Development of machine learning models for predicting energies of sodium-ion battery materials

KM Monareng[1], RR Maphanga[2,3] and PS Ntoahae[1]

[1]Department of Physics, University of Limpopo, Private Bag X 1106, Sovenga, 0727
[2]Next Generation Enterprices and Institutions, Council for Scientific and Industrial Research
[3]National Institute for Theoretical and Computational Sciences, NITheCS, Gauteng, 2000

E-mail: **mabelkmonareng@gmail.com**

Abstract. Machine learning methods have recently found applications in many areas of physics, chemistry, biology, and materials science, where large datasets are available. In this paper, machine learning regression techniques are applied to a large amount of density functional theory calculated data to develop machine learning models capable of accurately predicting the formation and total energy of sodium-ion battery cathode materials. Amongst various algorithms that were evaluated, the Bayesian ridge model was found to be the best model in predicting the formation energy with an accuracy of 0.99 and 0.01eV coefficient of determination and mean square error, respectively, and final energy with 0.98 and 0.03eV accuracy for the coefficient of determination and mean square error, respectively. The results show that the descriptors used to predict the energies have a predictive capacity with a high accuracy.

1. Introduction

The development of energy storage and conversion devices is essential to reduce the discontinuities and instability of renewable energy generation [1]. New eco-friendly energy sources are necessary in the current era, as they are expected to reduce greenhouse gas emissions and ultimately benefit human health. Over the past two decades, lithium-ion batteries (LIBs) have dominated the portable electronics industry and solid-state electrochemical research [2]. Due to the use of LIBs in portable electronics such as laptops, electric vehicles, and cell phones, lithium-ion batteries have received a lot of attention [3]. Among the most favourable characteristics of Li-ion batteries are their longer lifetime, higher energy, high efficiencies, and power densities. Despite their success, lithium-ion batteries are expensive to produce due to limited lithium resources in the Earth's crust and, in addition, large-scale energy storage is not possible with this technology.

Despite concerted efforts to develop novel materials for energy storage technologies, there is a continuous need for technologies that can push the limits on material properties [4]. Alkaline-ion batteries have developed rapidly in recent decades due to their high energy density and environmentally friendly properties [1], these batteries have gained a good reputation as alternatives to LIBs due to the high abundance of Na- and K-ions in the environment. Sodium-ion battery (SIB) technology has gained the privilege of enabling new and more demanding applications for large-scale energy storage systems than LIBs); this is due to the high abundance of sodium-ion resources present in the Earth's crust and seawater [5, 6]. Quantum mechanical methods have been shown to be successful in predicting and finding functional new materials, such as SIB materials, to replace LIB materials; however, the

calculations are computationally expensive and time-consuming. The lack of suitable electrodes and electrolyte materials has limited the development of sodium-ion batteries. In this context, the data-driven machine learning (ML) approach has recently found ways to address material discovery at a faster rate and with limited use of computational resources.

Previous studies have shown that a combination of density functional theory (DFT) and machine learning (ML) methods can accelerate the prediction of material properties and the discovery of novel materials [7, 8]. Machine learning models and algorithms are increasingly applied in battery materials research, with superior time efficiency and high accuracy in property prediction. Some examples include successful application to predict the properties of battery material properties [9-11] and discovery of new battery materials [12, 13]. In this paper, we develop machine learning models capable of accurately predicting the formation and total energy of sodium-ion battery cathode materials.

2. Methodology

Machine learning algorithms are typically expressed as a computer program that can learn from experience (E) with respect to some class of tasks (T) and the performance measure (P) [1]. Thus, ML is simply denoted as <P, T, E>. If its performance in tasks in T, measured by P, improves with experience E. It is not the purpose of this paper to discuss details on the theoretical background of machine learning methods, models, and their applications in materials science; however, more details can be found elsewhere in the literature [14]. With regard to materials science, the ML process consists of the three key steps summarized in the following subsections.

2.1. Sample construction

In this step, there are two important activities, namely data curation and feature engineering. The data are cleaned, pre-processed, while features are engineered to help improve the accuracy of property prediction. Pymatgen Materials Genomes was used to extract materials data from the Materials Project Database [15], which contains 7397 different sodium containing materials with different properties calculated using DFT. Chemical descriptors were used to construct machine learning features based on fundamental atomic properties, such as the chemical formula and atomic number. Feature vectors are then generated from details of the chemical formula. The features extracted from the Materials Project include *formation_energy_per_atom, final_energy_per_atom, energy, Fermi_energy, energy_above_hull, density, and bandgap.* To obtain the chemical name (X) of a material, an algorithm was developed to generate a set of chemical and physical descriptive attributes in which the energies (Y) were predicted. Descriptor attributes were used to predict formation and final energy in this study.

2.2. Model development

Model development is a black box that links the input data to the output data employing a set of functions, which can be either linear or non-linear based on the input data. The models were developed using a Scikit library machine learning module and a Python code. The data was divided into 70% train set and 30% test set. The tested models include the Bayesian ridge (BR), gradient boosting regressor (GBR), light gradient boosting machine (LGBM), extra trees regressor (ETR), orthogonal matching pursuit (OMP), and random forest regressor (RFR) among others. Detailed descriptions of all these models [16-18] and their application in materials science can be found elsewhere [14, 19].

2.3. Model evaluation

The model evaluation step involves evaluating and validating the performance accuracy of the developed models. A model is evaluated using different evaluation metrics to measure its performance. In this case, to evaluate the accuracy of the model, we compared the calculated DFT properties with the values predicted by the ML model employing K-fold cross validation. For the light gradient boosting machine, the hyperparameters used include the number of trees = 1000, the maximum depth = 3 and the regularization of L2 = 1 without cross validation and the number of trees = 350, maximum depth = 10

and L2 regularization = 10 with cross validation. The hyperparameter used for the Bayesian ridge model iteration = 350 without cross validation and iteration = 50 with cross validation. The following precision measures were used to evaluate the performance of the model: coefficient of determination (R^2) and mean square error (MSE).

Coefficient of determination is computed using the formula: $$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}, \quad (1)$$

where y_i, are the observed true DFT values, $\hat{y}_i$ are the ML predicted values, and $\bar{y}$ is the mean value of the *i-th* sample and the sample size of the testing set. It is recommended that an R^2 reading of 0.8 or higher is adequate for a good reading, indicating that the model is fitting well. The MSE provides an indication of the quality of the model and is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

where y_i are the observed true DFT values, $\hat{y}_i$ are the predicted ML values of the *i-th* sample and n is the number of data points. Zero mean square error shows that the model predicted 100% correct actual values; hence the model must achieve MSE closer to zero.

3. Results and discussion

3.1. Feature engineering

Figure 1 shows 18 x18 matrix correlation heatmap ranging from -1 to 1 with squares representing the relationship between variables to predict both the formation and total energy of sodium containing materials. When the correlation is close to 1 or -1, it implies that the variables have a strong relationship. In addition, a value closer to zero indicates that the two variables are not linearly related. Since all the diagonals are -1 (black) or 1(fawn), there is a perfect correlation. A larger number and darker or lighter colour indicate a stronger correlation between the two variables. In this study, 18 descriptors were considered and evaluated in order to determine the important descriptors in predicting the energies and average covalent radius and average single bond covalent radius were found to be the most important features, as can be seen on the heatmap by both correlations closer to one and the light fawn colour, both feature correlations range between 0.82 and 0.99.

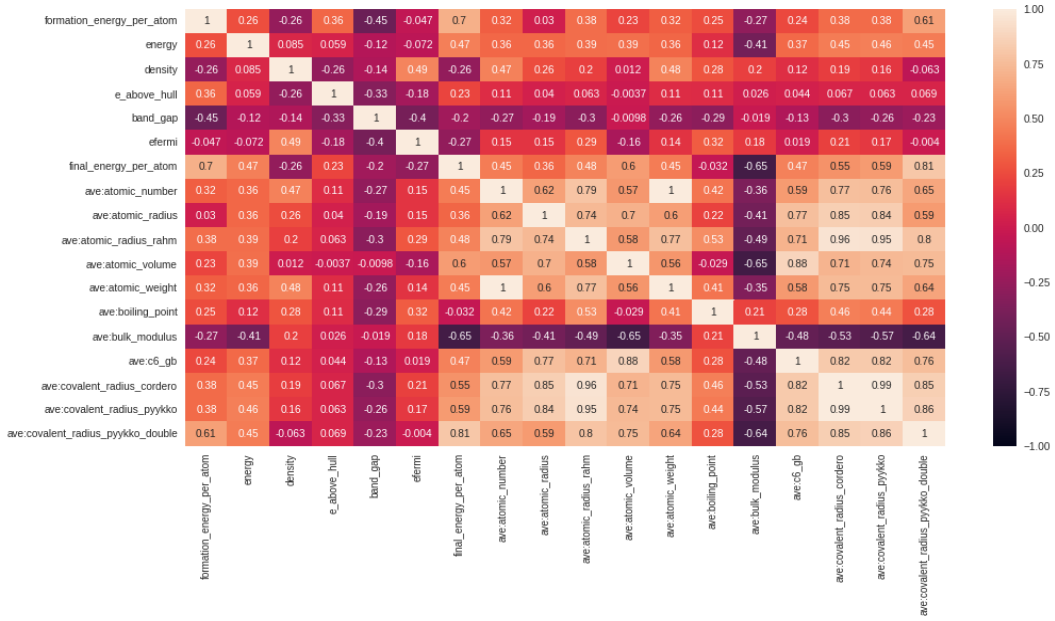

Figure 1. Correlation heatmap for predicting the formation energy and total energy

3.2. Model selection

The best model is selected on its capability to predict the target property, in this case formation and total energy. Figure 2 illustrates the measures of the predicted coefficient of determination for the formation (left) and final energy (right) as determined by various models while Figure 3 shows the measures of the predicted formation energy MSE (left) and final energy MSE (right) as determined by various models. During model tuning, data is pre-processed using statistical normalization, then hyperparameters are optimized using cross-validation score. The tuned algorithm is fitted to the training data, which comprised 70%, and the learned model is then applied to the test data, which comprised 30%. Various models are evaluated by resampling on data collected outside the sample (or, more precisely, the development process of the model). Amongst the developed models, the Bayesian ridge was found to be the best model based on its accuracy in predicting the energies.

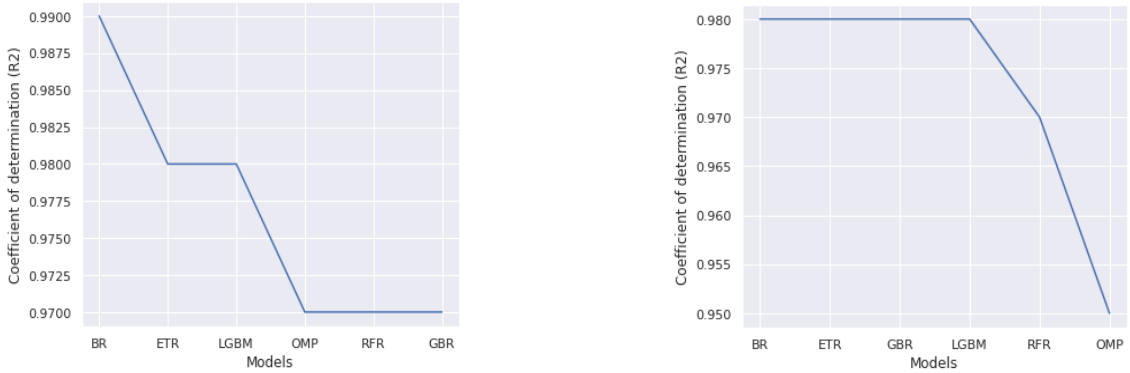

Figure 2. Measures of predicted formation energy coefficient of determination (left) and final energy coefficient of determination (right) as determined by various models

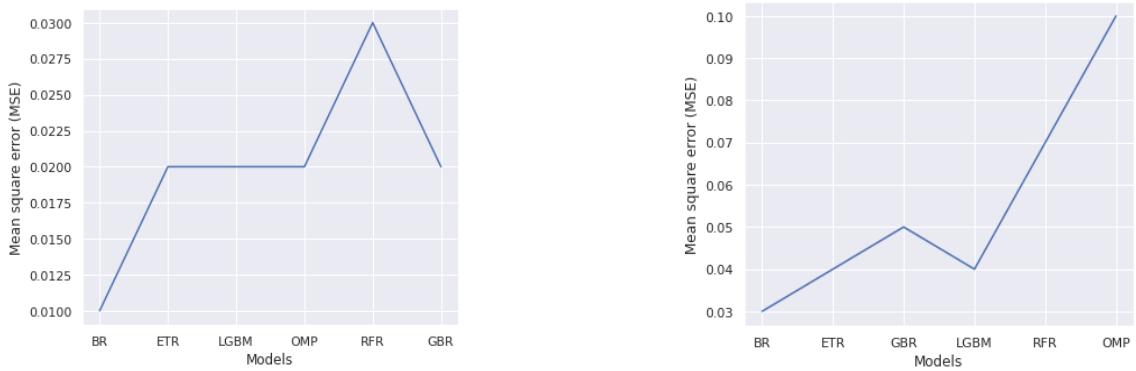

Figure 3. Measures of predicted formation energy mean square error (left) and final energy mean square error (right) as determined by various models

3.3. Model validation

Figure 4 shows graphical representations of model performance in the training set (left) and the test set (right), containing data points that reflect the predicted formation energy as a function of the DFT calculated formation energy obtained using Bayesian ridge regressor. Bayesian ridge regression is found to be the best performing model with the predicted formation energy achieving a coefficient of determination R^2 of 0.99 and an MSE of 0.01eV. LGBM predicted formation energy poorly with a coefficient of determination of 0.59 and a mean square error of 1.40 eV for the training set and the coefficient of determination under the testing set was found to be 0.55 and a mean square error of 1.51 eV. Due to this poor performance of LGBM, features that did not add value to the prediction of the formation energy were then removed in order to determine the role that feature vectors play on model

performance. The models were trained again with few feature vectors (we removed density and band gap) and found to improve the performance to 0.69 and 0.01 eV for R^2 and MSE, respectively.

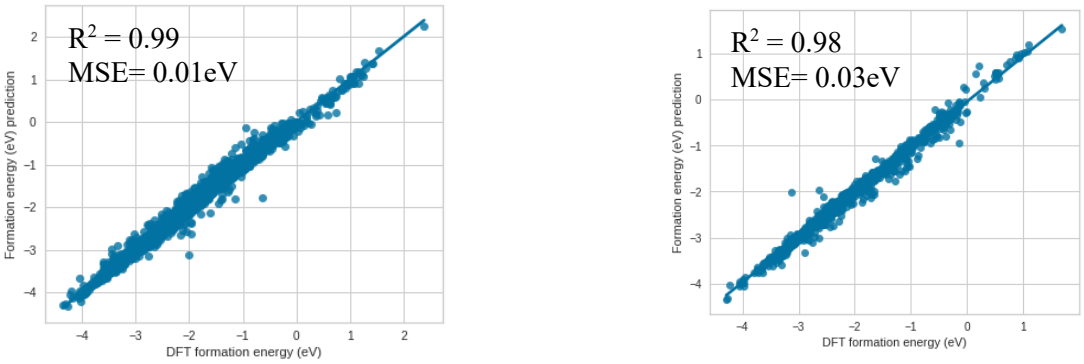

Figure 4. Parity plot of Bayesian ridge model predicted formation energy versus DFT formation energy model performance in train set (left) and test set (right)

Figure 5 shows a graphical representation of the model performance in training set (left) and test set (right), containing data points that reflect the predicted final energy as a function of DFT calculated final energy obtained from the Materials Project database. Bayesian ridge regression with elemental descriptors predicted the final energy, achieving a coefficient of determination R^2 of 0.98 and an MSE of 0.03eV. Later, the performance of the optimized algorithm gave an MSE of 0.04eV and an R^2 of 0.97 through model tuning. For the best advantage, all the points should pass through the diagonal regressed line, and the model Bayesian ridge resulted in a high regression score. Based on the small difference between the train and test-model performance results, it is evident that the models were not overfitted.

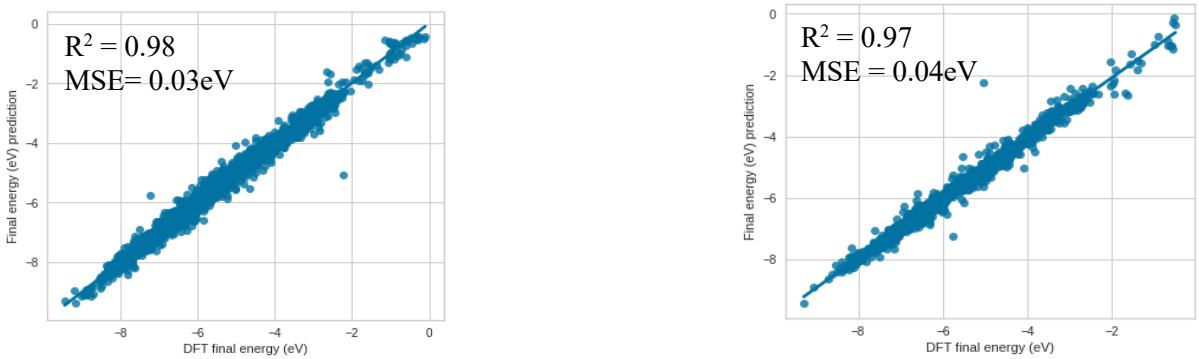

Figure 5. Parity plot of Bayesian ridge predicted final energy versus DFT final energy model performance in train set (left) and testing set (right)

4. Conclusion

The machine learning models were successfully developed, from which the formation and final energies of various sodium-ion battery materials were predicted. The average covalent radius and the average single-bond covalent radius were the most important features for predicting the formation and final energy of these materials. Several models were evaluated, and the best model was selected based on its accuracy in predicting the afore-mentioned energies. Amongst the various algorithms that were evaluated, Bayesian ridge model was found to be the best model with the following accuracy measures: coefficient of determination R^2 of 0.99 and 0.98 for formation and final energy respectively, mean square error of 0.01 and 0.03eV for formation energy and final energy respectively. The machine learning models were further validated, from which the DFT calculated properties were compared with their corresponding predicted machine learning values. There is good agreement between the model on the

train and the test set. Machine learning models can yield accurate material properties faster, making them useful in materials-properties prediction.

5. Acknowledgments

The study is funded by the Department of Science and Innovation-Inter-bursary Support (IBS) Programme of the Council for Scientific and Industrial Research, and the funding support is duly acknowledged.

References

[1] Kauwe S, Rhone T and Sparks T 2019 Crystal (Basel). **9** 1-9
[2] Peng L, Zhu Y, Chen D, Ruoff R and Yu G 2016 Adv. Energy Mater. **6** 1-21
[3] Takagishi Y, Yamanaka T and Yamaue T 2019 Batteries. **5** 1-14
[4] Chayambuka K, Mulder G, Danilov D and Notten P 2018 Adv. Energy Mater. **8** 1-49
[5] Haxel G, Hendrick J and Orris G 2002 US. Geological Survey. **87** 3529-614
[6] Hwang J, Myung S and Sun S 2017 Chem. Soc. Rev. **46** 3529-614
[7] Hautier G, Fischer C, Jain A, Mueller T, Ceder G 2010 Chem. Mater. **22** 3762–67
[8] Hautier B, Agrawal A, Kirklin S, Saal J, Doak J, Thompson A, Zhang K, Choudhary A and Wolverton C 2014 Phys. Rev. B. **89** 1-7
[9] Maphanga R, Mokoena T and Ratsoma R 2021 Mater. Today: Proceedings. **38** 773-8
[10] Joshi R, Eickholt J, Li L, Fornari M, Barone V and Peralta J 2019 ACS Appl. Mater. & Interfaces. **11** 18494–503
[11] Liu Y, Guo B, Zhou X, Li Y and Shi S 2020 Energy Stor. Mater. **31** 434-50
[12] Moses I, Joshi R, Ozdemir B, Kumar N, Eickholt J and Barone V 2021 ACS Appl. Mater. & Interfaces. **13** 53355–62
[13] Liu Y, Niu C, Wang Z, Gan Y, Zhu Y, Sun S and Shen T 2020 J. Mater. Sci. **57** 113–22
[14] Mueller T, Kusne A and Ramprasad R 2019 Rev. Comput. Chem. **29** 186-273
[15] https://materialsproject.org
[16] Fan J, Ma X, Wu L, Zhang F, Yu X and Zeng W 2019 Agric. Water Manag. **225** 105758
[17] Polamuri R, Srinivas K and Mohan A 2019 IIJRTE. **8** 2277–3878
[18] MacKay D 1992 CNS. **4** 415–47
[19] Schmidt J, Marques M, Botti S and Marques M 2019 npj Comput Mater. **5** 1-36