use of PCA for feature reduction.

4. In-depth Machine Learning Analysis

The overall processed materials dataset ready for machine learning analysis has 31,417 entries. We build two supervised regression models for predicting formation energy and bandgap. For both predictive machine learning models, we split the entire dataset into 70% training set and 30% testing set. We trained machine learning models using scaling, and three machine learning algorithms: random forest regressor, XGboost regressor and Gaussian process regressor, separately. We also used approach of PCA with linear regression to represent the baseline of performance of the machine learning model. For each algorithm, RandomizedSearchCV is employed for optimal hyper-parameter search, so that the predictions have the lowest root mean squared logarithmic error (RMSLE) and mean absolute error (MAE). The design of experiments (DOE) can be seen in Figure 6.

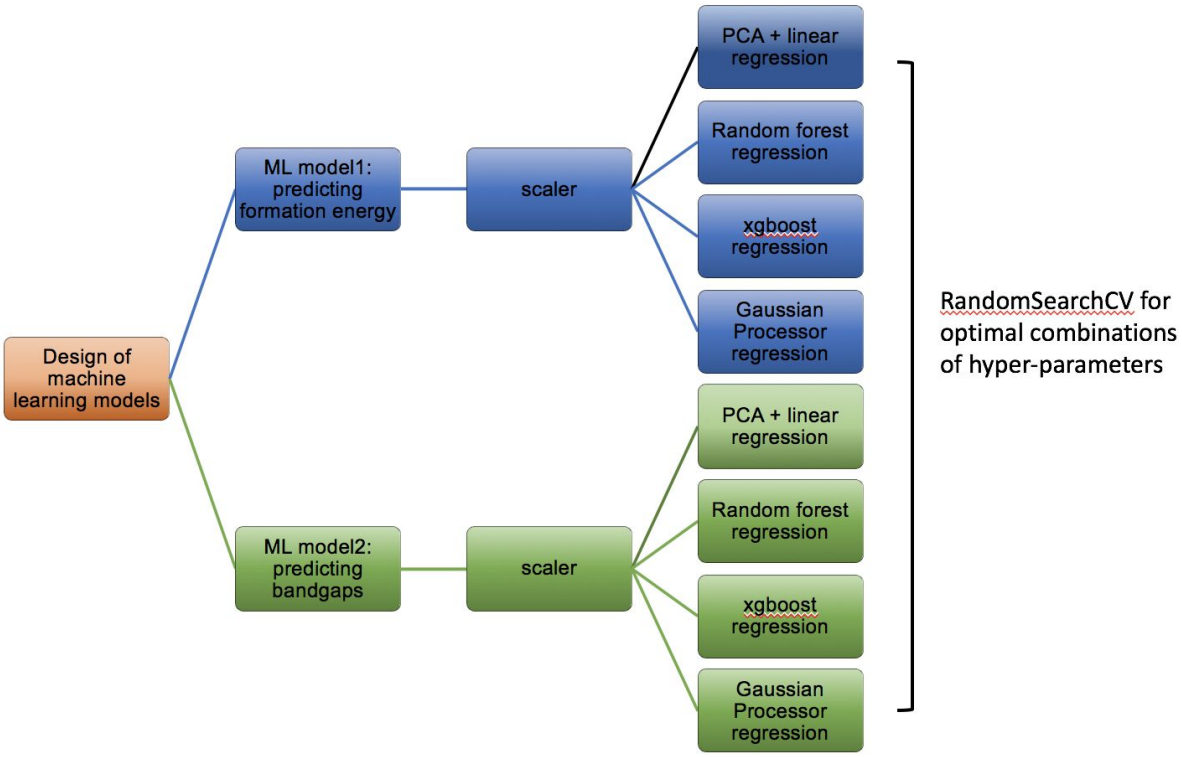


Figure 6. The Design of the experiments for building the predictive models of predicting formation energies and bandgaps

4.1 Machine Learning Algorithm 0: Linear regression with PCA

Linear regression with principal component analysis (PCA) was used to train the machine learning and its performance serves as our model performance baseline. We employed 5-fold RandomizedSearchCV to search for the optimal number of components for PCA.

```
k_fold = 5
n_iter_search = 50

pca = PCA()
scaler= StandardScaler()
linear_regr = LinearRegression()

n_components = [int(x) for x in np.linspace(10, 140, num = 10)]

pipe_linear = make_pipeline(scaler,pca,linear_regr)

randomsearch_linear =  RandomizedSearchCV(pipe_linear,
                                dict(pca__n_components=n_components),
                                n_iter = n_iter_search,
                                cv = k_fold,
                                random_state = 42,
                                verbose = 1,
                                n_jobs= -1)
```

The best number of principal components is 140, which means we used entire features to predict the model. The trained linear regression provides the cross validation score on test data of 0.837, with MAE of 0.32 eV/atom and RMSLE of 0.12 eV/atom. The predicted Gibbs formation energy vs. the actual Gibbs formation energy is shown in Figure 7.

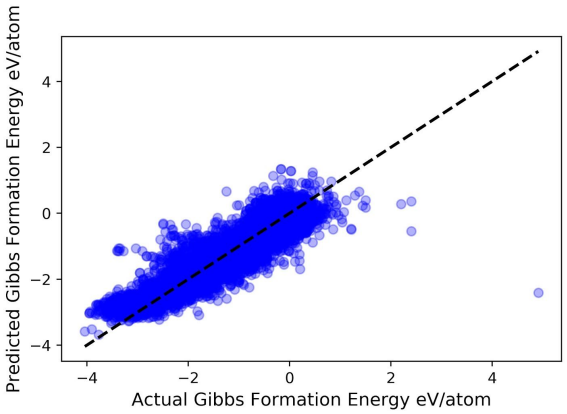


Figure 7. The predicted Gibbs formation energy vs. the actual Gibbs formation energy using linear regression

Using linear regression gives biased bandgap prediction, with MAE of 0.55 eV and RMSLE of 0.31 eV, which indicate the inadequacy of the model. The predicted bandgap vs. the actual bandgap is shown in Figure 8.

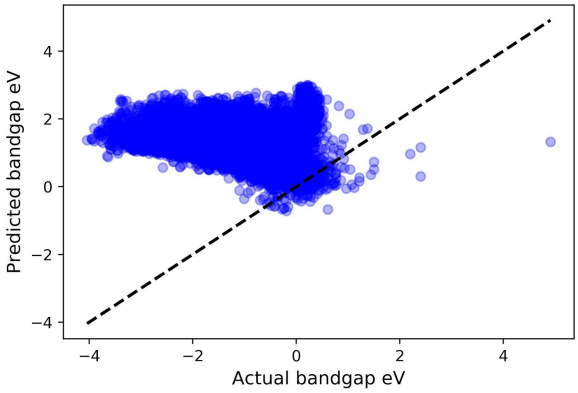


Figure 8. The predicted bandgap vs. the actual bandgap using linear regression

4.2 Machine Learning Algorithm 1: Random Forest Regressor

Although the performance of linear regression on predicting formation energies is somewhat decent, its prediction on bandgaps is quite biased. Therefore, more sophisticated and powerful machine learning models are needed. One of state-of-art machine learning algorithms is random forest. It is an ensemble technique with the capability of performing both regression and classification tasks. It combines multiple decision trees using bootstrap aggregation to determine the final output rather than relying on individual decision trees. Its advantages of overfitting reduction and less variance make random forest a more accurate model.

The hyper-parameters we tuned for random forest regressor are n_estimators, max_depth, max_features, max_depth, min_samples_split, min_samples_leaf, and bootstrap. The parameter grid can be seen below:

```
# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 10, stop = 200, num = 10)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
```

```
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]

# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
```

We conducted randomized 5-fold cross-validation search with 50 iterations, which leads to 250 searches in total.

```
k_fold = 5
n_iter_search = 50
rf = RandomForestRegressor()

randomsearch_rf = RandomizedSearchCV(estimator = rf,
                                     param_distributions = random_grid,
                                     n_iter = n_iter_search,
                                       verbose=2,
                                       random_state=42,
                                       cv = k_fold,
                                       n_jobs= -1)
```

The optimal hyper-parameters of predicting formation energies of materials are

```
RandomForestRegressor(bootstrap=False, criterion='mse', max_depth=70,
          max_features='sqrt', max_leaf_nodes=None,
          min_impurity_decrease=0.0, min_impurity_split=None,
          min_samples_leaf=1, min_samples_split=5,
          min_weight_fraction_leaf=0.0, n_estimators=157, n_jobs=None,
          oob_score=False, random_state=None, verbose=0, warm_start=False)
```

which grives the cross-validation accuracy score on test data of 0.981, MAE of 0.07 eV/atom and RMSLE of 0.04 eV/atom.

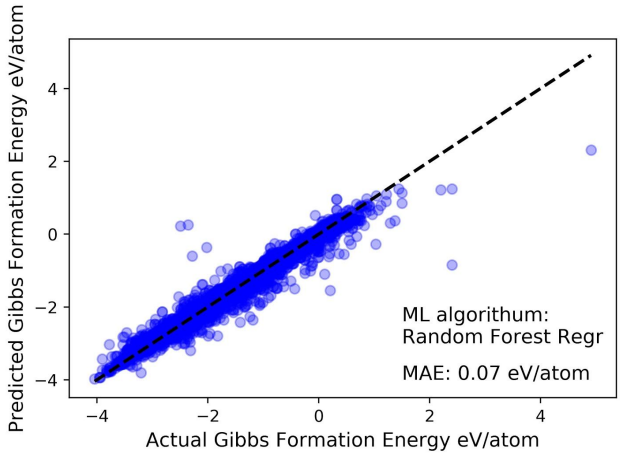


Figure 9. The predicted Gibbs formation energy vs. the actual Gibbs formation energy using random forest regressor

The predicted Gibbs formation energy vs. the actual Gibbs formation energy is shown in Figure 9, with the predicted values nicely align with the actual values. Similar approach has been applied for tuning the hyper-parameters of random forest regressor when predicting band gaps and the optimal hyper-parameters are the following:

```
RandomForestRegressor(bootstrap=False, criterion='mse', max_depth=30,
        max_features='sqrt', max_leaf_nodes=None,
        min_impurity_decrease=0.0, min_impurity_split=None,
        min_samples_leaf=1, min_samples_split=5,
        min_weight_fraction_leaf=0.0, n_estimators=200, n_jobs=None,
        oob_score=False, random_state=None, verbose=0, warm_start=False)
```

This set of optimal hyper-parameters provides 0.807 test cross validation score, MAE of 0.22 eV and RMSLE of 0.17 eV. Figure 10 shows the predicted bandgaps vs. the actual bandgaps. Random forest regressor also performances less well than it performs on predicting formation energies.

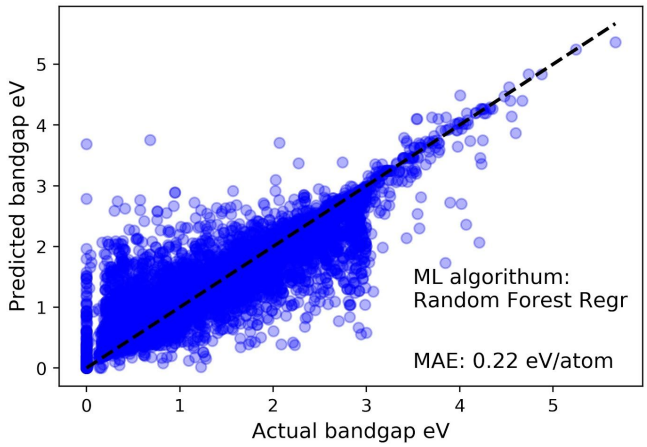


Figure 10. The predicted bandgap vs. the actual bandgap using random forest regressor

4.3 Machine Learning Algorithm 2: Xgboost Regressor

XGBoost (Extreme Gradient Boosting) belongs to a family of boosting algorithms, a sequential technique which works on the principle of an ensemble, combines a set of weak learners and delivers improved prediction accuracy, and is another popular machine learning algorithm besides random forest. It is also good for both regression and classification. It is faster than ensemble algorithms, parallelizable and usually more robust than other machine learning algorithms.

We employed the similar 5 fold cross-validation random search with 50 iterations to obtain the optimal hyper-parameters of Xgboost regressor. The hyper-parameters tuned are n_estimators, max_depth, learning_rate, gamma, colsample_bytree and subsample The parameter grid is set up as the following:

```
# Number of estimators
n_estimators = [int(x) for x in np.linspace(start = 10, stop = 200, num = 10)]
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
# learn rate
learning_rate = [float(x) for x in np.arange(0.01, 1.0, step = 0.02)]
# gamma
gamma = [float(x) for x in np.arange(0, 1.0, step = 0.1)]
# colsample_bytree
colsample_bytree = [float(x) for x in np.arange(0.2, 1.0, step = 0.2)]
# subsample
```

```
subsample = [float(x) for x in np.arange(0.2, 1.0, step = 0.2)]

# Create the random grid

k_fold = 5
n_iter_search = 50

tuned_parameters_xgb = {'n_estimators': n_estimators,
                        'max_depth': max_depth,
                        'learning_rate': learning_rate,
                        'gamma': gamma,
                        'colsample_bytree': colsample_bytree,
                        'subsample': subsample}

scaler= StandardScaler()
xgb_regr = XGBRegressor(objective="reg:linear", random_state=42)

randomsearch_xgb = RandomizedSearchCV(estimator = xgb_regr,
                                      param_distributions = tuned_parameters_xgb,
                                      n_iter = n_iter_search,
                                      cv = k_fold,
                                      random_state = 42,
                                      verbose = 1,
                                      n_jobs= -1)
```

The optimal hyper-parameters of predicting formation energies of materials give cross-validation score of 0.977, MAE(mean absolute error) of 0.10 eV/atom and RMSLE of 0.048 eV/atom. The predicted Gibbs formation energy vs. the actual Gibbs formation energy using xgboost regressor can be seen in Figure 11.

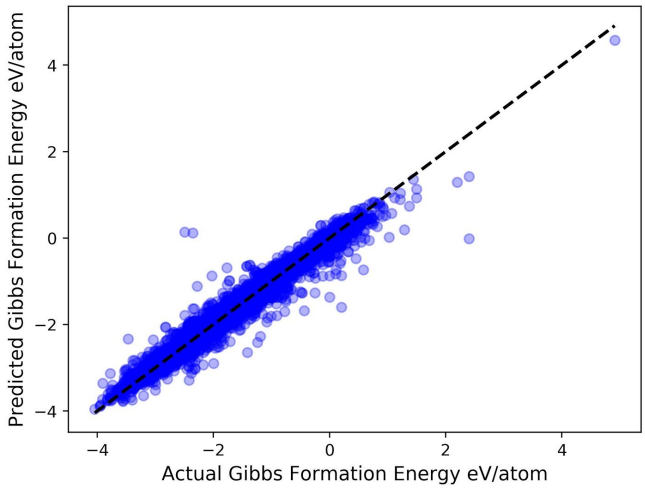


Figure 11. The predicted Gibbs formation energy vs. the actual Gibbs formation energy using xgboost regressor

Similar approach has been applied for tuning the hyper-parameters of xgboost regressor when predicting band gaps, which provides 0.801 test cross validation score, MAE of 0.26 eV and RMSLE of 0.18 eV. Figure 12 shows the predicted bandgaps vs. the actual bandgaps. Similar to random forest regressor, xgboost performances less well on predicting bandgaps than it performs on predicting formation energies.

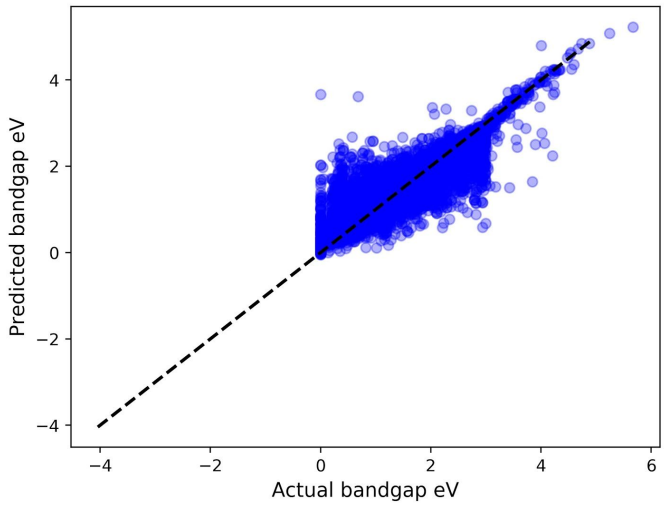


Figure 12. The predicted bandgap vs. the actual bandgap using xgboost regressor

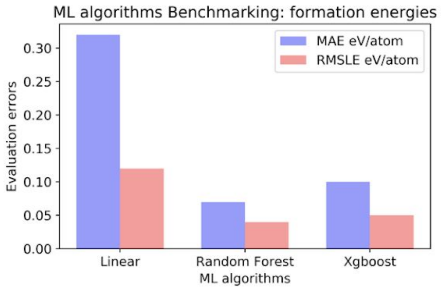
4.4 Machine Learning Algorithm 3: Gaussian Process Regressor

The initial experimental design also proposed to employ Gaussian process regressor to train the model as it provides the prediction uncertainties. However, we later realized that Gaussian process models are difficult to scale to large datasets.The basic complexity of Gaussian processes is $O(N^3)$ where N is the number of data points, which limits its application in the case where N ≈ 1000 or fewer are available. Our training data has the size of ~22,000, which is not a suitable case for using Gaussian Process regressor.

4.5 Benchmarking Performances of Three Machine Learning Models

The summarized MAE and RMSLE for predicting two targets: formation energies and bandgaps by using three machine learning regressors is shown in Table 3.

Predicting formation energies	MAE (eV/atom)	RMSLE (eV/atom)
Model 0: PCA + Linear Regression	0.32	0.12
Model 1: Random Forest Regression	0.07	0.04
Model 2: Xgboost Regression	0.10	0.05
Model 3: Gaussian Process Regression	N/A	N/A



Predicting bandgaps	MAE (eV)	RMSLE (eV)
Model 0: PCA + Linear Regression	0.55	0.31
Model 1: Random Forest Regression	0.22	0.17
Model 2: Xgboost Regression	0.26	0.18
Model 3: Gaussian Process Regression	N/A	N/A

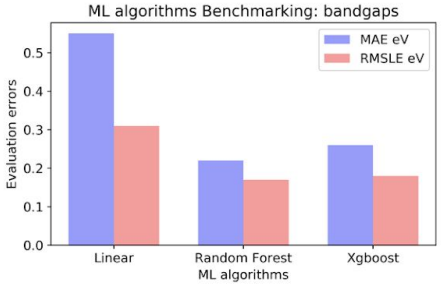


Figure 13. Benchmarking model's performances on predicting formation energies (upper) and bandgaps (lower)

Figure 13 has demonstrated that random forest regression performs slightly better than xgboost regressor for both predicting formation energies and bandgaps. Both model performance much better than the linear regression model.

4.6 Discussions on Machine Learning Predicted Formation Energies and Bandgaps

Theoretical calculations using DFT(density functional theory) provides 0.2 eV/atom and 0.5 eV prediction uncertainties predictions on formation energies and bandgaps, respectively. While its predictions on formation energies are decent, it usually underestimates bandgaps systematically as DFT only concerns ground states while computing bandgaps needs taking excited states into a consideration as well. Going beyond DFT such as using GW approximations leads to a

prediction error reduction to ~0.05 eV. However, this more sophisticated method is also computational expensive. By using machine learning techniques, we are able to get good predictions on both formation energies and bandgaps within a reasonable timeframe using less computing resources. One the other hand, we can extract chemistry knowledge from machine learning models by examining the feature importances. Figure 14 has showed the feature importance from random forest regression model when predicting formation energies. It demonstrates that materials' valence bands, ground state information, unfilled bands, electronegativities have significant impacts on formation energies. Further research can dig down into each of the important features and unravel the structure-properties relationships.

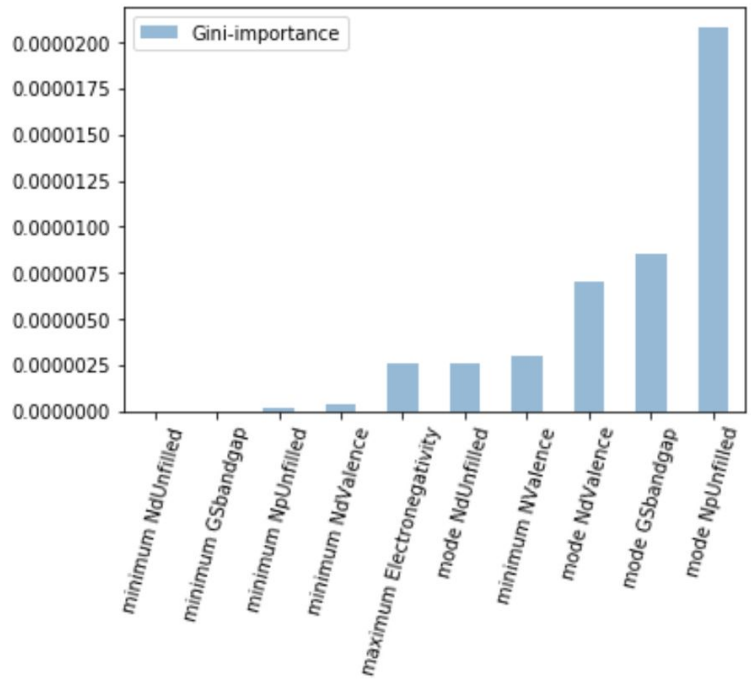


Figure 14. Feature importance from random forest regression model when predicting formation energies

5. Conclusions and Future Directions

Materials data are collected and processed from four data resources. Statistical analysis and visualizations are employed to facilitate a comprehensive overview of the entire dataset and identify the possible outliers and wrong data. We performed supervised machine learning on ~30,000 processed materials data to predict their formation energies and bandgaps, by using linear regression as a baseline, random forest regressor and xgboost regressor. Random forest regression performs slightly better than xgboost regressor for both predicting formation energies and bandgaps. Random forest regressor and xgboost regressor performance much better than the

linear regression model. Gaussian process regressor is not applicable for the datasets with its size larger than ~1000, therefore it is not suitable to be employed on our dataset.

For trained Random forest and xgboost models, we can further do analysis on its importance features, in order to unravel the structure-properties correlations of the promising materials candidates. To further improve the model performance, we could also employ a model stacking technique which ensembles multiple predictive models into a new model. By adding a 2nd-level model on top of the base model such as random forest regression or xgboost regression might produce a more accurate predictions.