Current Applied Science and Technology Vol. 23 No. 6 (November-December 2023)

Research article

Thermoelectric Prediction from Material Descriptors Using Machine Learning Technique

Pakawat Sungphueng and Kittiphong Amnuyswat*

College of Materials Innovation and Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

Received: 10 September 2022, Revised: 2 February 2023, Accepted: 5 April 2023

DOI: 10.55003/cast.2023.06.23.014

Keywords

thermoelectric;
machine learning;
neural network;
random forest

Abstract

In this work, we employed a machine learning framework to predict the thermoelectric power factors of materials based on their composition and structure. To generate a broad range of materials for analysis, we sourced an existing dataset from the Materials Project database. The electronic transport properties, which serve as the output variables, were obtained from the same database via a Boltzmann transport theory calculation beyond ab-initio method. These properties were used to generate input data, or material descriptors, which rely solely on atomic information and crystal structure without recourse to density functional theory calculations. The descriptors were transformed into numerical features using the open-source software Matminer. Non-linear machine learning regression models were trained and tested on the transformed datasets, and their performance was evaluated. The optimized random forest model produced the most accurate predictions, with a yield of 88%. The ultimate goals of this research were to develop material selection strategies that bypass the need for self-consumption in density functional theory calculations, and to demonstrate the potential of machine learning models to describe the thermoelectric properties of existing materials datasets.

1. Introduction

Thermoelectric materials have gained significant attention in recent years due to their potential for efficient energy conversion and waste heat recovery. These materials have the ability to convert heat into electrical energy and vice versa, making them suitable for various applications in the energy sector. Thermoelectric materials are classified based on their ability to generate a voltage (Seebeck coefficient), conduct electricity (electrical conductivity), and resist heat flow (thermal conductivity). These materials display a unique combination of high electrical conductivity and low thermal conductivity, making them ideal for thermoelectric applications.

*Corresponding author: Tel.: (+66) 869020207
E-mail: kittiphong.am@kmitl.ac.th

At present, the current state of thermoelectric materials and their usefulness in various applications were explored. One of the most significant applications of thermoelectric materials is in waste heat recovery. In industrial processes, a significant amount of energy is lost as waste heat. This waste heat can be captured and converted into electrical energy using thermoelectric materials [1]. This not only reduces energy waste but also improves the overall efficiency of the system. For example, thermoelectric materials have been used in automobiles to recover waste heat from the engine and convert it into electrical energy to power accessories. Another application of thermoelectric materials is in power generation. For example, they have been used in power plants to convert the waste heat generated during the power generation process into electrical energy. This improves the overall efficiency of the power plant and reduces energy waste.

Nowadays, finding practical novel materials is a difficult process which often relies on experimentation that usually takes a lot of trials to succeed. Development in density functional theory (DFT) has made it a high-performance discovery tool for a wide variety of material properties for various material types including metals, semiconductors, and organic and inorganic compounds. Because of its potential in many applications, DFT has attracted a lot of attention from scientists. It can enable the prediction of exactly the right properties with tunable composition and structure of materials. Although DFT is a widely accepted method, the DFT calculation involves large computational consumption when a material structure slightly changes. A huge DFT workflow still needs to be done before presenting a material discovery to obtain an accurate display, especially when dealing with a post DFT method to obtain specific properties that represent experimental value accurately. These fundamentally limited computational times required mean a DFT calculation can cause non-system errors to occur. Many scientists are accelerating data-driven material design as a key step in the discovery of high throughput materials [2-4]. A very recent development in materials research is to create predictive models using machine learning to automatically predict a wide variety of properties from existing materials databases. This active area of materials research provides alternative algorithms that save the cost and time consumed in DFT calculation with results of similar accuracy [5]. Nowadays, some fascinating aspects of the machine learning pipeline are being successfully deployed in an emerging area of research of various properties. This is because machine learning can unravel the previously unknown relationships between a material's properties and its material descriptors. Over the past decade, machine learning algorithms have been used as a new area of research in materials science [6-9].

In this article, the electronic power factor was defined as a thermoelectric property, and it is one of the properties of the material to be studied. Traditionally, the full Boltzmann transport equations (BTE) was used to calculate Seebeck and electrical conductivity, which are parameters for computing the power factor [10]. However, they can take a long time to calculate because a full Brillouin zone integration in the crystal structure is required. Therefore, a supervised machine learning algorithm was performed to predict the true electronic power factor. In 2018, high-performance thermoelectric materials were discovered by material screening [5]. In this work, we present a framework where only atomic information and their crystalline structurewere used as input datasets for a machine learning pipeline. The thermoelectric values predicted using this approach were systematically studied and discussed with a traditional DFT calculation.

2. Materials and Methods

The experiment was depicted through a flow chart to clearly illustrate the sequential steps taken, as shown in Figure 1. The material database is a crucial component of a machine learning pipeline aimed at achieving the desired outcomes. MaterialProject.org is an extensive, open-source database of material properties that is commonly employed in the design of materials [8]. This database

includes over 139,367 crystalline compounds, as calculated using density functional theory (DFT); however, only 8,061 of these compounds have a power factor provided; the factor is essential for predicting the target in machine learning algorithms. In this study, the dataset was obtained with the aid of an open-source Python library called Matminer, which serves as an intermediary between users and commonly used open-source databases. After a process of data filtering and cleaning, our dataset comprised 8,061 indexed rows and four columns including material_ID, pretty_formula, structure, and an output column labeled power factor. Each row corresponded to a specific material. The generated columns contained string objects that had to be transformed into numerical attributes for use as inputs into the machine learning model. In this study, two types of features were employed. The first type comprised direct features, which were numerical information that can be directly obtained from atomic information or crystal structure. The second type were transformed features, which were quantitative descriptors that had to be prepared from existing raw material information based on an existing module of matminer. In this study, the materials agnostic platform for informatics and exploration (MAGPIE) was utilized to compute elemental property attributes. This involved the inclusion of atomic orbitals, valence orbitals, stoichiometry, band center, and fraction of metal atom in the computation of elemental attributes. Additionally, the structural properties of materials were characterized using features such as heterogeneity, complexity of material structure, chemical ordering, and density. At this step, some of the transformations resulted in a "not a number" value, which had to be filtered out from the dataset. Following this filtering process, the final dataset comprised 7,654 indexed rows and 179 features. To identify an appropriate machine learning technique, we evaluated the performance of our models using the cross-validation technique. A nonlinear algorithm model [11] was chosen to learn a training dataset, then validated and tested to achieve optimal performance. The R-squared and root mean squared error metrics were used to compare the performance of the obtained models against the true output power factor. Finally, a simple test was conducted to estimate the performance of changes in crystal information of simple materials, which are commonly used in thermoelectric devices.

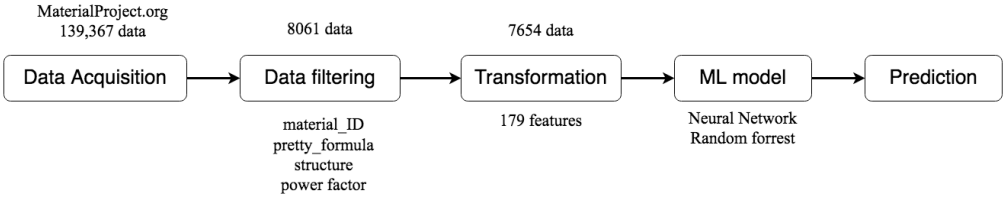

Figure 1. A graphical representation of the various steps involved in the process of building a machine learning model. It starts with collection of data from a database. Next, the data is preprocessed, cleaned, and transformed. The data is then fed into an algorithm to train the model and fine-tune its parameters.

3. Results and Discussion

Machine learning (ML) is a rapidly growing field that has already made significant contributions to various industries. In recent years, the use of machine learning in the field of thermoelectrics has been gaining traction due to its potential to predict power factors with high accuracy. In this article, we explore the importance of machine learning in predicting the power factor of thermoelectric materials. The power factor of a thermoelectric material is an important parameter that determines its efficiency in energy conversion. In the field of thermoelectrics, the power factor is defined as the

product of the Seebeck coefficient and the electrical conductivity of the material. Predicting the power factor of a thermoelectric material is a complex task as it depends on multiple parameters, including temperature, material composition, and crystal structure. Machine learning algorithms can analyze large datasets and identify complex relationships between parameters and the power factor.

Initially, the focus of this study was the development of a machine learning framework. The objective of the dataset preparation step was to generate a set of quantitative representations that uniquely define the composite of materials in the dataset and its influence on the relevant properties. To achieve this goal, multiple attribute sets were created using two types of transformation protocols. The first type involves directly obtaining information from the materials' structural properties. The second type encompasses large attribute sets that described the raw materials based on statistical information derived from their structural data. A total of 179 attributes were generated to establish the relationship between atomic and structural information and the power factor. These transformed quantiles were intended to be used to identify correlation patterns using the machine learning model, which could potentially recognize relationships from a set of these features automatically.

For the next stage of the machine learning framework, two nonlinear machine learning algorithms, the neural network (NN) and random forest (RF) regressors, were chosen to map these attributes to their corresponding power factor properties. The neural network model fits a multi-layer Perceptron regressor by utilizing the rectified linear unit function. The random forest model is trained on various sub-samples of the dataset by classifying decision trees and then averaging to produce a final predictive model. The performance of each algorithm is meticulously evaluated.

We evaluated the performance of the models using the R-squared value and the root mean squared error (RMSE). The R-squared value is a measure of how well the model fits the data, with values close to 100 indicating a good fit. The RMSE measures the difference between the predicted values and the actual values, with lower values indicating better performance. Figure 2 illustrates a comparison of the R-squared and root mean squared error (RMSE) performance metrics across a range of machine learning algorithms. The random forest (RF) model was found to be the best performing algorithm, with an R-squared value of 88% and a root mean squared error (RMSE) of 13.328. The neural network (NN) model, on the other hand, had an R-squared value of 7.1% and a significantly higher RMSE of 36.679 (Figure 2). These results demonstrated that the RF algorithm was superior in predicting the power factor of materials compared to the NN algorithm. The RF algorithm, in particular, can be a useful tool for researchers to predict the thermoelectric power factor of materials. Our results suggest that the RF algorithm is more accurate than the NN algorithm in predicting this property.

The power factor is a crucial parameter in the design and optimization of thermoelectric materials. However, the performance of these models can vary depending on the range of power factors in the dataset. It can be challenging to predict accurately as it depends on the complex interplay between the atomic and structural properties of the material. In this study, we developed a practical nonlinear random forest model to predict the power factor of thermoelectric materials. We systematically evaluated the performance of the model across a range of power factors from 0 to 100-800 GW/m.K^2. Figure 3 shows a different subset of the data points plotted within the range. The data points in each subplot are represented by dots and compared the calculated and predicted power factors. The dashed line in each subplot represents a perfect prediction by the model, where the predicted power factor equals the calculated power factor. It was shown that the predicted power factors follow the trend of diagonal dashed line, indicating acceptable agreement between the calculated and predicted power factors for the data limited under 200 GW/m.K^2. Our results show that the performance of the random forest model significantly improved when data points with a power factor above 200 GW/m.K^2 were filtered out. The initial RMSE of 13 decreased to 9, indicating an increase in the accuracy of the model. The majority of data points in the dataset lay

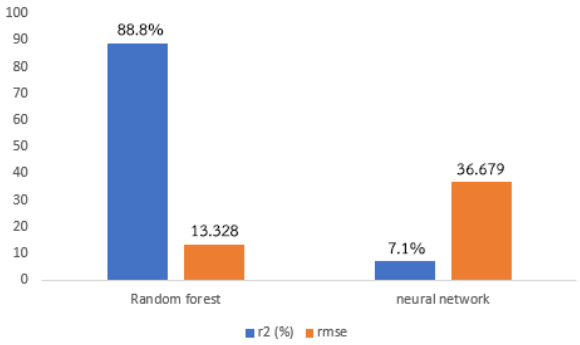

Figure 2. The evaluation metrics of R-squared and RMSE were utilized to compare the performance of the neural network (NN) and random forest (RF) models.

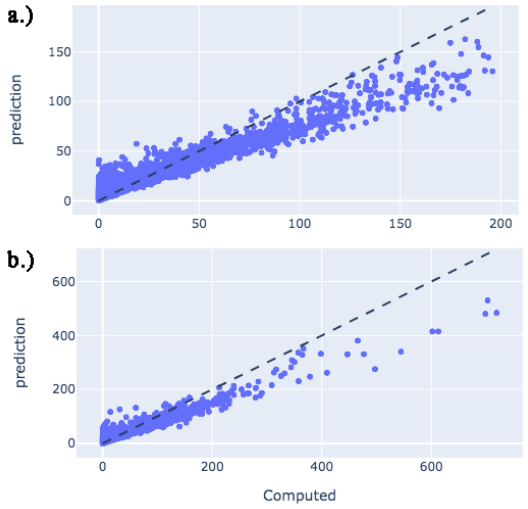

Figure 3. A scatter plot that compares the calculated and predicted power factors across a range of 0-200 GW/m.K². In panel (a), the plot displays a line that represents the relationship between the two factors, while in panel (b), all data points are plotted individually. The x-axis represents the calculated power factor, while the y-axis represents the predicted power factor.

under this limit, and the small number of data points with a higher power factor made it difficult for the model to predict with accuracy.

In order to investigate the effect of filtering the dataset on the gradient of the data, we compared the results obtained from the random forest model with and without filtering. Our analysis revealed a significant increase in the gradient of the data when the dataset was not filtered out, as shown in Table 1 and Figure 4. This indicates that the inclusion of a larger range of power factors in the dataset can decrease the overall performance of the model. This is due to the fact that the model needs to address the issue of missing data in a high-value range, and simply disregarding missing values in favor of achieving optimal performance can result in a more complex model that may be challenging to interpret.

Table 1. Comparison of the range of power factor with the performance of the random forest model and the number of data points when filtered.

Limiting Point	R-Squared	RMSE	Number of Data	Gradient of Data	Maximum Regression Coefficient
100	87.9	6.275	7371	0.7105	0.038315
200	88.3	9.292	7593	0.7246	0.053258
300	87.8	10.837	7629	0.7181	0.052526
400	89.1	11.240	7643	0.7301	0.045484
500	87.9	12.440	7648	0.7237	0.048442
600	89.0	12.063	7649	0.7217	0.048540
700	87.9	12.349	7652	0.7174	0.043270
No filtering	88.8	13.328	7654	0.7943	0.044670

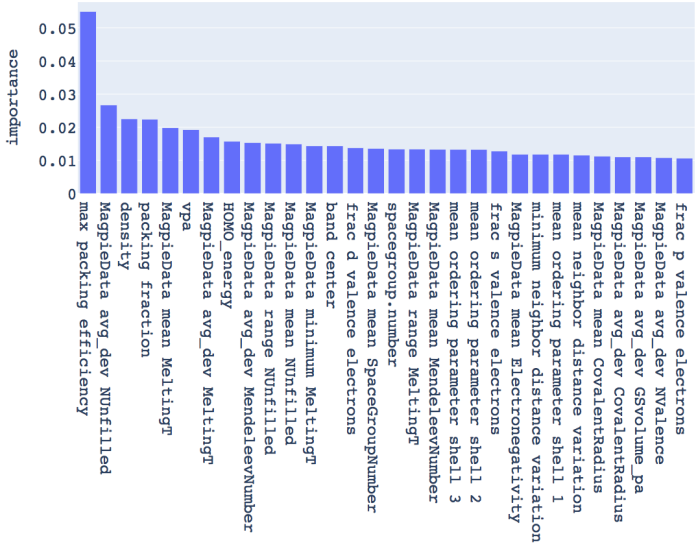

Figure 4. The significant features from the random forest model

The coefficient of a high importance feature is called regression coefficient. This term refers to the estimated value of the effect of a given feature on the outcome variable in a regression model. When a feature has high importance, it means that its regression coefficient is relatively large, indicating that the feature has a strong relationship with the outcome variable. Based on our analysis, we found that limiting the dataset to less than 200 GW/m.K² resulted in a higher regression coefficient compared to using the full range of data. Specifically, our results show that by limiting the dataset to less than 200 GW/m.K², we were able to collect 99% of all available data, resulting in a significantly higher gradient in the relationship between the power factor and the material attributes. This suggests that using a limited dataset with a carefully chosen range can lead to better performance in the predictive model and may be a useful approach in other materials applications. Further research is needed to determine the optimal range for different types of materials and applications.

To investigate the relationship between the power factor and material attributes, the estimated parameters of the random forest model were plotted. The results indicate that the attributes that best describe the power factor of materials are related to packing efficiency, which depends on

the type of structure. Specifically, the percentage of spaces filled by the particles in the unit cell always shows some void spaces in the unit cell that can be used to identify the material. The atomic properties described by the MAGPIE transformer were also found to be important in predicting the power factor of materials. These atomic properties include electronegativity, ionic radius, and atomic volume. Notably, the values of these properties were found to be quite similar across the materials. Overall, the findings suggest that a combination of structural and atomic properties are key predictors of the power factor of materials. These insights can be useful in the development of new materials with optimized thermoelectric properties.

The developed machine learning framework with a random forest model was applied to predict the power factor of various materials used in thermoelectric devices. The analysis was performed on a dataset with a length of less than 200 GW/m.K². The list of selected materials [12] used in the analysis is provided in Table 2 and Figure 5. The obtained power factor was acquired and assembled from the available database. Our predictive model was able to accurately predict the power factor of various novel materials with high precision. The model was found to be highly effective in predicting the power factor of materials based on their structure, through their packing efficiency, which was found to be the most significant attribute affecting the power factor.

Table 2. A comparison between the calculated and predicted power factors for different materials using the random forest model are present. The materials listed are novel materials that are being considered for use in thermoelectric devices.

Materials	Obtained Power Factor (GW/m.K²)	Predicted Power Factor (GW/m.K²)
ZnO	5.512	6.304
HfCoSb	9.365	11.801
$CoSb_3$	10.177	12.904
CuBiSeO	10.874	13.151
SnS	12.349	10.664
SiGe	13.102	18.578
Mg_2Si	17.814	17.088
Ag	19.007	19.289
Te	27.076	28.908
SnSe	31.303	22.898
TePb	33.368	41.292

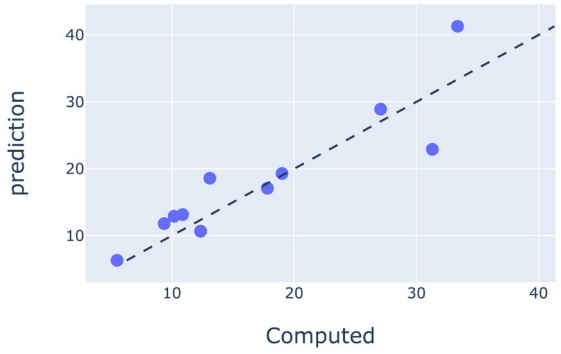

Figure 5. Comparing between obtained and predicted power factors from various materials for a thermoelectric device. The dashed line indicates a perfect prediction by model.

However, it is worth noting that the model's strength was destroyed when the data reached saturation limit, and the predicted values underestimated the actual computed values. Overall, the results suggest that the established machine learning framework has the potential to predict the power factors of materials used in thermoelectric devices provided that the materials can be effectively described using the established attribute sets.

4. Conclusions

We have shown that a machine learning algorithm can reliably predict the power factor of thermoelectric materials from a theoretically calculated data set. Our results show that a random forest model based on a simple machine learning algorithm with direct atomic and crystal structure information and their converter attribute can be used to predict the theoretically calculated values of power factor in the thermoelectric materials. Our results show that the RF algorithm was superior to the NN algorithm in predicting this property. The excellence performance was nearly 90% when predictions were compared to actual results. This can be extended to design excellent functional materials from an available dataset. The results of this study demonstrate the potential of machine learning algorithms in materials science research and can serve as useful tools for researchers when predicting material properties.

5. Acknowledgements

This work was partially supported by the College of Materials Innovation and Technology (CMIT), King Mongkut's Institute of Technology Ladkrabang.

References

[1] Ismail, I.B. and Ahmed, H.W., 2009. Thermoelectric power generation using waste-heat energy as an alternative green technology. *Recent Patents on Electrical Engineering*, 2(1), DOI: 10.2174/1874476110902010027.

[2] Ward, L., Agrawal, A., Choudhary, A. and Wolverton, C., 2016. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2, DOI: 10.1038/npjcompumats.2016.28.

[3] Curtarolo, S., Hart, G.L.W., Nardelli, M.B., Mingo, N., Sanvito, S. and Levy, O., 2013. The high-throughput highway to computational materials design. *Nature Material*, 12, 191-201, DOI: 10.1038/nmat3568.

[4] Xi, L., Pan, S., Li, X., Xu, Y., Ni, J., Sun, X., Yang, J., Luo, J., Xi, J., Zhu, W., Li, X., Jiang, D., Dronskowski, R., Shi, X., Snyder, G.F. and Zhang, W., 2013. Discovery of high-performance thermoelectric chalcogenides through reliable high-throughput material screening. *Journal of the American Chemical Society*, 140, 10785-10793, DOI: 10.1021/jacs.8b04704.

[5] Ye, W., Chen, C., Dwaraknath, S., Jain, A., Ong, S. and Persson, K., 2018. Harnessing the materials project for machine-learning and accelerated discovery. *MRS Bulletin*, 43(9), 664-669, DOI: 10.1557/mrs.2018.202.

[6] Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. and Kim, C., 2017. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3, DOI: 10.1038/s41524-017-0056-5.

[7] Meredig, B., Agrawal, A., Kirklin, S., Saal, J.E., Doak, J.W., Thompson, A., Zhang, K., Choudhary, A. and Wolverton, C., 2014. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B*, 89, DOI: 10.1103/PhysRevB.89.094104.

[8] Lu, S., Zhou, Q., Ouyang, Y., Guo, Y., Li, Q. and Wang, J., 2018. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nature Communication*, 9, 1-8, DOI: 10.1038/s41467-018-05761-w.

[9] Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G. and Persson, K.A., 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), DOI: 10.1063/1.4812323.

[10] Madsen, G.K.H., Carrete, J. and Verstraete, M.J., 2018. BoltzTraP2, a program for interpolating band structures and calculating semi-classical transport coefficients. *Computer Physics Communications*, 231, 140-145, DOI: 10.1016/j.cpc.2018.05.010.

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E., 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830, DOI: 10.48550/arXiv.1201.0490.

[12] Gaultois, M.W., Sparks, T.D., Borg, C.K.H., Seshadri, R., Bonificio, W.D. and Clarke, D.R., 2013. Data-driven review of thermoelectric materials: Performance and resource considerations. *Chemistry of Materials*, 25, 2911-2920, DOI: 10.1021/cm400893e.